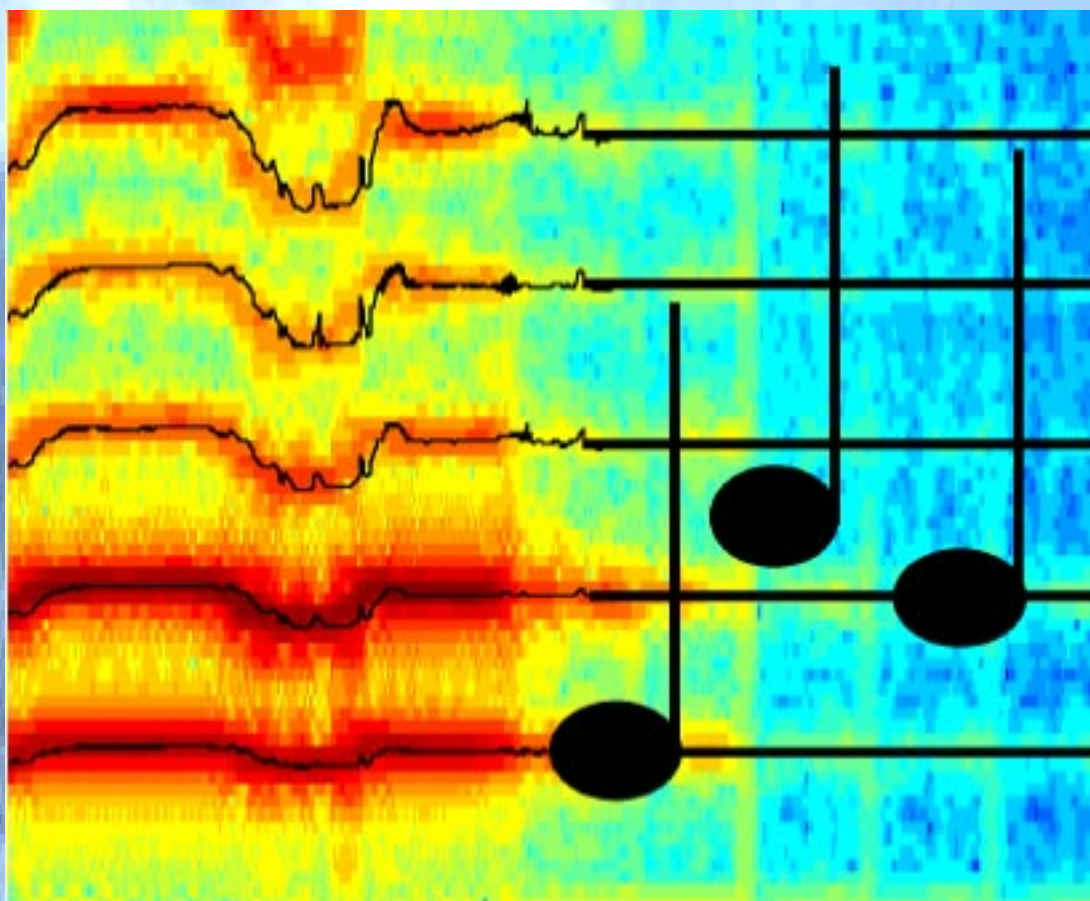


# Proceedings of the Mosart Midterm Meeting

Esbjerg, August 29-31, 2002



# MOSART

Musical Orchestration Systems in Algorithmic Research and Technology  
An EU IHP project, funded for three years

# Proceedings of the Mosart Midterm Meeting

**Esbjerg, August 29-31, 2002**

MOSART is a major network project, funded by the Brussels Increasing Human Potential Program. The network consists of thirteen nodes, all active in the cross field of technology and musical activity. The network theme is Musical Orchestration Systems in Algorithmic Research and Technology (MOSART), and it invites young researchers to participate in the three year network project program.

During three days, August 29 - 31, a meeting will be held in Esbjerg in Denmark, in order to assess the progress of the work done in the network. All participants will be there together with the Brussels officials for a formal review of the project. We furthermore has invited Ph.D. students and their supervisors and related seniors to attend the meeting as well.

The joint program of work is broken down into the following seven tasks, T1 .. T7:

- T1 Analysis of Current Methods for Music Orchestration Systems.
- T2 Musical Sound Understanding and Timbre Modeling.
- T3 Control and Virtualisation of Instruments.
- T4 Detection of Human Motion and Interactive Musical Performance.
- T5 Symbolic Recognition of Musical Patterns and Recomposition.
- T6 Computer Music Composition Tools.
- T7 Technology transfer and Industrial exploitation.

Each tasks has a task leader, and a number of associated partners, as detailed in the MOSART Contract Annex 1.

The MOSART Midterm Meeting consist of an scientific part, where Senior partners and invited young researchers present their scientific work, and an administrative part, where the networking aspects of the project are detailed. Please find below the program of the meeting, and the papers submitted to the MOSART Midterm Meeting. In addition, the final panels from the Barcelona MOSART Meeting are included in the proceedings.

**Jens Arnspang**, Aalborg University Esbjerg, Coordinator of MOSART for Brussels

**Kristoffer Jensen**, University of Copenhagen, Editor of the Proceedings of the MOSART Midterm Meeting



# Table of Content

<b>MOSART Midterm Meeting Program</b>	Page 1
---------------------------------------	--------

## **Task 1 progress report.** Task leader Xavier Serra.

Since the Task 1 is finished, there is no progress report on this task.  
Please see the Barcelona meeting<sup>1</sup> for the final report on task 1.

<b>Task 2 progress report.</b> Task leader Giovanni de Poli	Page 9
---	--------

<b>Overview on timbre perception and modeling</b> Hanna Järveläinen, Giovanni De Poli	Page 10
--	---------

<b>Automatic Classification of Musical Instrument Sounds</b> Perfecto Herrera-Boyer, Geoffroy Peeters, Shlomo Dubnov	Page 17
---	---------

<b>Timbre Modeling and Analysis-Synthesis of Sounds</b> Kronland-Martinet, Richard. Guillemain, Philippe. Ystad, Sølvi.	Page 38
--	---------

<b>The Timbre Model - Discrimination and Expression</b> Kristoffer Jensen	Page 49
--	---------

<b>Directional patterns and recordings of musical instruments in auralizations</b> Felipe Otondo, Jens Holger Rindel, Claus Lynge Christensen	Page 65
--	---------

<b>Toward an objective method for the timbre analysis of sound-objects in the electro-acoustic music repertoire</b> Sergio Canazza, Chiara Marini	Page 68
--	---------

<b>Abstract musical timbre and physical modeling</b> Giovanni De Poli, Davide Rocchesso	Page 75
--	---------

<b>Importance of phase in sound modeling of acoustic instruments</b> Tue Haste Andersen, Kristoffer Jensen	Page 96
---	---------

<b>A Hybrid Re-Synthesis Model for Hammer-Strings Interaction of Piano Tones</b> J. Bensa, K. Jensen, and R. Kronland-Martinet	Page 105
---	----------

<b>Harmbal: a program for calculating steady-state solutions to non-linear physical models of wind and string instruments</b> S. Farner	Page 115
--	----------

---

<sup>1</sup> <http://www.iua.upf.es/mtg/mosart/>

<b>Task 3 progress report.</b> Task leader Jan Tro	Page 123
<b>Control and Virtualisation of Musical Instruments.</b> Jan Tro	Page 124
<b>A classification of controllers for parameters highlighting</b> Gabriele Boschi	Page 128
<b>Musical Instruments Control and Expression</b> Kristoffer Jensen	Page 134
<b>From Sounds to Music: Different Approaches to Event Piloted Instruments</b> P. Gobin R. Kronland-Martinet G.-A. Lagesse, T. Voinier, S. Ystad	Page 143
<b>New Method for the Directional Representation of Musical Instruments in Auralizations</b> Felipe Otondo, Jens Holger Rindel	Page 169
<b>Directional representation of a clarinet in a room</b> Felipe Otondo and Jens Holger Rindel	Page 173
<b>Task 4 progress report.</b> Task leader Jens Arnspang	Page 181
<b>Detection of Human Motion and Interactive Musical Performance</b> Jens Arnspang, Kristoffer Jensen, Declan Murphy	Page 182
<b>Gesture Recognition for Conductor and Dance Interpretation</b> Volker Krüger	Page 187
<b>A Smart Analog to MIDI Interface</b> Gabriele Boschi	Page 191
<b>Building A Hand Posture Recognition System From Multiple Video Images: A Bottom-Up Approach</b> Declan Murphy	Page 193
<b>The Votion Project</b> B. Stang, E. Tind, D. Murphy, J. Arnspang, K. Jensen, A.-M. Bach Jensen, C. Beyer, M. Gugliemi	Page 227
<b>An Improved Edge Detection and Ranking Technique</b> Declan Murphy	Page 232
<b>Extracting Arm Gestures for VR using EyesWeb</b> Declan Murphy	Page 240
<b>Task 5 progress report.</b> Task leader Gerhard Widmer.	Page 247
<b>Patterns in music and music performance</b> Gerhard Widmer	Page 248

<b>Task 6 progress report.</b> Task leader Barry Eaglestone	Page 257
<b>Requirements specification for a composition tools system</b> Barry Eaglestone, Guy Brown, Nigel Ford, Adrian Moore, Ralf Nuhn,	Page 258
<b>Task 7 progress report.</b> Task leader Antonio Camurri.	
Since task 7 is just starting, a progress report is unavailable at present.	
<b>Summary of the Panel Discussions from Workshop on Current Research Directions in Computer Music</b>	Page 297
<b>Panel on Future Directions in Music Sound Modeling</b> Enric Guaus (editor)	Page 298
<b>Panel on Future Directions in Music Interfaces</b> Álvaro Barbosa (editor)	Page 302
<b>Panel on Future Directions in Music Performance</b> Maarten Grachten (editor)	Page 306
<b>Empirical Music Performance Research: ÖFAI's Position</b> Gerhard Widmer, Simon Dixon, Werner Goebel, Efstathios Stamatatos, Asmir Tobudic	Page 311
<b>Music Performance Panel: Position Statement: KTH Group</b> Johan Sundberg, Anders Friberg, Roberto Bresin	Page 314
<b>Music Performance Panel: NICI / MMM Position Statement</b> Peter Desain, Henkjan Honing and Renee Timmers	Page 318
<b>Panel on Future Directions in Music Generation</b> Rubén Hinojosa Chapel (editor)	Page 323



# Meeting Agenda

**Thursday, August 29**

**9:00-9:15 Raymond Monk**, Eu representative Overview

**9:15-10:00 Jens Arnspang**, Mosart Network Activities Overview

**10:00-10:20 Völker Krüger**, Gesture Recognition for Conductor and Dancer Interpretation

Our principle goal is to design, develop and evaluate a novel system using digital cameras for the detection and interpretation of human motions for interaction with musical performances, e.g. for the interpretations of conductors, the control of musical instruments or the automatic choreographic assistances for dancers. Such a system will provide 3-D measurements of human movements of the whole bodies, body parts and joints. The availability of these descriptions will allow us to describe, manipulate and interpret the human movements according to our goals.

**10:20-10:40 Kristoffer Jensen**, Timbre Model Research at DIKU

This presentation will give an overview of the research at DIKU music informatics laboratory. In particular, the cooperative nature of the research is outlined, the dissemination and publications are given. Piano modeling, sound quality, expression and control, gesture capture are some related topics with relevance to the timbre research.

**10:40-10:50 Jan Tro** (NTNU, Norway)

- Discussing the research activities primarily connected to the Task 3 (Control and Virtualization of Instruments) concerning research strategy and status;
- Activities at the NTNU-Acoustics Group related to T3.

**10:50-11:10 Gabriele Boschi**, Performing Control Parameters

Research activities at the NTNU stay:

- Memo on performing control parameters;
- Study on SoC implementation of controller for contemporary music;
- Multiple Digital recordings (40) of a flute piece in different acoustical environment.

**11:10-11:40 Coffe break**

**11:40-11:50 Johan Sundberg**, KTH activities overview

**11:50-12:10 Erwin Schoonderwald**, Vibrato in violin performance



Erwin Schoonderwaldt has analysed vibrato in violin performance. He summarised his results in terms of performance rules that have been implemented in the Director Musices performance grammar. Applying these rules seems to substantially reduce the highly artificial quality of violin synthesis typically obtained even from high quality synthesizers.

### **12:10-12:30 Kjetil Falkenberg Hansen, Scratching overview**

A number of experiments and analyses involving scratching (using a turntable and a mixer as a musical instrument) were done. The main goal of the first experiments was to record a professional musician (DJ) and acquire useable data to analyse and get an overview on the matter. From the initial experiments a report was written and published. Following the more general recordings, the subject DJ visited TMH again to perform explicit playing techniques and short musical phrases. The analyses from these built the ground for making a virtualisation of scratching using Miller Puckette's Pd (Pure data) application. All experiments were performed on the same equipment, one turntable and one scratch-mixer. For the first experiment only audio signal was recorded, for the later experiments also the result of the hand movements (on the vinyl against the needle and on the crossfader) were recorded. The analyses show that the turntable/mixer should not be treated as a conventional instrument with traditional tonal and rhythmical aspects, and they make clear that even small variations in hand movements have impact on the sound produced. A crucial issue is the use of the mixer's crossfader in shaping the sound onsets and offsets.

With MOSART, I got the opportunity to get me started with a research topic that is both new and uncommon, and during the six months I learned a lot. These experiences helped me continue the research in the SOb (Sounding Objects) European project. Some of the work in the MOSART project is naturally reflected in SOb and vice versa.

#### Publications

Hansen, K F. and Bresin, R. (2002) Scratching: from analysis to modeling. In Models and Algorithms for Control of Sounding Objects , Deliverable 8, EU-IST Project no. IST-2000-25287, 15-48 (<http://www.soundobject.org/papers/deliv8.pdf>)

Hansen (2001). Playing the turntable. An introduction to scratching. TMH-QPSR, Speech Music and Hearing Quarterly Progress and Status Report, Volume 42/2001, Stockholm, 69-79. JNMR (submitted).

### **12:30-14:00 Lunch break**

### **14:00 - 14:20 Esben Skovenborg, (DAIMI) A Computational Measure of Sensory Consonance.**

A current model of pitch perception, based on cochlear filtering and periodicity detection, is extended to characterise the sensory consonance of pitch intervals. This auditory model is implemented, and a simple scalar measure of sensory consonance is developed. To evaluate the measure as a perceptually related feature, the consonance is studied for musical

pitch intervals.

**14:20 - 14:30 Jens Holger Rindel**, DTU young researchers project work and network experience

**14:30-14:50, Felipe Otondo**, Directional representation of a clarinet in a room

This overview presents a study of the directional characteristics of a clarinet in the context of a real performance. Anechoic measurements of the directivity of a Bb clarinet have been done in the horizontal and vertical planes for isolated frequency tones. Results are discussed comparing the particular directivity of tones and the averaged directivity over the whole frequency range of the instrument. Room acoustic simulations with the measured and averaged directivities have been carried out for a concert hall as an example of a realistic application. Further developments will consider measurements with other instruments as well as auralizations and tests with an alternative sound radiation representation.

**14:50-15:00 Barry Eaglestone**, USFD Mozart research overview

This presentation will provide a brief overview of research at the University of Sheffield on the Mozart Task 6 (Composition Systems) and specifically, the contribution made by the two (Mozart) young researchers, Ralf Nuhn (German) and Kevin Dahan (French). Also, detailed at the associated dissemination and networking activities.

**15:00-15:20 Kevin Dahan**, Requirement for Electroacoustic music composition

This presentation will develop ideas relating to the validation of the set of requirements for electroacoustic music composition systems, derived in the first phase of the Mozart Task 6. These requirements have been derived through an in-depth study of composers at work, conducted mainly at the University of Sheffield, but involving composers in the UK, Spain and Sweden, using qualitative research methods. Specifically, the presentation will describe a planned prototype composition system which will address many of those requirements, and in particular support for creative (divergent) activities, wholistic (Gestalt) approaches, and the ability to move seamlessly between different levels of abstraction.

**15:20 - 15:30 Henkjan Honing**, NICI contribution to Mid-term review meeting

Presentation of NICI contributions to T1, T4 and T5, Network experience, and report on project work on vibrato research (Stéphane Rossignol) and quantization (Torsten Anders) [presented by Henkjan Honing, site coordinator].  
Followed by a report on recent music performance research

**15:30-15:50, Renée Timmers**, Research on grace notes timing and timing of melodies

This talk outlines three studies that were done at the NICI, University of Nijmegen, as part of a PhD project that led to the publication of the book "Freedom and constraints in timing and ornamentation". The common aim of the studies was to investigate factors that influence the commonality and diversity between performances of classical piano music. More in detail, the first study sought to answer the question whether diversity between performances is constrained by the musical structure. The second study aimed to demonstrate that performances have intrinsic rules; rules that do not lead to commonality between performances, but ask for consistency in expression. And the third study investigated the timing of grace notes and the principles that may commonly underlie different interpretations. The formalization of these principles led to a model of grace note timing, which is demonstrated with a web-demo that allows easy exploration of original recorded performances and model-generated performances. The generated performances add a grace note to an original recorded performance without grace notes. This addition causes a shift in the surrounding notes and attributes a duration to the grace notes in accordance with certain settings of the parameters.

**15:50 - 16:00 Gerhard Widmer**, ÖFAI young researchers project work and network experience

I will briefly present recent and ongoing work at our institute that is related to the MOSART project, in particular, task T5 ("Symbolic Recognition of Musical Patterns"). At ÖFAI, MOSART research is embedded in a long-term research program that investigates expressive music performance with Artificial Intelligence methods. Large databases of performance measurements have been compiled, and these data are studied with intelligent data analysis methods (machine learning, data mining). Within MOSART, a young researcher (post-doc) worked on the problem of automatic performer identification based on a set of global performance features, with surprising results. In addition, a number of cooperations with other MOSART partners will be briefly reported on.

**16:00:16:20 Efstathios Stamatatos**, Learning to quantify the differences between music performers

In this study we deal with the problem of automatic music performer recognition given a set of performances of the same musical piece by a number of skilled performers. Following a comparative study of features for representing the performer's style, we introduce the norm-based features which are obtained from the comparison of a given performance with the average performance of a set of reference performances of the piece.

We show empirically that the norm-based features are more accurate and stable in comparison to the score-based features. Moreover, the combination of different features based on machine learning techniques are examined and the resulting classifiers have given promising results in multi-class automatic music performer recognition experiments, unlikely to be matched by human listeners under similar conditions.

**16:20-16:50 Invited Speaker: Marc Leman**, IPEM - Dept. of Musicology, Ghent University

### Tendencies, Perspectives, and Opportunities of Musical Audio-Mining

Content-based music information retrieval and associated data-mining opens a number of perspectives for music industry and related multimedia commercial activities. Due to the great variability of musical audio, its non-verbal basis, and its interconnected levels of description, musical audio-mining is a very complex research domain that involves efforts from musicology, signal processing, and statistical modeling. This paper gives a general critical overview of the state-of-the-art followed by a discussion of musical audio-mining issues which are related to bottom-up processing (feature extraction), top-down processing (taxonomies and knowledge-driven processing), similarity matching, and user analysis and profiling.

## **Friday, August 30**

**9:00-9:10 Giovanni de Poli**, DEI young researchers project work and network experience

**9:10 - 9:20 Antonio Camurri**, DIST young researchers project work and network experience

**9:20 - 9:30 Leonello Tarabella**, CNUCE young researchers project work and network experience

**9:30-9:50, Laura Hyland**, The human voice as a controller

Laura Hyland is currently investigating the use of the human voice as a real time controller for a sound synthesis engine.

This is a two-part project; the first part concerns the design and implementation of voice analysis algorithms which will extract useful information from the voice.

The second part concerns design and implementation of a sound synthesis engine which generates sounds that complement the timbre of the voice."

**9:50-10:10 Declan Murphy**, Gesture Capture for Manipulation of Music

This work focuses on capturing hand gestures, and their subsequent use in manipulating music. Gestures are captured by multiple digital video cameras, and later serve both as a means of constructing/re-arranging musical structure for composition, and as a means of real-time control of musical parameters for performance.

The approach pursued started out with the Handel system from cART, CNR, Pisa, and considered other approaches in the literature. An

accurate physical model is constructed beforehand. For each video frame, the outline of the hand and edges from knuckles and overlapping fingers are extracted. Finally these processed edges are reconciled with the model by considering their inverse projection.

**10:10-10:20 Richard Kronland-Martinet**, Research at the CNRS-LMA

Richard Kronland-Martinet will briefly present the activity of the CNRS-LMA Mosart node and the collaborations with the Mosart partners.

**10:20-10:40 Snorre Farner**, The Digital Clarinet

As the piano player has the digital keyboard which gives a world of new ways to play music, we search to give for instance the clarinet player a digital clarinet with corresponding possibilities. As a first step towards this goal, it is important to understand the control of the clarinet and connect this to its sound.

The harmonic balance method is well suited for finding steady-state solutions of suggested models also for other wind and string instruments. It takes playing parameters as input and returns the steady-state spectrum, i.e. the characteristics of the sound, which in turn can be compared to real instrumental sound. The present work has been to make a computer program that makes these calculations, and thus a systematic study of suggested models can be performed.

**10:40 - 10:50 Xavier Serra**, UPF young researchers project work and network experience

**10:50-11:40 Coffe break**

**11:40-12:10 Invited Speaker: Cynthia M. Grund**, University of Southern Denmark  
Leader of the Danish network NTSMB

A BRIEF INTRODUCTION TO NTSMB

NTSMB is the abbreviation for \*Netværk for tværvideenskabelige studier af music og betydning\*, the English version of which is \*The Danish Network for Cross-Disciplinary Studies of Music and Meaning.\* NTSMB is an open research network. It was officially founded September 1, 2001 on a grant from the Danish Research Council for the Humanities (SHF), and currently has over 140 members. The purpose of NTSMB is to establish and maintain contact among geographically and institutionally dispersed

researchers within Denmark and the broader international research community who have interests relating to the study of music and meaning, and to enable ongoing discussions and on-line publication of research results.

The conviction that music is in some way meaningful raises a number of basic questions related to the concept of meaning, and to human cognition in general. Theoretical reflection over the concept of meaning has so far primarily come from within philosophy of language, or taken language as its point of departure. We are seeing, however, a growing interest in extending theoretical reflection over the concept of meaning to include issues regarding non-linguistic meaning in general and musical meaning in particular.

By disseminating knowledge through the network's web site and by organizing public conferences, seminars, and lectures, NTSMB is working to increase awareness of national and international research within the field of music and meaning. A related objective of the network is to promote cross-disciplinary research in issues regarding non-verbal meaning in general.

**12:10-12:40 Invited Speaker: Carol Krumhansl**, Department of Psychology, Cornell University, USA

The dynamics of musical experience

In a series of subsequent experiments, I have used the Nielsen (1983) methodology in a variety of studies on music cognition and emotion. The interplay between expectations and the sounded events is hypothesized to play a central role in creating musical tension and relaxation. The research to be presented investigates how this dynamic aspect of musical emotion relates to the cognition of musical structure. Consistent with the earlier results, tension covaries with both surface features (such as melodic contour, tone density, dynamics) as well as deeper levels of musical structure (harmony, tonality, form). Musical emotions change over time in intensity and quality, and these covary with psychophysiological measures, such as heart and respiration rate, skin conductance and temperature, and blood pressure. Emotion-specific patterns of emotion physiology are suggested. A schema of temporal organization is described that relates episodes of tension and relaxation to musical form and expressive aspects of musical performance, specifically tempo. In addition, some results suggest that the expression of emotion in music shares properties with emotion in speech and dance. Finally, a number of experiments have been conducted recently investigating the dynamics of tonality induction, and the music theoretic account of musical tension proposed by Lerdahl's (2001) Tonal Pitch Space model.

**12:40-14:00 Lunch Break**

**14:00-14:10 Kristoffer Jensen**, Proceedings of the Mozart Midterm Meeting

**14:10-14:20 Xavier Serra**, Task 1 progress and networking remarks

**14:20-14:30 Giovanni de Poli**, Task 2 progress and networking remarks

**14:30-14:40 Jan Tro**, Task 3 progress and networking remarks

**14:40-14:50 Jens Arnsfang**, Task 4 progress and networking remarks

**14:50-15:00 Gerhard Widmer**, Task 5 progress and networking remarks

**15:00-15:30 Coffee break**

**15:30-15:40 Barry Eaglestone**, Task 6 progress and networking remarks

**15:40-15:50 Antonio Camurri**, Task 7 progress and networking remarks

**15:50-16:00 Jens Arnsfang**, Final MOSART Project remarks

**IHP Network HPRN-CT-2000-00115 MOSART**  
**Music Orchestration System in Algorithmic Research and Technology**

**MOSART Task 2:**  
**Timbre Modeling**

**Edited and compiled by**  
**Giovanni de Poli**

Deliverables D22.

Report on significant progress in experiments on **Timbre modeling**

**Table Of Content**

<b>Overview on timbre perception and modeling</b> Hanna Järveläinen, Giovanni De Poli	Page 10
<b>Automatic Classification of Musical Instrument Sounds</b> Perfecto Herrera-Boyer, Geoffroy Peeters, Shlomo Dubnov	Page 17
<b>Timbre Modeling and Analysis-Synthesis of Sounds</b> Kronland-Martinet, Richard. Guillemain, Philippe. Ystad, Sølvi.	Page 38
<b>The Timbre Model - Discrimination and Expression</b> Kristoffer Jensen	Page 49
<b>Directional patterns and recordings of musical instruments in auralizations</b> Felipe Otondo, Jens Holger Rindel, Claus Lynge Christensen	Page 65
<b>Toward an objective method for the timbre analysis of sound-objects in the electro-acoustic music repertoire</b> Sergio Canazza, Chiara Marini	Page 68
<b>Abstract musical timbre and physical modeling</b> Giovanni De Poli, Davide Rocchesso	Page 75
<b>Importance of phase in sound modeling of acoustic instruments</b> Tue Haste Andersen, Kristoffer Jensen	Page 96
<b>A Hybrid Re-Synthesis Model for Hammer-Strings Interaction of Piano Tones</b> J. Bensa, K. Jensen, and R. Kronland-Martinet	Page 105
<b>Harmbal: a program for calculating steady-state solutions to non-linear physical models of wind and string instruments</b> S. Farner	Page 115



## EVALUATION REPORT OF EXPERIMENTS WITH TIMBRE MODELING MOSART DELIVERABLE D22

*Hanna Järveläinen*

DEI – University of Padova  
hjarvela@cc.hut.fi

*Giovanni De Poli*

DEI – University of Padova  
depoli@unipd.dei.it

### ABSTRACT

This is an introduction to the perception and modelling of timbre, contributing to the MOSART deliverable d12: Definition of timbre modelling. Timbre is a complex multidimensional sensation, which is related to the spectral, temporal, and spectro-temporal attributes of sound. Simplified models of timbre perception are utilized in sound synthesis, automatic identification and classification of sounds, and the building of timbre spaces for different groups of sounds.

### 1. INTRODUCTION

The MOSART network (Music Orchestration Systems in Algorithmic Research and Technology) aims at understanding aspects of the use of computers in music analysis, digital music representation and computer assisted musical composition and performance. Specifically, the task T2, Musical Sound Understanding and Timbre Modelling, explores the perceptual dimensions of timbre perception and studies how this knowledge could be utilized in the analysis and synthesis of musical sounds. This report defines the main lines of timbre modelling and draws together the advances and applications in this field in the context of MOSART network. The activity in this task is connected with task T3, Control and Visualization of Instruments, specially regarding the control aspects of sound modelling and algorithms for virtual instruments, and task T6, Computer Music Composition Tools, specially on the musical use of timbre models for composition and as a composing metaphor. The report is organized on a first part on general issues and definition of timbre modelling with contribution and point of views from different partners; then a second part presents specific research aspects developed during the first 18 months of activity of Mosart network, related to timbre modelling.

### 2. DEFINITION OF TIMBRE MODELLING

Giving a definition of timbre modelling is a complicated task. The meaning of the term "timbre" in itself is somewhat unclear. The American National Standards Institute (ANSI) defines timbre as "... that attribute of auditory sensation in terms of which a listener can judge that two sounds, similarly presented and having the same loudness and pitch, are different". This obviously leaves lots of space for imagination, stating only that anything which is not loudness or pitch belongs under the title "timbre". Sometimes timbre is referred to as "tone color". Many previous, informal definitions agree that timbre is something which gives a musical instrument its identity, so that it may be distinguished from other instruments or instrument families. Another aspect of timbre is tone quality: even though a sound maintains its identity under varying

conditions, its quality may change in many ways. For instance, the voice of the same speaker sounds different heard acoustically from a short distance than heard over the telephone line.

The attributes of sound that make up a timbre are numerous and partly unknown. The desire to control timbre by a small number of parameters has led to a hunt for the most important dimensions that contribute to the identity of a sound. This brings us near to the idea of modelling: learning to understand a complex system through a simplification. The modelling of timbre is focused around two viewpoints: modelling of perception and sound modelling. The perceptual viewpoint aims at understanding the mechanisms of our hearing system that are responsible for the timbre sensation. The sound modelling viewpoint includes analysis, automatic identification and classification of sounds as well as sound synthesis.

The afore mentioned identity and quality aspects relate timbre to the sound source. This is important for the control of the timbre of synthesized sounds. Once the identity of a musical instrument, for instance, is created by a sound model, its timbre can be controlled by a limited number of parameters in the range typical of that instrument class. However, virtual synthetic sounds don't belong to any well-defined timbre class. For the analysis of the timbre of these sounds, new schemes should be developed to create the concept of sound object which is not related to any physical sound source.

Another aspect to be considered is the musical, or compositional one. This calls for a higher-level cognitional view of timbre as an information carrier. Timbre has a special meaning for musical sounds, produced both acoustically and electronically.

### 3. PERCEPTION OF TIMBRE

Timbre is described as anything which is not pitch, loudness or direction. It is the only one of the four perceptual characteristics of sound that cannot be described by numbers. The further definition of timbre has been subject to speculations. One of the frequently given is that timbre is the feature of sound that allows us to identify it and distinguish it from other sounds with the same pitch and loudness [1]. Along with identifying the sound source, timbre specifies its quality: for instance, a single speaker can create many different vocal timbres. The ambiguity of timbre as a concept creates many methodological problems for measuring timbre perception. For a presentation of these issues, see [2].

What results as timbre is a combination of several perceptual dimensions, whose number and quality are partly unknown. The main factors are obviously related to the spectral content. The number and organization of the spectral components decides whether a sound is perceived as tonal or harmonic, for instance. The spec-

tral envelope, i.e. the relative amplitudes of the components, is another important factor. Also dynamic (time-variant) attributes seem to be salient.

Along with spectral characteristics, it has been shown that also temporal events and modulations have a great effect on timbre. The attack transients of musical instrument sounds contribute to the characteristic timbre of each type of instrument so much that if they are absent, it is hard to recognize the instrument any more, see [3] for discussion. The same goes for instance for vibrato in singing voice synthesis. Even though the temporal events do not create a certain kind of timbre themselves, connected to the spectral characteristics they seem to reinforce or bring out the unique timbre of the sound source. It seems, however, that the attacks are mostly important for the identification of sound sources, while the attributes that are present throughout the sound help us judge the timbre in a more general way [3].

The uniqueness of the timbre of a certain type of instrument, for instance, leaves another open question in timbre perception. Even though the spectro-temporal characteristics vary as a function of pitch and loudness, the identity of the sound source is unchanged. In this sense, timbre should be considered in relation to a sound source, a feature of a sound object whose other features are pitch, loudness, and direction. This way the interaction between each of the features could be captured in source-specific models. This kind of object-based viewpoint should become attractive with the development of the structured representations of sound [4], [5], [6].

Along with the physical characteristics that cause a perceived timbre, we should consider the information-carrying capacity of timbre, i.e., the different aspects of timbre perception. The most important one was mentioned earlier, the uniqueness of timbre for different types of sound sources, which allows identifying the sound or placing it to a certain category. Changes in timbre can also carry musical information by means of affecting musical tension which seems to be related to roughness, a perceptual measure for the effects of beats [7], [8]. Another example are vowel sounds in speech. Timbral effects can also help us make conclusions about the shape of the surrounding space, such as a concert hall. These examples seem to engage also top-down processing, meaning that our brain is sensitive to the kind of spectral patterns that are common in our environment and can recognize those [9].

### 3.1. Timbre space

During the past 20 years, there has been remarkable interest to establish a kind of perceptual space for timbre, i.e., to find a small number of relevant perceptual dimensions of timbre. The features of timbre are presented in a low-dimensional space whose orthogonal dimensions represent the most prominent components. A timbre space is constructed by a perceptual study of the desired type of sounds, for instance musical instrument sounds. Listeners will judge the perceived similarity of each timbre against all others, and from the similarity matrix derived from the ratings, the euclidean distances between each of the timbres can be calculated and presented in what is called the timbre space. A short distance between two timbres corresponds to high similarity and a long distance to high dissimilarity. The method of measuring perceptual distances is called multidimensional scaling (MDS) [10], and it is an efficient way of reducing the dimensionality of the timbre features.

The results obtained by the multidimensional scaling technique differ to a certain extent at the number and quality of the

perceptual dimensions. For instance, for orchestral instruments a three-dimensional timbre space was found, whose axes were spectral energy, synchronization of the transients, and a temporal attribute related to the beginning of the sound [11]. Expanding this material by hybrid sounds brought more diverse findings. One of these timbre spaces included the dimensions brightness, steepness of the attack, and the offset between the rise of the high and the low frequency harmonics [12]. Later on, the spectral flux, i.e., the variation of the spectral content with time, was also found to be important [13], but this wasn't confirmed by a later study [14].

An important point is that the studies mentioned above used almost entirely string and wind instruments. When the set of sounds was extended to percussive instruments [15], it was found that although two or three common dimensions could be found related to the spectral centroid and rise time, the results were generally context-dependent. This shows that a unitary timbre space is hardly likely to be found even for the timbres of musical instruments.

For this reason the instrument sound description in the context of MPEG-7 [16] is based on separate timbre spaces for harmonic and percussive sounds [17]. A number of physical attributes, or signal descriptors, that best correlate with the similarity ratings and the qualitative axes of the timbre space, were derived according to the results of Krumhansl [13] and McAdams [14] for harmonic sounds: Log-attack-time, spectral centroid, spectral spread, spectral variation, and spectral deviation. For percussive sounds, three descriptors were found based on the work of Lakatos [15]. They are the Log-attack-time, temporal centroid and spectral centroid.

### 3.2. Quantization of timbre sensations

A draw-back of the MDS technique is that it is non-metric. However, the timbre sensation can also be described on many metric levels. Semantic bipolar rating scales are commonly used to acquire subjective descriptions of timbre. Typically, the same timbre is judged on a number of scales, for instance dull-brilliant, cold-warm, pure-rich, dull-sharp, full-empty, and compact-scattered. After an analysis, some of the scales may be found not to be independent, but correlated with others. Usually the timbre sensation can be described by a few most important scales [2]. The problem is that the semantic scales are hard to relate directly to certain physical attributes of the sound.

Between the semantic descriptions and numeric spectral measures there are a number of psychoacoustic indices, developed for quantitative description of sound quality and perceptual response to sounds. Measures like sharpness, roughness, or fluctuation strength have been derived by matching the results of listening experiments to the physical characteristics of the sound. The sharpness measure, for instance, is connected to the perceived loudness measure suggested by Zwicker [18], which integrates the energy over the critical bands. Rather than for describing all kinds of timbres, these measures are used in specific tasks to compose more complex indices for instance for sound quality in vehicles or consumer products.

The physical parameters that have been found reveal that many perceptual timbre features correspond to physical characteristics other than spectral envelope. Various temporal and spectro-temporal measures are used to characterize timbre. For instance, the attack time and the different decay times between harmonics contribute to the timbre of musical instrument sounds. Measures related to the interaural time and level differences (ITD and ILD) or the pro-

portion of reflections and direct sound, etc. are used to compose indices of concert hall acoustics, such as brightness, warmth, or envelopment.

An interesting development is the search for the representation of timbre in the primary auditory cortex (AI) [19]. It was found that some of the cells in the AI are specialized in extracting the spectro-temporal features, which in psychoacoustic studies have been found important to timbre perception, such as the local bandwidth and asymmetry of spectral peaks, and their onset and offset transition rates.

## 4. SOUND MODELLING

By sound modelling we wish to learn to understand the physical attributes of sounds that create certain kinds of timbre sensations. Practical applications include the identification and classification of sounds as well as sound synthesis. However, sound modelling is a wider concept than timbre modelling, covering all kinds of sounds and synthesis methods and not only natural instruments. An example is the FM synthesis of abstract sounds: even if we focus on traditional musical instruments, sound modelling can refer to any kind of sounds that a composer wants to create.

### 4.1. Synthesis of natural instruments

There is a strong synthesis-viewpoint in sound modelling. However, not all sound synthesis techniques aim at direct control of timbre. During the latest years, the techniques have been developed in two main categories: physical modelling and spectral modelling. Timbre modelling is mostly related to the latter. The former attributes a physical model to the sound source and emulates the generation and behavior of sound in real musical instruments [20], [21], [22]. The benefits of physical modelling include the easy generation of natural sounding timbres by evoking the model by the recorded impulse response of a real instrument. On the other hand, control and arbitrary transformation of the timbre is not feasible beyond the typical performance of the original instrument.

The spectral modelling technique is based on splitting the spectral representation of the sound into its deterministic (sinusoidal) and stochastic (noise) components [23], [24]. In this way the emulation of natural timbres is more elaborate, because the time-varying behavior of each partial has to be controlled individually. On the other hand, the method gains control over the spectral envelope, which is mostly responsible of the alterations in timbre.

The control of timbre while pitch and loudness are kept constant could be regarded as timbre modelling. In addition to the sound synthesis technique described above, the applications of timbre modelling include the automatic classification and identification of sounds, and definitions of the typical timbre spaces for different types of sounds.

However, a great question is, how to isolate timbre effects from pitch and even loudness effects. In natural sound sources, these are usually interconnected and related to the same physical attributes [1]. We should find a way to view sounds as objects with a set of characteristic timbre features, whose prominence at certain instants of time depend on the state of the sound source, such as loudness and pitch. This kind of approach seems especially suitable for musical instrument sounds [25], which form a group of fairly similar sound sources whose behavior is well understood.

#### 4.1.1. Timbre models of musical instruments

Musical instrument sounds have long been a specific field of interest for timbre modelling. This is maybe because timbre perception is closely related to musical sounds, and because the synthesis of musical instrument sounds is well-developed. There is thus a need for the classification, identification, and resynthesis of musical timbres. Another field of application is data reduction in sound synthesis.

The timbre models apply a variable degree of parametrization and perceptual knowledge. The least parametric technique in this sense is spectral modelling, which is mainly used for the resynthesis of musical instrument sounds [26]. The technique extracts at first the harmonics of an existing sound and then generates a set of time-varying sinusoids accordingly. To preserve all the remaining features of the original sound, the sinusoids are subtracted from it to create a residual noise signal. The sinusoids together with the noise represent the original sound. Whole instruments or instrument families can be modelled by collecting a database that includes the whole timbre space of the instrument type, and synthesizing sounds accordingly.

The synthesizer presented by Jehan [27] models and predicts the timbre of acoustic instruments based on a small number of perceptual features. The parameters pitch, loudness, and brightness are extracted from the audio stream of an instrument. This perceptual sound data is combined to a timbre model, based on the spectral decomposition of the audio signal. The timbre of the instrument or an entire instrument family contained in the spectral model, the perceptual control parameters extracted from a new stream of audio are used as input to the system. The output is a vector of spectral data in real time, used to generate the deterministic part of the synthesized signal by additive synthesis. The noise part is generated by a stochastic noise process.

The synthesizer thus predicts the spectral representation of the audio data by controlling a timbre model by perceptual parameters. Possible application areas include the synthesis of musical instruments, but also cross-synthesis and morphing are enabled. In cross-synthesis, the instrument model is driven with control data from another instrument, for instance a timbre model of the violin could be controlled by parameters extracted from the clarinet. A mixed timbre morphed between the violin and the clarinet could be generated by running two simultaneous predictions and combining the output components according to a chosen weight (morphing) parameter. Applications of this kind of system could also be found in transmission and compression of audio.

Jensen [28] proposes a method for the modeling and classification of timbre that is based on three different levels of parametrization. This model is also based on the spectral presentation, from which a great number of perceptual parameters are extracted that fully define the timbre of the instrument. The parameters are related to the spectral envelope (the amplitudes of the partials at one instant of time), time envelope (time-varying partial amplitudes), and irregularity, which can be either spectral or temporal. The timbre model is composed of timbre attributes which are derived from these basic parameters. For instance, the parameters analyzed from the spectral envelope include brightness, the odd/even coefficient, the tristimulus and the irregularity. The first two levels are High Level Attributes (HLA), in which each partial is parametrized individually, Minimum Definition Attributes (MDA), in which the evolution of the attributes is defined in relation to each of the harmonics. On the Instrument Definition Attributes level (IDA), the

attributes are collected for one instrument, with varying pitch, intensity, or other performance parameters.

Many temporal effects, for instance the attack transients and vibrato, are not considered in this model, which makes practical sound synthesis problematic. However, the high degree of parametrization makes the method more suitable for timbre classification and identification than the techniques that are more strictly related to spectral modeling. Moreover, the whole performance of an instrument, i.e., the different playing techniques related to musical interpretation, are presented in parametric form. For instance, playing in high or low tempo is found to affect the attack and release time parameters.

#### 4.2. Creation of new timbre classes

Each sound synthesis algorithm can be thought of as a computational model for the sound itself and for the timbre it generates. Though this observation may seem quite obvious, its meaning for sound synthesis is not so straightforward. As a matter of fact, modelling sounds is much more than just generating them, as a computational model can be used for representing and generating a whole class of sounds, depending on the choice of control parameters. In the same sense an algorithm can define a new timbre class. The idea of associating a class of sounds to a digital sound model is in complete accordance with the way we tend to classify natural musical instruments according to their sound generation mechanism. For example, strings and woodwinds are normally seen as timbre classes of acoustic instruments characterized by their sound generation mechanism.

It should be clear that the degree of compactness of a class of sounds is determined, on one hand, by the sensitivity of the digital model to parameter variations and, on the other hand, the amount of control that is necessary to obtain a certain desired sound. As an extreme example we may think of a situation in which a musician is required to generate sounds sample by sample, while the task of the computing equipment is just that of playing the samples. In this case the control signal is represented by the sound itself, therefore the class of sounds that can be produced is unlimited but the instrument is impossible for a musician to control and play. An opposite extremal situation is that in which the synthesis technique is actually the model of an acoustic musical instrument. In this case the class of sounds that can be produced is much more limited (it is characteristic of the mechanism that is being modelled by the algorithm), but the degree of difficulty involved in generating the control parameters is quite modest, as it corresponds to physical parameters that have an intuitive counterpart in the experience of the musician.

It can be noticed that the generality of the class of sounds associated to a sound synthesis algorithm is somehow in contrast with the "playability" of the algorithm itself. One should remember that the "playability" is of crucial importance for the success of a specific sound synthesis algorithm as, in order for a sound synthesis algorithm to be suitable for musical purposes, the musician needs an intuitive and easy access to its control parameters during both the sound design process and the performance. Such requirements often represents the reason why a certain synthesis technique is preferred to others.

#### 4.3. Sound models and music

From a mathematical viewpoint, the musical use of sound models opens some interesting issues: description of a class of models that

are suitable for the representation of musically-relevant acoustic phenomena; description of efficient and versatile algorithms that realize the models; mapping between meaningful acoustic and musical parameters and numerical parameters of the models; analysis of sound signals that produces estimates of model parameters and control signals; approximation and simplification of the models based on the perceptual relevance of their features; generalization of computational structures and models in order to enhance versatility.

It is important to notice, however, that the analysis and synthesis of acoustical instruments will produce means of resynthesizing the type of music which is based on pitch. This excludes electronic music almost totally. Since the sounds in electronic music don't necessarily originate from identifiable musical instruments, it is clear that the concept of sound objects is also more abstract. To analyze the timbre of these sounds is a vast task, firstly because there are no restrictions regarding the type of sound source, which results in a huge diversity of timbres, and secondly because the other attributes of the sound objects are ill-specified. For instance, should pitch be treated as a independent perceptual dimension, if there is no clear pitch sensation.

Another point is that, concentrating on the timbre of single isolated sounds, we ignore the effects of musical context. Articulation, i.e., the way that sequential sounds connect to each other and form musical passages, can affect timbre in the same sense as loudness or pitch. The difference is that sound attributes derived from articulation depend on a larger time scale and cannot be based on instantaneous spectral and temporal measures. In this sense modelling of the musical performance and the player lead sound modelling to a higher level towards music modelling.

### 5. CONTRIBUTED ARTICLES

Many aspects of timbre modeling are covered within the activity of MOSART network. The collection of contributed papers gives an overview to the ongoing research in the field. The first part of the report gives a general introduction to three sub-topics of timbre modeling: Extraction of timbre features from natural sounds, the concept of abstract timbre, and the importance of directional issues in timbre modeling and acoustic rendering.

#### 5.1. Extraction of timbre features

The relevant dimensions of the timbre of natural musical instruments are being studied both from a theoretical and practical point of view. The search for a low-dimensional description of timbre has lead to a number of psychoacoustic studies relating subjective ratings to physical attributes. Specific applications include automatic recognition and classification of musical instruments and accurate resynthesis of natural instrument sounds.

##### 5.1.1. Timbre recognition

Herrera et al. [29] present a review on automatic classification of musical instrument sounds. The strongest motivation for automatic sound classification is the need to label existing audio recordings or samples of isolated instrument sounds, or to locate a certain instrument in a mixture of musical sounds. The two approaches discussed in the paper include a perceptual method, where psychoacoustic similarity functions are used for timbre clustering and the search and retrieval of sounds. The taxonomic ap-

proach finds signal-related indices for labeling sounds on a culture- or user-dependent basis.

The descriptive features for both methods are either found directly from the audio signal, or through transformations such as the Fourier or the Wavelet transform, or by use of a signal model or an auditory model. The relevant features for the perceptual method are selected on basis of similarity ratings of different timbres given by listeners. Descriptors like spectral centroid, the logarithm of the attack time, and spectral irregularity have been found to correlate with the subjective ratings. The taxonomic approach uses more general spectral methods for sound classification, such as Linear Predictive Coding or Cepstral Coefficients with Mel scaling (MCC), or energy measures. More specialized measures are also used, such as brightness, inharmonicity, and spectral envelope. A common problem to both methods is the selection of a small number of relevant descriptors from a high-dimensional space. Multidimensional Scaling (MDS) is often utilized in the perceptual method for reducing dimensionality. The taxonomic approach uses Principal Component Analysis, Genetic Algorithms, or Artificial Neural Networks for the purpose. The paper by Herrera et al. gives examples related to each of these techniques.

### 5.1.2. Analysis-synthesis

The contribution of Kronland-Martinet et al. [30] gives another viewpoint to timbre modeling. While the objective in automatic classification is to find a set of signal-based parameters that allow placing a given audio signal to a class of (musical instrument) sounds, the aim of sound modeling is to produce a synthetic audio signal that sounds identical to the original (musical instrument) sound. The process of reconstructing a natural sound is called analysis-synthesis. The main concern is how to produce a relevant set of control parameters from the original sound. A somewhat different approach is needed for each synthesis technique. The paper presents several sound synthesis methods related to signal modeling, physical modeling, and hybrid model synthesis, and discusses the extraction and estimation of control parameters for each of these models.

### 5.1.3. Timbre attributes

Jensen's [31] contributed paper tackles the question of parameterization of synthesis models of musical instruments. More specifically, he describes the timbre model, which aims at resynthesis of musical sounds by controlling the time-varying amplitudes and frequencies of the harmonics by a perceptually relevant set of attributes. These include spectral envelope for modeling the brightness and resonances of the instrument body, frequency envelope for reproducing the correct pitch and amount of inharmonicity, amplitude envelope for the control of start, attack, sustain, release, and end segments, and the amplitude and frequency irregularity (shimmer and jitter).

Jensen's model also extends the concept of timbre from the identity of an instrument to features that are more related to expression and playing style, such as intensity, vibrato, tremolo, legato, and staccato. The model catches the behavior of each timbre attribute corresponding to different styles.

## 5.2. Abstract timbre

### 5.2.1. Timbre in electroacoustic music

The analysis-synthesis of natural instruments produces means of resynthesizing voiced or percussive sounds but excludes electronic music almost entirely. Producing imaginary timbres by electronic means and creating music based on other attributes than pitch has a history of more than 50 years, but until today the timbre of electronic musical sounds has not received much attention. A reason for this might be that not even the timbre of natural instruments has been deeply understood.

The work of Canazza and Marini [32] takes the first steps into this unexploited area. The starting conditions for a study of timbre in electronic music are perhaps more difficult than those for natural instruments, because many of the "instruments", scores and composer's notes from the early years of electroacoustic music have disappeared. The study of Canazza and Marini can be seen as analysis-synthesis or modeling of timbre in the historical electroacoustic music repertoire. Their aim is to define a methodology for reconstructing the composing process and considering it in relation to its time. The work consists of spectral analysis and resynthesis of a selection of timbres taken originally from old recordings.

### 5.2.2. Generation of abstract timbres

The work of Rocchesso and De Poli [33] discusses the discordance between the degree of abstraction and effective control of sound models. A highly abstract model is free from instrument classes and the physical behavior of a given sound source, thus having no restrictions regarding pitch range, timbre space, evolution of timbre over pitch and playing styles, or identity. On the other hand, such a model might be hard to control, because no relations would exist between synthesis parameters. What comes to good "playability", it involves simple ways of controlling the synthesis model in a musical context.

Physical models offer a realistic environment for effective control schemes, although research is still required in this field. Control of the synthesis model should also be reached from higher levels of abstraction, i.e., musical gesture. In addition, a future direction is the creation of abstract timbres on basis of the general computational structures. This way new timbre classes could be created with a relation to existing ones that are well-described mathematically and acoustically, optimizing between a high degree of abstraction and effective control.

## 5.3. Importance of direction

The papers by Otondo et al. [34] [35] brings directional issues into the discussion of exact resynthesis of sounds. The directional pattern of a musical instrument in a given space varies in time, and the technique that aims at (re-)creating a realistic directional listening experience is called auralization. The aim of the study is to model the instrument-room interaction and to optimize the presentation of the directional characteristics of musical instruments.

Phase is often not considered an important factor of perceptual quality of synthesized sounds. However, it is known to affect the localization of the sound source. A study is reported by Andersen [36], in which the importance of the phase information for the perception of natural musical sounds was investigated. Two results are described in the paper – the importance of phase in binaural recorded sounds and also in monaural conditions both during

the transient and steady part of the sound.

#### 5.4. Research and applications

The second section of the present report gives more concrete insight into the relevant research areas in the context of MOSART network. A number of papers are related to the physical modeling technique. The joint contribution of CNRS-LMA and DIKU by Bensa et al. [37] describes the analysis-synthesis of the piano sound by a hybrid model that consists of a waveguide-based resonant part and an excitation part produced by subtractive synthesis. Another work concerning the piano sound is reported in the paper by Avanzini et al. [38] The compilation of a complete real-time physical model of the piano is described from sound generation principles to sound radiation issues and other effects. The paper of Farner [39] introduces the Harmbal program to calculate the steady-state spectrum of dynamic nonlinear systems, applied on physical modeling of self-sustained musical instruments such as the clarinet.

Examples of the extraction of timbre features are given in three contributed papers. Automatic recognition of percussion instruments is reviewed by Gouyon and Herrera [40]. Novel applications of higher-level spectral models are presented in the paper by Amatriain et al. [41]. It is shown how a set of basic and more specific digital sound effects can be created based on a musically meaningful parameterization of the spectral model. The effects include pitch transposition, vibrato, harmonization, gender change, and pitch discretization. A progress report is given by Marentakis and Jensen [42] on the Timbre Engine, a software synthesizer based on the principles presented in a more general way in Jensen's paper in the first part of this report.

Timbre perception is considered in two papers. The work by Ottaviani et al. [43] goes deeper into acoustic rendering, presenting a study of perception of shape of 3-D resonators. The objective of the study is to create shape controlled resonator models that could be used for shape morphing. The question of how we perceive physical dimensions of sounding objects, such as size, shape, or material, is the starting point. The perception of synthesized musical sounds in relation to model parameterization is considered in the paper by Järveläinen [44]. The study aims at perceptual guidelines for parameter selection and quantization in sound modeling.

In the paper of Otondo and Rindel [45] a method is proposed for the directional representation of musical instruments in auralizations. When musical instruments are used as (virtual) sound sources in auralizations, the time-varying character of their directional patterns must be taken into account.

An example of modeling the higher level expressive musical parameters is given in the paper by Schoonderwaldt and Friberg [46]. A rule-based model is proposed for the control of vibrato in relation to expressivity. The model can be used to provide vibrato control for a violin synthesizer.

#### 6. REFERENCES

- [1] A. Houtsma, "Pitch and timbre: Definition, meaning, and use," *J. New Music Research*, vol. 26, pp. 104–115, 1997.
- [2] J. M. Hajda, R. A. Kendall, E. C. Carterette, and M. L. Hashberger, "Methodological issues in timbre research," in *Perception and Cognition of Music* (J. Delige and J. Sloboda, eds.), ch. 12, pp. 253–306, East Sussex, UK: Psychology Press, 1997.
- [3] P. Iverson and C. Krumhansl, "Isolating the dynamic attributes of musical timbre," *J. Acoust. Soc. Am.*, vol. 94, no. 5, pp. 2595–2563, 1993.
- [4] ISO/IEC, "ISO/IEC IS 14496-3 Information Technology – Coding of Audiovisual Objects, Part 3: Audio," 1999.
- [5] B. Vercoe, W. G. Gardner, and E. D. Scheirer, "Structured audio: Creation, transmission, and rendering of parametric sound representations," *Proc. IEEE*, vol. 86, no. 5, pp. 922–940, 1998.
- [6] E. D. Scheirer and J.-W. Yang, "Synthetic and SNHC audio in MPEG-4," *Signal Processing: Image Communication*, vol. 15, pp. 445–461, 2000.
- [7] D. Pressnitzer, S. McAdams, S. Winsberg, and J. Fineberg, "Roughness and musical tension of orchestral timbres," in *Proc. 4th International Conference on Music Perception and Cognition*, pp. 85–90, 1996.
- [8] D. Pressnitzer, S. McAdams, S. Winsberg, and J. Fineberg, "Perception of musical tension for non-tonal orchestral timbres and its relation to psychoacoustic roughness," *Perception and Psychophysics: Human Perception and Performance*, vol. 62, pp. 66–80, 2000.
- [9] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, Massachusetts: MIT Press, 1990.
- [10] J. B. Kruskal, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, pp. 115–129, 1964.
- [11] J. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1270–1277, 1977.
- [12] D. Wessel, "Timbre space as musical control structure," *Computer Music J.*, vol. 3, no. 2, pp. 45–52, 1979.
- [13] C. Krumhansl, "Why is musical timbre so hard to understand?," in *Structure and perception of electroacoustic sound and music* (S. Nielzenand and O. Olsson, eds.), Amsterdam: Elsevier, 1989.
- [14] S. McAdams, S. Winsberg, G. de Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes," *Psychological research*, vol. 58, pp. 177–192, 1995.
- [15] S. Lakatos, "A common perceptual space for harmonic and percussive timbres," *Perception and psychophysics*, vol. 62, pp. 1426–1439, 2000.
- [16] MPEG-7, "Overview of MPEG-7," [www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm](http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm).
- [17] G. Peeters, S. McAdams, and P. Herrera, "Instrument sound description in the context of MPEG-7," *Proc. Int. Computer Music Conf.*, Berlin, Germany, 2000.
- [18] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*. Springer Verlag, New York, 1990.
- [19] P. Ru and A. Shamma, "Representation of musical timbre in the auditory cortex," *J. New Music Research*, vol. 26, no. 2, 1997.
- [20] J. O. Smith, "Principles of digital waveguide models of musical instruments," in *Applications of Digital Signal Processing to Audio and Acoustics* (M. Kahrs and K. Brandenburg, eds.), ch. 10, pp. 417–466, Kluwer, 1998.

- [21] J. Smith, "Physical modeling using digital waveguides," *Computer Music J.*, vol. 16, no. 4, pp. 74–91, 1992.
- [22] M. Karjalainen, V. Välimäki, and T. Tolonen, "Plucked-string models: from karplus-strong algorithm to digital waveguides and beyond," *Computer Music J.*, vol. 22, no. 3, pp. 17–32, 1998.
- [23] X. Serra and J. Smith, "Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music J.*, vol. 14, no. 4, pp. 12–24, 1990.
- [24] T. Verma and T. Meng, "Extending spectral modeling synthesis with transient modeling synthesis," *Computer Music J.*, vol. 27, no. 2, pp. 47–59, 2000.
- [25] T. Tolonen, *Object-based sound source modeling*. PhD thesis, Helsinki University of Technology, 2000.
- [26] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing* (C. Roads, S. Pope, A. Piccilli, and G. D. Poli, eds.), Zwets & Zeitlinger Publishers, 1997.
- [27] T. Jehan and B. Schoner, "An audio-driven perceptually meaningful timbre synthesizer," in *Proc. Int. Computer Music Conf.*, (Havana, Cuba), 2001.
- [28] K. Jensen, "The timbre model," in *Proc. Workshop on current research directions in computer music*, (Barcelona, Spain), pp. 174–186, 2001.
- [29] P. Herrera-Boyer, G. Peeters, and S. Dubnov, "Automatic classification of musical instrument sounds." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [30] R. Kronland-Martinet, P. Guillemin, and S. Ystad, "Timbre modeling and analysis-synthesis of sounds." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [31] R. Kronland-Martinet, P. Guillemin, and S. Ystad, "The timbre model - discrimination and expression." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [32] S. Canazza and C. Marini, "Toward an objective method for the timbre analysis of sound-objects in the electro-acoustic music repertoire." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [33] G. De Poli and D. Rocchesso, "Abstract musical timbre and physical modeling." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [34] F. Otondo, J. H. Rindel, C. Lyng, and C. Ørsted, "Directional patterns and recordings of musical instruments in auralizations." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [35] F. Otondo and J. H. Rindel, "Directional patterns of a clarinet in a room." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [36] T. H. Andersen and K. Jensen, "Importance of phase in sound modeling of acoustic instruments." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [37] J. Bensa, K. Jensen, and R. Kronland-Martinet, "A hybrid re-synthesis model for hammer-strings interaction of piano tones." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [38] F. Avanzini, B. Bank, G. Borin, G. De Poli, F. Fontana, and D. Rocchesso, "Musical instrument modeling: the case of the piano." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [39] S. Farner, "Harmbal: a program for calculating steady-state solutions to non-linear physical models of wind and string instruments." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [40] F. Gouyon and P. Herrera, "Exploration of techniques for automatic labeling of audio drum tracks instruments." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [41] X. Amatriain, J. Bonada, A. Loscos, and X. Serra, "Spectral modeling for higher-level sound transformation." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [42] G. Marentakis and K. Jensen, "The timbre engine: progress report." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [43] L. Ottaviani, F. Fontana, D. Rocchesso, and M. Rath, "Sounds from shape morphing of 3-d resonators." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [44] H. Järveläinen, "Applying perceptual knowledge to string instrument synthesis." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [45] F. Otondo and J. H. Rindel, "New method for the directional representation of musical instruments in auralizations." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.
- [46] E. Schoonderwaldt and A. Friberg, "Toward a rule-based model for violin vibrato." Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.

# Automatic Classification of Musical Instrument Sounds

Perfecto Herrera-Boyer (1)

Geoffroy Peeters (2)

Shlomo Dubnov (3)

(1) Universitat Pompeu Fabra, Pg. Circumval·lació 8, 08003 Barcelona, Spain  
[perfecto.herrera@iua.upf.es](mailto:perfecto.herrera@iua.upf.es) <http://www.iua.upf.es/mtg>

(2) IRCAM, 1Pl. Igor Stravinsky, 75004 Paris, France  
[peeters@ircam.fr](mailto:peeters@ircam.fr) <http://www.ircam.fr/equipes/analyse-synthese/peeters/>

(3) The Hebrew University, Edmond Safra Campus, Givat Ram, Jerusalem, Israel  
[dubnov@cs.huji.ac.il](mailto:dubnov@cs.huji.ac.il) <http://cse.cse.bgu.ac.il/~dubnov/>

## Abstract

We present an exhaustive review of research on automatic classification of sounds from musical instruments. Two different but complementary approaches are examined, the perceptual approach and the taxonomic approach. The former is targeted to derive perceptual similarity functions in order to use them for timbre clustering and for searching and retrieving sounds by timbral similarity. The latter is targeted to derive indexes for labeling sounds after culture- or user-biased taxonomies. We review the relevant features that have been used in the two areas and then we present and discuss different techniques for similarity-based clustering of sounds and for classification into pre-defined instrumental categories.

## 1 Introduction

The need for automatic classification of sounds arises in different contexts: biology (e.g. for identifying animals belonging to a given species, or for cataloguing communicative resources) (Fristrup & Watkins, 1995; Mills, 1995; Potter, Mellinger & Clark, 1994), medical diagnosis (e.g. for detecting abnormal conditions of vital organs) (Shiyong, Zehan, Fei, Li & Shouzong, 1998; Buller & Lutman, 1998; Schön, Puppe & Manteuffel, 2001), surveillance (e.g. for recognizing machine-failure conditions) (McLaughling, Owsley & Atlas, 1997), military operations (e.g. for detecting an enemy engine approaching or for weapon identification) (Gorman & Sejnowski, 1988; Antonic & Zagar, 2000; Dubnov & Tishby, 1997), and multimedia content description (e.g. for helping video scene classification or object detection) (Liu, Wang & Chen, 1998; Pfeiffer, Lienhart & Effelsberg, 1998). Speech, sound effects, and music are the three main sonic categories that are combined in multimedia databases. Describing multimedia sound therefore means describing each one of those categories. In the case of speech, the main description concerns speaker identification and speech transcription. Describing sound effects means determining the apparent sound source, or clustering similar sounds even though they have been generated by different sources. In the case of music, description calls for deriving indexes in

order to locate melodic patterns, harmonic or rhythmic structures, musical instrument sets, usage of expressivity resources, etc. As we are not concerned here with discrimination between speech, music and sound effects, we recommend interested readers consult the work by Zhang and Kuo (1998b; 1999a). Provided that we are interested in a music-only stream of audio data, one of the most important description problems is the correct identification of the musical instruments present in the stream. This is a very difficult task that is far from being solved. The practical utility for musical instrument classification is twofold:

- First, to provide labels for monophonic recordings, for “sound samples” inside sample libraries, or for new patches created with a given synthesizer;
- Second, to provide indexes for locating the main instruments that are included in a musical mixture (for example, one might want to locate a saxophone “solo” in the middle of a song);

The first problem is easier than the second, and it seems clearly solvable given the current state of the art, as we will see later in this paper. The second is tougher, and it is not clear if research done on solving the first one may help.

Common sense dictates that a reasonable approach to the second problem would be the initial separation of the sounds corresponding to the different sound sources, followed by the segmentation<sup>i</sup> and



classification<sup>ii</sup> on those separated tracks. Techniques for source separation cannot yet provide satisfactory solutions although some promising approaches have been developed (Casey & Westner, 2001; Ellis, 1996; Bell & Sejnowski, 1995; Varga & Moore, 1990). As a consequence, research on classification has concentrated on working with isolated sounds under the assumption that separation and segmentation have been previously performed. This implies the use of a sound sample collection (usually isolated notes) consisting of different instrument families and classes. The general classification procedure can be described as follows:

- Lists of features are selected to describe the samples.
- Values for these features are computed.
- A learning algorithm that uses the selected features to discriminate between instrument families or classes is applied.
- The performance of the learning procedure is evaluated by classifying new sound samples (cross-validation).

Note that there is a very important tradeoff in endorsing this isolated-notes strategy: we gain simplicity and tractability, but we lose contextual and time-dependent cues that can be exploited as relevant features for classifying musical sounds in complex mixtures. It is also important to note that the implicit assumption that solutions for isolated sounds can be extrapolated to complex mixtures should not be taken for granted, as we will discuss in the final section. Another implicit assumption that should not be taken for granted is that the arbitrary taxonomy that we use is optimal or, at least, good for the task (see Kartomi

(1990)) for issues regarding arbitrary taxonomies of musical instruments).

An alternative approach to the whole problem is to shift focus from the traditional *transcription* concern to that of *description* or *understanding* (Scheirer, 2000). This is what some Computational Auditory Scene Analysis systems have addressed (Ellis, 1996; Kashino & Murase, 1997a). We will return to this distinction later but for the moment a clarifying practical example of this different focus can be provided with an “instrument browser” as the one depicted in figure 1. In order to develop this kind of application, we only need to detect the instrument *boundaries*. The boundaries can surround individual instruments or classes of instruments (Aucouturier & Sandler, 2001). For example, note how the “soprano singer” instrument has been drawn separately whereas the other instruments are grouped into classes. In Figure 1, the string section subsumes the phrases played by violins, violas and cellos. The goal of this approach is not to separate into distinct tracks each of the instrumental voices but, rather, to label their locations within the context of the musical work. Thus, the user, when clicking on one of the labels would not hear an isolated instrument; instead, the user would be taken to part of the piece where the desired instrument or instrument family can be clearly heard. Manipulating the source file to bring to the foreground the selected instrument(s) is a possible enhancement of this boundary-based approach. In order to develop that kind of application we *only* need to detect the instrument boundaries.

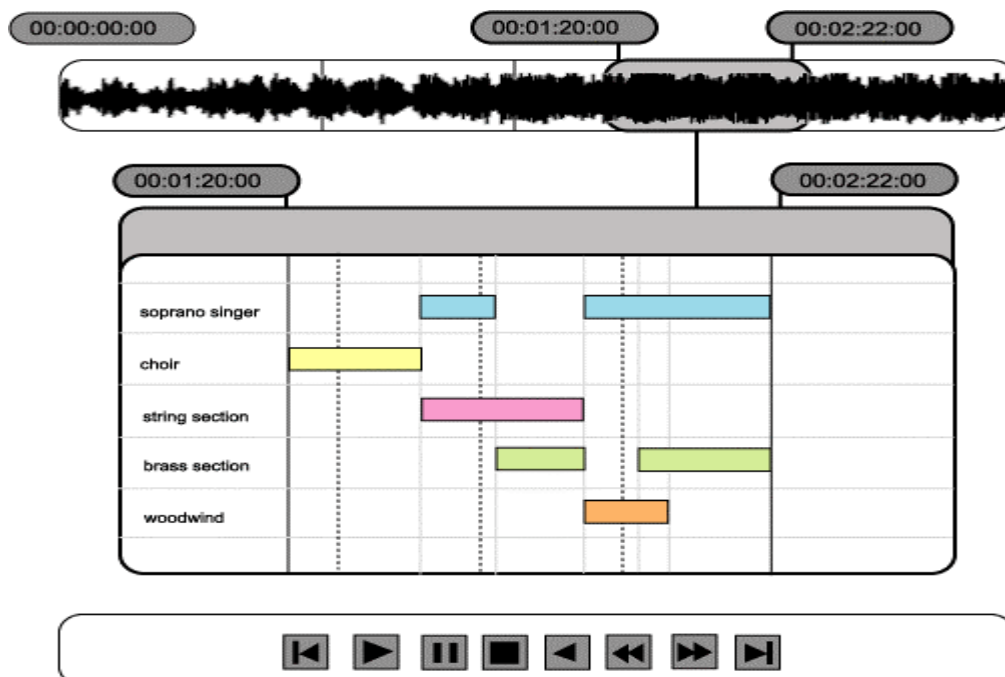


Figure 1. An imaginary instrument browser adapted from Smoliar and Wilcox ( Smoliar & Wilcox, 1997).

A very different type of classification arises when our target is not an instrument class but a cluster of sounds that can be judged to be perceptually similar. In that case, classification does not rely on culturally shared labels but on timbre similarity measures and distance functions derived from psychoacoustical studies (Grey, 1977; Krumhansl, 1989; McAdams, Winsberg, de Soete & Krimphoff, 1995; Lakatos, 2000). This type of perceptual classification or clustering is addressed to provide indexes for retrieving sounds by similarity, using a query by example strategy.

In the next sections we are going to review the different features (perceptual-based or taxonomic-based) that have been used for musical sound classification, and then the techniques that have been tested for classification and clustering of isolated sounds. We have purposely refrained from writing mathematical formulae in order to facilitate the basic understanding to casual readers. It is our hope that the comprehensive list of references at the end of the chapter will compensate this lack, and will help in finding the complementary technical information that a thorough comprehension requires.

## 2 Perceptual description versus taxonomic classification

Perceptual description departs from taxonomic classification in that it tries to find features that explain human perception of sounds, while the latter is interested in assigning to sounds some label from a previously established taxonomy (family of musical instruments, instruments names, sound effects category...). Therefore, the latter may be considered deterministic while the former is derived from experimental results using human subjects or artificial systems that simulate some of their perceptual processes.

Perception of sounds has been studied systematically since Helmholtz. It is now well accepted that sounds can be described in terms of their pitch, loudness, subjective duration, and something called "timbre". According to the ANSI definition (American National Standards Institute, 1973), timbre refers to the features that allow one to distinguish two sounds that are equal in pitch, loudness, and subjective duration. The underlying perceptual mechanisms are rather complex but they involve taking into account several perceptual dimensions at the same time in a possibly complex way. Timbre is thus a multi-dimensional sensation that relies among others, on spectral envelope, temporal envelope, and on variations of each of them. In order to understand better what the timbre

feature refers to, numerous experiments have been performed (Plomp, 1970; Plomp, 1976; Wedin & Goude, 1972; Wessel, 1979; Grey, 1977; Krumhansl, 1989; McAdams, Winsberg, de Soete, & Krimphoff, 1995; Lakatos, 2000).

In all of these experiments, people were asked for a dis-similarity judgment on pairs of sounds. Multidimensional Scaling (MDS) analysis<sup>iii</sup> was used to process the judgments, and to represent the sound stimuli in a low-dimensional space revealing the underlying attributes used by listeners when making the judgments. Researchers often refer to this low-dimensional representation as a "Timbre Space" (see Figure 2).

Grey (1977) performed one of the first experiments under this paradigm. Using 16 instrument sounds from the orchestra (string and wind instruments), he derived from MDS a timbre space with 3 dimensions corresponding to the main perceptual axes. A qualitative description of these axes allowed him to assign one dimension to the spectral energy distribution, another to the amount of synchronicity of the transients and amount of spectral fluctuation, and the last one to the temporal attribute of the beginning of the sound.

Wessel's experiments (Wessel, 1979) used the 16 sounds from Grey (1977) plus 8 hybrid sounds (in order to use non-existing sounds that avoided the class recognition effects and also for getting

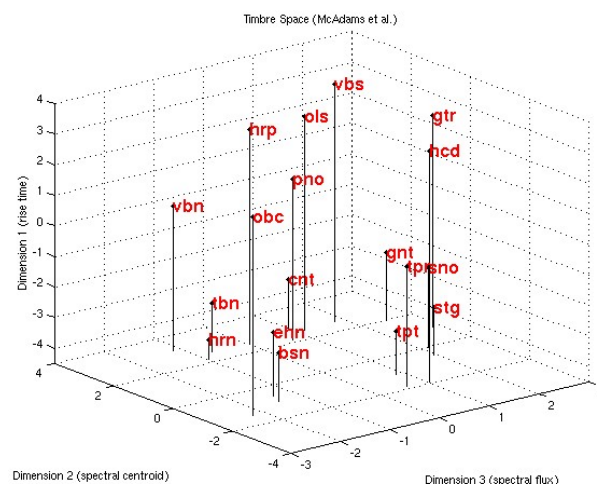


Figure 2. Timbre Space coming from McAdams et al. (1995) experiment. It was derived from dissimilarity ratings on 18 timbres by 88 subjects with specificities and five latent subject classes. Acoustic correlates of the three dimensions: rise time, spectra

intermediate "timbral steps" between sounds). This research yielded a 2-dimensional space with one dimension assigned to the "brightness" of the sustained part of the sound, and the other to the steepness of the attack and the offset between the beginnings of the high frequency harmonics to the low frequency ones.

Krumhansl (1989) used 21 FM-synthesis sounds from Wessel, Bristow & Settel (1987), mainly sustained harmonic sounds. She found the same results as Grey, but assigned the third dimension to something called “spectral flux” that was supposed to be related to the variations of the spectral content along time. McAdams et al. (1995), also used also these 21 FM-synthesis sounds in a new experiment and tested a new MDS technique that estimates the latent classes of subjects, instrument specificity values, and separate weights for each class. Compared to Krumhansl’s results, they confirmed the assignment of one dimension to the attack-time, another to the spectral centroid, but they did not confirm the “spectral flux” for the last dimension.

Lakatos’ experiment (2000) used 36 natural sounds from the McGill University sound library, both wind and string (17) and percussive (18) sounds. The goal of this experiment was to extend the timbre space to percussive and mixed percussive/sustained sounds. This yields a two dimensional space and a three dimensional space. The conclusion of the experiment is that, except for spectral centroid and rise time, additional perceptual dimensions exist but their precise acoustic correlates are context dependent and therefore less prominent.

An interesting practical application of this similarity-based research is that of setting up a given orchestration with some set of reference sound samples and then substituting some of them without radically changing the orchestration. Practical reasons for doing query-by-similarity of sound samples could include performance rights or copyrights issues, sample format compatibility, etc. Working examples of the timbre similarity approach are, for example, the *Soundfisher* system developed by Muscelfish<sup>iv</sup>, and the *Studio On Line* developed by IRCAM<sup>v</sup>. *Soundfisher*, recently incorporated as a plug-in into a commercial video-logger called Virage, is designed to perform the classification, indexing and search of sounds in general, though it can be used in a music context. The initial versions of *Soundfisher*<sup>vi</sup> (Keislar, Blum, Wheaton & Wold, 1995) did not yield an explicit class decision but, rather, generated a list of mathematically similar sounds. Some kind of class decision procedure, however, seems to have been recently implemented (Keislar, Blum, Wheaton & Wold, 1999). The *Soundfisher* system implicitly implements the assumption that what is mathematically similar can be also considered perceptually similar; in other words, that the computed features accurately represent perceptual dimensions, an assumption that contradicts most empirical studies. In contrast, *Studio On Line* computes similarity by using features that have been extracted under the paradigm of the above-

cited perceptual similarity psychoacoustical experiments. The interested reader can find in Peeters, McAdams & Herrera (2000) a recent validation of the psychoacoustical approach in the context of MPEG-7.

### 3. Relevant features for classification

#### 3.1 Types of features

The term *feature* denotes a quantity or a quality<sup>vii</sup> describing an object of the world. In the realm of signal processing and pattern recognition, objects are usually described by using vectors or lists of features. Features are also known as *attributes* or *descriptors*. Audio signal features are usually computed directly from the signal, or from the output yielded by transformations such as the Fast Fourier Transform or the Wavelet Transform. These audio signal features are usually computed every few milliseconds, for a very short segment of audio samples, in order to grasp their micro-temporal evolution. Macro-temporal evolution features can also be computed by using a longer segment of samples (e.g. attack time, vibrato rate...), or by summarizing micro-temporal values (e.g. averages, variances...).

A systematic taxonomy of features is outside the scope of this paper; nevertheless we could distinguish features at least according to four points of view:

1. The steadiness or dynamicity of the feature, i.e. the fact that the features represent a value extracted from the signal at a given time, or a parameter from a model of the signal behavior along time (mean, standard deviation, derivative or Markov model of a parameter);
2. The time extent of the description provided by the features: some description applies to only part of the object (e.g. description of the attack of the sound), whereas other apply to the whole signal (e.g. loudness);
3. The “abstractness”, i.e. what does the feature represent (e.g. cepstrum and linear prediction are two different representation and extraction techniques for representing spectral envelope, but probably the former one can be considered as more abstract than the latter)
4. The extraction process of the feature. According to this point of view, we could further distinguish:
  - Features that are directly computed on the waveform data as, for example, zero-crossing rate (the rate that the waveform changes from positive to negative values);
  - Features that are extracted after performing a transform of the signal (FFT, wavelet...) as, for example, spectral

centroid (the “gravity center” of the spectrum);

- Features that relate to a signal model, for example the sinusoidal model or the source/filter model;
- Features that try to mimic the output of the ear system (bark or erb bank filter output).

### 3.2. Relevant features for perceptual classification

For each of the “timbre” experiments, people have tried to *qualify* the dimensions of these timbre spaces, the perceptual axes, in terms of “brightness”, “attack”, etc. Only recently attempts have been made to *quantitatively* describe these perceptual axes, i.e. relate the perceptual axes to variables or descriptors directly derived from the signal (Grey, 1978; Krimphoff, McAdams & Winsberg, 1994; Misdariis, Smith, Pressnitzer, Susini & McAdams, 1998).

This quantitative description is done by finding the signal features that best explain the dis-similarity judgment. This is usually done using regression or multiple-regression between feature values and sound positions in the “timbre” space, and keeping only the features that yield the largest correlation. This makes the perceptual description framework different from taxonomic classification, since in the latter we’re not looking at features that “best explain” but at features that allow to “best discriminate” (between the considered classes).

In the Grey and Gordon (1978) experiment, only one dimension correlated significantly with a perceptual dimension of their “timbre” space: the spectral centroid. Krimphoff et al. (1994) worked with Krumhansl’s space (1989) trying to find the quantitative parameters corresponding to its qualitative features and found, as Grey did, significant correlations with the spectral centroid, but also with the logarithm of the attack time and what they called the “spectral irregularity”, which is the average departure of the spectral harmonic amplitudes from a global spectral envelope. Krumhansl (1989) had labelled this dimension as “spectral flux”. Misdariis, Smith, Pressnitzer, Susini & McAdams, (1998) combined results coming from the Krumhansl (1989) and McAdams et al. (1995) experiments. They found the same features as Krimphoff did plus a new one that explained one dimension of McAdams et al. (1995) experiment: spectral flux defined here as the average of the correlation between amplitude spectra in adjacent time windows.

Peeters et al. (2000) considered also the two above-cited experiments by Krumhansl and McAdams et al., called here “sustained harmonic sound space” as opposed to the “percussive sound space” coming from Lakatos (2000) experiment. Two methods

were used for the selection of the features, a “position” method, which tries to explain from the feature values the position of the sound in the timbre space, and a “distance” method, which tries to explain directly the perceived distance between sounds from a difference of feature values. From this study the following features, now part of the MPEG-7 standard, have been derived to describe the perceived similarity. For the “harmonic sustained sounds”: log-attack time, harmonic spectral centroid, harmonic spectral spread (the extent of the spectrum’s energy around the spectral centroid), harmonic spectral variation (the amount of variation of the spectrum energy distribution along time), and harmonic spectral deviation (the deviation of the spectrum harmonic from a global envelope). For the “percussive sounds”: log-attack time, temporal centroid (the temporal centre of gravity of the signal energy), and spectral centroid (the centre of gravity of the power spectrum of the whole sound).

Another approach is the one taken by the company Muscle Fish in the development of the *Soundfisher* system (Wold, Blum, Keislar & Wheaton, 1966). In this case the selected features are not derived from experiments but they constitute a set that is similar to the one discussed above: loudness (rms value in dB), pitch, brightness (spectral centroid), bandwidth (spread of the spectrum around the spectral centroid), harmonicity (amount of energy of the signal explained by a periodic signal model)... In order to capture the temporal trend of the features, it is proposed to store their average, variance and auto-correlation values along time.

### 3.3. Relevant features for taxonomic classification

Mel-Frequency Cepstrum Coefficients (hence MFCCs) are features that have proved useful for such speech processing tasks as, for example, speaker identification and speaker recognition (Rabiner & Juang, 1993). MFCCs are computed by taking the log of the power spectrum of a windowed signal, then non-linearly mapping the spectrum coefficients in a perceptually-oriented way (inspired by the Mel scale). This mapping is intended to emphasize perceptually meaningful frequencies. The Mel-weighted log-spectrum is then compacted into cepstral coefficients through the use of a discrete cosine transform. This transformation reduces the dimensionality of the representation without losing information (typically, the power spectrum may contain 256 values, whereas the MFCCs are usually less than 15). MFCCs provide a rather compact representation of the spectral envelope and are probably more musically meaningful than other common representations like Linear Predictive Coding coefficients or curve-fitting approximations to

spectrum. Despite these strengths, MFCCs by themselves can only convey information about static behavior and, as a consequence, temporal dynamics cannot be considered. Another important drawback is that MFCCs do not have an obvious direct interpretation, though they seem to be related (in an abstract way) with the resonances of instruments. Despite these shortcomings Marques (1999) used MFCCs in a broad series of classification studies. Eronen and Klapuri (2000) used Cepstral Coefficients (without the Mel scaling) and combined these features with a long list (up to 43) of complementary descriptors. Their list included, among others, centroid, rise and decay time, FM/AM rate and width, fundamental frequency and fundamental-variation-related features for onset and for the remainder of the note. In a more recent study, using a very large set of features (Eronen, 2001), the most important ones seemed to be the MFCCs, their standard deviations, and their deltas (differences between contiguous frames), the spectral centroid and related features, onset duration, and crest factor (specially for instrument family discrimination). There are ways, however, for adding temporal information into a MFCCS classification schema. For example, Cosi, De Poli & Prandoni (1994) created a Kohonen Feature Map<sup>viii</sup> (Kohonen, 1995) using both note durations and the feature coefficients. The network then clustered and mapped the right temporal sequence into a bi-dimensional space. As a result, sounds were clustered in a human perceptual-like way (i.e. not into taxonomic classes but into timbrally similar conglomerates). Brown (1999) used cepstral coefficients from constant-Q transforms instead of taking them after FFT-transforms; she also clustered feature vectors in a way that the resulting clusters seemed to be coding some temporal dynamics.

One of the most commonly used descriptors for musical, as well as non-musical, sound classification is energy. In (Kaminskyj & Materka, 1995), Root Mean Square (RMS) energy was used for classifying 4 different types of instruments with a neural network. In an additional, but apparently unfinished extension of this work (Kaminskyj & Voumard, 1996), the authors also included brightness, spectral onset asynchrony, harmonicity and MFCCs. In a more recent and comprehensive work (Kaminskyj, 2001) the main author used the RMS envelope, the Constant-Q frequency spectrum, and a set of spectral features derived from Principal Component Analysis (PCA from now on). PCA is commonly used to reduce dimensionality of complex data sets with a minimum loss of information. In PCA data is projected into abstract dimensions that are contributed with different –but partially related– variables. Then PCA calculates which projections, amongst all possible, are the best for representing the structure of data. The projections are chosen so that the maximum

variability of the data is represented using the smallest number of dimensions. In this specific research, the 177 spectral bins of the Constant-Q were reduced, after PCA, to 53 “abstract” features without any significant loss in discriminative power.

Martin and Kim (Martin & Kim, 1998) exemplified the idea of testing very long lists of features and then selecting only those shown to be most relevant for performing classifications. Martin and Kim worked with log-lag correlograms to better approximate the way our hearing system processes sonic information. They examined 31 features to classify a corpus of 14 orchestral wind and string instruments. They found the following features to be the most useful: vibrato and tremolo strength and frequency, onset harmonic skew (i.e., the time difference of the harmonics to arise in the attack portion), centroid related measures (e.g., average, variance, ratio along note segments, modulation), onset duration, and select pitch related measures (e.g., value, variance). The authors noted that the features they studied exhibited non-uniform influences, that is, some features were better at classifying some instruments and instrument families and not others. In other words, features could be both relevant and non-relevant depending on the context. The influence of non-relevant features degraded the classification success rates between 7% and 14%. This degradation is an important theoretical issue (Blum & Langley, 1997) that unfortunately has been overlooked by the majority of studies we have reviewed. It should be noted that there are some classification techniques that also provide some indication about the relevance of the involved features. This is the case with Discriminant Analysis (see section 4.2.3). Using this technique “backward” deletion and “forward” addition of features can be used in order to settle into a good (though sometimes suboptimal) set. Agostini, Longari, and Pollastri (2001) have used this method for reducing their original set of eighteen features to the eight ones that best separate the groups. The best features were: inharmonicity mean, centroid mean and standard deviation, harmonicity energy mean, zero-crossing rate, bandwidth mean and standard deviation, and standard deviation of harmonic skewness.

Spectral flatness is a feature that has been recently used in the context of MPEG-7 (Herre, Allamanche & Hellmuth, 2001) for robust retrieval of song archives. It is a “newcomer” in musical instrument classification but can be quite useful because it indicates how flat (i.e. “white-noisy”) the spectrum of a sound is. Our current work indicates that it can also be a good descriptor for percussive sound classification (Herrera, Yeterian & Gouyon, 2002).

Jensen and Arnspang (1999) used amplitude, brightness, tristimulus, amplitude of odd partials, irregularity of spectral envelope, shimmer and jitter measures, and inharmonicity, for studying the classification of 1500 sounds from 7 instruments.

Jensen (1999), using PCA, had earlier identified these features as the most relevant from an initial set of 20 and indicated 3 relevant dimensions that could summarize the most important features. He labeled these, in decreasing order of importance, “spectral envelope”, “(temporal) envelope”, and “noise”. Kashino and Murase (1997b) applied PCA to the instrument classification problem: 41 features were reduced to 11. PCA, in the context of sound classification, can be also found in the works of Sandell and Martens (1995), and Rochebois and Charbonneau, (1997). Less compact representations for temporal or spectral envelopes can be found in Fragoulis, Avaritsiotis, and Papaodysseus (1999), who used the slope of the first five partials, the time delay in the onset of these partials, and the high-frequency energy. Cemgil and Gürgen (1997)) also used a set of harmonics (the first twelve) as discriminative features in their neural networks study. Apart from PCA, another useful method for reducing the dimensions of the feature selection problem is the application of Genetic Algorithms (GAs). GAs are modeled on the processes that drive the evolution of gene populations (e.g., crossover, mutation, evaluation of fitness, and selection of the *best adapted*). GAs have a property called *implicit search*, which means that near-optimal combinations of genes can be found without explicitly evaluating all possible combinations. GAs have been used in other musical contexts (e.g., sound synthesis and music composition) but the only known application to sound classification has been that of Fujinaga, Moore, and Sullivan (1998) where GAs were used to discover the best feature set. From an initial set of 352 features, their GA determined that the centroid, fundamental frequency, energy, standard deviation and skewness of spectrum, and the amplitudes of the first two harmonics were the best features to achieve a successful classification rate. In a more recent work (Fujinaga & MacMillan, 2000), two additional significant features were reported: spectral irregularity and a modified version of tristimulus. Unfortunately, the selection of best features was heavily instrument-dependent. This problematic dependence has been also noted by other studies. The intensive study of feature selection performed by Kostek (1998) represents another interesting approach. Kostek thoroughly examined approximately a dozen features. Examined features include, for example, energy of fundamental and of sets of partials, brightness, odd/even partials ratio, tristimulus-like features, and time delays of partials with respect to the fundamental. Kostek also explored, in other studies, the use of features derived from Wavelet Transforms instead of FFT-derived features. She found that the latter provided slightly better results than the former. One of the more interesting aspects of Kostek’s work is her use of *rough sets* (Pawlak, 1982; Pawlak,

1991). Rough sets are a technique that was developed in the realm of knowledge-based discovery systems and data mining. Rough sets are implemented with the aim of classifying objects and then evaluating the relevance of the features used in the classification process. An elementary introduction to rough sets can be found in (Pawlak, 1998). We will return later with a fuller explication of rough sets.

Applications of the rough sets technique to different problems, including those of signal processing, can be found in (Czyzewski, 1998). Polkowski and Skowron (1998) present a thoughtful discussion of software tools implementing this kind of formalisms. Several studies by Kostek and her collaborators (Kostek, 1995; Kostek, 1998; Kostek, 1999; Kostek & Czyzewski, 2001), and by Wieczorkowska (1999b), used rough sets for reducing a large initial set of features for instrument classification. Wieczorkowska’s study provides the clearest example of set reduction using rough sets. She found that a starting set of sixty-two spectral and temporal features describing attack, steady state, and release of sounds could be further reduced to a set of sixteen features. Examples of the more significant features include: tristimulus, energy of 5<sup>th</sup>, 6<sup>th</sup> and 7<sup>th</sup> harmonics, energy of even partials, energy of odd partials, the most deviating of the lower partials, mean frequency deviation for low partials, brightness, and energy of high partials. Temporal differences between values of the same feature have been rarely used in the reviewed studies. *Soundfisher*, the commercial system mentioned earlier, incorporates temporal differences alongside such basic features as loudness, pitch, brightness, bandwidth, and MFCCs (Wold, Blum, Keislar & Wheaton, 1999). The Fujinaga or Eronen studies (cited above) have also incorporated temporal differences.

To summarize this section, there are two inter-related factors that influence the success of feature-based identification and classification tasks. First, one must determine, and then select, the most discriminatory features from a seemingly infinite number of candidates. Second, one must reduce the number of applied features in order to make the resultant calculations tractable. We might intuitively conclude that using more than fifteen or twenty features seems to be a non-optimal strategy for attempting automatic classification of musical instruments. In order to settle into a short feature list, reliable data reduction techniques should be used. PCA and some types of Discriminant Analysis (both explained below) are robust and relatively easy to compute. Other techniques such as Kohonen maps, Genetic Algorithms, Rough Sets, etc., might yield better results when appropriate parameters and data are selected, but are inherently more complex. It is also clear that

there are some features that are discriminative only for certain types of instruments, and that not only temporal and spectral features, but also their temporal evolution, should be considered.

## 4. Techniques for sound classification

### 4.2. Perceptual-based clustering and classification

Retrieving sounds from a database by directly selecting signal features as those cited in the previous section is not a friendly task. As a consequence, exploiting relationships between them and high-level descriptions such as class or property (roughness, brightness) is required. A different way of retrieving sounds is by providing examples that are similar to what we are searching for; this is known as “query by example”. A specific kind of “query by example” is the one based on similarity of perception of sounds, instead of being based on sound categories. Leaving pitch, loudness and duration apart, this points directly to the notion of timbre and therefore to “timbre similarity”.

Several authors have proposed a measure of timbre similarity that has been derived from psycho-acoustical experiments (see section 2). This measure allows one to approximate the average judgment of perceived similarity obtained from people’s dissimilarity judgments between pairs of sounds. In order to do that, features or combinations of them, are used, with a possible weighting, to position the sound into a multi-dimensional space. Giving two sounds, a measure of timbre similarity can be approximated. Therefore, for a given target sound, it is possible to find in a database the one that “sounds” the closest to the target.

Misdariis et al. (1998) derived such a similarity measure approximation from Krumhansl (1989) and McAdams et al. (1995) experiments. Its formulation uses four features: log-attack-time, spectral centroid, spectral irregularity and spectral flux. Use of the similarity measure proposed by Misdariis et al. (1998) can be found, for example, in the search engine of IRCAM’s “Studio On Line” sound database. Peeters et al. (2000) proposed a new approximation adding the new feature “spectral spread”. They also proposed an equivalent approximation for percussive sounds derived from the Lakatos (2000) experiment. This latter uses the log-attack time, the spectral centroid and the temporal centroid.

A still remaining problem concerns the applicability of such a timbre similarity measure for sounds belonging to different families (as for example comparing a sustained harmonic sound –

i.e. an oboe sound- with a percussive sound –i.e. a snare sound-). Current research is trying to construct a meta-timbre-space allowing such comparison between sounds belonging to different sound classes.

Another kind of approach is that of Feiten and Günzel (1994), Cosi, De Poli, and Lauzzana (1994), or Spevak and Polfreman (2000). Signal features used in these works try to take into account the properties of human perception: MFCCs, Loudness critical-band rate time patterns, Lyon’s cochlear model, Gamma filter banks, etc. These features are then used in order to construct, automatically, what is called a “physical timbre space”. The “physical timbre space” aims at being the equivalent to usual timbre spaces but derived from signal features instead of from dissimilarity judgments yielded by human subjects in experimental conditions.

A “physical timbre space” can be derived from signal features using various techniques: Hierarchical Clustering, Multi-Dimensional Scaling analysis (see section 2), Kohonen Feature Maps (a.k.a. Self Organizing Maps, see note 8), or Principal Component Analysis (see section 3.3).

Prandoni (1994) and De Poli and Prandoni (1997) used a combination of MFCCs, Self-Organized Maps, and PCA analysis. The authors applied this framework to the sounds of Wessel et al. (1987) and found that brightness and spectral slope are the features that best explain two of its “physical timbre space” axes. Prandoni (1994) used the barycentre of the representation of each sound family in a feature (MFCCs) space. Using MDS and Hierarchical Clustering analysis he found similar results than Grey did, and assigned the first two axes of his space to brightness and to something called “presence”, which is a measure of the energy inside the 800 Hz. region. In these two studies the obtained spaces were compared to usual timbre spaces coming from human experiments such as the above cited (sections 2 and 3.2).

In Feiten and Günzer (1994), and Spevak and Polfreman (2000)), the obtained spaces are used to make a temporal model of the sound evolution. The former authors define two sound feature maps (SFM). The first SFM is derived directly from a Kohonen Feature Map training using the MFCCs. This SFM, called the Steady State SFM, represents the steady parts of the sounds. Each sound is then represented by a trajectory between the states of the Steady State SFM. A Dynamic State SFM is then computed from these trajectories. The latter authors, on the other hand, make a comparison between different feature sets (Lyon’s cochlear model, Gamma Tone filterbank and MFCCs), considering their abilities to represent clear and separated trajectories in the SFM. They conclude that the best feature set is the Gamma Tone

filterbank combined with Meddis's inner hair cell model.

### 4.3. Taxonomic classification

In this section we are going to present different techniques that have been used for learning to classify isolated musical notes into instrument or music family categories. Although we have focused on the testing phase success rate as a way for evaluating them, we have to be cautious because other factors (number of instances used in the learning phase, number of instances used in the testing phase, testing procedure, number of classes to be learned, etc.) may have a large impact on the results.

#### K-Nearest Neighbors

The *K-Nearest Neighbours* (K-NN) algorithm is one of the most popular algorithms for instance-based learning. It first stores the feature vectors of all the training examples and then, for classifying a new instance, it finds a set of  $k$  nearest training examples in the feature space, and assigns the new example to the class that has more examples in the set. Traditionally, the Euclidean distance measure is used to determine similarity. Although it is an easy algorithm to implement, the K-NN technique has several significant drawbacks:

- As it is a lazy algorithm (Mitchell, 1997), it requires having all the training instances in memory in order to yield a decision for classifying a new instance.
- It does not provide a generalization

mechanism (because it is only based on local information).

- It is highly sensitive to irrelevant features that can dominate the distance metrics.
- It may require a significant computational load each time a new query is processed.

A k-NN algorithm classified 4 instruments with almost complete accuracy in Kaminskyj and Materka (1995), but the small size of the database (with restricted note range to one octave, although including different dynamics) was a drawback for taking this result as robust. In recent years Kaminskyj (2001) has reported hit rates of 82% for a database of 517 sounds and 19 instrumental categories. Some interesting features of this study are the use of PCA for reduction of data obtained after applying a Constant Q Transform and the use of a "reliability" estimation that can be extracted from confusion matrices.

Martin and Kim (1998) developed a classification system that used a k-NN on a database of 1023 sounds with 31 features extracted from cochleagrams (see also Martin (1999)). Their study included a hierarchical procedure consisting of:

- An initial discrimination of *pizzicati* from continuous notes.
- A discrimination between different "families" (e.g., sustained sounds further divided into strings, woodwind, and brass),
- A final classification of sounds into instrument categories.

When no hierarchy was used, Martin and Kim achieved a 87% classification success rate at the family level and a 61% rate at the instrument level.

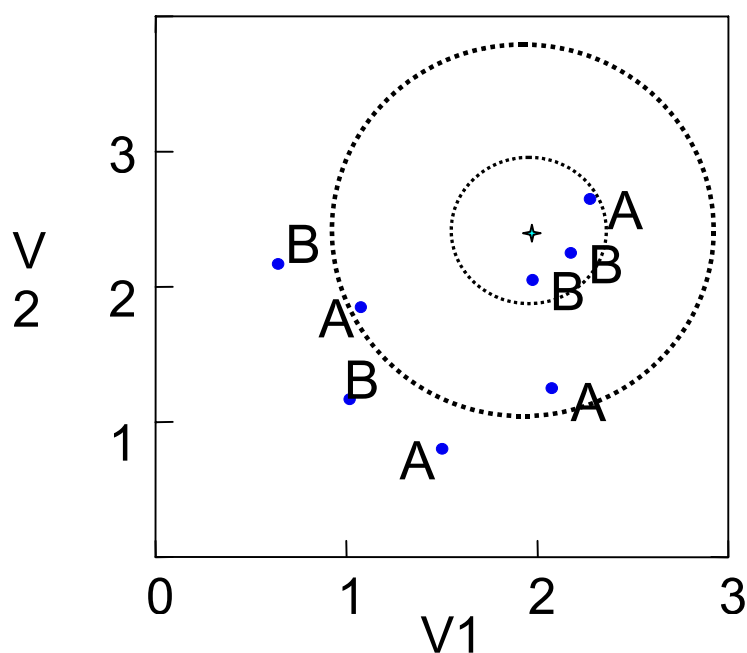


Figure 3. An illustration of the K-NN technique. The point marked with a star would be classified as belonging to category "B" when  $K=3$  (as two out of its 3 neighbours are from class "B"); but note that in case of using  $K=5$  classification would be "A".



Use of the hierarchical procedure increased the accuracy at the instrument level to 79% but it degraded the performance at the family level to 79%. In the case of not including the hierarchical procedure, performance figures were lower than the ones they obtained with a Bayesian classifier. Similar results (65% for 27 instrument classes; 77% for a two-level 6-element hierarchy) were reported by Agostini et al. (2001). In this report, the k-NN technique compared unfavorably against Discriminant Functions and also against Support Vector Machines.

Eronen and Klapuri (2000) used a combination of k-NN and a Gaussian classifier (which was only used for rough discrimination between pizzicati and sustained sounds) for classifying 1498 samples into specific instrumental families or specific instrument labels. Using a system architecture very similar to Martin and Kim's hierarchy—wherein sounds are first classified in broad categories and then the classification is refined inside that category—they reported success rates of 75% in individual instrument classification and 94% for family classification. They also reported a small accuracy improvement by only using the best features for each instrument and no hierarchy at all (80%). A quite surprising result is the extreme degradation of performance results (35%) that has been reported in a more recent paper (Eronen, 2001). The explanation may be found in several facts: they used a larger and more varied database (5286 sounds coming from different collections) and more restrictive cross-validation methods (the test phase used sounds that were completely excluded from the learning set).

A possible enhancement of the K-NN technique, which includes the weighting of each feature according to its particular relevance for the task, has been used by the Fujinaga team (Fujinaga et al., 1998; Fujinaga, 1998; Fraser & Fujinaga, 1999; Fujinaga & MacMillan, 2000). In a series of three experiments using over 1200 notes from 39 different instruments, the initial success rate of 50%, observed when only the spectral shape of steady-state notes was used, increased to 68% when tristimulus, attack position, and features of the dynamically changing spectrum envelope (i.e., the change rate of the centroid) were added. In their last paper, a real-time version of this system was reported.

The k-NN literature—including the works of such research leaders as Martin and Fujinaga—consistently reports accuracy rates around 80%. Provided that the feature selection has been optimized with genetic or other optimization techniques, one can thus interpret the 80% accuracy value as an estimation of the limitations of the K-NN algorithm. Therefore, more powerful techniques should be explored.

## Naive Bayesian Classifiers

A *Naive Bayesian Classifier* (NBC) incorporates a learning step in which the probabilities for the classes and the conditional probabilities for a given feature and a given class are estimated. Probability estimates for each of these are based on their frequencies as found in a collection of training data. The set of these estimates corresponds to the learned hypothesis, which is formed by simply counting the occurrences of various data combinations within the training examples. Each new instance is classified based upon the conditional probabilities calculated during the learning phase. This type of classifier is called *naive* because it assumes the independence of the features.

Brown (1999) used the NBC technique in conjunction with 18 Cepstral Coefficients computed after a constant Q transform. After clustering the feature vectors with a K-means algorithm, a Gaussian mixture model from their means and variances was built. This model was used to estimate the probabilities for a Bayesian classifier. It then classified 30 short sounds of oboe and sax with an accuracy rate of 85%. In a more recent paper (Brown, Houix & McAdams, 2001) she and her collaborators reported similar hit rates for four classes of instruments (oboe, sax, clarinet and flute); these good results were replicated for different types of descriptors (cepstral coefficients, bin-to-bin differences of the constant-Q spectrum, and autocorrelation coefficients).

Martin (1999) enhanced a similar Bayesian classifier with context-dependent feature selection procedures, rule-one-out category decisions, beam search, and Fisher discriminant analysis, to estimate the maximum *a priori* probabilities. In (Martin & Kim, 1998), performance of this system was better than that of a K-NN algorithm at the instrument level with a 71% accuracy rate and equivalent to it at the family level with 85% accuracy rate.

Kashino and his team (1995) have also used a Bayesian classifier in their CASA system. Their implementation is reported to be able to classify, and even separate, five different instruments: clarinet, flute, piano, trumpet and violin. Unfortunately, no specific performance data are provided in their paper.

## Discriminant Analysis

Classification using categories or labels that have been previously defined can be done with the help of *Discriminant Analysis* (DA), a technique that is related to multivariate analysis of variance (MANOVA) and multiple regression. DA attempts to minimize the ratio of within-class scatter to the

between-class scatter and builds a definite decision region between the classes. It provides linear, quadratic or logistic functions of the variables that "best" separate cases into two or more predefined groups. DA is also useful for determining which are the most discriminative features and the most similar/dissimilar groups. Surprisingly there have been very few studies using these techniques. Martin and Kim (1998)) made limited use of this method when they used a linear DA to estimate the mean and variance of the Gaussian distributions of each class to be fed into an enhanced naive Bayesian classifier.

More recently Agostini et al. (2001) have found that a set of quadratic discriminant functions outperformed even Support Vector Machines (93% versus 70% hit rates) in classifying 1007 tones from 27 musical instruments with a very small set of descriptors. In our laboratory we carried out, some time ago, an unpublished study with 120 sounds from 8 classes and 3 families in which we got a 75% accuracy using also quadratic linear discriminant functions in two steps (sounds were first assigned to a family, and then they were specifically classified). As the features we used were not optimized for instrument classification but for perceptual similarity classification, it would be reasonable to expect still better results when including other more task-specific features. In a more recent work (Herrera et al., 2002) that used a database of 464 drum sounds (kick, snare, hi-hat, tom, cymbals) and an initial set of more than thirty different features, we got hit rates higher than 94% with four canonical Discriminant functions<sup>x</sup> that combined 18 features comprising some MFCCs, attack and decay descriptors, and relative energies in some selected bands.

## Higher Order Statistics

When signals have Gaussian density distributions, we can describe them thoroughly with such second order measures as the autocorrelation function or the spectrum. In the case of noisy signals such as engine noises or sound effects, the variations in the spectral envelope do not allow a good signal characterisation and matching. A method to match signals using a variant of matched filter using polyspectral matching was presented in (Dubnov & Tishby, 1997), and it could be specifically useful for the classification of sounds from percussive instruments. There are some authors who claim that musical signals, because they have been generated through non-linear processes, do not fit a Gaussian distribution. In that case, using *higher order statistics* or polyspectra, as for example skewness of bispectrum and kurtosis of trispectrum, it is possible to capture all information that could be lost if using a simpler Gaussian model. With these

techniques, and using a Maximum Likelihood classifier, Dubnov, Tishby, and Cohen (1997) have showed that discrimination between 18 instruments from string, woodwind and brass families is possible. Unfortunately the detailed data that is presented there comes from a classification experiment that used machine and other types of non-instrumental sounds. Acoustic justification for differences in kurtosis among families of instruments was provided in (Dubnov & Rodet, 1997). The measure of kurtosis was shown to correspond to the phenomenon of phase coupling, which implies coherence in phase fluctuations among the partials.

## Binary trees

*Binary Trees*, in different formulations, are pervasively used for different machine learning and classification tasks. They are constructed top-down, beginning with the feature that seems to be the most informative one, that is, the one that maximally reduces entropy. Branches are then created from each one of the different values of this descriptor. In the case of non-binary valued descriptors, a procedure for dichotomic partitioning of the value range must be defined. The training examples are sorted to the appropriate descendant node, and the entire process is then repeated recursively using the examples of one of the descendant nodes, then with the other. Once the tree has been built, it can be pruned to avoid overfitting and to remove secondary features. Although building a binary tree is a recursive procedure, it is order of times faster than, for example, training a neural network.

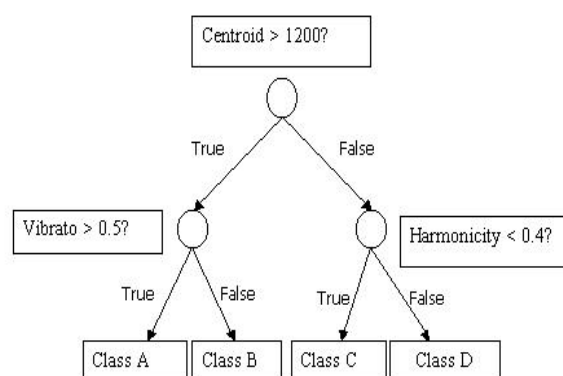


Figure 3: An imaginary binary tree for classification of sounds into 4 different classes.

Binary trees are best suited for approximating discrete-valued target functions but they can be adapted to real-valued features. Jensen and Arnspang's binary decision tree (1999) exemplifies this approach to instrument classification. In their

system, the trees are constructed by asking a large number of questions designed in each case to divide the data into two sets (e.g., “Is attack time longer than 60 ms?”). Goodness of split (e.g., average entropy) is calculated and the question that renders the best goodness is chosen. Once the tree has been built using the learning set, it can be used for classifying new sounds because each leaf corresponds to one specific class. The tree can also be used for making explicit rules about which features better discriminate one instrument from another. Unfortunately, detailed results regarding the classification of new sounds have not yet been published. Consult Jensen’s thesis (1999), however, for his discussion of log-likelihood classification functions.

Wieczorkowska (1999a) used a binary tree approach, called the *C4.5* algorithm (Quinlan, 1993), to classify a database of 18 classes and 62 features. Accuracy rates varied between 64% and 68% depending on the test procedure applied. In our above-mentioned drum sounds classification study (Herrera et al., 2002) we obtained slightly better figures (83% of hit rates) using the *C4.5* algorithm for classifying nine different classes of instruments.

A final example of a binary tree for audio classification, although not specifically tested with musical sounds, is that of Foote (1997). His tree-based approach uses MFCCs and supervised vector quantization to partition the feature space into a number of discrete regions. Each split decision in the tree involves comparing one element of the vector with a fixed threshold that is chosen to maximize the mutual information between the data and the associated human-applied class labels. Once the tree is built, it can be used as a classifier by computing histograms of frequencies of classes in each leaf of the tree; histograms are similarly generated for the test sounds then compared with tree-derived histograms.

## Artificial Neural Networks

An *Artificial Neural Network* (ANN) is an information processing structure that is composed of a large number of highly interconnected processing elements—called neurons or units—working in unison to solve specific problems. Neurons are grouped into layers (usually called *input*, *output*, and *hidden*) that can be interconnected through different connectivity patterns. An ANN learns complex mappings between *input* and *output* vectors by changing the weights that interconnect neurons. These changes may proceed either *supervised* or *unsupervised*. In the supervised case, a teaching instance is presented to the ANN, it is asked to generate an output, this out is then compared with an expected

“correct” output, and the weights are consequently changed in order to minimize future errors. In the unsupervised case, the weights “settle” into a pattern that represents the collection of input stimulus.

A very simple feedforward network with a backpropagation training algorithm was used in (Kaminskyj & Materka, 1995). The network (a system with 3 input units, 5 hidden units, and 4 output units) learned to classify sounds from 4 very different instruments—piano, marimba, accordion and guitar—with an accuracy rate as high as 97%. Slightly better results were obtained, however, using a simpler K-NN algorithm.

A three-way evaluative investigation involving a multilayer network, a time-delayed network, and a hybrid self-organizing network/radial basis function (see note 5) can be found in (Cemgil & Gürgen, 1997). Although very high success rates were found (e.g., 97% for the multilayer network, 100% for the time-delay network, and 94% for the self-organizing network) it should be noted that the experiments used only 40 sounds from 10 different classes with the pitch range limited to one octave.

Implementations of self-organizing maps (Kohonen, 1995) can be found in (Feiten & Günzel, 1994; Cosi, De Poli & Lauzzana, 1994; Cosi et al., 1994; Toiviainen et al., 1998). All these studies used some kind of human auditory pre-processing simulation to derive the features that were fed to the network. Each then built a map and evaluated its quality by comparing the network clustering results to those human-based sound similarity judgments (Grey, 1977; Wessel, 1979). From their maps and their comparisons they advance timbral spaces to be explored, or confirm/reject theoretical models that explain the data. We must note, however, that the classification we get from self-organizing maps has not traditionally been directly usable for instrument recognition, as the maps are not provided with any *a priori* label to be learned (i.e., no instrument names). Nevertheless, there are several promising mechanisms being explored for associating the output clusters to specific labels (e.g., the radial basis function used by Cemgil, (see above). The ARTMAP architecture (Carpenter, Grossberg & Reynolds, 1991) is another means to implement this strategy. ARTMAP has a very complex topology including a couple of associative memory subsystems and also an “attentional” subsystem. Fragoulis et al. (1999) successfully used an ARTMAP for the classification of 5 instruments with the help of only ten features: slopes of the first five partials, time delays of the first 4 partials relative to the fundamental, and high frequency energy. The small 2% error rate reported was attributed to neglecting different playing dynamics in the training phase.

Kostek’s (1999) is the most exhaustive study on instrument classification using neural networks.

Kostek's team has carried out several studies (Kostek & Krolikowski, 1997; Kostek & Czyzewski, 2000; Kostek & Czyzewski, 2001) on network architecture, training procedures, and number and type of features, although the number of classes to be classified has been always too small. They have used a feedforward NN with one hidden layer. Initially their classes were instruments with somewhat similar sounds: trombone, bass trombone, English horn and contrabassoon. In last papers more categories (double bass, cello, viola, violin, trumpet, flute, clarinet...) have been added to the tests. Accuracy rates higher than 90% were achieved for different sets of four classes, although the results varied depending on the types of training and descriptors used.

Some ANN architectures are capable of approximating any function. This attribute makes neural networks a good choice when the function to be learned is not known in advance, or it is suspected to be nonlinear. ANN's do have some important drawbacks, however, that must be considered before they are implemented: the computation time for the learning phase is very long, adjustment of parameters can be tedious and prohibitively time consuming, and data over-fitting can degrade their generalization capabilities. It is still an open question whether ANN's can outperform simpler classification approaches. They do, however, exhibit one strong attribute that recommends their use: once the learning phase is completed, the classification decision is very fast when compared to other popular methods such as k-NN.

## Support Vector Machines

SVMs are based on statistical learning theory (Vapnik, 1998). The basic training principle underlying SVMs is finding the optimal linear hyperplane such that the expected classification error for unseen test samples is minimized (i.e., they look for good generalization performance). According to the structural risk minimization inductive principle, a function that classifies the training data accurately, and which belongs to a set of functions with the lowest complexity, will generalize best regardless of the dimensionality of the input space. Based on this principle, a SVM uses a systematic approach to find a linear function with the lowest complexity. For linearly non-separable data, SVMs can (nonlinearly) map the input to a high dimensional feature space where a linear hyperplane can be found. This mapping is done by means of a so-called *kernel* function (denoted by  $\phi$  in Figure 4).

Although there is no guarantee that a linear solution will always exist in the high dimensional space, in practice it is quite feasible to construct a

working solution. In other words, it can be said that training a SVM is equivalent to solving a quadratic programming with linear constraints and as many variables as data points. Anyway, SVM present also some drawbacks: first, there is a risk of selecting a non-optimal kernel function; second, when there are more than two categories to classify, the usual way to proceed is to perform a

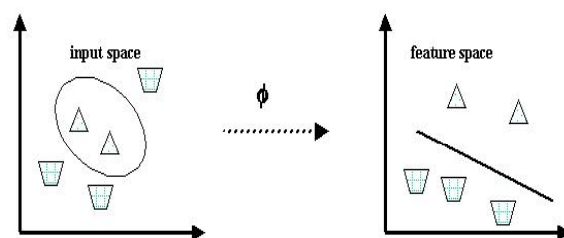


Figure 4. In SVM's the Kernel function  $f$  maps the input space (where discrimination of the two classes of instances is not easy to be defined) into a so-called feature space, where a linear boundary can be set between the two classes

concatenation of two-class learning procedures; and third, the procedure is computationally intensive.

Marques (1999) used an SVM for the classification of 8 solo instruments playing musical scores from well-known composers. The best accuracy rate was 70% using 16 MFCCs and 0.2 second sound segments. When she attempted classification on longer segments an improvement was observed (83%). There were, however, two instruments found to be very difficult to classify: trombone and harpsichord. Another noteworthy feature of this study was the use of truly independent sets for the learning and for the testing consisting mainly of "solo" phrases from commercial recordings.

Agostini et al. have reported quite surprising results (Agostini et al., 2001). In their study an SVM performed marginally better than (Linear) Canonical Discriminant functions and also better than k-NN's, but not nearly as good as a set of Quadratic Discriminant Functions (see section 4.2.3).

Some promising applications of SVM that are related to music classification but are not specific to music instrument labelling can be found in Li & Guo (2000), Whitman, Flake and Lawrence (2001), Moreno and Rifkin (2000), or Guo, Zhang, and Li (2001).

## Rough Sets

*Rough sets* are a novel technique for evaluating the relevance of the features used for description and classification. These are similar to, but should not

be confused with, *fuzzy sets*. In rough set theory, any set of similar or *indiscernible* objects is called an elementary set and forms a basic granule of knowledge about the universe; on the other hand, the set of *discernible* objects are considered rough (i.e., imprecise or vague). Vague concepts cannot be characterized in terms of information about their elements; however, they may be replaced by two precise concepts, respectively called the *lower approximation* and the *upper approximation* of the vague concept (see figure 5 for a graphical illustration of these ideas). The lower approximation consists of all objects that surely belong to the concept whereas the upper approximation contains all objects that could possibly belong to the concept. The difference between both approximations is called the *boundary region* of the concept.

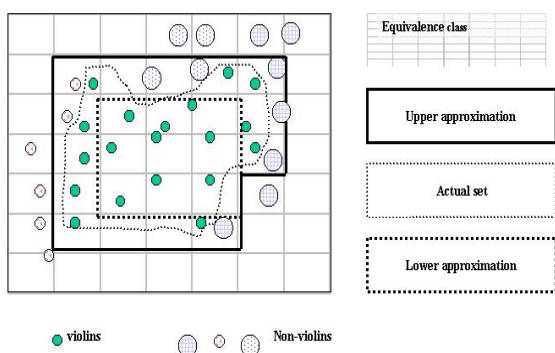


Figure 5. An illustration of rough sets concepts.

The assignment of an object to a set is made through a membership function that has a probabilistic flavour. Once data are conveniently organized into information tables, this technique is used to assess the degree of vagueness of the concepts and the interdependency of attributes; it therefore is useful for reducing complexity in the table without reducing the information it provides. Information tables regarding cases and features can be interpreted as conditional decision rules of the form **IF {feature  $x$ } is observed, THEN {is\_a\_Y\_object}**, and consequently they can be used as classifiers. When applied to instrument classification, (Kostek, 1998) reports accuracy rates higher than 80% for classification of the same 4 instruments mentioned in the ANN's section. While both useful and powerful, the use of rough sets does entail some significant costs. The need for feature value quantization is the principal and non-trivial cost associated with rough sets. Furthermore, the choice of quantization method can affect output results. In the context of instrument classification, different quantization methods have been discussed in (Kostek & Wieczorkowska, 1997), (Kostek, 1998), and (Wieczorkowska, 1999b). When compared to neural networks or fuzzy sets rules, rough sets are

computationally less expensive while at the same time yielding results similar to those obtained with the other two techniques.

## Hidden Markov Models

Hidden Markov Models (HMMs), as the name implies, contain two components: a set of hidden variables that can not be observed directly from the data, and a Markov property that is usually related to some dynamical behaviour of the hidden variables.

A HMM is a generative model that assumes that a sequence of measurements or observations is produced through another sequence of hidden states  $s_1, \dots, s_T$ , so that the model generates, in each state, a random measurement drawn from a different (finite or continuous) distribution. Thus, given a sequence of measurements and assuming a certain sequence of hidden states, the HMM model specifies a joint probability distribution.

$$p(s_1..s_T, x_1..x_T) = p(s_1)p(x_1 | s_1) \prod_{t=2}^T p(s_t | s_{t-1})p(x_t | s_t)$$

The HMM paradigm is used to solve three main tasks: classification, segmentation and learning. Learning is the first problem that needs to be solved in order to use a HMM model, unless the parameters of the model are externally specified. It means estimating the parameters of the models, usually iteratively done by the EM algorithm (Dempster, Laird & Rubin, 1977). The tasks of segmentation and classification are accomplished via forward-backward recursions, which propagate information across the Markov state transition graph. The segmentation problem means finding the most likely sequence of the hidden states given an observation  $x_1..x_T$ . Given several candidate HMM models that represent different acoustic sources (musical instruments in our case), the classification problem computes the probability that the observations came from these models. The model that gives the highest probability is chosen as the likely source of the observation.

HMMs have been used to address musical segmentation problems by several researchers (Raphael, 1999; Aucouturier & Sandler, 2001). These works dealt with segmentation of a sound into large-scale entities such as complete notes or sections of musical recordings, with the purpose of performing tasks such as score following or identification of texture changes in a musical piece.

Works that address the classification problem usually take a simpler view that discards the Markovian dynamics. Based on a work by Reynolds on speaker identification (Reynolds &

Rose, 1995), several researchers considered a Gaussian Mixture Model (GMM) for computer identification of musical instruments (Brown, 1999; Marques, 1999). GMMs consider a continuous probability density of the observation, and model it as a weighted sum of several Gaussian densities. The hidden parameters in GMM are the mean vector, covariance matrix and mixture weight of the component densities. Parameter estimation is performed using an EM procedure or k-means. Using a GMM in an eight-instrument classification task, Marques reported an overall error rate of 5% for 32 Gaussians with MCCs as features. Brown performed a two-instrument classification experiment where she compared machine classification results with human perception for a set oboe and saxophone sounds. She reported a lower error rate for the computer than humans for oboe samples and roughly the same for the sax samples. Eronen and Klapuri (2000) also compare a GMM classifier to other classifiers for various features.

In the HMM model for sound clips presented by Zhang and Kuo (1998a; 1999b) they use a continuous observation density probability distribution function (pdf) with various architectures of the Markov transition graphs. They also incorporate an explicit State Duration model (semi-markov model, (Rabiner, 1989) for modelling the possibility that  $d$  consecutive observations belong to the same state. Denote a complete parameter set of HMM as  $\lambda = (A, B, D, \pi)$ , with  $A$  for the transition probability,  $B$  for the GMM parameters,  $D$  for duration pdf parameters and  $\pi$  for initial state distribution. In this model, two types of information are represented in the HMM: timbre and rhythm. Each kind of timbre is modelled by a state, and rhythm information is denoted by transition and duration parameters. The authors arrive at a three step learning procedure that first uses GMM for estimating  $B$ , then  $A$  is calculated from statistics of the state transitions and eventually  $D$  is estimated state by state, assuming a Gaussian density for the durations. This simplified procedure is not a strict HMM learning process and it is used to simplify the computational load of the learning stage. They report over 80% accurate classification rate for 50 sound clips, with misclassifications reportedly happening with classes of perceptually similar sounds, such as applause, rain, river and windstorm. The timbre of sound is described primarily by the frequency energy distribution that is extracted from short time spectrum. In their experiments, Zhang and Kuo employ a rather naive feature set for description of the timbre, that consist of log amplitude from a 128-point FFT vector (thus obtaining a 65 dimensional feature vector), calculated at

approximately 9 msec intervals. Depending on the type of sound that is analyzed, a partial or complete HMM models is employed. The simplest ones are single state sounds, and sounds that omit duration and transition information. These are used when every timbral state in the model can occur anywhere in time and for any duration. Second model includes transition probabilities, but without durations. The third (complete case) includes sounds such as footsteps and clock ticks, which carry both transition and duration information. An improvement to the timbral description was recently suggested by Casey and Westner (2001). Instead of using magnitude FFT, they suggest reduced rank spectra as a feature set for HMM classifier. After FFT analysis, singular value decomposition (SVD) is used to estimate a new basis for the data and, by discarding basis vectors with low eigenvalues, a data-reduction step is performed. Then the results are passed to independent component analysis (ICA<sup>x</sup>), which imposes additional constraints on the output features. The resulting representation consists of a projection of a data into a lower-dimensional space with marginal distributions being approximately independent. They report a success rate of 92.65% for reduced-rank versus 60.61% for the full-rank spectra HMM classifier.

Another variant of Markov modelling, but this time using explicit (not hidden) observations with arbitrary length Markov modelling was used by Dubnov and Rodet (1998). In this work a universal classifier is constructed using a discrete set of features. The features were obtained by clustering (vector quantization of) cepstral and cepstral derivative coefficients. The motivation for this model is a universal sequence classification method of Ziv-Merhav (Ziv & Merhav, 1993) that performs matching of arbitrary sequences with no prior knowledge of the source statistics and having an asymptotic performance as good as any Markov or finite-state model. Two types of information are modelled in their work: timbre information and local sound dynamics, which are represented by cepstral and cepstral derivative features (observables). The long-term temporal behaviour is captured by modelling innovation statistics of the sequence, i.e. a probability to see a new symbol given the history of that sequence (for all possible length prefixes). By clever sampling of the sequence history, only most significant prefixes are used for prediction and clustering. The clustering method was tested on a set of 20 examples from 4 musical instruments, giving a 100% correct clustering.

## Conclusions

We have examined the techniques that have been used for classification of isolated sounds and the

features that have been found as more relevant for the task. We have also reviewed the perceptual features that account for clustering of sounds based on timbral similarity. Regarding the perceptual approach, we have presented empirical data for defining timbral spaces that are spanned by a small number of perceptual dimensions. These timbral spaces may help users of a music content-processing system to navigate through collections of sounds, to suggest perceptually based labels, and to perform groupings of sounds that capture similarity concepts. Regarding the taxonomic classification, we have discussed a variety of techniques and features that have provided different degrees of success when classifying isolated instrumental sounds. All of them show advantages and disadvantages that should be balanced according to the specifics of the classification task (database size, real-time constraints, learning phase complexity, etc.).

An approach yet to be tested is the combination of perceptual and taxonomic data in order to propose mixtures of perceptual and taxonomic labels (i.e. *bright snare-like tom* or *nasal violin-like flute*). It remains unclear, however, whether taxonomic classification techniques and features can be applied directly and successfully to the task of complex mixtures' *segmenting-by-instrument*. Additionally, because many of these techniques assume *a priori* isolation of input sounds, they would not accomplish the requirements outlined by Martin (1999) for real-world sound-source recognition systems. Anyway, we have been lately focusing in a special type of sound mixtures, so-called "drum loops", where some dual and ternary combinations of sounds can be found, and we have obtained very good classification results adopting the isolated sounds approach (Herrera, Yeterian, Gouyon, 2002). We have elsewhere (Herrera, Amatriain, Batlle & Serra, 2000) suggested some strategies for overcoming this limitation and for guiding some forthcoming research.

## Acknowledgments

The writing of this paper was partially made possible thank to funding received for the project CUIDADO from the European Community IST Program. The first author would like to express gratitude to Eloi Batlle and Xavier Amatriain for their collaboration as reviewers of preliminary drafts for some sections of this paper. He would also like to point out that large parts of this text have benefited from the editorial corrections and suggestions made by Stephen Downie, as editor of an alternative version (focused only on taxonomic classification) to be published elsewhere (Herrera, Amatriain, Batlle & Serra, 2002), and by other three anonymous reviewers. The second author

would also like to express gratitude to Stephen McAdams. Finally, thanks to Alex Sanjurjo for the graphical design of figure 1.

## References

- Agostini, G., Longari, M., & Pollastri, E. (2001). Musical instrument timbres classification with spectral features. IEEE Multimedia Signal Processing, IEEE.
- American National Standards Institute. (1973). American national psychoacoustical terminology. S3.20. New York: American Standards Association.
- Antonic, D., & Zagar, M. (2000). Method for determining classification significant features from acoustic signature of mine-like buried objects. 15th World Conference on Non-Destructive Testing, Rome .
- Aucouturier, J. J., & Sandler, M. (2001). Segmentation of musical signals using hidden markov models. AES 110th Convention.
- Bell, A. J., & Sejnowski, T. J. (1995). An information maximisation approach to blind separation and blind deconvolution. Neural Computation, 7, (6), 1129-1159.
- Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97, 245-271.
- Brown, J. C. (1999). Musical instrument identification using pattern recognition with cepstral coefficients as features. Journal of the Acoustical Society of America, 105, (3), 1933-1941.
- Brown, J. C., Houix, O., & McAdams, S. (2001). Feature dependence in the automatic identification of musical woodwind instruments. Journal of the Acoustical Society of America, 109, (3), 1064-1072.
- Buller, G. & Lutman, M. E. (1998). Automatic classification of transiently evoked otoacoustic emissions using an artificial neural network. British Journal of Audiology, 32, 235-247.
- Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1991). ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organising neural network. Neural Networks, 4, 565-588.
- Casey, M. A., & Westner, A. (2001). Separation of mixed audio sources by independent subspace analysis. Proceedings of the International Computer Music Conference, ICMA.
- Cemgil, A. T., & Grgeç, F. (1997). Classification of musical instrument sounds using neural networks. Proceedings of SIU97. Bodrum, Turkey.
- Cosi, P., De Poli, G., & Lauzzana, G. (1994). Auditory modelling and self-organizing neural networks for timbre classification. Journal of New Music Research, 21, (1), 71-98.

- Cosi, P., De Poli, G., & Prandoni, P. (1994). Timbre characterization with Mel-cepstrum and neural nets. Proceedings of the 1994 International Computer Music Conference, (pp. 42-45). San Francisco, CA: International Computer Music Association.
- Czyzewski, A. (1998). Soft processing of audio signals. In L. Polkowski & A. Skowron (Eds.), Rough Sets in Knowledge Discovery: 2: Applications, Case Studies and Software Systems. (pp. 147-165). Heidelberg: Physica Verlag.
- De Poli, G., & Prandoni, P. (1997). Sonological models for timbre characterization. Journal of New Music Research, 26, (2), 170-197.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, (34), 1-38.
- Dubnov, S., & Rodet, X. (1997). Statistical Modelling of Sound Aperiodicities. Proceedings of International Computer Music Conference, International Computer Music Association.
- Dubnov, S., & Rodet, X. (1998). Timbre recognition with combined stationary and temporal features. Proceedings of 1998 International Computer Music Conference. San Francisco, CA: International Computer Music Association.
- Dubnov, S., & Tishby, N. (1997). Analysis of sound textures in musical and machine sounds by means of higher order statistical features. Proceedings of the International Conference on Acoustics Speech and Signal Processing.
- Dubnov, S., Tishby, N., & Cohen, D. (1997). Polyspectra as measures of sound texture and timbre. Journal of New Music Research, 26, (4), 277-314.
- Ellis, D. P. W. (1996). Prediction-driven computational auditory scene analysis. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- Eronen, A. (2001). Comparison of features for musical instrument recognition. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'01), IEEE.
- Eronen, A., & Klapuri, A. (2000). Musical instrument recognition using cepstral coefficients and temporal features. Proceedings of the ICASSP. Istanbul, Turkey: IEEE.
- Feiten, B. and Günzel, S. (1994). Automatic indexing of a sound database using self-organizing neural nets. Computer Music Journal, 18, (3), 53-65-.
- Foote, J. T. (1997). A similarity measure for automatic audio classification. Proceedings of the AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora. Stanford, CA: AAAI Press.
- Fragoulis, D. K., Avaritsiotis, J. N., & Papaodysseus, C. N. (1999). Timbre recognition of single notes using an ARTMAP neural network. Proceedings of the 6th IEEE International Conference on Electronics, Circuits and Systems. Paphos, Cyprus.
- Fraser, A., & Fujinaga, I. (1999). Toward real-time recognition of acoustic musical instruments. Proceedings of the 1999 International Computer Music Conference, (pp. 175-177). San Francisco, CA: International Computer Music Association.
- Fristrup, K. M., & Watkins, W. A. . Marine animal sound classification. Journal of the Acoustical Society of America, 97, (5), 3369-3370.
- Fujinaga, I. (1998). Machine recognition of timbre using steady-state tone of acoustical musical instruments. Proceedings of the 1998 International Computer Music Conference, (pp. 207-210). San Francisco, CA: International Computer Music Association.
- Fujinaga, I., & MacMillan, K. (2000). Realtime recognition of orchestral instruments. Proceedings of the 2000 International Computer Music Conference, (pp. 141-143). San Francisco, CA: International Computer Music Association.
- Fujinaga, I., Moore, S., & Sullivan, D. S. (1998). Implementation of exemplar-based learning model for music cognition. Proceedings of the International Conference on Music Perception and Cognition, (pp. 171-179).
- Gorman, R. P., & Sejnowski, T. J. (1988). Learned classification of sonar targets using a massively parallel network. IEEE Transactions on Acoustics, Speech and Signal Processing, 36, (7), 1135- 1140.
- Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. Journal of the Acoustical Society of America, 61, (5), 1270-1277.
- Grey, J. M. (1978). Timbre Discrimination in Musical Patterns. Journal of the Acoustics Society of America, 64, (2), 467-472.
- Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. Journal of the Acoustical Society of America, 63, (5), 1493-1500.
- Guo, G. D., Zhang, H. J., & Li, S. Z. (2001). Boosting for Content-based Audio Classification and Retrieval: An Evaluation. IEEE International Conference on Multimedia and Expo.
- Herre, J., Allamanche, E., & Hellmuth, O. (2001). Robust matching of audio signals using spectral flatness features. 2001 IEEE Workshop on



- Applications of Signal Processing to Audio and Acoustics (WASPAA'01), IEEE.
- Herrera, P., Amatriain, X., Batlle, E., & Serra, X. (2000). Towards instrument segmentation for music content description: A critical review of instrument classification techniques. International Symposium on Music Information Retrieval.
- Herrera, P., Amatriain, X., Batlle, E., & Serra, X. (2002). A critical review of automatic musical instrument classification. In D. Byrd, J. S. Downie, & T. Crawford (Eds.), Recent Research in Music Information Retrieval: Audio, MIDI, and Score. Kluwer Academic Press.
- Herrera, P., Yeterian, A., & Gouyon, F. (2002). Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. International Conference on Music and Artificial Intelligence. Edinburgh, United Kingdom.
- Jensen, K. (1999). Timbre models of musical sounds. Unpublished doctoral dissertation, University of Copenhagen, Copenhagen, Denmark.
- Jensen, K., & Arnsfang, J. (1999). Binary decision tree classification of musical sounds. Proceedings of the 1999 International Computer Music Conference. San Francisco, CA: International Computer Music Association.
- Kaminskyj, I. (2001). Multi-feature Musical Instrument Sound Classifier. Australasian Computer Music Conference.
- Kaminskyj, I., & Materka, A. (1995). Automatic source identification of monophonic musical instrument sounds. Proceedings of the IEEE International Conference On Neural Networks, 1, (pp. 189-194).
- Kaminskyj, I., & Voumard, P. (1996). Enhanced automatic source identification of monophonic musical instrument sounds. Proceedings of the 1996 Australian New Zealand Conference on Intelligent Information Systems, (pp. 76-79).
- Kartomi, M. (1990). On Concepts and Classification of Musical Instruments. Chicago: The University of Chicago Press.
- Kashino, K., & Murase, H. (1997a). A music stream segregation system based on adaptive multi-agents. Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97), 2, (pp. 1126-1131).
- Kashino, K., & Murase, H. (1997b). Sound source identification for ensemble music based on the music stream extraction. Working Notes of the IJCAI-97 Computational Auditory Scene Analysis Workshop, (pp. 127-134).
- Kashino, K., Nakadai, K., Kinoshita, T., & Tanaka, H. (1995). Application of Bayesian probability network to music scene analysis. Proceedings of the 1995 International Joint Conference on Artificial Intelligence, (pp. 52-59). Montreal, Canada.
- Keislar, D., Blum, T., Wheaton, J., & Wold, E. (1995). Audio analysis for content-based retrieval. Proceedings of the 1995 International Computer Music Conference, (pp. 199-202). San Francisco, CA: International Computer Music Association.
- Keislar, D., Blum, T., Wheaton, J., & Wold, E. (1999). A content-ware sound browser. Proceedings of the 1999 International Computer Music Conference. San Francisco, CA: International Computer Music Association.
- Kohonen, T. (1995). Self-organizing maps. Berlin: Springer-Verlag.
- Kostek, B. (1995). Feature extraction methods for the intelligent processing of musical sounds. AES 100th convention, Audio Engineering Society.
- Kostek, B. (1998). Soft computing-based recognition of musical sounds. In L. Polkowski & A. Skowron (Eds.), Rough Sets in Knowledge Discovery. Heidelberg: Physica-Verlag.
- Kostek, B. (1999). Soft computing in acoustics: Applications of neural networks, fuzzy logic and rough sets to musical acoustics. Heidelberg: Physica Verlag.
- Kostek, B., & Czyzewski, A. (2000). An approach to the automatic classification of musical sounds. AES 108th convention. Paris: Audio Engineering Society.
- Kostek, B., & Czyzewski, A. (2001). Representing musical instrument sounds for their automatic classification. Journal of the Audio Engineering Society, 49, (9), 768-785.
- Kostek, B., & Krolikowski, R. (1997). Application of artificial neural networks to the recognition of musical sounds. Archives of Acoustics, 22, (1), 27-50.
- Kostek, B., & Wieczorkowska, A. (1997). Parametric representation of musical sounds. Archives of Acoustics, 22, (1), 3-26.
- Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caractérisation du timbre des sons complexes. II: Analyses acoustiques et quantification psychophysique. Journal de Physique, 4, 625-628.
- Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In S. Nielzenand & O. Olsson (Eds.), Structure and perception of electroacoustic sound and music (pp. 43-53). Amsterdam: Elsevier.
- Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. Perception and Psychophysics, Submitted for publication.
- Li, S. Z., & Guo, G. (2000). Content-based audio Classification and retrieval using SVM

- learning. Proceedings of the First IEEE Pacific-Rim Conference on Multimedia. Sidney, Australia: IEEE.
- Liu, Z., Wang, Y., & Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. Journal of VLSI Signal Processing, 20, (1/2), 61-79.
- Marques, J. (1999). An automatic annotation system for audio data containing music. Unpublished master's thesis, Massachusetts Institute of Technology, Cambridge, MA.
- Martin, K. D. (1999). Sound-source recognition: A theory and computational model. Unpublished doctoral dissertation, MIT, Cambridge, MA.
- Martin, K. D., & Kim, Y. E. (1998). Musical instrument identification: A pattern-recognition approach. Proceedings of the 136th meeting of the Acoustical Society of America.
- McAdams, S. & Windsberg, S. (in preparation). A meta-analysis of timbre space. I: Multidimensional scaling of group data with common dimensions, specificities, and latent subject classes.
- McAdams, S., Winsberg, S., de Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychological Research, 58, 177-192.
- McAdams, S., Susini, P., Krimphoff, J., Peeters, G., Rioux, V., Misdariis, N. & Smith, B. (in preparation). A meta-analysis of timbre space. II: Psychophysical quantification of common dimensions.
- McLaughling, J., Owsley, L. M. D., & Atlas, L. E. (1997). Advances in real-time monitoring of acoustic emissions. Proceedings of the SAE Aerospace Manufacturing Technology and Exposition, (pp. 291-297). Seattle, Washington.
- Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine learning, neural and statistical classification. Chichester: Ellis Horwood.
- Mills, H. Automatic detection and classification of nocturnal migrant bird calls. Journal of the Acoustical Society of America, 97, (5), 3370-3371.
- Misdariis, N., Smith, B., Pressnitzer, D., Susini, P., & McAdams, S. (1998). Validation and multidimensional distance model for perceptual dissimilarities among musical timbres. Proc. of Joint meeting of the 16th congress on ICA, 135th meeting of ASA.
- Mitchell, T. M. (1997). Machine learning. Boston, MA: McGraw-Hill.
- Moreno, P.J., & Rifkin, R. (2000). Using the Fisher Kernel method for web audio classification. Proceedings of the International Conference on Acoustics, Speech and Signal Processing.
- Pawlak, Z. (1982). Rough sets. Journal of Computer and Information Science, 11, (5), 341-356.
- Pawlak, Z. (1991). Rough sets: Theoretical aspects of reasoning about data. Dordrecht: Kluwer.
- Pawlak, Z. (1998). Rough set elements. In L. Polkowski & A. Skowron (Eds.), Rough Sets in Knowledge Discovery. Heidelberg: Physica-Verlag.
- Peeters, G., McAdams, S., & Herrera, P. (2000). Instrument sound description in the context of MPEG-7. Proceedings of the 2000 International Computer Music Conference. San Francisco, CA: International Computer Music Association.
- Pfeiffer, S. R., Lienhart, R., & Effelsberg, W. (1998). Scene determination based on video and audio features (TR-98-020). University of Mannheim, Mannheim, Germany.
- Plomp, R. (1970). Old and new data on tone perception (IZF1970-14).
- Plomp, R. (1976). Aspects of Tone Sensation: A Psychophysical Study. London: Academic Press.
- Polkowski, L., & Skowron, A. (1998). Rough sets in knowledge discovery. Heidelberg: Physica-Verlag.
- Potter, J. R., Mellinger, D. K., & Clark, C. W. (1994). Marine mammal call discrimination using artificial neural networks. Journal of the Acoustical Society of America, 96, (3), 1255-1262.
- Prandoni, P. (1994). An analysis-based timbre space.
- Quinlan, J. R. (1993). C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77, (pp. 257-286) IEEE.
- Rabiner, L. R., & Juang, B. H. (1993). Fundamentals of speech recognition. New York: Prentice-Hall.
- Raphael, C. (1999). Automatic segmentation of acoustic musical signals using hidden Markov models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21, (4), 360-370.
- Reynolds, D. A., and Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing, January 1995.
- Rochebois, T., & Charbonneau, G. (1997). Cross-synthesis using inverted principal harmonic sub-spaces. In M. Leman (Ed.), Music, Gestalt and Computing (pp. 221-244). Berlin: Springer.
- Sandell, G. J., & Martens, W. L. (1995). Perceptual evaluation of principal-component-based synthesis of musical timbres. Journal of the Acoustical Society of America, 43, (12), 1013-1028.

- Scheirer, E. D. (2000). Music-listening systems. Unpublished doctoral dissertation, Massachusetts Institute of Technology.
- Schön, P. C., Puppe, B., & Manteuffel, G. (2001). Linear prediction coding analysis and self-organizing feature map as tools to classify stress calls of domestic pigs (*Sus scrofa*). Journal of the Acoustical Society of America, 110, (3), 1425-1431.
- Shiyong, Z., Zehan, C., Fei, G., Li, F., & Shouzhong, X. (1998). The knowledge-based signal analysis for a heart sound information system. Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 20, (pp. 1622-1624). Piscataway, NJ: IEEE Computer Society Press.
- Smoliar, S. W., & Wilcox, L. D. (1997). Indexing the content of multimedia documents. Proceedings of the Second International Conference on Visual Information Systems, (pp. 53-60). San Diego, CA.
- Spevak, C., & Polfreman, R. (2000). Analyzing auditory representations for sound classification with self-organizing neural networks. COST G-6 Conference on Digital Audio Effects (DAFX-00).
- Toivainen, P., Tervaniemi, M., Louhivuori, J., Saher, M., Huottilainen, M., & Näätänen, R. (1998). Timbre similarity: Convergence of neural, behavioral, and computational approaches. Music Perception, 16, (2), 223-241.
- Vapnik, V. N. (1998). Statistical learning theory. New York: Wiley.
- Varga, A. P., & Moore, R. K. (1990). Hidden Markov model decomposition of speech and noise. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, (pp. 845-848).
- Wedin, L., & Goude, G. (1972). Dimension analysis of the perception of instrumental timbre. Scandinavian Journal of Psychology, (13), 228-240.
- Wessel, D. (1979). Timbre space as a musical control structure. Computer Music Journal, 3, (2), 45-52.
- Wessel, D., Bristow, D., & Settel, Z. (1987). Control of phrasing and articulation in synthesis. International Computer Music Conference, (pp. 108-116). San Francisco: International Computer Music Association.
- Whitman, B., Flake, G., Lawrence, S. (2001). Artist detection in music with Minnowmatch. Proceedings of the 2001 IEEE Workshop on Neural Networks for Signal Processing. 559-568.
- Wieczorkowska, A. (1999a). Classification of musical instrument sounds using decision trees. Proceedings of the 8th International Symposium on Sound Engineering and Mastering. ISSEM'99, (pp. 225-230). Gdansk, Poland.
- Wieczorkowska, A. (1999b). Rough sets as a tool for audio signal classification. In Z. W. Ras & A. Skowron (Eds.), Foundations of Intelligent Systems: Proceedings of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS-99) (pp. 367-375). Berlin: Springer-Verlag.
- Wish, M., & Carroll, J. D. (1982). Multidimensional scaling and its applications. In P. R. Krishnaiah & L. N. Kanal (Eds.), Handbook of statistics: 2. (pp. 317-345). Amsterdam: North-Holland.
- Wold, E., Blum, T., Keislar, D., & Wheaton, J. (1999). Classification, search and retrieval of audio. In B. Furth (Ed.), Handbook of Multimedia Computing (pp. 207-226). Boca Raton, FLA: CRC Press.
- Wold, E., Blum, T., Keislar, D., & Wheaton, J. (1996). Content-based classification, search and retrieval of audio. IEEE Multimedia, 3, 27-36.
- Zhang, T., & Jay Kuo, C.-C. (1998a). Content-based classification and retrieval of audio. SPIE's 43rd Annual Meeting - Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII, 3461, (pp. 432-443). San Diego, CA.
- Zhang, T., & Jay Kuo, C.-C. (1998b). Hierarchical system for content-based audio classification and retrieval. SPIE's Conference on Multimedia Storage and Archiving Systems III, 3527, (pp. 398-409). Boston: SPIE.
- Zhang, T., & Jay Kuo, C.-C. (1999a). Heuristic approach for generic audio data segmentation and annotation. ACM Multimedia Conference, (pp. 67-76). Orlando, FLA.
- Zhang, T., & Jay Kuo, C.-C. (1999b). Hierarchical classification of audio data for archiving and retrieving. IEEE International Conference On Acoustics, Speech, and Signal Processing, 6, 3004. Phoenix, AR.
- Ziv, J., & Merhav, N. (1993). A measure of relative entropy between individual sequences with application to universal classification. IEEE Transactions on Information Theory, (July), 1270-1279.

## NOTES

---

- <sup>i</sup> Segmentation can be defined as the process of breaking up an audio stream into temporal segments by means of applying a boundary detection criterion as, for example, texture, note, instrument, rhythm pattern, overall structure, etc. The same audio stream can be segmented in different ways by recurrently applying different criteria.
- <sup>ii</sup> Once an audio stream has been segmented, labels have to be attached to the segments. Two different families of algorithms can be used for learning labels: in the case we know in advance the labels to be used, *pattern recognition*, *discrimination*, or *supervised learning* techniques are the logical choice; when we do not know beforehand the labels and they will have to be inferred from the data, then the right choice is some *unsupervised* learning or *clustering* technique. See {Michie, Spiegelhalter, et al. 1994 109 /id} for more details.
- <sup>iii</sup> Multidimensional Scaling is a technique for discovering the number of underlying dimensions appropriate for a set of multidimensional data and for locating the observations in a low-dimensional space (Wish & Carroll, 1982).
- <sup>iv</sup> <http://www.musclefish.com>
- <sup>v</sup> <http://www.ircam.fr/produits/technologies/sol/index-e.html>
- <sup>vi</sup> <http://www.soundfisher.com>
- <sup>vii</sup> In this paper we will only consider the quantitative approach.
- <sup>viii</sup> A Kohonen or Self Organized Feature Map is a type of neural network that uses a single layer of interconnected units in order to learn a compact representation (i.e. with reduced features) of similar instances. It is very useful to cluster objects or instances that share some type of similarity because it preserves the inner space topology.
- <sup>ix</sup> A canonical Discriminant function uses standardized values and Mahalannobis distances instead of raw values and Euclidean distances.
- <sup>x</sup> Independent component analysis (ICA) tries to improve upon the more traditional Principal Component Analysis (PCA) method of feature extraction by performing an additional linear transformation (rotating and scaling) of the PCA features so as to obtain maximal statistical independence between the feature vectors. One must note that PCA arrives at uncorrelated features, which are independent only when the signal statistics are Gaussian. It is claimed by several researchers that both in vision and sound the more "natural" features are the ICA vectors. The motivation for this claim is that ICA features are better localized in time (or space, in the case of vision) [Bell and Sejnowsky 1996, 1997], and arrive at a more sparse representation of sound, that is, requiring less features, at every given instant of time (or space) in order to describe the signal. (One should

---

note, though, that the total number of features needed to describe the whole signal is not changed). A serious study of the utility of ICA for sound recognition still needs to be carried out, especially in view of the computational overhead that needs to be "paid" for ICA processing, vs. the improvement in recognition rates.

# Timbre Modeling and Analysis-Synthesis of Sounds

R. Kronland-Martinet<sup>(1)</sup>, Ph. Guillemain<sup>(1)</sup> & S. Ystad<sup>(1)(2)</sup>

- (1) CNRS – LMA 31 chemin J. Aiguier 13402 Marseille Cedex 20 France  
(2) N.T.N.U. (Norges Teknisk-Naturvitenskapelige Universitet)  
Department of Telecommunications N-7034 Trondheim Norway

## Abstract

Timbre modeling consists in designing synthesis methods to generate sounds under perceptual constraints. The process of Analysis-Synthesis consists in reconstructing a given natural sound using algorithmic techniques. This paper deals with the problem of designing methods to extract the parameters corresponding to timbre models so that the generated sound is similar to a given natural sound. We start by presenting an overview of sound and timbre models, including signal, physical and hybrid models. We then discuss the analysis problem and focus on time-frequency methods, which are well adapted to the analysis of musical sounds. We further deal with the analysis-synthesis problem for each class of sound modeling, including additive synthesis, non-linear synthesis and digital waveguide synthesis. As a conclusion, analysis-synthesis approaches based on hybrid models are described.

## 1 Scope of the article

Analysis-synthesis is a set of procedures to reconstruct a given natural sound. Different methods can be used, and the success of each method depends on their adaptive possibilities and the sound effect to be produced. The «direct» analysis-synthesis process consists in reconstructing a sound signal by inversion of an analysis procedure. This is a useful process that uses an invertible analysis method to get a representation of a sound and permits various sound transformations by acting on the parameters between the analysis and the synthesis process. Nevertheless, the result of such processes is not always intuitive since the modification of the analysis parameters is generally not a valid operation from a mathematical point of view. In this article we specially pay attention to the analysis-synthesis process associated to sound and timbre modeling. Here, the representations obtained from the analysis provide parameters corresponding to given synthesis models. This brings us to the concept of algorithmic sampler which consists in simulating natural sounds through a synthesis process which is well adapted to algorithmic and real time manipulations. The resynthesis and the transformation of natural sounds are then part of the same concept, allowing the morphing of natural sounds in a way which is similar to the modification of sounds using an algorithmic synthesizer.

The paper is organised as follows. We first present the most commonly used synthesis methods as timbre generators. Then analysis methods such as time-frequency and wavelet transforms are described, as well as algorithms for separating and characterizing spectral components. We conclude by showing how the analysis of real sounds can be used to estimate the synthesis parameters corresponding to different classes of sound models. Most of these techniques have been developed in our group «Modeling, Synthesis and Control of Audio and Musical Signals» (acronym S2M) in Marseille, France. This paper is an updated concentrate of the previously published paper [1].

## 2 Sound and timbre Synthesis

Digital synthesis uses methods of signal generation that can be divided into two classes:

- signal models aimed at reconstructing a perceived effect without being concerned with the specific source that made the sound.
- physical models aimed at simulating the behavior of existing or virtual sound sources.

### 2.1 Signal Model Synthesis

Signal models use a purely mathematical description of sounds. They are numerically easy to implement, and they guarantee a close relation between the synthesis parameters and the resulting sound. These methods are

similar to shaping and building structures from materials, and the three principal groups can be classified as follows

- additive synthesis
- subtractive synthesis
- non-linear or global synthesis

### **2.1.1 Additive Synthesis**

A complex sound can be constructed as a superposition of elementary sounds, generally sinusoidal signals modulated in amplitude and frequency. For periodic or quasi periodic sounds, these components have average frequencies that are multiples of one fundamental frequency and are called harmonics. The periodic structure leads to electronic organ sounds if one does not consider the micro variations that can be found through the amplitude and frequency modulation laws of the components of any real sound. These dynamic laws must therefore be very precise when one reproduces a real sound. The advantages of these synthesis methods are essentially the possibilities of intimate and dynamic modifications of the sound. Granular synthesis can be considered as a special kind of additive synthesis, since it also consists in summing up elementary signals (grains) localized in both the time and the frequency domains [2].

### **2.1.2 Subtractive Synthesis**

Like sculptor removes unwanted parts from his stone, a sound can be constructed by removing undesired components from an initial, complex sound such as a noise. This synthesis technique is closely linked to the theory of digital filtering [4] and can be related to some physical sound generation systems like for instance the speech signal [5], [6]. The advantage of this approach (if we omit the physical aspects which will be discussed when describing synthesis models by physical modeling) is the possibility of uncoupling the excitation source and the resonance system. The sound transformations related to these methods often use this property in order to make hybrid sounds or crossed synthesis of two different sounds by combining the excitation source of a sound and the resonant system of another [7][8].

### **2.1.3 Non-linear or Global Synthesis**

Like modeling different objects from a block of clay, a simple and "inert" signal can be dynamically modeled using global synthesis models. This method is non-linear since the operations on the signals are not simple additions and amplifications. The most well-known example of global synthesis is audio Frequency Modulation (FM) updated by John Chowning [9]. The advantage of this method is that it calls for very few parameters, and that a small number of operations can generate complex spectra. This simplifies the numerical implementation and the control. However, it is difficult to control the shaping of a sound by this method, since the timbre is related to the synthesis parameters in a non-linear way and the continuous modification of these parameters may give discontinuities in the sound. Other related methods have proved to be efficient for signal synthesis, such as the waveshaping technique [10].

## **2.2 Physical Model Synthesis**

Unlike signal models using a purely mathematical description of sounds, physical models describe the sound generation system through physical considerations. Such models can be constructed either from the equations describing the behavior of the waves propagating in the structure and their radiation in air, or from the behavior of the solution of the same equations. The first approach is costly in terms of calculations and is generally used only in connection with research work [11]. Synthesis by simulation of the solution of the propagation equation has led to the digital waveguide synthesis models [12], which have the advantage of being easy to construct with a behavior close to that of a real instrument. Thus such synthesis methods are well adapted to the modeling of acoustical instruments. Synthesis models related to a particular digital filter are known as digital waveguide models. They can be used to simulate many different systems, such as for instance the tube representing the resonant system in wind instruments [13].

## **2.3 Synthesis by Hybrid Models**

It is possible to combine physical and signal models into a so-called hybrid model [14], [15]. Such a model takes advantage of the positive aspects of both of the previous methods. As already mentioned the physical part of the model makes it possible to take into account physical characteristics of the sound producing system so that a physical interpretation can be linked to the model's parameters. The signal model part makes it possible to

simulate systems that normally would be too time demanding or complicated to be simulated by physical models. For a great number of musical instruments this implies that a signal model should model the excitation part while a physical model generally can model the resonator part. This means that although the physics of musical instruments often is too complicated for a purely physical model to be applied, physically meaningful parameters can be included in the model, since parts of it are constructed by physical modeling. This facilitates for instance the control of the model with an interface and makes it possible to make sounds from virtual instruments with physical exaggerated characteristics like for instance gigantic strings on a violin.

The coupling between the physical and signal models is of importance for the complexity of the hybrid model. In most cases the two models will interact reciprocally like in the flute case where a non-linear signal model can model the excitation, while a digital waveguide model can model the resonator. The interaction between these two models will in this case lead to a non-linear waveguide synthesis technique [16].

### 3 Sound Analysis

The analysis of natural sounds calls for several methods giving a description or a representation of pertinent physical and perceptual characteristics of the sound [17]. Even though the spectral content of a sound is often of great importance, the time course of its energy is at least as important. This can be shown by artificially modifying the attack of a percussive sound in order to make it "woolly", or by playing the sound backwards. In these cases the sound will be radically changed while the power spectral density remains the same. The time and frequency evolution of each partial component is also essential for the perceived characteristics of the sound. Effects like vibrato and tremolo can be analyzed from these evolutions (the time and frequency evolution). Such information also gives access to another perceptually important aspect of the sound, namely the different decay times of the partials of transient sounds such as sounds from plucked vibrating strings. To solve this general analysis problem of signals, a collection of methods called joint representations has been designed.

The analysis methods of signals can be divided into two principal classes: parametric methods and non-parametric methods. The parametric methods require a priori knowledge of the signal, and consist in adjusting the parameters of a model. The non-parametric models do not need any knowledge of the signal to be analysed, but they often require a large number of coefficients. We shall in this article focus on this last analysis class, since it generally corresponds to representations with physically and/or perceptually meaningful parameters.

#### 3.1 Spectral Analysis

The best known representation is the spectral representation obtained through the Fourier transform. The signal is in this case associated with a representation giving the energy distribution as a function of frequency. As mentioned earlier, this representation is not sufficient for characterizing the timbre and the dynamic aspects of a sound. Actually, even though all the information concerning the sound is contained in its Fourier transform, it is not obvious to give a sense to the phase of the transform. That is why only the modulus of the spectrum is generally considered, leading to an average energy information. The time is dramatically missing.

#### 3.2 Time-frequency and time-scale techniques

In what follows we describe the joint time-frequency representations considering both dynamic and frequency aspects. The time-frequency transformations distribute the total energy of the signal in a plane similar to a musical score in which one of the axes corresponds to the time and the other to the frequency. Such representations are to the sound what the musical scores are to the melodies. They can be obtained in two different ways depending on whether the analysis acts on the energy of the signal or on the signal itself. In the first case the methods are said to be non-linear (or bilinear), giving for instance representations from the so-called "Cohen's class". The best known example of transformations within this class is the Wigner-Ville distribution [18]. In the second case the representations are said to be linear, leading to the Fourier transform with sliding window, the Gabor transform, or the wavelet transform. The linear methods have, at least as far as sound signals are concerned, a great advantage compared to the non-linear methods since they always make the resynthesis of signals possible and ensure that no spurious terms cause confusion during the interpretation of the analysis. We will therefore focus on the linear time-frequency methods.

By decomposing the signal into a continuous sum of elementary functions having the same properties of localization both in time and in frequency, linear representations are obtained. These elementary functions correspond to the impulse response of bandpass filters. The central frequency of the analysis band is related to a frequency parameter for time-frequency transforms and to a scaling parameter for wavelet transforms. The choice of the elementary functions gives the shape of the filter.

##### 3.2.1 Gabor transform

The elementary functions of the Gabor transform, also called time-frequency atoms, are all generated from a mother function (window) translated in time and in frequency. The mother function is chosen to be well-localized in time and frequency and to have finite energy (for instance a Gaussian function) (figure 1).

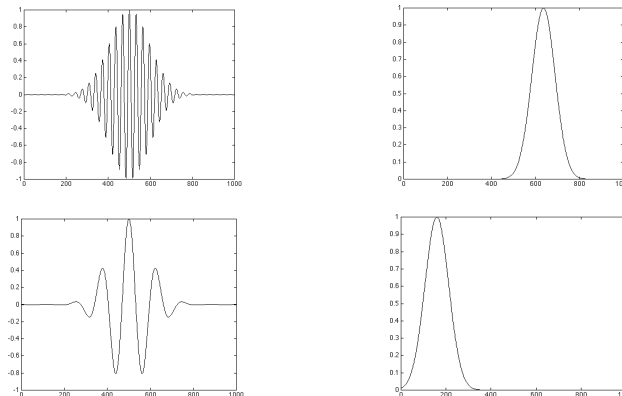


Figure 1: Two Gabor functions in the time domain (left), and their Fourier transform (right).

Each value of the transform in the time-frequency plane is obtained by comparing the signal to a time-frequency atom. This comparison is mathematically expressed by a scalar product. Each horizontal line of the Gabor transform then corresponds to a filtering of the signal by a band-pass filter centered at a given frequency with a shape that is constant as a function of frequency. The vertical lines correspond to the Fourier transform of a part of the signal, isolated by a window centered on a given time. The transform obtained this way is generally complex-valued, since the atoms themselves are complex-valued, giving two complementary images [19]. The first one is the modulus of the transform and corresponds to a classical spectrogram, the square of the modulus being interpreted as the energy distribution in the time-frequency plane. The second image corresponds to the phase of the transform and is generally less used, even though it contains a lot of information. This information mainly concerns the "oscillating part" of the signal (figures 2 and 3). Actually, the time derivative of the phase has the dimension of a frequency and leads to the frequency modulation law of the spectral components of the signal [20].

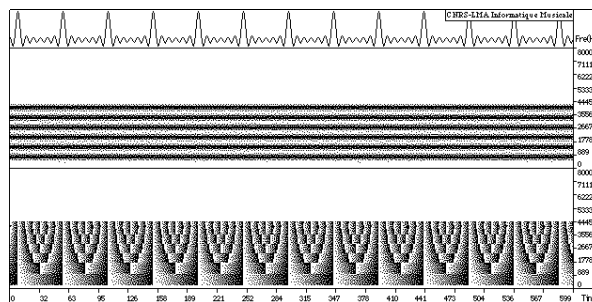


Figure 2

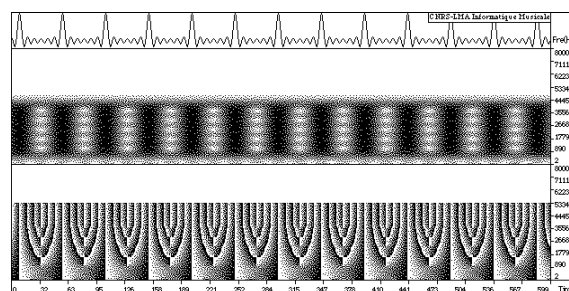


Figure 3: Gabor transform of the sum of six harmonic components analyzed with two different windows. The horizontal axis is time. The vertical axis is frequency. The upper picture is the modulus, the lower is the phase, represented modulo-2p. In the first picture (figure 2), the window is well localized in frequency, allowing the



resolution of each component. In the second picture (figure 3), the window is well localized with respect to time, leading to a bad separation of the components in the frequency domain, but showing impulses in time. In both pictures, the phase behaves similarly, showing the periodicity of each component.

### 3.1.2 Wavelet transform

The wavelet transform follows a principle close to that of the Gabor transform. Again the horizontal lines of the wavelet transform correspond to a filtering of the signal, but in this case the shape of the filter is independent of the scale while the bandwidth is inversely proportional to the scale. The analysis functions are all obtained from a mother wavelet by translation and change of scale (dilation) (figure 4).

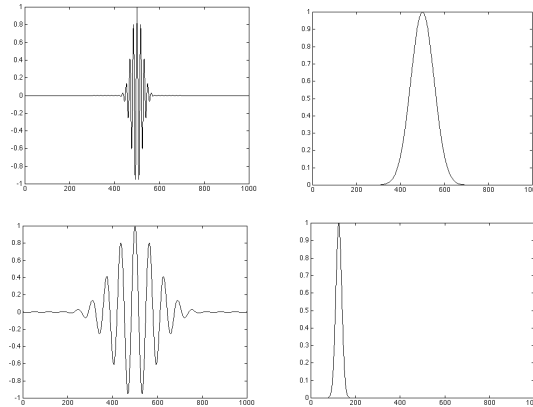


Figure 4: Two wavelets in the time domain (left), and their Fourier transform (right). In the wavelet representation, all the filters are obtained through dilation of a mother function in time, yielding a constant relative ( $\Delta w/w$ ) bandwidth analysis.

The mother wavelet is a function with finite energy and zero mean value. These "weak" conditions offer great freedom in the choice of the mother wavelet. One can for example imagine a decomposition of a speech signal in order to detect the word "bonjour" pronounced at different pitches and with different duration. By using a mother wavelet made of two wavelets separated for example by an octave, one can detect octave chords in a musical play [3]. This corresponds to a matched filtering at different scales. One important aspect of the wavelet transform is the localization. By acting on the dilation parameter, the analyzing function is automatically adapted to the size of the observed phenomena (figure 5). A high frequency phenomenon should be analyzed with a function that is well-localized in time, whereas a low-frequency phenomenon requires a function well-localized in frequency. This leads to an appropriate tool for the characterization of transient signals [20]. The particular geometry of the time-scale representation, where the dilation is represented according to a logarithmic scale (in fraction of octaves) enables the transform to be interpreted like a musical score associated to the analysed sound.

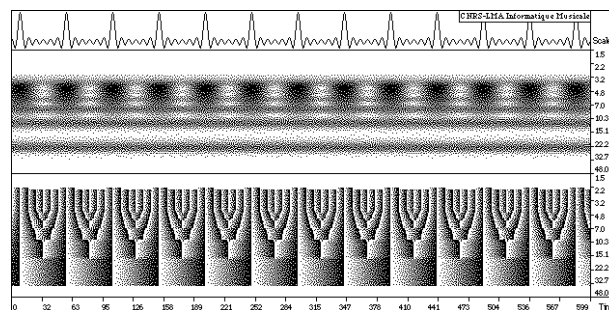


Figure 5: Wavelet transform of the same sum of six harmonic components. In contrast to pictures 2 and 3 obtained through the Gabor transform, the wavelet transform privileges the frequency accuracy at low frequencies (large scales) and the time accuracy at high frequencies (small scales).

## 4 Parameter extraction

The parameter extraction method makes use of the qualitative information given by the time-frequency and the time-scale transform in order to extract quantitative information from the signal. Even though the representations are not parametric, the character of the extracted information is generally determined by the supposed characteristics of the signal and by future applications. A useful representation for isolated musical instrument sounds is the additive model. It describes the sound as a sum of elementary components modulated in amplitude and in frequency, which is relevant from a physical and a perceptive point of view (figure 6).

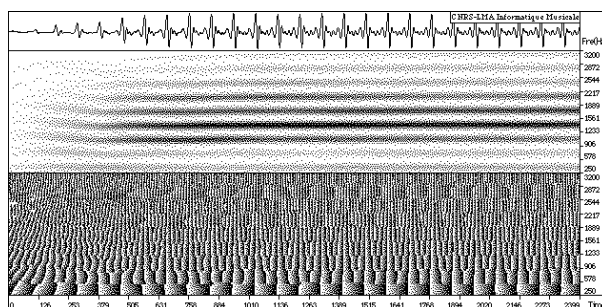


Figure 6: Gabor representation of the first 75ms of a trumpet sound. Many harmonics with different time dependencies are visible on the modulus picture. The phase picture shows different regions, around each harmonic, where the phase wraps regularly at the time period of each harmonic.

Thus, to estimate parameters for an additive resynthesis of the sound, amplitude and frequency modulation laws associated to each partial should be extracted from the transform. Of course, this process must be efficient even for extracting components that are very close to each other and have rapidly changing amplitude modulation laws. Unfortunately all the constraints for constructing the representation make this final operation complicated. The justification is of the same nature as the one given in the introduction in connection with the sound transformation by modifying the representations. Absolute accuracy both in time and in frequency is impossible because of a mathematical relation between the transform in a point of a time-frequency plane and the close vicinity of this point. Human hearing follows a rather similar "uncertainty" principle: to identify the pitch of a pure sound, it must last for a certain time. The consequences of these limitations on the additive model parameter estimation are easy to understand. A high-frequency resolution necessitates analysis functions that are well-localized in the frequency domain and therefore badly localized in the time domain. The extraction of the amplitude modulation law of a component from the modulus of the transform on a trajectory in the time-frequency plane smooths the actual modulation law. This smoothing effect acts in a time interval with the same length as the analysis function. Conversely, the choice of well-localized analysis functions in the time domain generally yields oscillations in the estimated amplitude modulation laws, because of the presence of several components in the same analysis band. It is possible however to avoid this problem by astutely using the phase of the transform to precisely estimate the frequency of each component and by taking advantage of the linearity in order to separate them, without a hypothesis on the frequency selectivity of the analysis. The procedure uses linear combinations of analysis functions for different frequencies to construct a bank of filters with a quasi-perfect reconstruction. Each filter specifically estimates a component while conserving a good localization in the time domain. Different kinds of filters can be designed, which permit an exact estimation of amplitude modulation laws locally polynomial on the time support of the filters [20]. The strict limitations of the wavelet transform or of the Gabor transform can be avoided by optimizing the selectivity of the filter as a function of the vicinity of the frequency components.

Another important aspect of the musical sound is the frequency modulation of the components, in particular during the attack of the sound. Here the judicious use of the time derivative of the transform phase offers the possibility of developing iterative algorithms tracking the modulation laws, thus precluding the computation of the whole transform. These algorithms use frequency-modulated analysis functions, the modulations of which are automatically matched to the ones of the signal [20].

## 5 Fitting synthesis parameters

The extraction techniques using the time-frequency transforms directly provide a group of parameters, which permit the resynthesis of a sound with the additive model. In addition, they can be used for identification of

other synthesis models. The direct parameter identification techniques for the non-linear models are difficult. Generally they do not give an exact reproduction of a given sound. The estimation criteria can be statistical (minimization of non-linear functions) [21] or psychoacoustic [22],[15]. The direct estimation of physical or subtractive model parameters requires techniques like linear prediction, used for instance in speech synthesis [23]. Another solution consists in using parameters from the additive synthesis model to estimate another set of parameters corresponding to another synthesis model. In what follows we shall see how this operation can be done for the most common models.

### 5.1 Parameter estimation for signal model synthesis

The parameter estimation for the additive model is the simplest one, since the parameters are determined in the analysis. The modeling of the envelopes can greatly reduce the data when one uses only perceptual criteria. The first reduction consists in associating each amplitude and frequency modulation law to a piecewise linear function [24] (figure 7). This makes it possible to automatically generate, for example, a Music V or a Csound score associated to the sound.

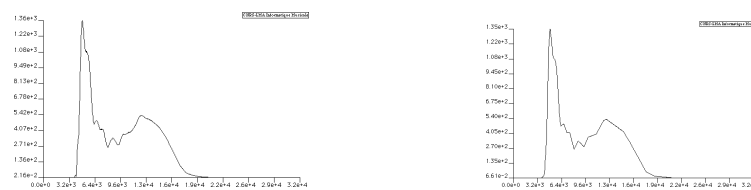


Figure 7: Original and modeled envelopes of a saxophone sound. The modeled curve is defined with 35 breakpoints and linear interpolation between them, while the original is defined on 32000 samples.

Another possible reduction consists in grouping the components from the additive synthesis (group additive synthesis) [25], [26]. This can be done by statistical methods, like principal component analysis, or by following an additive condition defined as the perceptual similarity between the amplitude modulations of the components [27]. This method offers a significant reduction in the number of synthesis parameters, since several components with a complex waveshape have the same amplitude modulation laws (figures 8 and 9).

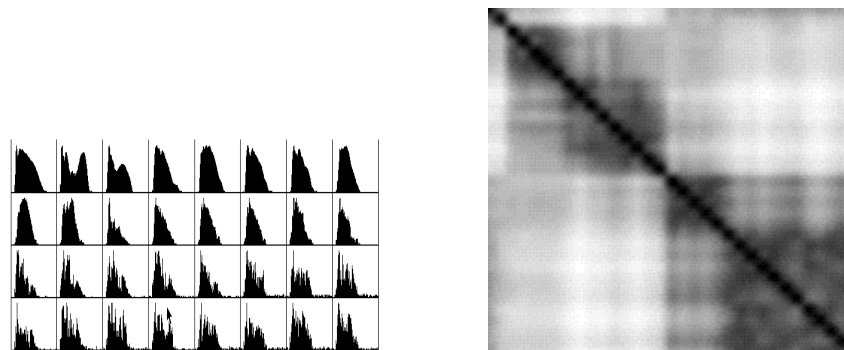


Figure 8: A whole set of envelopes of a violin sound, and the matrix showing the correlation between them. The dark regions around the diagonal correspond to curves that look similar and that correspond to components that are close in the frequency domain.

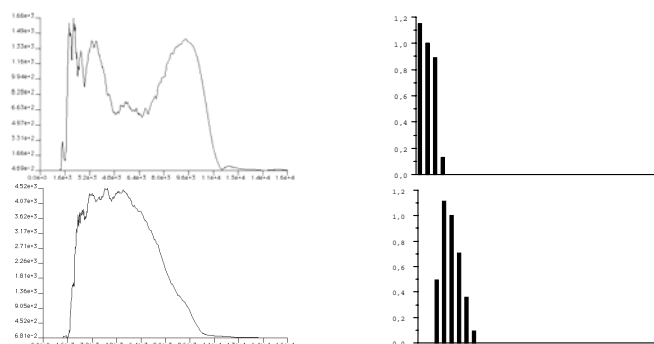


Figure 9: Two main envelopes of the group additive synthesis model, with the spectrum of their associated waveform. Psychoacoustic criteria can be used to generate a perceptively similar spectrum with non-linear techniques.

## 5.2 Subtractive synthesis

An evolving spectral envelope can be built by creating intermediate components obtained from the modulation laws of the additive modeling. Their amplitude modulation laws are obtained by interpolation of the envelopes of two adjacent components in the frequency domain (figure 10). These envelopes can then be used in order to "sculpt" another sound (crossed synthesis). As we already mentioned, physical modeling is sometimes close to subtractive synthesis. This aspect will be developed later.

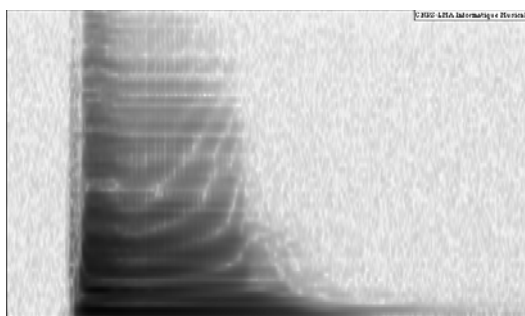


Figure 10: Spectral envelope of a saxophone sound built from the additive synthesis parameters. This envelope can be used to "sculpt" the modulus of the Gabor transform of another sound in order to perform a crossed synthesis.

## 5.3 Waveshaping and frequency modulation synthesis

From the parameters corresponding to the group additive synthesis (complex waves and their associated amplitude laws), one can deduce non-linear synthesis parameters [27]. The technique consists in approaching each complex wave shape by an elementary non-linear module. In the case of waveshaping, the knowledge of the complex wave allows the calculation of an exact distortion function. In the case of FM, the spectral components should be grouped, not only according to a perceptive criterion, but also according to a condition of spectral proximity. This condition is meaningful because real similarities between envelopes of neighbouring components are often observed. To generate the waveform corresponding to a group of components by an elementary FM oscillator, the perceptive approach is best suited. In that case, one can consider the energy and the spectral extent of the waveforms, which are directly related to the modulation index. Other methods based on the minimization of non-linear functions by the simulated annealing or genetic algorithms have also been explored [21]. Attempts at direct estimation of the FM parameters by extraction of frequency modulation laws from the phase of the analytic signal related to the real sound have led to interesting results [28], [29].

Recently, optimization techniques using perceptual criteria have been developed. Here, rather than attempting to reconstruct a signal similar to the original, one looks for a perceptually similar sound. Criterion such as the Tristimuli [30] can be used to define a perceptual distance. This distance is then minimized using classical methods [15] leading to the optimal value of the synthesis parameters.

## 5.4 Parameter estimation for physical model synthesis

The digital waveguide synthesis parameters are of a different kind. They characterize both the medium where the waves propagate and the way this medium is excited. From a physical point of view, it is difficult to separate these two aspects: the air jet of a wind instrument causes vortex sheddings interacting with the acoustic pressure in the tube [31]; the piano hammer modifies the characteristics of a string while it is in contact with it [32]. These source-resonator interactions are generally non-linear and often difficult to model physically. However, a simple, linear digital waveguide model often gives satisfactory sound results. In a general way, the study of linear wave propagation equations in a bounded medium shows that the response to a transient excitation can be written as a sum of exponentially damped sine functions. The inharmonicity is related to the dispersive characteristics of the propagation medium, the decay times are related to the dissipative characteristics of the

medium, and the amplitudes are related to the spectrum of the excitation. In the same way, the impulse response of the simple digital waveguide model can be approximated by a sum of exponentially damped sinusoids whose frequencies, amplitudes, and damping rates are related in a simple way to the filter coefficients [14]. Thanks to the additive synthesis parameters one can, for the percussive sound class, determine the parameters of the waveguide model, and also recover the physical parameters characterizing the instrument [33] (figure 11). For sustained sounds, the estimation problem of the exciting source is crucial and necessitates the use of deconvolution techniques. This approach is entirely non parametric, but it is also possible to use parametric techniques. Indeed, the discrete time formulation of the synthesis algorithm corresponds to a modeling of the so-called ARMA type (AutoRegressive Moving Average).

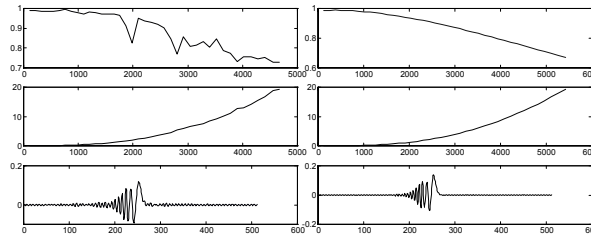


Figure 11: Parameter estimation for the digital waveguide model. Pictures on the left show the data from the estimation. Pictures on the right show the data from the movement equation of a stiff string. Respectively, from top to bottom: modulus (related to losses during the propagation); phase derivative (related to the dispersion law of the propagation medium) of the Fourier transform of the filter inside the loop; impulse response of the loop filter.

## 5.5 Parameter estimation for hybrid model synthesis

Hybrid models being a combination of signal and physical models, one can estimate their parameters using most of the previous techniques. Actually, starting with a transient response of the instrument (sound generated from a rapidly closed fingerhole for example) one can estimate the parameters of the resonator's model. Then, using a deconvolution technique, one can extract the source signal from the natural sound. This source can then be modeled using signal models such as waveshaping. Two examples of such methods can be found in [14] for the flute case and [34] for the piano case.

This approach recently improved by considering a more general model constituted by a loop system including both a linear filter and a non-linear element. This model is very general and can be considered as a "non-linear digital waveguide". Thank to the previous method, one can estimate the parameters of such a model, even though it's non-linear behavior makes the analysis-synthesis process very complicated [35].

## 6 Acknowledgments

This work was partly supported by the Norwegian Research Council

## References

- [1] Kronland-Martinet R., Guillemain, Ph & Ystad S. 1997, "Modeling of natural sounds by time-frequency and wavelet representations", *Organised Sound*, 2(3) : 179-191.
- [2] Roads, C. Automated granular synthesis of sound. *Computer Music Journal*, 1978, 2, No. 2, 61-62.
- [3] Kronland-Martinet, R. The use of the wavelet transform for the analysis, synthesis and processing of speech and music sounds. *Computer Music Journal*, 1989, 12 n° 4, 11-20 (with sound examples on disk).
- [4] Rabiner L.R., Gold B., *Theory and Application of Digital Signal Processing*, 1975, Prentice Hall.
- [5] Atal B.S. & Hanauer S.L 1971, "Speech analysis and synthesis by linear prediction of the speech wave" *J. Acoust. Soc. Amer.*, 50, p. 637-655.
- [6] Flanagan J.L., Coker C.H., Rabiner P.R., Schafer R.W., & Umeda N. , *Synthetic voices for computer*. I.E.E.E Spectrum, 1970 , 7, p. 22-45.

- [7] Makhoul, Linear prediction, a tutorial review. Proceedings of Institute of Electrical and Electronics Engineers, 1975, 63, 561-580.
- [8] Kronland-Martinet R., Digital subtractive synthesis of signals based on the analysis of natural sounds. In "Etat de la Recherche Musicale (au 1er janvier 1989)" 1988, Ed. A.R.C.A.M. , Aix en Provence.
- [9] Chowning, J. The synthesis of complex audio spectra by means of frequency modulation. Journal of the Audio Engineering Society, 1973, 21, 526-534.
- [10] Le Brun, M. Digital waveshaping synthesis. Journal of the Audio Engineering Society, 1979, 27, 250-266.
- [11] Chaigne, A. Trends and Challenges in Physical Modeling of Musical Instruments. Proceedings of the ICA, Volume III, p.397-400 Trondheim, Norway 26-30 june 1995.
- [12] Smith, J. Physical modeling using digital waveguides. Computer Music Journal, 1992, 16 n° 4, 74-91.
- [13] Cook, P.R. A meta-wind-instrument physical model controller, and a meta-controller for real-time performance control. Proceedings of the 1992 International Music Conference, San Francisco: Computer Music Association, 1992, 273-276.
- [14] Ystad, S. 1998. "*Sound Modeling Using a Combination of Physical and Signal Models.*" Ph. D. Thesis from the University of Aix-Marseille II.
- [15] Ystad, S. 2000. "Sound Modeling Applied to Flute Sounds" *Journal of the Audio Engineering Society*, 48(9) ; 810-825.
- [16] Ystad, Voinier, ICMCCUBA or MOSART2001
- [17] Risset J.C, & Wessel D.L, Exploration of timbre by analysis and synthesis, in D. Deutsch ed., The Psychology of Music, Academic Press 1982, 26-58.
- [18] Flandrin P. Temps-fréquence. 1993, Hermès, Traité des nouvelles technologies, série traitement du signal.
- [19] Kronland-Martinet, R., Morlet, J., & Grossman, A. Analysis of sound patterns through wavelet transforms. Intern. Journal of Pattern Recognition and Artificial Intelligence, 1987, 11 n° 2, 97-126.
- [20] Guillemain, Ph., & Kronland-Martinet, R., Characterisation of Acoustics Signals Through Continuous Linear Time-frequency Representations Proceedings of the IEEE, Special Issue on Wavelets, 1996, Vol. 84 n°4, 561-585.
- [21] Horner, A. Double-Modulator FM Matching of Instruments Tones, Computer Music Journal, 1996, V. 20 (2), pp 57-71
- [22] Beauchamp, J. W. Analysis and synthesis of cornet tones using non-linear interharmonic relationships. Journal of the Audio Engineering Society, 1975, 23, 778-795.
- [23] Markel, J. D, & Gray, A. H. Linear Prediction of Speech. Springer-Verlag, Communication and Cybernetics (12), 1976.
- [24] Horner, A., & Beauchamp, J. Piecewise-linear approximation of additive synthesis envelopes: a comparison of various methods. Computer Music Journal, 1996, 20 n° 2, 72-95.
- [25] Oates S., & Eagleston B. Analytic Methods for Group Additive Synthesis, Computer Music Journal, 1997, V. 21 (2), 21-39.
- [26] Kleczkowski, P. Group additive synthesis. Computer Music Journal, 1989, 13 n° 1, 12-20.
- [27] Kronland-Martinet, R., & Guillemain, P. Towards non-linear resynthesis of instrumental sounds. Proceedings of the 1993 International Music Conference, San Francisco: Computer Music Association, 1993, 86-93.
- [28] Justice, J. Analytic signal processing in music computation. IEEE Transactions on Speech, Acoustics and Signal Processing, 1979, ASSP-27, 670-684.
- [29] Delprat, N., Guillemain, P., & Kronland-Martinet, R. Parameter estimation for non-linear resynthesis methods with the help of a time-frequency analysis of natural sounds. Proceedings of the 1990 International Music Conference, Glasgow: Computer Music Association, 1990, 88-90.
- [30] Pollard H.F. and Jansson E.V. 1982. "A Tristimulus Method for the Specification of Musical Timbre." *Acoustica*, Vol. 51 ;162-171.

- [31] Verge M.P. 1995. "Aeroacoustics of Confined Jets with Applications to the Physical Modeling of Recorder-Like Instruments." PhD thesis, Eindhoven University.
- [32] Weinreich, G. Coupled piano strings. *Journal of the Acoustical Society of America*, 1977, 62, 1474-1484.
- [33] Guillemain Ph., Kronland-Martinet R., & Ystad S.. Physical Modelling Based on the Analysis of real Sounds. *Proceeding of Institute of Acoustics, Edinburgh 1997, Vol 19, ISMA 97*, 445-450
- [34] J. Bensa, K. Jensen, R. Kronland-Martinet, S. Ystad "Perceptual and Analytical Analysis of the effect of the Hammer Impact on the Piano Tones", *proceedings of the International Computer Music Conference ICMC 2000*, pp.58-61, Berlin (Germany) 27 août au 1 September 2000.
- [35] Ystad, S & Voinier, Th. Analysis-Synthesis of Flute Sounds with a Looped Non-linear Model. *MOSART workshop, Barcelona, November 17.-20 2001*

# The Timbre Model - Discrimination and Expression

Kristoffer Jensen

Music Informatics Laboratory  
Department of Datalogy, University of Copenhagen  
Universitetsparken 1, 2100 Copenhagen Ø, Denmark  
[krist@diku.dk](mailto:krist@diku.dk), <http://www.diku.dk/~krist>

## Abstract

This paper presents the timbre model, a signal model which has been built to better understand the relationship between the perception of timbre and the musical sounds most commonly associated with timbre, and the initial results from a research project involving the discrimination sensibility of parameter changes of the timbre model. In addition, an extension to the timbre model incorporating expressions is introduced. The presented work therefore has relation to a large field of science, including auditory perception, signal processing, physical models and the acoustics of musical instruments, music expression, and other computer music research. The timbre model is based on a sinusoidal model, and it consists of a spectral envelope, frequencies, a temporal envelope and different irregularity parameters. The paper is divided into the following parts: an overview of the research done on the perception of timbre, an overview of the signal processing aspects dealing with sinusoidal modeling, the timbre model, a preliminary study on the discrimination of the timbre model parameters, and an introduction of some expressive extensions to the timbre model.

## 1 Introduction

The sound of the musical instrument can be qualified by the timbre (or the identity) of the sound and the expressions caused by gestures. Expressions associated with musical instruments are well defined by common musical terms, such as note, intensity, tempo or style. Timbre seems to be a multi-dimensional quality. Generally, timbre is separated from the expression attributes pitch, intensity, and length of a sound. Furthermore, research has shown that timbre consists of the spectral envelope, an amplitude envelope function, which can be attack, decay, etc., the irregularity of the amplitude of the partials, and noise.

The additive parameters constitute a good analysis/synthesis model of voiced sounds, but it has a very large parameter set, which is not always easily manipulated. This paper presents an approach to modeling the additive parameters in an easily controllable, intuitive model, whose parameters are closely related to the timbre dimensions as proposed in timbre research

The timbre model models each partial in a few pertinent parameters: the spectral envelope, the mean frequencies, the amplitude envelope, and the noise, or irregularity parameters. Furthermore, the rate and extend of the vibrato and tremolo are modeled.

The timbre model has a fixed parameter size, dependent only on the number of partials, and most of the parameters of the model have an intuitive perceptive quality, due to their relation with timbre perception. Furthermore, by interpolating between

parameter sets obtained from different sounds, morphing is easily implemented.

The timbre model can be used to resynthesize the sound, with some or all of the parameters of the model. In this way, the validity and importance of each parameter of the model can be verified.

This paper starts with an overview of the perceptual research, which forms the basis for the model, then an overview of the additive analysis is given, and the timbre model is detailed. Next a novel set of expression additions to the timbre model is presented, other applications to the model are outlined and a conclusion is given.

## 2 Perceptual Research

The timbre model is derived from conclusions extracted from the auditory perception research. Several methodologies have been used in this research, and even though the results are given in many formats, the conclusions are generally the same. The research suffers from a lack of comparable sound material, and in particular, the lack of noisy, non-harmonic or percussive sounds in the experiments. In addition, different pitches are generally not used.

This is an overview of the research on which the timbre model is based. For a larger overview of timbre research, see for instance [38], [66].

In a larger scope, [7] presents some aspects of timbre used in composition. Timbre research is of course related to auditory perception [103] and psycho-acoustics [106] research. An initial study of the Just Noticeable Difference (JND) of many of the timbre attributes presented here can be found in [48].



The mpeg community, which defined the popular mp3 compression standard, are currently defining mpeg 7 [63], which defines the content of sound and music. The outcome of this standardization process may be very useful in finding common terms for objectively describing music and music sounds. It differentiates between noise and harmonic sounds, substituting spectrum measures in the first case with harmonic measures in the second case.

## 2.1 Timbre Definition

Timbre is defined in ASA [2] as that quality which distinguishes two sounds with the same pitch, loudness and duration. This definition defines what timbre is not, not what timbre is.

Timbre is generally assumed to be multidimensional, where some of the dimensions has to do with the spectral envelope, the amplitude envelope, etc. The difficulty of timbre identity research is often increased by the fact that many timbre parameters are more similar for different instrument sounds with the same pitch, than for sounds from the same instrument with different pitch. For instance, many timbre parameters of a high pitched piano sound are closer to the parameters of a high-pitched flute sound than to a low-pitched piano sound. Nevertheless, human perception or cognition generally identifies the instrument correctly. Unfortunately, not much research has dealt with timbre perception for different pitches. Greg Sandell has made a list [85] of different peoples definition of the word timbre.

## 2.2 Verbal Attributes

Timbre is best defined in the human community outside the scientific sphere by its verbal attributes (historically, up to and including today, by the name of the instrument that has produced the sound). von Bismarck [99] had subjects rate speech, musical sounds and artificial sounds on 30 verbal attributes. He then did a multidimensional scaling on the result, and found 4 axes, the first associated with the verbal attribute pair dull-sharp, the second compact-scattered, the third full-empty and the fourth colorful-colorless. The dull-sharp axis was further found to be determined by the frequency position of the overall energy concentration of the spectrum. The compact-scattered axis was determined by the tone/noise character of the sound. The other two axes were not attributed to any specific quality.

## 2.3 Dissimilarity Tests

The dissimilarity test is a common method of finding proximity in the timbre of different musical instruments. Asking subjects to judge the dissimilarity of a number of sounds and analyzing the results is the essence of the dissimilarity tests. A multidimensional scaling is used on the

dissimilarity scores, and the resulting dimensions are analyzed to find the relevant timbre quality.

Grey [33] found the most important timbre dimension to be the spectral envelope. Furthermore, the attack-decay behavior and synchronicity were found important, as were the spectral fluctuation in time and the presence or not of high frequency energy preceding the attack.

Iverson & Krumhansl [44] tried to isolate the effect of the attack from the steady state effect. The surprising conclusion was that the attack contained all the important features, such as the spectral envelope, but also that the same characteristics were present in the steady state. The resulting timbre space was similar no matter if the full tones, the attack only or the remainders only were examined.

Later studies Krimphoff *et al.* [54] refined the analysis, and found the most important timbre dimensions to be brightness, attack time, and the spectral fine structure.

Grey & Gordon [35], Iverson & Krumhansl [44] and Krimphoff *et al.* [54] compared the subject ratings with calculated attributes from the spectrum. Grey & Gordon [35] found that the centroid of the bark [90] domain spectral envelope correlated with the first axis of the analysis. Iverson & Krumhansl [44] also found that the centroid of the spectral envelope, here calculated in the linear frequency domain (brightness), correlated with the first dimension. Krimphoff *et al.* [54] also found the brightness to correlate well with the most important dimension of the timbre. In addition, they found the log of the rise time (attack time) to correlate with the second dimension of the timbre, and the irregularity of the spectral envelope to correlate with the third dimension of the timbre. McAdams *et al.* [66] further refined this hypothesis, substituting spectral irregularity with spectral flux.

The dissimilarity tests performed so far do not indicate any noise perception. Grey [33] introduced the high frequency noise preceding the attack as an important attribute, but it was later discarded in Iverson & Krumhansl [44]. The lack of indications of noise discrimination might be explained by the fact that no noisy sounds were included in the test sounds. McAdams *et al.* [66] promises a study with a larger variety of test sounds, including noisy sounds. It can also be explained by the fact that the most common analysis methods doesn't permit the analysis of noise, which then cannot be correlated with the ratings.

## 2.4 Auditory Stream Segregation

An interesting way of examining the qualities of timbre that can be related to its perception is the auditory stream segregation [12].

The auditory stream segregation is referring to the tendency to group together (i.e. relate them to the same source) sounds with components falling in similar frequency ranges. This phenomenon is called fusion or coherence. The opposite

phenomenon when the sounds are separated into several groups, as coming from different sound sources, is called fission or segregation. For intermediate frequency separations between successive tones in a rapid sequence the percept may be ambiguous.

Experimenters have taken advantage of auditory stream segregation to identify timbre qualities related to timbre perception. Using a three tone repeated sequence, by means of adjusting spectral characteristics of the middle tone they tried to investigate how the fusion/fission boundary is affected in relation to the value it assumes for the monotimbral (no change in the timbre of the middle tone) case.

Singh and Bregman [92] presented monotimbral and bitimbral sequences of complex sounds to listeners where for the bitimbral case there were changes in the attack time and the number of harmonics in the middle sound. The results indicate that the effect of the change was highly significant for both the fission and the fusion boundary fundamental frequency value. The boundaries assumed lower values than for the monotimbral case in all changes and the order of the impact on the results was higher for changes in the number of partials than for the envelope changes. This method seems promising when searching for an absolute value for timbre changes, since any timbre change provoking fusion/fission can be related to the corresponding pitch change when there is no timbre change.

## 2.5 Discrimination

Several studies asked subjects the rate the differences between original (unmodified) sounds and modified sounds, in order to evaluate the perceptual importance of the modification. One recent such study is McAdams *et al.* [65], in which the additive parameters of seven sounds are modified in different ways, and the importance of the modifications are asserted. The sounds, clarinet, flute, oboe, trumpet, violin, harpsichord and marimbe were normalized to E4-flat (311.1 Hz), 2 secs and equal subjective loudness. The participants were asked to discriminate between the original (the normalized E4-flat, 2secs) and modified sounds, and some of the results of this research are:

The effect of musical training is weak but the effect of instrument is strong. Therefore each instrument is analyzed individually whereas the participants are grouped.

The most important results from the effect of simplification are that most simplifications are perceptible and that accumulated simplifications are equivalent to the most potent constituent (Most salient feature dominate in combined simplifications).

The order of the simplifications are: spectral envelope smoothing, spectral flux (amplitude envelope coherence) (very good discrimination), forced harmonic frequency variations, frequency

variations smoothings, frequency micro-variations and amplitude micro-variations (moderate to poor discriminations).

In conclusion, several research methods have been used to determine the dimensions of timbre. Although no clear consensus has emerged, the most common dimensions seem to be spectral envelope, temporal envelope and irregularities.

## 3 Additive Analysis/Synthesis

The additive model has been chosen as the basis for the timbre model for the known analysis/synthesis qualities and the perceptually meaningful parameters of this model. Many analysis/synthesis systems using the additive model exist today, including sndan [8], SMS [91], the loris program [25] and the additive program [82].

### 3.1 Choice of model

In order to perform analysis by synthesis, the timbre model must be based on an analysis/synthesis model. The choice of underlying model is the additive (sinusoidal) model, for its well-understood parameters (time, amplitude and frequency) and for its proven analysis methods. Other methods investigated include the physical model [45], the granular synthesis [97], the wavelet analysis/synthesis [55], the atomic decomposition [14], [36] and the modal distribution analysis [73]. In the additive analysis, [91] added a stochastic part to the sound, making the model non-homogenous. Other non-homogenous additions include the transients [98].

### 3.2 Additive Model

The additive analysis consists in associating a number of sinusoidal with a sound, and estimating the time-varying amplitudes ( $a_k(t)$ ) and frequencies ( $f_k(t)$ ) of the  $N$  sinusoids (partials) from the sound. The sound can then be resynthesized, with a high degree of realism, by summing the sinusoids,

$$s(t) = \sum_{k=1}^N a_k(t) \cdot \sin(2\pi \int_{\tau=0}^t f_k(\tau) d\tau). \quad (1)$$

In order to have a good resynthesis quality, the absolute phases are generally necessary [3].

### 3.3 Additive Analysis

Several methods exist for determining the time-varying amplitudes and frequencies of the partials. Already in the last century, musical instrument tones were divided into their Fourier series [40], [74]. Early techniques for the time-varying analysis of the additive parameters are presented by Matthews *et al.* [64] and Freedman [28]. Other more recent techniques for the additive analysis of musical signals are the proven heterodyne filtering

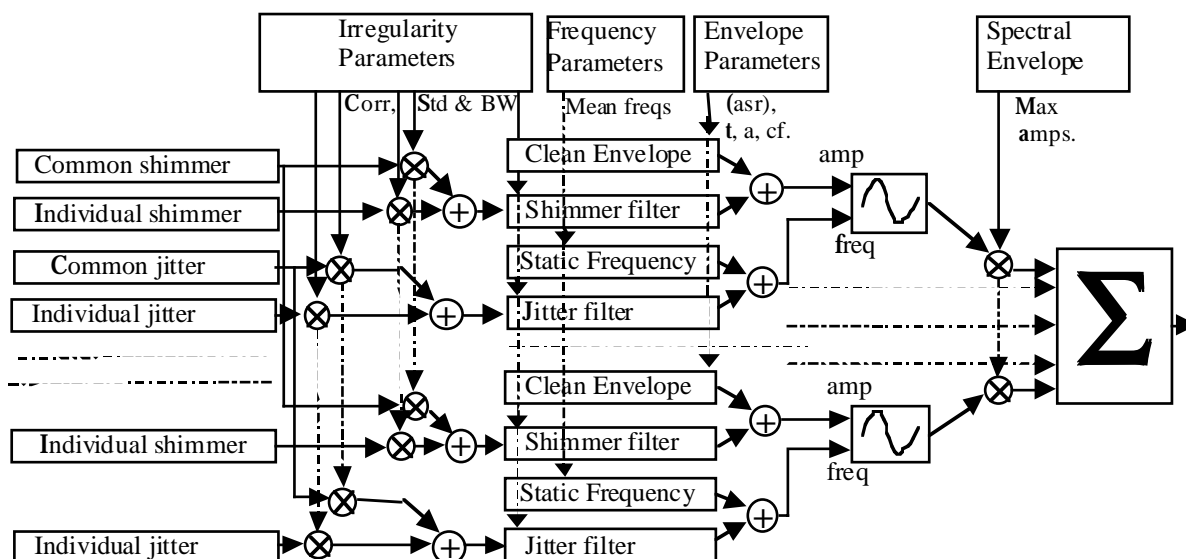


Figure 1. Timbre model diagram. The model consists of a number of sinusoids, with amplitude as a sum of the clean envelope and the shimmer multiplied with the spectral envelope, and the frequency as a sum of the static frequency and the jitter. The shimmer and jitter are a sum of common (correlated) and individual gaussian white noise filtered and scaled by the standard deviation.

[34], the much-used FFT-based analysis [67], and the linear time/frequency analysis [37]. The time and frequency reassignment method [4] has recently gained popularity [11], [23]. Ding and Qian [19] has presented an interesting method, fitting a waveform by minimizing the energy of the residual, improved and dubbed adaptive analysis by Röbel [79]. A great deal of research has been put into understanding and improving [100], [17] the windows [39] of the short term Fourier transfer [1]. Finally, Marchand [60] used signal derivatives to estimate the amplitudes and frequencies.

### 3.4 Analysis/synthesis evaluation

Not many objective listening tests (outside the compression community) have been performed in the music community to evaluate analysis/synthesis methods. [95] evaluated the spectral/time envelope model with listening tests. [34] compared analysis/synthesis and different data-reductions, and [85] evaluated the PCA-based data reduction with listening tests.

Recently, however, the additive analysis/synthesis has been tested in several well-controlled listening test experiments. The listening tests performed in connection with this work have been inspired by the listening tests performed for the evaluation of speech and music compression. The method used is called double blind triple stimulus with hidden reference [75]. [52] found the linear time-frequency additive analysis/synthesis to have a sound quality between imperceptible and perceptible, but not annoying using short musical sounds. [13] found comparable results using several additive analysis systems and longer musical sequences. In addition [3] found that including the phase gave significantly better results. Other evaluation methods include estimating the time resolution,

[52] found that the LTF analysis method [37] has a time resolution that is twice as good as the FFT-based analysis, and error estimations. [11] found the mean error in the amplitude and frequency estimation to be significantly better for the frequency reassignment method, than the peak interpolation, or phase difference methods.

In this work, the FFT-based analysis is used. The additive parameters are saved for each half-period of the fundamental, and only quasi-harmonic partials are estimated. The window used is the kaiser window, and the block size is 2.8 periods of the fundamental.

## 4 The Timbre Model

The timbre model is inspired by the perceptual research, but it has been derived from the analysis of musical sounds using the analysis by synthesis method [78] and by literature studies into related research. By these methods, the model has been defined to consist of a spectral envelope, associated with brightness and resonances of the sounds, a frequency envelope, associated with pitch and inharmonicity. It also consists of an amplitude envelope consisting of five segments, start, attack, sustain, release and end, each with individual start and end relative amplitude and time, and curve form, and the amplitude and frequency irregularity (shimmer and jitter). The shimmer and jitter are modeled as a low-pass filtered gaussian with a given standard deviation and bandwidth. The amplitude envelope is associated with important timbre attributes, such as attack time, and sustained/percussive quality, and the irregularity is associated with both additive noise, but also slow irregularities, giving life to the sounds.

Other methods of modeling the additive parameters include the Group Additive Synthesis [53], [22],

[15], where similar partials are grouped together to improve efficiency. [96] use envelope time points to morph between different musical sounds. Marchand has proposed the Structured Additive Synthesis [59].

Schaeffer proposes a classification of attack genres (among others things) in his monumental work [87].

The base of the timbre model is the additive parameters, the amplitude of which is controlled by the spectral envelope, amplitude envelope and shimmer parameters, and the frequency of which is controlled by the mean frequency and the jitter parameters.

The timbre model diagram can be seen in figure 1. It consists of a number of sinusoidals (partials), whose amplitude is the sum of a clean envelope (attack-sustain/decay-release) and irregularity (shimmer) multiplied with the spectral envelope value, and whose frequency is the sum of a static value and irregularity (jitter). The timbre model parameters are (from right to left): Max amplitudes, envelope model times, amplitudes and curve form coefficients, mean frequencies, irregularity correlation, standard deviation and bandwidth.

## 4.1 The Spectral Envelope

The spectral envelope is very important for the perceived effect of the sound; indeed, the spectral envelope alone is often enough to distinguish or recognize a sound. This is especially true for the recognition of vowels, which are entirely defined by the spectral envelope. As was seen earlier, the spectral envelope is indeed one of the most important timbre dimensions. Nevertheless, the spectral envelope alone is not enough to recreate any sound with realism. Many methods exist to model the spectral envelope, including the linear predictive coding (lpc), cepstrum, etc. [89]. Back in 1966 [95] synthesized wind instruments with a combination of spectral and temporal envelopes. [81] use spectral envelopes as a filter with different source models, including the additive model.

[71] introduced the discrete summation formulas in sound synthesis, which are here called the brightness creation function [52]. There exists an easy calculation and recreation of brightness with these formulas [52].

The spectral envelope is defined in this work as the maximum amplitude of each partial, denoted  $\hat{a}_k$ . As it is difficult to estimate the spectral envelope outside the discrete frequency points in voiced sounds, the spectral envelope model using the partial amplitudes is judged the most reliable.

## 4.2 Frequencies

The static frequencies are modeled as the weighted mean of the frequency for the sustain part of each partial, denoted  $\hat{f}_k$ .

Most sustained instruments are supposed to be perfectly harmonic, i.e.  $\hat{f}_k = k\hat{f}_0$ . The frequencies

are therefore best visualized divided by the harmonic partial index. The piano, in contrast, has inharmonic partial frequencies due to the stiffness of the strings [26]. Therefore, the piano partial frequencies are slightly higher than the harmonic frequencies.

## 4.3 Amplitude Envelopes

The envelope of each partial is modeled in five segments, start and end segments, supposedly silent, and attack, sustain and release segments. Thus, there are 6 amplitude/time split points, where the first is (0,0) and the last amplitude also is zero, since all partials are supposed to start and end in silence. The amplitudes are saved as a percentage of the maximum of the amplitude (the spectral envelope), and the times are saved in msec. Furthermore, the curve form of each segment is modeled by a curve, which has an appropriate exponential or logarithmic form. The resulting concatenated clean amplitude envelopes are denoted  $\tilde{a}_k(t)$ . The formula for one segment can be seen in equation (13).

The envelope split-points and curve form coefficients are found by a method inspired by the scale-space community in image processing [57], in which the split-points are found on the very smoothed time-derivative envelopes, and then followed gradually to the unsmoothed case. The smoothing is performed by convoluting the envelope with a gaussian,

$$env_\sigma(t) = a_k(t) * g_\sigma(t), \quad g_\sigma(t) = \frac{1}{2\pi\sigma} e^{-\frac{t^2}{2\sigma^2}}. \quad (2)$$

The derivative of the amplitude has been shown to be important in the perception of the attack [32]. The middle of the attack and release are now found by finding the maximum and minimum of the time derivative of the smoothed envelope,

$$\frac{\max}{\min} L_{t,\sigma}(t), \quad L_{t,\sigma}(t) = \frac{\partial}{\partial t} env_\sigma(t) \quad (3)$$

The start and end of the attack and release are found by following  $L_{t,\sigma}$  forwards and backwards (in time) until it is close to zero (about one tenth of the maximum derivative for the attack and end of release, and the double for the start of release). This method generally succeeds in finding the proper release time for the decay-release piano sound. A further development of the envelope analysis and model can be found in [52], [47].

## 4.4 Irregularities

Although the clean recreated envelopes have the general shape of the original envelope, there is a great deal of irregularity left, which is not modeled in the clean envelopes. The same holds true for the frequencies. The irregularities are divided into periodicity and non-periodic noises. The noise on the amplitude envelope is called shimmer, and the

noise on the frequency is called jitter. Shimmer and jitter are modeled for the attack, sustain and release segments. It is supposed to have a Gaussian distribution; the amplitude of the noise is then characterized by the standard deviation. The frequency magnitude of the noise is assumed low-pass and modeled with the 3dB bandwidth, and the correlation between the shimmer and jitter of each partial and the fundamental is also modeled.

Other noise models of musical sounds include the residual noise in the FFT [91], [70] and the random point process model of music [77] or speech noise [76].

Models of noise on sinusoidals include the narrow band basis functions (NBBF) in speech models [62]. In music analysis, [24] introduced the bandwidth enhanced sinusoidal modeling. Both models model only jitter, not shimmer.

Other analysis of the noise, and irregularity of the music sounds include the analysis of aperiodicity [68], [88], and the analysis of higher order statistics [20], [21].

The shimmer and jitter are calculated normalized with the clean amplitudes and the mean frequencies respectively,

$$shimmer_k = \frac{a_k(t) - \tilde{a}_k(t)}{\hat{a}_k} \quad (4)$$

$$jitter_k = \frac{f_k - \hat{f}_k}{\hat{f}_k} \quad (5)$$

The jitter and shimmer is then assumed to be stochastic, with a gaussian distribution, and modeled by its standard deviation and bandwidth (and later recreated using a single-tap recursive filter).

#### 4.5 Vibrato and Tremolo

Although assumed to be part of the expression of the sound, and therefore not necessary in the timbre model, some periodicity is nevertheless often found in the time-varying amplitudes and frequencies of the partials. This periodicity is modeled by its rate, strength, and phase. The frequency periodicity is here called vibrato, and the amplitude periodicity is called tremolo. There are individual vibrato and tremolo parameters for each partial.

The vibrato and tremolo parameters are found by searching for the max in the absolute value of the FFT of the time-varying frequency (subtracted by the mean frequency of each partial) or amplitude (subtracted by the clean amplitude curve). This provides an estimate of the strength, rate and phase of the periodic information in the signal, if there is any. [41] uses a comparable method for the vibrato estimation.

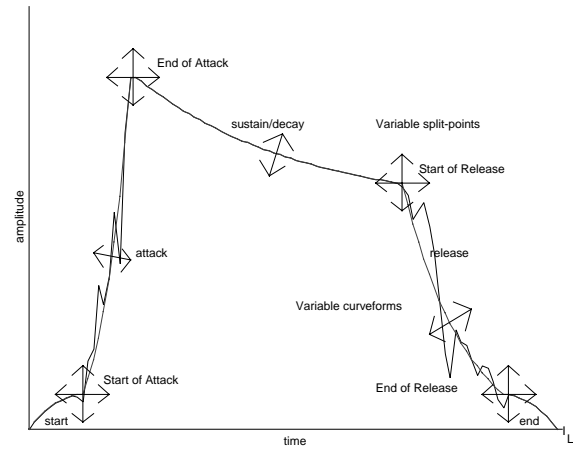


Figure 2. Total amplitude envelope for one partial. The five segments (start, attack, sustain, release and end) have individual split-points and curve-forms. The attack has high bandwidth shimmer, and the release has low bandwidth shimmer.

With the addition of irregularities, the timbre model is finished. An example of the resulting amplitude envelope for one partial can be seen in figure 3. The envelope is slightly decaying, and it has high bandwidth shimmer at the attack, and low bandwidth shimmer at the release.

#### 4.6 Visualization

The timbre model parameters have now been presented, and an overview of the parameter estimation methods has been given. A more throughout presentation of the timbre model is given in [52]. This section presents a proposed visualization of the timbre attributes. Many of them are best plotted logarithmically, and all the attributes are plotted with fixed axes, to facilitate comparisons between sounds.

There are 12 different timbre attributes, some of which have values for more than one segment (attack, decay, sustain, release or end). In order to have an easy overview, all the timbre attributes are collected in one plot. All the start and end (which are assumed to be silent) parameters are omitted. In total, there are 12 attributes, which can be plotted in one figure in 6 rows and 2 columns. The left column has from the top to the bottom the spectral envelope, the normalized frequencies divided by the partial index and fundamental, and envelope timing (attack and release), the envelope percents, the envelope curve forms, and the vibrato and tremolo rate. The right column has from the top to the bottom the shimmer standard deviation, the jitter standard deviation, the shimmer filter bandwidth, the jitter filter bandwidth, the shimmer and jitter correlation and the tremolo/vibrato strength.

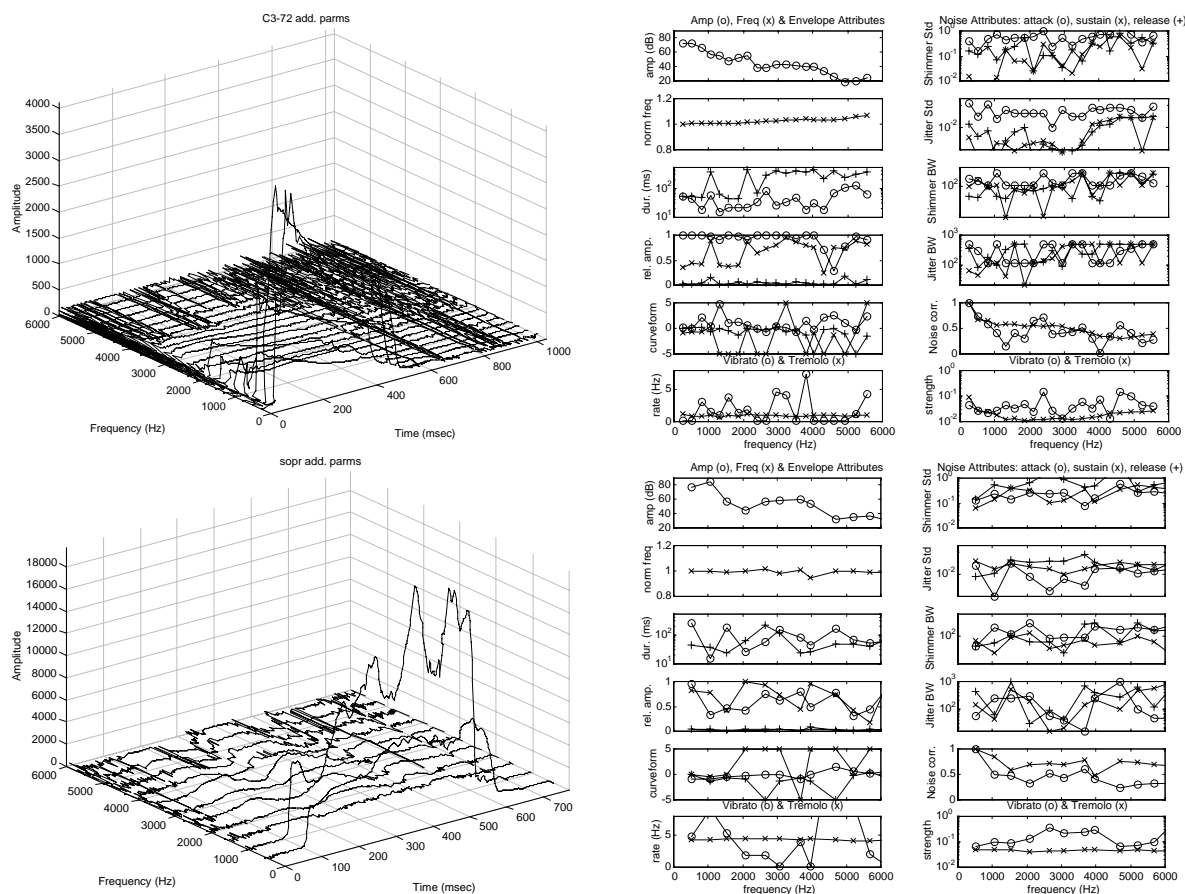


Figure 3. Additive parameters (left) and corresponding timbre attributes (right), Piano (top) and Soprano (bottom). Since all the axes in the timbre attribute plots are fixed, it is easy to compare the sounds. The amplitude values are denoted (o) and the frequency values (x). Attack is (o), sustain (x) and release (+). The piano and soprano have resonances in the spectral envelope, the piano is slightly inharmonic, and it has a faster attack. The piano is definitely percussive, since the attack relative amplitude is higher than the release. The soprano has a definite vibrato at around 4 Hz. Both the piano and the soprano have tremolo on some partials. The piano has much band-pass jitter in the attack. The jitter is more correlated than the shimmer for both sounds.

## 5 Timbre Model Parameter Discrimination

These experiments are performed to get an initial idea of the just noticeable differences of the parameters of the timbre model. This information will be used when creating a user interface for synthesis manipulations, such as the Timbre Engine [61]. This work has been made in collaboration with Georgios Marentakis and it was previously published in [48]. A good range for each parameter of the timbre model needs to have sufficient resolution. This can be determined by finding the just noticeable difference of each parameter. This is work in progress, and more results are expected to be available.

### 5.1 Default Sound

In order to test the sensibility of changes of the timbre model parameters, a series of informal listening tests has been performed. Since it is unclear how the perception changes as a function of

pitch, intensity, etc, a number of different sounds have been prepared for testing the sensibilities. These are

- Pitches (100, 300, 1000 Hz)
- Relative Intensities (-20, 0 +20 dB)
- Durations (100, 300, 1000 msec)
- Brightness (dark, normal, bright)

In addition, the changes on the timbre model parameters can be done on individual partials, on uniform scaling, or exponential scaling (first partials doesn't scale, following partials scale more and more). We have chosen the following modification situations:

- Individual partials (1<sup>st</sup>, 5<sup>th</sup>, 10<sup>th</sup>)
- Uniform scaling (all partials scale by the same amount)
- Exponential scaling (high order partials scale more)

The following parameters are used in these experiments

- Amplitude
- Frequency
- Attack and Release durations
- Release amplitude
- Attack curve form coefficients
- Sustain Shimmer Standard Deviation
- Sustain Shimmer Bandwidth
- Sustain Jitter Standard Deviation
- Sustain Jitter Bandwidth
- Sustain Jitter and Shimmer Correlation

The default sounds have harmonic frequencies, 20 msec attack and release durations, 100 % release amplitude (full sustain), and no shimmer or jitter. The 100 Hz sounds have 100 partials, the 300 Hz sounds have 30 partials, and the 1000 Hz sound have 10 partials.

## 5.2 Experimental method

In order to assess the perception of the parameters, a series of pairs of sounds with gradually varied parameter value are played sequentially. Starting with the default sound, the timbre model parameter is gradually distanced from the default value and the participant is asked to indicate when there is a noticeable difference. Both the upward and downward differences are asserted. The participant can change the range of values, and listen to the sequences as many times as necessary. The sounds are played through a headphone (Beyer-Dynamic DT 990) at a relative high level. This method has some drawbacks, in particular it seems necessary to point out to the participants that it is necessary to zoom out on each parameter, to be sure to listen to the effect under test. A better method would be, for instance, the up/down method [56].

## 5.3 Participants

Only a few participants have been used, and it has not been deemed necessary to perform any statistical analysis, except the mean, on the results. Five participants, between 22 to 37 years, male, non-musicians were used in the tests.

## 5.4 Preliminary results

The tests so far have been done only with the middle pitch (300 Hz), intensity (0 dB), duration (300 msec) and brightness (normal). No results as to how the perception changes with these parameters are yet available, although this is work in progress.

Of course, the sensibility of changes of these parameters depends largely on the default values of the parameters. These experiments have been done with default values, which are believed to be common in musical instruments. More experiments are necessary, however, to determine the effect of changing the default values.

All of the following results are valid only for the uniform scaling. No results are available, as yet, for the exponential, or the individuals partial scaling.

### Attack time

The default attack time was 60 msec (although comparable results were obtained with a default attack time of 20 msec). The attack time was varied between 0 and 120 msec, and the discrimination were determined to be between 10 and 80 (mean 50) msec when scaling all partials uniformly. This attribute being the first under test, it is possible that it has suffered from the adaptation of the participants to the test.

### Release time

The default release time was set to the same value as the default attack time, 60 msec. The sensibility to the release time seemed lower than the attack time sensibility, and the JND of the release time were determined to be between 15 and 80 (mean 30) msec.

### Attack curve form coefficient

The attack curve form coefficient decides which shape the attack has (exponential, linear or logarithmic). The formula for the curve form is

$$\hat{a}(t) = v_0 + (v_1 - v_0) \frac{e^{n \cdot t} - 1}{e^n - 1} \quad (6)$$

where  $t$  is the time,  $v_0$  is the start value,  $v_1$  is the end value and  $n$  is the curve form coefficient.

The default shape is linear ( $n=1$ ). The JND of the curve form coefficient for the attack segment was determined to be around 0.6 and 6 respectively (for 60 msec attack time duration).

### Release Amplitude

The release amplitude decides whether the sound is percussive, or sustained. The default value is 0 dB, which is a sustained sound. When this value is decreased, the sound becomes more percussive, and when it is increased, the sound gets a not natural, growing sound. The just noticeable difference for the relative release amplitude is situated between 1.2 and 6 (mean 3) dB, with similar values for the downward and upward cases.

### Shimmer Std

The shimmer (irregularity on the amplitude evolution of the partials) adds a quality to the sound, which can be both liveliness, and also additive noise. The shimmer is added as a filtered gaussian noise to the partial amplitudes. No shimmer makes the sound dead, unnatural, and much shimmer increases the noise and irregularity of the sound. The default value of the shimmer std is 0.2 (the default BW is 10 Hz, and the default correlation is 0.8), and the just noticeable difference of the shimmer std has been determined to be

between 0.6 and 3 dB. The mean downward JND was 1 dB and the mean upward JND was 1.9 dB.

### **Shimmer BW**

The bandwidth decides the frequency content of the irregularity of the shimmer. This changes the effect of the sound going from rumbling to additive white noise. The default shimmer BW is 20 Hz. The JND for the downward case is around 5 Hz, and it is around 100 Hz for the upward case.

### **Jitter Std**

The jitter (irregularity on the frequencies of the partials) adds a different quality than the shimmer to the sound. The jitter gives more low-frequency random pitch variations (with low bandwidth) or it adds roughness (with high bandwidth, going to an almost screaming quality when the correlation is high. The default std of the jitter is 0.02 (the default BW is 10 Hz, and the default correlation is 0.8) and the downward and upward perception threshold is about 0.07 and 0.3 dB respectively.

### **Jitter BW**

The jitter bandwidth changes the effect of the sound from low-frequency random pitch variations to adding roughness to the sound. The default jitter bandwidth is 10 Hz, and the downward and upward perception threshold is 10 Hz and 50 Hz respectively.

### **Shimmer Correlation**

The shimmer correlation changes the sound from a rich irregular sound to a fluctuating sound. The default shimmer correlation is 0.8 (the default std is 0.4 and the default bandwidth is around 1 Hz). The upward JND is around 0.1 and the downward JND is around 0.4.

### **Jitter Correlation**

The jitter correlation effect is quite dramatic, changing the sound from a calm sound with pitch variations to a sound with increasing roughness, going towards a screaming effect when the correlation is high. The default jitter correlation is 0.8 (the default std is 0.04 and the default BW is around 1 Hz). The downward and upward perception threshold is around 0.25 and 0.08.

### **Frequencies**

The uniform scaling of the frequencies changes the pitch of the sound. Of course, much research has been done on the perception of pitch changes, but this research has not been used here. Both the upward and the downward perception threshold have been found to lie between 1.5 and 3 Hz for the default sound used. This corresponds well with the 2 Hz just noticeable variation in frequency in [106] (page 182).

### **Amplitudes**

The uniform scaling of the amplitudes is changing the level of the sound. The JND of the change of loudness lies between 3/4 dB and 3 dB, slightly higher than the 1 dB 'element' mentioned in [106] (page 175).

## **5.5 Conclusion**

Preliminary experiments on the perception of complex sounds using the timbre model shows promising results. A methodology has been developed, which is easy and fast to use, although more precise methods probably exist. More (and better-defined) experiments are necessary, both for assessing the perception as a function of pitch, loudness, and other parameters, but also for assuring that the results are trustworthy.

## **6 Expressive Additions to the Timbre Model**

The timbre model has proven its value in a number of applications, including the analysis of different expression styles [52]. This analysis will here serve as the basis for the inclusion of a number of parameters, which govern the behavior of the timbre model attributes when used in a synthesis context. The expressions are treated in an analysis/synthesis context, and adapted for real-time synthesis [61], where possible.

The expressive modes introduced to the timbre model include variants, pitch, intensity, vibrato and tremolo, and other expressions, such as legato/staccato.

The attempt to find expressive parameters that can be included in the timbre model is an initial study. Both additional literature studies and field test using the timbre engine [61] must be performed in order to assert the validity of the expression model. Of course, the music performance studies [31] gives a lot of information, which can also be gathered from research dealing with the acoustics [5], [9] or the physics [27] of musical instruments. In addition, appropriate gestures must be associated with the expression parameters [51], [101], [102] [43]. The vibrato problem [18] is an interesting topic, stating, for instance, whether to add cycles (vibrato) or stretch (glissando) a time-scaled continuous expression parameter. This is not a problem in the timbre model, because of the division into clean envelopes, where only the sustain part is scaled, and periodic and non-periodic irregularities, which are not scaled.

### **6.1 Variants**

In this work, the sound of a musical instrument is divided into an identity part, and an expression part, where the identity is the neutral sound, or what is common in all the expression styles of the instrument, and the expression is the



change/additions to the identity part introduced by the performer. The expression can be seen as the acoustic outcome of the gesture manipulation of the musical instrument. This division, however, is not simple in an analysis/synthesis situation, since the sound must be played to exist, and it thereby, by definition, always contains an expression part. One attempt at finding the identity is by introducing the notion of variants, which is assumed to be the equivalent of the sounds coming from several executions of the same expressive style.

The variants are calculated by introduced an ideal curve to the different timbre attributes. The deviation from this ideal curve is then assumed to be stochastic, with a given distribution, and for each new execution, a new instance of the deviation is created, giving the sound a clearly altered timbre. Some of the timbre attributes have ideal curves corresponding to some perceptual or physical reality, such as the brightness creation function [52] for the spectral envelope,

$$a_k = a_0 \left( \frac{B}{B-1} \right)^{-k} \quad (7)$$

where  $a_k$  is the amplitude of the partial  $k$ ,  $a_0$  is the fundamental amplitude, and  $B$  is the estimated brightness [6], see equation (9), and the equation for the ideal stiff string for the frequencies [26],

$$f_k = kf_0 \sqrt{1 + \beta k^2} \quad (8)$$

where  $\beta$  is the inharmonicity coefficient. Studies of the discrimination of inharmonicity can be found in, for instance, [46] and [80].

Most timbre attributes, however, are fitted with a simple exponential curve,

$$c_k = v_0 \cdot e^{v_1 k} \quad (9)$$

where  $v_0$  is the fundamental value and  $v_1$  is the exponential coefficient. This curve can model both almost linear curves with small  $v_1$ , but also exponential behaviors.

The parameters of the curves are found by minimizing the lms error using the Levenberg-Marquardt algorithm [72], except for the spectral envelope curve, which is created from the estimated brightness [6],

$$B = \left( \sum_{k=1}^N k a_k \right) / \sum_{k=1}^N a_k \quad (10)$$

The deviations from the ideal timbre attributes parameters are now calculated as,

$$d_k = c_k - \hat{c}_k \quad (11)$$

where  $c_k$  are the estimated timbre attribute parameters (amplitude, frequency, or other parameter), and  $\hat{c}_k$  are the ideal parameters and  $d_k$  is the deviation (assumed to be white gaussian noise).

The deviation  $d_k$  between the clean exponential curve and the estimated attributes is assumed to be

related to the execution of the sound, and the error can, if modeled properly, introduce new executions of the same sound, i.e. of the same instrument, player and style, in the same environment.

Although the clean curves generally fit well with the estimated parameters, there can be discrepancies caused by bad parameter estimation, correlated deviations between attributes, or inadequate modeling. These discrepancies generally do not make artifacts in the original timbre attribute generated sounds, but they sometimes make the variant sounds too different. One way of minimizing this phenomenon is by using weighted curve-fits, which does remove some of the discrepancies. However, since the heavily different variants may be desired, the variants influence is scaled,

$$\tilde{c}_k = \hat{c}_k + v \cdot \hat{d}_k + (1-v) \cdot d_k \quad (12)$$

where a variant scaling ( $v$ ) of zero gives the original timbre attributes, and a scaling of one gives entirely new timbre attribute deviations ( $\hat{d}_k$ ) for each execution. The total deviations can additionally be weighted [61], permitting more, or less, deviations from the identity of the sound.

## 6.2 Pitch

The modeling of the pitch evolution of the timbre attributes is important when executions for different notes are not available. This is the case, for instance, when creating new, or altered, timbre model parameters sets. Obviously, it's impossible to make pitch rules that encompasses all possible musical instruments, so instead, a simple interpolation scheme has been devised, which assures the proper modification of at least the important spectral envelope. In this scheme, all the timbre attributes are assumed to have the partial frequencies in the x axis, and the timbre attributes for the new pitch is found by interpolating between the neighboring values for each partial frequency. The values outside the original frequencies are found by using the extreme values. Although this scheme is rather simple, it has the double advantage of handling the most important timbre attribute correctly, and assuring continuous variations in the resulting sound, as the pitch is altered. In addition, the new timbre attribute values should be interpolated after each pitch change, thereby allowing for subtle effects, caused by for instance resonances. When adding vibrato, the sound would have a continuously varying timbre, thereby adding more life to the execution.

The scheme does not handle sound level, however. In the work on generative rules for music performance, [29] suggest a rule, which raises the sound level 3 dB/octave. An initial study have shown that this effect is present in many musical instruments, although ranging from below 3 (violin) to around 3 (piano, clarinet) to 10 (flute), and 15 dB/octave (soprano). This effect is therefore

parameterized (N dB/octave) and included into the expressive model.

### 6.3 Intensity

The intensity, or velocity, is another important expression attribute. In general, when increasing the velocity, blowing force, or bow velocity, etc., two things happen, the sound emitted gets louder, and it gets brighter. [10] showed that the increase in amplitude and brightness is asymptotic, i.e. the value changes less as velocity grows. In addition, it was shown that the change of brightness in the piano when changing the velocity of the hammer is governed by a straight line in the Hz/dB domain. Therefore, this is the model chosen for the intensity. The amplitude and spectral tilts (slope of the straight line) have an exponential form,

$$val = (v_M - v_m \cdot e^{-\beta \cdot v}), \quad (13)$$

where  $val$  can be either amplitude or spectral tilt [10], and  $v_M$  and  $v_m$  defines the maximum and minimum values,  $\beta$  the sensibility and  $v$  is the velocity.

The values of  $v_M$ ,  $v_m$  and  $\beta$  can be determined from the instrument, if enough velocity executions are available [10], or it can be user defined. In particular, the upper spectral tilt slope is a useful expression parameter, since it defines how bright the sound becomes, when increasing the velocity to a maximum. This model is also consistent with the analysis of piano executions performed in [52], which showed that the change of velocity (of the piano hammer) only affected the spectral envelope, except for an as yet unexplained change in the shimmer and jitter correlations.

### 6.4 Duration

The duration is of course also a very important expression parameter. Since the timbre model encompasses both percussive and sustained sounds, a general strategy for modifying the length of the sound is necessary. This strategy could be found by following the clean curve of the sustain/decay part of each partial, with the given curve form,

$$\tilde{a}_k(t) = a_0 + (a_T - a_0) \frac{c^{t/T} - 1}{c - 1}, \quad (14)$$

where  $c$  is the curve form coefficient,  $T$  is the segment duration, and  $a_0$  and  $a_T$  are the start and end values. Unfortunately, the curve form is sometimes fitted to part of the release segment, or sometimes only part of the sustain/decay segment is used, therefore the curve form is error prone. Instead, the decay is assumed to be logarithmic, and modeled as a straight line in the dB domain,

$$\hat{a}^{dB}(t) = \hat{a}_0^{dB} + bt, \quad (15)$$

where  $\hat{a}_0$  and  $b$  are found by fitting to the known split-point amplitude/time values. If the execution is lasting (played a long time), and the second

split-point has a higher amplitude than the first one ( $b > 0$ ), then clipping will eventually occur. The perceptual effect of this has been judged to be interesting enough to not prevent this happening. Instead, the amplitude of each partial is limited to a value at a user-defined percentage above the maximum value of the partial. Obviously, if this limit is set to 100%, then no crescendo effect is possible.

[29] tentatively suggests a modification to the attack and decay in durational contrasts. This is an interesting inclusion, but this effect has not been found [47] in relation to this work (these durational contrasts were not part of the sound material under test), and it is not included in the model.

### 6.5 Vibrato and Tremolo

The vibrato and tremolo are important expression parameters with specific values defined by the musical context in which the execution is produced. Therefore, a generative rule for the addition of vibrato and tremolo is not desirable in this work. Some vibrato or tremolo effects are, however, part of the identity of the sound, and these effects should not be user-controlled, but inherent in the sound. In particular, this is the case for the amplitude modulation caused by the beating of different modes or strings in, for instance, the piano [104].

The vibrato and tremolo are generally defined by three parameters, the rate (speed), the strength and the initial delay. [31] reviews several vibrato studies, and reports the vibrato rate of professional singers to be between 5.5 to 8 Hz, with a strength between one-tenth of a full tone to one full tone, averaging 0.5 to 0.6 of a whole-tone step.

[83] models the vibrato with the sum of a number of sinusoidals with time-varying amplitudes and phases. The phase of the vibrato/tremolo is necessary, if the resulting sound should be perceptually identical to the original one. Care must be taken to model the resonances correctly when adding vibrato [69]. In addition, the perceived pitch of tones with vibrato is also an important research field [16].

In order to assert whether a sound contains a vibrato or tremolo expression, or whether it contains periodic vibrations in its identity, two things can be examined. First, if the partials are heavily correlated, secondly, if the rate and strength values are correlated, then the chances of it containing expressive vibrato/tremolo is great. If neither of the two cases occur, periodicity is assumed to be part of the identity of the sound, and not controlled by the performer. If expression periodicity is found, it is removed from the sound, and only added back, if and when the performer is signifying it.

## 6.6 Other expressions

The expressions can be any kind of acoustic change resulting from manipulation of the music instrument.

The other expression parameters used in classical music include mainly styles (legato/staccato, for instance) and tempi. Since some of these expressions are controlled continuously by the performer, they are not easily integrated into the timbre model. In particular, no good gesture capture device has been available to perform tests. In addition, not much timbre attribute changes have been found when analyzing executions of different styles [52]. Therefore, the conclusion must be that the styles are mainly a matter of duration, which is easily controlled in this model.

Another important expression is the transition [94]. Since the transition is the time-varying amplitude, fundamental frequency, and timbre, it should not be too difficult to create timbre attribute sets with appropriate values for different transition.

Finally, another possible expression is the generic timbre navigation [105], [84]. In this case, the timbre is manipulated using sensors and various mapping strategies [51], [102].

## 7 Other Applications

The timbre model has been successfully used in many applications, in particular in classification and sound manipulation, including morphing.

### 7.1 Classification

Classification is an important topic today, for instance with the growing number of on-line sound samples.

The timbre model was used as the basis for sound classification [52], in which a subset of the timbre attributes (16 attributes from each sound) was used to classify 150 sounds in 5 instrument classes with no errors. Binary tree classification, using approximately the same data set, was presented in [49], giving much information about the importance of the timbre attributes in the classification.

A real-time learning classification scheme was presented in [30]. For a recent overview of musical instrument classification methods, see [42].

### 7.2 Sound manipulation

Morphing and sound manipulation is another application in which the timbre model has shown its value. Since the timbre model parameter set has a fixed size (except for the number of partials), it is easy to morph between the sounds by simply interpolating different timbre sets. The interpolated timbre model parameters can also be used to manipulate the additive parameters [52]. A similar strategy, but with an unequal number of features was presented in [96]. [84] interpolates the additive parameters directly, and [89] uses the spectral

envelope to modify sounds. Other morphing strategies include the Karaoke impersonator [58]. [93] uses results from auditory scene analysis to warp and interpolate between two sounds. Finally, a simplified timbre model [52] was presented in [50] for use in prototyping of musical sounds.

## 8 Conclusion

The paper gives an overview of the perceptual research and analysis/synthesis of the additive parameters. It presents the timbre model, a signal model, which models most musical sounds with a high fidelity. The timbre model is based on perceptual research and analysis by synthesis. It consists of a spectral envelope, frequencies, a five-segment amplitude envelope model with individual split-point times, values and curve form, and amplitude and frequency irregularities. The model parameters are perceptually relevant, and the model has previously been shown to have a high sound quality. In addition, a preliminary research result shows the just noticeable differences of most of the timbre model parameters. These results are interesting in designing the user interface.

An important addition to the model, introduced in this paper, is the expression model, which enhances the use of the model in a musical context. The expression model models the behavior of the timbre attributes, when playing a different note, intensity, duration, or other style. It also introduces the vibrato and tremolo into the model. A major contribution is the notion of variants, in which the timbre is divided into an identity part, and an expression part. The variant is then a stochastic model of the deviation from the identity part of the sound, with a new instance at each new execution. Therefore, an interesting novelty is added to each new sound emitted, even if it's played with the same pitch, loudness and duration.

## References

- [1] Allen, J. B. *Short term spectral analysis, synthesis and modification by discrete fourier transform*. IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 3, June. 1977.
- [2] American Standard Association. *Acoustical Terminology*, New York, 1960.
- [3] Andersen, T. H., K. Jensen. *Phase modeling of instrument sounds based on psycho acoustic experiments*, Workshop on current research directions in computer music, Barcelona, Spain, 2001.
- [4] Auger F., P. Flandrin. *Improving the Readability of Time Frequency and Time Scale Representations by the Reassignment Method*, IEEE Transactions on Signal Processing, vol. 43, pp. 1068-1089, 1995.

- [5] Backus, J. *The acoustical foundation of music*. John Murray Ltd. London, 1970.
- [6] Beauchamp, J. *Synthesis by spectral amplitude and "Brightness" matching of analyzed musical instrument tones*. J. Acoust. Eng. Soc., Vol. 30, No. 6. 1982.
- [7] Barrière, J-P (editor). *Le timbre, métaphore pour la composition*, C. Bourgeois Editeur, IRCAM, 1991.
- [8] Beauchamp, J. W. *Unix workstation software for analysis, graphics, modifications, and synthesis of musical sounds*, 94<sup>th</sup> AES Convention, preprint 3479, Berlin, Germany. 1993.
- [9] Benade, A. H. *Fundamentals of musical acoustics*. Dover publications inc. New York. 1990.
- [10] Bensa J., K. Jensen, R. Kronland-Martinet, S. Ystad. *Perceptual and Analytical Analysis of the effect of the Hammer Impact on the Piano Tones*, Proceedings of the ICMC, Berlin, Germany. 2000.
- [11] Borum, S., K. Jensen. *Additive Analysis/Synthesis Using Analytically Derived Windows*, Proceedings of the DAFX, Trondheim, Norway, 1999.
- [12] Bregman, A. S. *Auditory Scene Analysis*, The MIT Press, Massachusetts. 1990.
- [13] Chaudhary, A. S. *Perceptual Scheduling in Real-time Music and Audio Applications*, Doctoral dissertation, Computer Science, University of California, Berkeley, CA. 2001.
- [14] Chen, S. S., D. L. Donoho, M. A. Saunders. *Atomic decomposition by basis pursuit*. Dept. of Statistics Technical Report, Stanford University, February, 1996.
- [15] Cheung, N-M., A. B. Horner. *Group synthesis with genetic algorithms*. J. Audio Eng. Soc. Vol. 44, No. 3, March. 1996.
- [16] d'Alessandro, C., M. Castellengo. *The pitch of short-duration vibrato tones*, J. Acoust. Soc. Am, 95(3), pp. 1617-1630, March. 1994.
- [17] Depalle Ph., T. Hélie. *Extraction of spectral peak parameters using a short time Fourier transform modeling and no sidelobe windows*. Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, Oct. 1997.
- [18] Desain, P., H. Honing. *Time functions function best as functions of multiple times*. Computer Music Journal, 16(2), 17-34, 1992.
- [19] Ding, Y., X. Qian. *Processing of musical tones using a combined quadratic polynomial-phase sinusoid and residual (QUASAR) signal model*, J. Audio Eng. Soc. Vol. 45, No. 7/8, July/August 1997.
- [20] Dubnov, S., N. Tishby, D. Cohen. *Investigation of frequency jitter effect on higher order moments of musical sounds with application to synthesis and classification*. Proc of the Int. Comp. Music Conf. 1996.
- [21] Dubnov, S., X. Rodet. *Statistical modeling of sound aperiodicity*. Proc of the Int. Comp. Music Conf. 1997.
- [22] Eaglestone, B., S. Oates. *Analytic tools for group additive synthesis*. Proc. of the ICMC, 1990.
- [23] Fitz, K., L. Haken, P. Christensen. *Transient Preservation Under Transformation In an Additive Sound Model*, Proceedings of the International Computer Music Conference, Berlin, Germany. August 2000,
- [24] Fitz, K., L. Haken. *Bandwidth enhanced modeling in Lemur*. Proc. of the ICMC, 1995.
- [25] Fitz, K. Project: *Loris*. <<http://sourceforge.net/projects/loris>>, 08/10 2001.
- [26] Fletcher, H. *Normal vibrating modes of a stiff piano string*, J. Acoust. Soc. Am., Vol. 36, No. 1, 1964.
- [27] Fletcher, N. H., T. D. Rossing. *The physics of musical instruments*, Springer-Verlag. 1990.
- [28] Freedman, M. D. *Analysis of musical instrument tones*. J. Acoust. Soc. Am. Vol. 41, No. 4, 1967.
- [29] Friberg, A. *Generative rules for music performance: A formal description of a rule system*. Computer Music Journal, Vol. 15, No. 2, summer 1991.
- [30] Fujinaga, I., K. MacMillan. *Realtime recognition of orchestral instruments*. Proc. of the ICMC. 2000.
- [31] Gabrielson, A. *The performance of music, in The psychology of music*, D. Deutsch (editor), pp. 501-602, AP press, San Diego, USA, 2<sup>nd</sup> edition, 1999.
- [32] Gordon, J. W. *The perceptual attack time of musical tones*. J. Acoust. Soc. Am. 82(2), July 1987.
- [33] Grey, J. M. *Multidimensional perceptual scaling of musical timbres*. J. Acoust. Soc. Am., Vol. 61, No. 5, May 1977.
- [34] Grey, J. M., J. A. Moorer. *Perceptual evaluation of synthesized musical instrument tones*, J. Acoust. Soc. Am., Vol. 62, No. 2, August 1977.
- [35] Grey, J. M., J. W. Gordon. *Perceptual effects of spectral modification on musical timbres*. J. Acoust. Soc. Am. Vol. 63, No. 5, May 1978.
- [36] Gribonval, R., P. Depalle, X. Rodet, E. Bacry, S. Mallat. *Sound signal decomposition*

- using a high resolution matching pursuit. Proc. ICMC, 1996.
- [37] Guillemain, P. *Analyse et modélisation de signaux sonores par des représentations temps-frequence linéaires*. Ph.D. dissertation. Université d'Aix-Marseille II, 1994.
- [38] Hajda, J. M., R. A. Kendall, E. C. Carterette, M. L. Harschberger. *Methodological issues in timbre research*, in *The psychology of music*, D. Deutsch (editor), pp. 253-306, AP press, San Diego, USA, 2<sup>nd</sup> edition, 1999.
- [39] Harris, F. J. *On the use of windows for harmonic analysis with the discrete Fourier transform*. Proc IEEE, Vol. 66, No. 1, January 1978.
- [40] Helmholtz, H. *On the Sensations of Tone*, reprinted in 1954 by Dover, New York, 1885.
- [41] Herrera, P., J. Bonada. *Vibrato extraction and parameterization in the spectral modeling synthesis framework*, Proc. DAFx, Barcelona, Spain, 1998.
- [42] Herrera, P., X. Amatriain, E. Batlle, X. Serra. *Towards Instrument Segmentation for Music Content Description: a Critical Review of Instrument Classification Techniques*, Proceedings of International Symposium on Music Information Retrieval, 2000.
- [43] Hunt A., M. Wanderley. *Interactive Systems and Instrument Design in Music Working Group*, <<http://www.notam.uio.no/icma/interactive/systems/wg.html>>, October 4, 2001.
- [44] Iverson, P., C. L. Krumhansl. *Isolating the dynamic attributes of musical timbre*. J. Acoust. Soc. Am. 94(5), November 1993.
- [45] Jaffé, D. A., J. O. Smith. *Extension of the Karplus-Strong plucked-string algorithm*. Computer Music Journal, Vol. 7, No. 2, summer 1983.
- [46] Järvinen, H., V. Välimäki, M. Karjalainen. *Audibility of inharmonicity in string instrument sounds, and implications to digital sound synthesis*. ICMC Proceedings, Beijing, China, 359-362. 1999.
- [47] Jensen, K. *Envelope Model of Isolated Musical Sounds*, Proceedings of the DAFX, Trondheim, Norway, 1999.
- [48] Jensen, K., G. Marentakis, *Hybrid Perception*, Papers from the 1<sup>st</sup> Seminar on Auditory Models, Lyngby, Denmark, 2001.
- [49] Jensen, K., J. Arnsfang. *Binary Tree Classification of Musical Instruments*, Proceedings of the ICMC, Beijing, China, 1999.
- [50] Jensen, K. *Pitch Independent Prototyping of Musical Sounds*, Proceedings of the IEEE MMSP Denmark, 1999.
- [51] Jensen, K. *The Control of Musical Instruments*, Proceedings of the NAM. Helsinki, Finland. 1996.
- [52] Jensen, K. *Timbre Models of Musical Sounds*, PhD. Dissertation, DIKU Report 99/7, 1999.
- [53] Kleczowski, P. *Group Additive Synthesis*. Computer Music Journal 13(1), 1989.
- [54] Krimphoff, J., S. McAdams, S. Winsberg. *Caractérisation du timbre des sons complexes. II Analyses acoustiques et quantification psychophysique*. Journal de Physique IV, Colloque C5, Vol. 4. 1994.
- [55] Kronland-Martinet, R. *The wavelet transform for analysis, synthesis, and processing of speech and music sounds*. Computer Music Journal, 12(4), 1988.
- [56] Lewitt, H. *Transformed Up-Down Methods in Psychoacoustics*. The Journal of the Acoustical Society of America, 49/29, 467-477, 1970.
- [57] Lindeberg, T. *Edge detection and ridge detection with automatic scale selection*, CVAP Report, KTH, Stockholm, 1996.
- [58] Loscos, A., P. Cano, J. Bonada, M. de Boer, X. Serra. *Voice Morphing System for Impersonating in Karaoke Applications*, Proceedings of International Computer Music Conference 2000.
- [59] Marchand, S. *Musical sound effects in the sas model*, Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99), NTNU, Trondheim, December 9-11, 1999.
- [60] Marchand, S. *Improving Spectral Analysis Precision with Enhanced Phase Vocoder using Signal Derivatives*. In Proc. DAFX98 Digital Audio Effects Workshop, pages 114--118, Barcelona, November 1998.
- [61] Marentakis, G., K. Jensen. *Timbre Engine: Progress Report*, Workshop on current research directions in computer music, Barcelona, Spain, 2001.
- [62] Marques, J. S., A. J. Abrantes. *Hybrid harmonic coding of speech at low bit-rates*, Speech Communication 14, 1994.
- [63] Martínez J. M. *Overview of the MPEG-7 Standard (version 5.0)*, <<http://mpeg.telecomitalia.com/standards/mpeg-7/mpeg-7.htm>>, Singapore, March 2001.
- [64] Matthews, M. V., J. E. Miller, E. E. David. *Pitch synchronous analysis of voiced speech*. J. Acoust. Soc. Am. Vol. 33, No. 2, February 1961.

- [65] McAdams, S., J. W. Beauchamp, S. Meneguzzi. *Discrimination of musical instruments sounds resynthesized with simplified spectrotemporal parameters*, JASA 105(2), 1999.
- [66] McAdams, S., S. Winsberg, S. Donnadiou, G. de Soete, J. Krimphoff. *Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes*. Psychological Research, 58, pp. 177-192. 1992.
- [67] McAulay, R. J., T. F. Quatieri. *Speech analysis/synthesis based on a sinusoidal representation*, IEEE Trans. on Acoustics, Speech and Signal Proc., vol. ASSP-34, No. 4, August 1986.
- [68] McIntyre, M. E., R. T. Schumacher, J. Woodhouse. *Aperiodicity in bowed-string motion*, Acustica, Vol. 49, 1981.
- [69] Mellody, M., G. H. Wakefield. *A model distribution study of the violin vibrato*. Proc ICMC, 1997.
- [70] Møller, A. *Akustisk guitar syntese*. Master Thesis, Computer Science Department, University of Copenhagen, 1996.
- [71] Moorer, J. A. *The synthesis of complex audio spectra by means of discrete summation formulas*. J. Audio. Eng. Soc. Vol. 24, No. 9, November 1976.
- [72] Moré, J. J. *The Levenberg-Marquardt algorithm: Implementation and theory*. Lecture notes in mathematics, Edited by G. A. Watson, Springer-Verlag, 1977.
- [73] Pielemeier, W. J., G. H. Wakefield, M. H. Simoni. *Time-frequency analysis of musical signals*. Proc. of the IEEE, Vol. 84, No. 9. September 1996.
- [74] Rayleigh, J. W. S. *The theory of sound*. reprinted in 1945 by Dover Publications. MacMillan company 1896.
- [75] Recommendation ITU-R 85/10. *Methods for the subjective assessment of small impairments in audio systems, including multichannel sound systems*. International Telecommunication Union, Geneva, Switzerland. 16 March 1994.
- [76] Richard, G., C. d'Allesandro. *Analysis, Synthesis and modification of the speech aperiodic component*, Speech Communication 19, 1996.
- [77] Richard, G., C. d'Allesandro, S. Grau. *Musical noises synthesis using random formant waveforms*. Stockholm Music Acoustic Conference, Pub. of the royal Academy of Music, Stockholm, Sweden 1993.
- [78] Risset, J-C. *Timbre analysis by synthesis: Representation, imitation and variants for musical composition*. in Representations of musical signals. G. de Poli, A. Piccialli, C. Roads, Editors. The MIT Press, London 1991.
- [79] Röbel. A. *Adaptive Additive Synthesis of Sound*, Proceedings of the ICMC, Berlin, Germany, 1999.
- [80] Rocchesso, D., F. Scalcon. *Bandwidth of perceived inharmonicity for physical modeling of dispersive strings*. IEEE Transactions on Speech and Audio Processing, 7(5):597-601, September 1999.
- [81] Rodet, X., P. Depalle, G. Poiret. *Speech analysis and synthesis methods based on spectral envelopes and voiced/unvoiced functions*. European Conference on Speech Technology, Edinburgh, September 1987.
- [82] Rodet, X. *The additive analysis-synthesis package*, <<http://www.ircam.fr/equipes/analyse-synthese/DOCUMENTATIONS/additive/index-e.html>>, 08/10 2001.
- [83] Rossignal, S. *Segmentation et indexation des signaux sonores musicaux*, Thèse de doctorat de l'université Paris VI, Paris, France 2001.
- [84] Rován, J. B., M. M. Wanderley, S. Dubnov, Ph. Depalle. *Instrumental Gestural Mapping Strategies as Expressivity Determinants in Computer Music Performance*, Kansei- The Technology of Emotion Workshop - Genova - Italia, Oct. 3/4, 1997.
- [85] Sandell, G. *Definitions of the word "Timbre"*, <<http://sparky.parmly.luc.edu/sandell/sharc/timbred.html>>, 08/10 2001.
- [86] Sandell, G. J., W. L. Martens. *Perceptual evaluation of principal-component-based synthesis of musical timbres*. J. Acoust. Soc. Am. Vol. 43, No. 12, December 1995.
- [87] Schaeffer, P. *Traité des objets musicaux*, Editions de Seuil, 1966.
- [88] Schumacher, R. T., C. Chafe. *Characterization of aperiodicity in nearly periodic signals*. J. Acoust. Soc. Am. Vol. 91 No. 1, January 1992.
- [89] Schwarz, D., X. Rodet. *Spectral Envelope Estimation and Representation for Sound Analysis-Synthesis*, Proc. ICMC, pp 351-354, Beijing, China, 1999.
- [90] Sekey, A., B. A. Hanson. *Improved 1-bark bandwidth auditory filter*. J. Acoust. Soc. Am. Vol. 75, No. 6, 1984.
- [91] Serra, X., J. Smith. *Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic*

- decomposition*, Computer Music Journal, vol. 14, No. 4, winter 1990.
- [92] Singh P. G., A. S. Bregman. *The Influence of different timbre attributes on the perceptual segregation of complex-tone sequences*, Journal of the Acoustical Society of America 1,02 1997.
- [93] Slaney, M., M. Covell, B. Lassiter. 1996. *Automatic audio morphing*, Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 2, 1001-1004, 1996.
- [94] Strawn, J. *Orchestral Instruments: Analysis of performed transitions*, J. Audio Eng. Soc, Vol. 34, No. 11, pp. 867-880. November 1986.
- [95] Strong, W., M. Clark. *Synthesis of wind-instrument tones*. J. Acoust. Soc. Am. Vol. 41, No. 1, 1967.
- [96] Tellman E., L. Haken, B. Holloway. *Timbre morphing of sounds with unequal numbers of features*. J. Audio Eng. Soc. Vol. 43, No. 9, September 1995.
- [97] Truax, B. *Discovering inner complexity: Time shifting and transposition with a real-time granulation technique*. Computer Music Journal. 18(2), summer 1994.
- [98] Verma, T. S., S. N. Levine, T. H. Y. Meng. *Transient Modeling Synthesis: A flexible analysis/synthesis tool for transient signals*, Proceedings of the ICMC, 1997.
- [99] von Bismarck, G. *Timbre of steady sounds*. Acustica, Vol. 30, 1974.
- [100] Vos, K., R. Vafin, R. Heusdens, W. B. Kleijn. *High quality consistent analysis-synthesis in sinusoidal coding*, AES 17th Conference High Quality Audio Coding, Florence, Italy, 1999.
- [101] Wanderley M., M. Battier (eds). *Trends in Gestural Control of Music*. Ircam - Centre Pompidou, 2000.
- [102] Wanderley, M. *Interaction musician-instrument: Application au contrôle gestuel de la synthèse sonore*. Thèse de doctorat de l'université Paris VI, Paris, France 2001.
- [103] Warren, R. M. *Auditory Perception: A new analysis and synthesis*. Cambridge university press, Cambridge, 1999.
- [104] Weinreich, G. *Coupled piano strings*, J. Acoust. Proc. Am., Vol 62, No. 6, pp. 1474-1484, December 1977.
- [105] Wessel, D. *Timbre space as a musical control structure*, Computer Music Journal 3(29, pp.45-52, 1979.
- [106] Zwicker, E., H. Fastl. *Psycho-acoustics: Facts and models*, 2<sup>nd</sup> edition, Springer-verlag, Heidelberg, Germany, 1999.

# Directional patterns and recordings of musical instruments in auralizations

Felipe Otondo

Jens Holger Rindel

Claus Lynge Christensen

Ørsted DTU, Acoustic Technology, Technical University of Denmark

{fo,jhr,clc}@oersted.dtu.dk, <http://www.dat.dtu.dk>

## Abstract

This paper outlines the developing ideas of the investigation started the 1<sup>st</sup> of September 2001. The problem of the spatial representation of sound sources that vary their directional pattern in time in auralizations is introduced. Musical instruments are used as a reference for the discussion of the traditional representations with assumed fixed directional characteristics. A new method for representation of the spatial sound contributions in time is proposed using multiple-channel recordings and virtual sources in the simulations. Further developments and applications of the solution are outlined.

## 1 Introduction

Auralization is the analogous term to visualization introduced to describe rendering audible (imaginary) sound fields. The aim of a room auralization is to simulate as accurately as possible the binaural listening experience at a given position in a modeled space [1]. The directional characteristics of the sound source as well as the acoustical environment give important clues in auralization. Musical instruments, as any other acoustical sources, create a particular acoustical behavior in a room. The aim of this investigation is to study in deep this behavior and improve the auralization by optimizing the representation of the directional characteristics of musical instruments in order to model better the room/instrument interaction. It is also the goal of this work to make recordings and measurements of musical instruments that could be useful within the MOSART network.

## 2 Directional characteristics of musical instruments

Musical sounds require a complex acoustical analysis due to their particular features. Analyzing sounds produced by musical instruments involves a great deal of information such as harmonic structure, spectra, time transients, noise components, directional attributes and others. Like the sonic spectrum, the directional attributes of a musical instrument change with the different notes played on the instrument [2], the different performing intensities [3], the different techniques and also with the different performers of

the same instrument. These changes, due to the complexities of the musical instrument itself as a multi resonating system and other more complex reasons, are different for the diverse families and types of musical instruments [2]. Figure 1 shows the measured directional characteristics of an alto saxophone at 1000 Hz for two different notes played in the same octave by the same player [4].

## 3 Auralization with musical instruments as sources

In order to consider musical instruments as sound sources for auralization one needs to include the directional characteristics of the source to be able to specify the source radiation characteristics. As we have seen, musical instruments are sound sources which have a complex radiation pattern that is difficult to describe with accuracy in a real performance case where there will always be directivity changes in time. If we assume a fixed directional characteristic for each of the frequency bands, like it would be the case of a loudspeaker, the representation of the directivity of the source would be very poor and inaccurate. The real directional characteristics would be changing in time and we would be having the wrong directional pattern most of the time, emphasizing or diminishing the level for certain frequencies of the particular spectra. Therefore, a better representation of the sound intensity changes in time is needed.



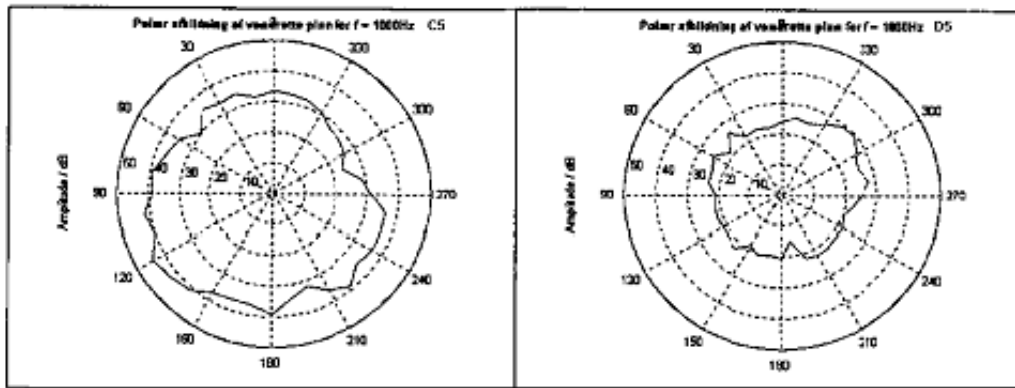


Figure 1 Polar diagram for 1000 Hz of an alto saxophone playing a C5 (left) and a D5 (right).

## 4 Improvement of the spatial representation of sound

A better description of the spatial sonic contributions of a musical instrument, or any source that changes its directivity in time, could be offered by considering the contributions of the acoustical intensity. One way to do this without referring to a determined fixed directional characteristic is to use several virtual sources in the simulation with fixed and neutral directional characteristics that do not overlap each other. The new source (all the virtual sources) should radiate a multi-channel recording of the source by each of the virtual sources simultaneously. This recording should be done in such a way that the different microphones achieve the sound from the source in different directions, as shown in Figure 3 for a 4-track recording. Also each of the virtual sources should reproduce the multi-track anechoic recordings corresponding to the orientation relative to the original source. That is, if for example we divide the radiation semi-sphere of the upper part of the instrument in 4 (assuming we have made a 4-channel anechoic recording), we can then simulate 4 discrete sources with a directional characteristic equal to a quarter of the semi-sphere in the direction of each anechoic recording. Figure 3 shows a room acoustic simulation using the software Odeon where an auralization considering 4 virtual sources was done, each source with an omnidirectional characteristic of a quarter of a semi-sphere and radiating in 0, 90, 180 and 270 degrees [5]. In this case the anechoic recordings include the sound pressure variations in time of the source, which were radiated by the virtual sources in their discrete sectors. The new source (constituted by the 4 virtual sources) radiates in a distinctive way in

each of the 4 directions following changes in level, asymmetries and orientation of the original source.

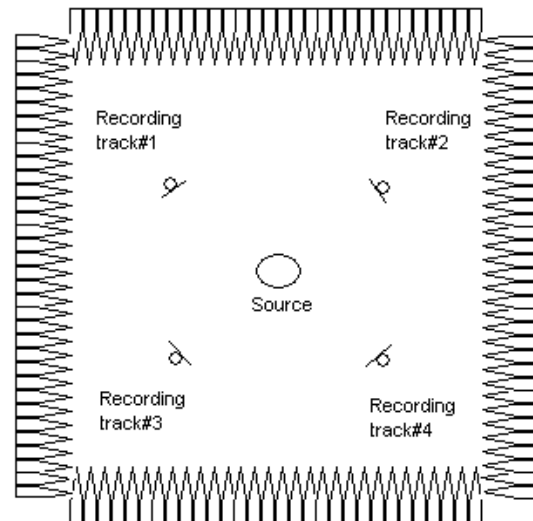


Figure 2 Setup for a 4-track anechoic recording of a source.

## 5 Further developments

In the near future we have planned to study in deep the relationship between the directional characteristics of musical instruments and auralizations applying the multi-channel recording method under different circumstances. The influence of the room acoustics characteristics, the kind of source used for the auralizations and the way the recordings are done are some of the points to be considered in our next steps. We would also like to expand the investigation to the study of auralizations with large sources (multiple musical instruments) using virtual sources in different positions.

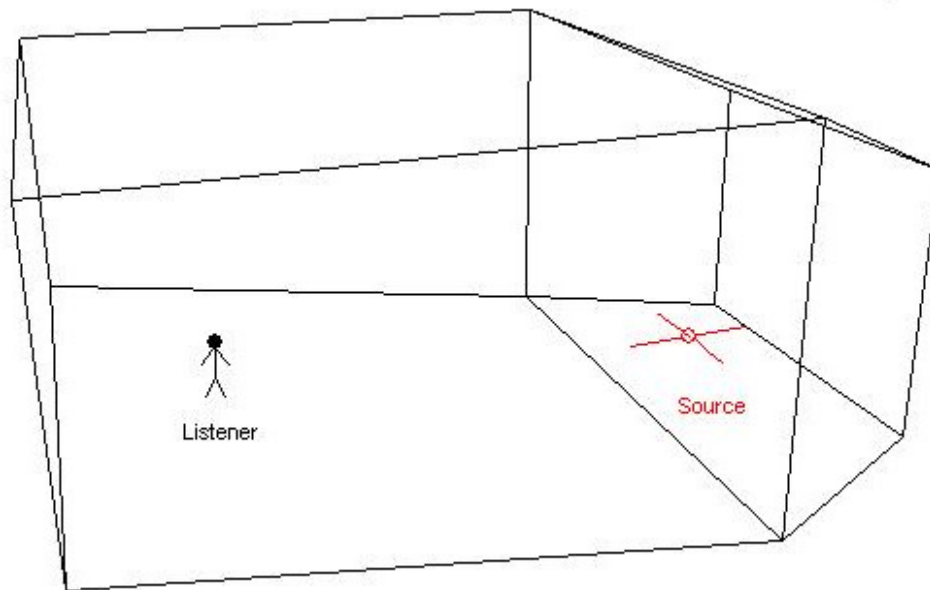


Figure 3 Room acoustic simulation with 4 virtual sources, each with a directional pattern, radiating in 4 different directions.

## References

- [1] Kleiner, M., Dalenback, B. I., and Svensson, P. 1993. Auralization-An Overview. *Journal of the Audio Engineering Society*, 41(11), pp. 861-874.
- [2] Meyer, J. 1978. *Acoustics and the performance of music*. Verlag. pp. 75-102.
- [3] Rossing, T. 1990. *The science of sound*. Second Edition, Addison-Wesley (chapter 11, pp. 231).
- [4] Lisa-Nielsen, M. 2001. *Rumakustik Simulering Blæseinstrumenter*. Master Thesis Report, Technical University of Denmark.
- [5] "The Odeon home page." <http://www.dat.dtu.dk/~odeon>

# TOWARD AN OBJECTIVE METHOD FOR THE TIMBRE ANALYSIS OF SOUND-OBJECTS IN THE ELECTRO-ACOUSTIC MUSIC REPERTOIRE

*Sergio Canazza*

Centro di Sonologia Computazionale (CSC)  
University of Padova  
canazza@dei.unipd.it

*Chiara Marini*

Mirage  
University of Udine  
chiaramarini@yahoo.it

## ABSTRACT

The world of 50's electro-acoustic music is extremely fascinating but presents many problems about music analysis because of the lack of data and references about it. This study wants to be a first step towards a classification of the different sounds and an analysis of their timbres. The method that has been used is the creation of a common database of different sound-objects, then a frequency analysis was carried out. Finally, an attempt was made to re-synthesize them using mathematical algorithms in order to reach a better understanding of compositive process in the field of electro-acoustic music repertoire.

## 1. INTRODUCTION

Electronic music proved to be the ideal medium for the application of the unifying principles of total serialism. All musical parameters - pitch, duration, volume, timbre and the location of sounds in space - could be precisely delineated and controlled. The pitch of a sound could be specified as an exact frequency, in cycles per second, rather than the label by which a note was named - B flat, C sharp or whatever; duration could be measured down to a tiny fraction of a second and the volume of a sound could be enumerated in decibels. But perhaps most importantly, new sounds could be composed from scratch by the fusion of sine-waves, the 'atoms' from which sounds are constructed; sounds that had never been heard before, and for which there were no names, came to life in the studio.

In the western music history the various parameters that characterize the sound were treated in different periods. This is a simplification that aims to simplify the complexity of the evolution of music itself. After a detailed examination of the problems relative to the correlation between pitches, done with the creation of melodies and increasingly complex polyphonies the attention was pointed on the rhythm. This was accomplished by focusing on the complex poly-rhythms of the musical piece which structures are partially derived from extra-European cultures. The last considered parameter, probably also for its complex and difficult to measure structure, was timbre. One of the first example of attention towards timbre was the timbre research

by Debussy and the experience of the "Klangfarbenmelodie" by Schoenberg. An example of composition "for timbres" is given by Arnold Schoenberg's "Farben", which is based upon a formal articulation of timbres that follows one another on fixed chords in order to create particular acoustic images. In this work the instrumental attacks are neutralized for creating a sort of moving sound blur. The ear, in this case, is no more able to clearly distinguish the transition between the various instruments of the orchestra whose attacks are executed in "pianissimo": this kind of operas will be defined by Schoenberg as "Klangfarbenmelodie" (melodies of sounding colors). But it is only with the futurist's music that the musical possibilities is enclosed non-harmonic sounds (noises).

Around mid-nineties composers like Edgard Varèse began to look for the possibility to compose music with machines. The opportunity came out when artists like John Cage realized that with some equipments for measurements, like discs with recorded sinusoidal sounds (usually employed to test the audio systems), it is possible to create new sonorities.

After those, firsts experiences in the comparison of timbre and pitch has been started with the appearance of the three schools of Paris and Colonia first, and then Milan. The idea came from the author's consciousness that, as Stockhausen said [1], the disc, the tape recorder and the radio had deeply changed the relation between music and the listener, but managers and radio producers tried only to reproduce a music which in the past had been written for concert halls and theatres and not to be "frozen" and preserved for being executed by loud speakers. The electronic music is not recorded on tape for being preserved and reproduced but it born from electro-acoustic machines or manipulations of natural sounds. This kind of music can exist only because it is composed on tape and can be listened only with loud speakers. The electro-acoustic music uses tape not just to reproduce but to produce. Stockhausen wrote that when the timbre space will be extended to the continuum of frequencies and to the sounds with non-defined pitch, then the music with "sounds", in the common sense, become only one musical opportunity. It is just this the focal point of the problem related to the analysis of electro-acoustic music: it is the sound-object itself that becomes the cornerstone of the composition, its processing and its bonding with its source which had inspired the composers that give a sense to the composition.

## 2. TIMBRE AND SOUND OBJECT IN ELECTRO-ACOUSTIC MUSIC

At this point we must deal with the strong acousmatic nature which often characterizes the electro-acoustic music and the great variety of sound possibilities, which range from the instrumental or concrete sounds to the pure synthesized ones passing by sounds variously denaturalized and interpolated [2]. This acousmatic nature leads to consider the comprehension of sound objects as the first step toward a complete understanding of an electro-acoustic work and of its structure. It is not by chance that many, if not every, analysis methods proposed until now take the sound object study as a very important step. We can roughly pick out three kinds of approach regarding sound objects. The first one is fundamentally empirical and is typical of the concrete school of Paris where time is the most modified parameter of sound.

The aim is to record on tape existing sounds, like train noise, human scream to traditional instruments, and processing them in order to obtain recorded (i.e. concrete) sounds. In this school the main ways of processing sounds were given by filtering, tape reversal, tape loops, speed changes, tape splices. It is clear that the parameters related to time are the most studied and processed sound parameters. With recorded sounds you can break, recompose and control the time in order to make the sound source unrecognizable, for creating envelopes, durations and speeds never experienced before. The obtained new sonorities were used as basic material for new works. Schaeffer, in his "Traité des objets musicaux" [3] tried to create a table for describing sound objects and recognizing some of their typologies in order to use them to new compositions. This table however is too devoted to the compositional process and, even if this is the first attempt to create an appropriate lexicon for electro-acoustic music, it is sometimes redundant and complex to compile [4].

The second kind of approach is mainly scientific and the electro-acoustic processes are used in order to study natural sounds, for instance by filtering or by reversing the tape, to discover their rules and use the obtained information to create totally synthesized sounds where the source is entirely hidden. In this sense, technology allows to the composer to have a deeper understanding about the sound nature. This approach starts from the Fourier Series. In this case there are no natural sounds, every object must be created by additive synthesis starting from sinusoids with exact pitch, phase and duration.

The composer becomes a sort of "chemist" which reduces sounds to the simples elements and reconstruct, by means of these elements, a completely synthesized sound. It is clear that when the purpose of such manipulation is artistic, the result is "artistic" in the Greek sense, the sum of manual skill and art (in its common sense).

Another chance given by additive synthesis is to have the total control of the sound parameters. Now the composer is able to carry out his ideas without being subject to the rules imposed by the physical characteristics of natural sounds; starting from this, it

was possible "to compose timbre and (to compose) with timbre" [5].

The maximum expression of the composer desire to control all parameters is given by Stockhausen's "Studie II", where pitch, duration and amplitude are all determined through multi-parametric series. The result of those sound processing was square waves assembled in a sort of "chord of sinusoids".

After this work it was understood that it was not sufficient to handle the different timbre parameters to guarantee the hoped variety. This goal can be obtained only by a dynamic correlation of the parameters.

The third kind of approach starts from the timbre researches focused on the intrinsic peculiarities of sound sources, both textural and gesture. In this case the simultaneous presence both of natural sounds and of their processing is preferred. This approach became the synthesis between the previous approaches. A particularly exhaustive example of this kind of opéra is Luciano Berio's "Tema - Omaggio a Joyce": in this opera the language is totally dissolved in sounds: the idea is to cancel the meaning, the semantic part of words, in order to make emerge only the musical properties of the materials for emphasizing those qualities. Destruction and reconstruction of the original sound objects become a single process, in fact while text lose its linguistic meaning, its capabilities of convey structured information, the sounding material as the acoustic residual of language, is structured into sounds and, only in this way, acquires a musical sense.

Smalley differentiated two kind of "source bonding": "mimetic" when the listener perceives the sound closely related to the natural source, and "aural" when the attention of the listener is caught by the intrinsic aspects of the sounding material leaving aside natural source. There is another category besides these ones defined by Schaeffer as "neutral listening" or rather a type of listening which doesn't consider such important the source.

## 3. POSSIBILITIES OF MANIPULATING SOUNDS THROUGH ELECTRO-ACOUSTIC EQUIPMENTS

From this brief (and not exhaustive) outline it is simple to understand that in the electro-acoustic music it is not important the link with the source but a new quality of timbres, something not yet listened, something especially created for recorded music, something that could not exist if not recorded on tape and broadcasted by loud speakers. Afterwards timbre in electro-acoustic music is research considered both as study of natural sounds and sonority, with their spectral characteristics, both as research of new sonorities taken out and inspired by any audible, traditional sound every day life noise and synthesized sounds.

After this brief introduction upon the timbre in electro-acoustic music it is necessary to examine the possibilities offered by electro-acoustic means.

The first opportunity given by tape is to control the "time" by changing the playing speed, obtaining a slower sound (with lower

pitch), or a faster one, with higher pitch. In this case the original sound and its source are often still recognizable. Another possibility is to read the tape reversed. In this case the spectrum and pitch don't change but source is often unrecognizable; in this case it is very difficult, if not impossible, to reconstruct the source. The last possibility of changing time is to cut the tape, to mix the parts of one or more sounds and to paste them in order to obtain particular hybrid sounds or a granular synthesis often losing the link with the source bounding as in figure 1 (bubblings). Another example of electro-acoustic process is given by filters: as said above, it is possible to filter a complex sound and obtain a very simple one: for instance, you can obtain a sinusoidal sound from white noise, or filter some parts of different sounds both natural and synthesized and remix them with many recording in succession or simply with cut and paste. The new technology had a very important role upon the imaginary of music in the Fifties. The experimentation with those new possibilities processing which sometimes preserve the link with source, but several times create new sounds which can stimulate the composer's ideas and the listener's fantasy and attention.

#### 4. CASE STUDIES

In particular, we present the analysis of a Bruno Maderna's "Continuo", a completely electro-acoustic work, in which are used many typical sounds of this kind of works useful to create the "database" for the study of this type of music.

The other analyzed work is "Portrait of Erasmo", in which electro-acoustic music is used to create particular atmospheres giving a sense of indefinite, of obscurity and tension using sounds which our ear perceive as known but can not recognize. It is also used to elaborate demon's voices by distortion and by varying the playing speed of the tape, to underline the contraposition between the idea of a man without freedom, outlined in Calvino's oration, and the absolute freedom of demons which symbolizes the humanistic idea that gives up determinism to devote herself to the freedom of research.

Unfortunately it is not well known how those works, both completely electro-acoustic and part narration part electro-acoustic music, were realized. The poverty of information is primary due to absence of scores [6], work notes, tapes fragments useful to rebuild the processing phases of the sound materials. The lack of score is justified by the difficulty to mark simultaneously many parameters, as frequency, time and amplitude, and at the same time gives information about which kind of processing is used. The lack of work notes is probably due to the behavior of the composer that had an empirical approach with the devices they used, the various electro-acoustic works are often the results of a "random processing" done simultaneously with the recording, then the lack of documentation is another aspect that confuse the analysis process.

The main problem is to develop a method for the study of this type of music with the aim of build an analysis standard that permits a correct identification of electronic sound nature and their synthesis methods. There are some attempts to create such method [7], [8], [9] and [10]. This work is aimed to be the first step for the creation of this standard with the identification of different sound timbers, common to the various works, and the creation of a database, necessary for the retrieval and comparison of different works sound elements.

#### 5. METHODOLOGY

There are some steps you must do in order to make a philological analysis: the first is the identification of the original support which is very important considering the possible presence of degradations in production tapes and too aggressive restorations which can substantially alter the frequency content (and timbre) of electro-acoustic works. Another problem to deal with is the identification of the original work's support that, in the case of a magnetic tape, is vital for the analysis in order to obtain information such as the presence of cuts or tape inversions. But the identification of sound types without information regarding the nature and the methods used in their creation is not sufficient and effective so the second step consists of a detailed and rigorous identification of those timbres based on the study of equipment and sound processing possibilities available to the composers at the time of the compositions.

Our work consists on selecting different kinds of representative timbres, recognized by perceptive analysis. Then at first time we analyze them by means of frequency analysis in order to have a rough idea about how the frequencies change in time; this is obviously not an exhaustive kind of analysis but it may be a good starting point. Then make an hypothesis on how this sound could be created and finally try a sound synthesis, with different mathematical models, of the sound types; its comparison with the original sounds validate the correctness of the models and consequently permits to obtain valuable information on the nature of those sounds (analysis-by-synthesis method). In this way the analysis will be a means to obtain a reconstruction of the sounds of synthesis, of the programs for processing the sounds. This study will involve the analysis of the systems available at the time of the composition [11] and the study of related papers. The aim is to define a methodology that will allow the analysis of the composing process of a piece of music, on the base of the compositional process of the time as well as on any paper document referred to this or similar compositional process (notes, reflections, schemes, scores). Such a procedure would go through the hypothetical reconstruction and the phases of the synthesis of the piece, comparing the results with the original piece, and producing a critical analysis of the compositional process. Going beyond the arbitrary reconstruction of the sound we aim at historically positioning the piece taking into account its style, the composer style and the whole contemporary production.

## 6. ANALYSIS

The following sounds had been created and processed by the composer, by means of these equipments: oscillators; white noise generator; filters band pass (2-10 Hz); amplitude and ring modulator; reverberation chamber; alteration of the velocity of tape recording with or without changing the pitch; cut and paste [11].

### 6.1. Different kinds of sounds taken from "Continuo"

The sounds in fig. 1 were identified perceptively and defined as follow:

1. Frizzling. These are short and strident sounds at high frequencies with very thin bands of frequencies separated by few Hz.
2. Bubbling. These are short groups of sounds at different frequencies repeated very fast in a random way. They give the sensation of a bubbling liquid.
3. Cu-cu. A sound formed by two short sounds, the first more high than the second.
4. Sinusoidal sounds. These are sounds created by applying a band pass filter to white noise.

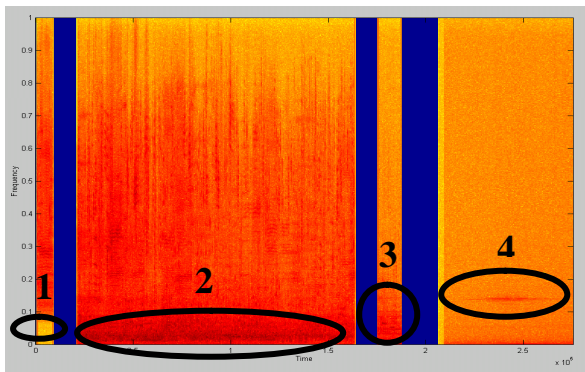


Figure 1. Four timbres of sounds which are repeated in different parts of the composition.

Bubbings: those are very complex sounds. The most probable way of synthesize this kind of sound is the granular synthesis. Nowadays it is easy enough to create such sounds but the problem is how to obtain it only with simple operations of editing as those of the Fifties.

The most probable way was to elaborate some sound with cut and paste of very small pieces of tape with noise or concrete sounds (for example clapping). This would explain both the complexity of the sound and the presence of an impulsive noise, a sort of click, maybe caused by cuts and then masked by reverb. In the second part of the figure 2 there is the our attempt of re-synthesize the sound.

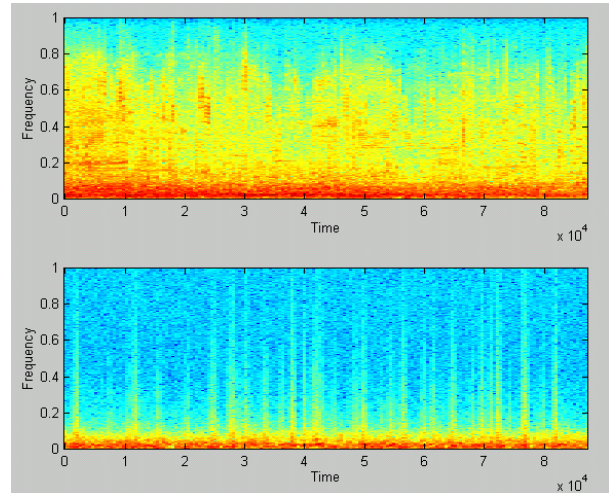


Figure 2. Comparison between a "bubbling" sound taken from "Continuo" and a synthesized one.

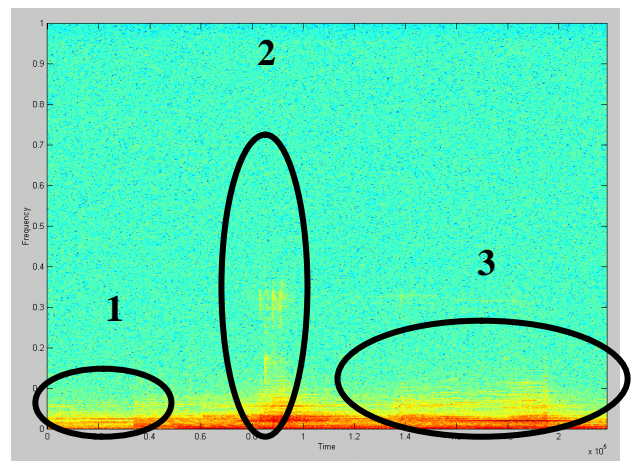


Figure 3. Three particular sound-objects.

Fig. 3 shows some sound-objects obtained by concrete sounds:

1. Sound that if accelerated by 7 semitones seems an organ.
2. Slowed percussions.
3. Sinusoidal sound.

### 6.2. Examples of use of the voice as a sound object

In "Portrait of Erasmo" Maderna uses also processed demon's voices by distortion and by varying the playing speed of the tape, to underline the contraposition between the idea of a man without freedom and the absolute freedom of demons (see fig. 4 and 5).

There is a surreal sound environment created by various electronics and noise sounds, into which enters the processed voice, speaking and singing. The voice is used like sound-object, thanks to their particular timbre.

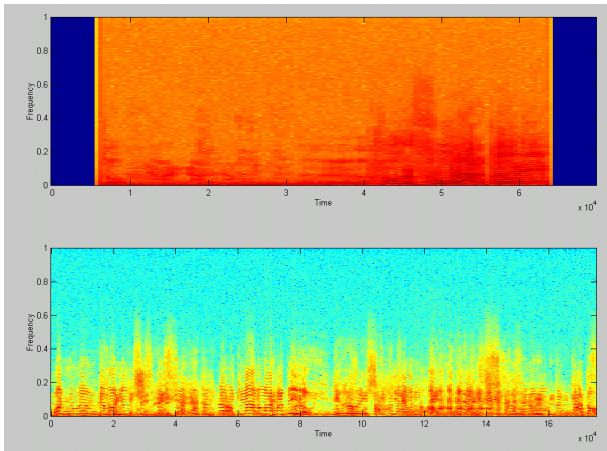


Figure 4. Up: *Voices of the actors behind the scenes* (“no, no, no...ho sbagliato”). Bottom: *Descriptions of Demons used as background to the narration.*

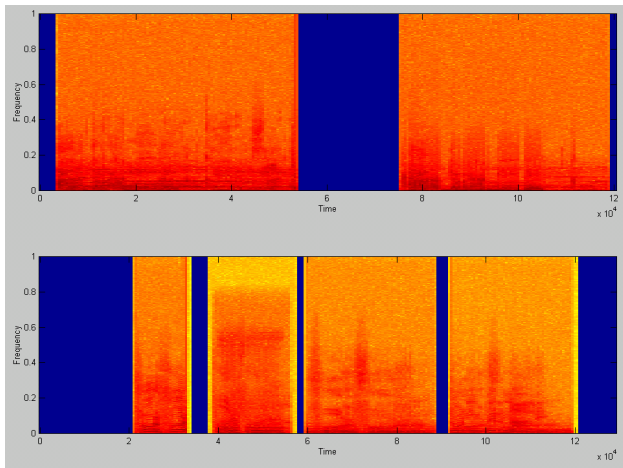


Figure 5. Processed voices. Up: "Abramelec" (two times). Bottom: "Furfur" (4 times)

### 6.3. Comparison between a sound taken from "Continuo" and two audio restoration of this original sound

In the nineteenth music, in particular in the electro-acoustic repertoire, the importance given to timbre it presents many difficulties of analysis, writing and restoration. The sense as such of the compositions is sometimes contained in some shades too often neglected and ignored.

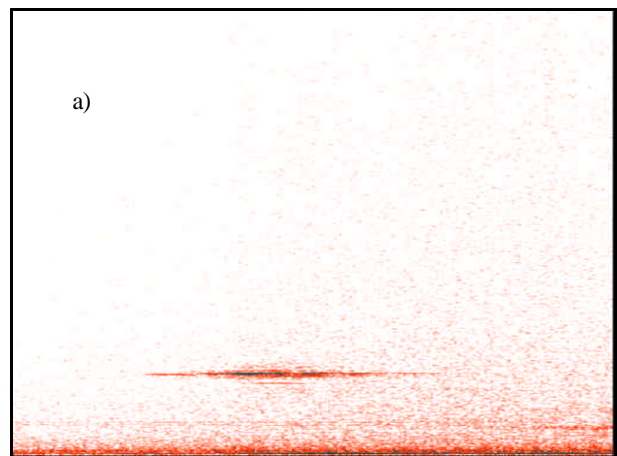
In the light of those considerations it is clear that the commercial restore, too aggressive, has deeply damaged the original sound. In this case the original was a greatly reverberated sound which seemed to come from the “mists of times” and the restored one sounds as a whistle deprived of any aesthetic meaning.

The restoration works, in our opinion, must be inspired by the following concept: to give back to the recording a sound quality

considered optimal according to the standard procedures of the time, without transcending the technological level historically reached by the Studios. What we wanted to reproduce, therefore, is a copy restored according to a degree of audio quality technologically comparable with that which could have been realized by using the equipment and techniques of the time in ideal ways and conditions.

This methodology aims at preserving the results of the creative process which are specific to the electro-acoustic assembling technology of the fifties, such as the tape junctions (provided that they are carried out well) and the mixing and editing procedures. It follows that the characterization of the degradations cannot derive only from the analysis of the digitalized signal, but also implies some external knowledge which includes musicological notions and, more precisely, the history of electronic composing techniques and the history of audio recording and reproducing systems.

From fig. 6 it can be noticed as a restoration carried out with commercial aims can tend to eliminate some aesthetic elements, removing a “noise patina” which was been approved by the artist. Sometimes, an aggressive restoration, finalized to match the aesthetic tastes of the modern audience (adapted to the modern recordings quality), can also remove some sound-objects (bubbling, cu-cu, frizzling, for instance) used by the composer as musical instruments.



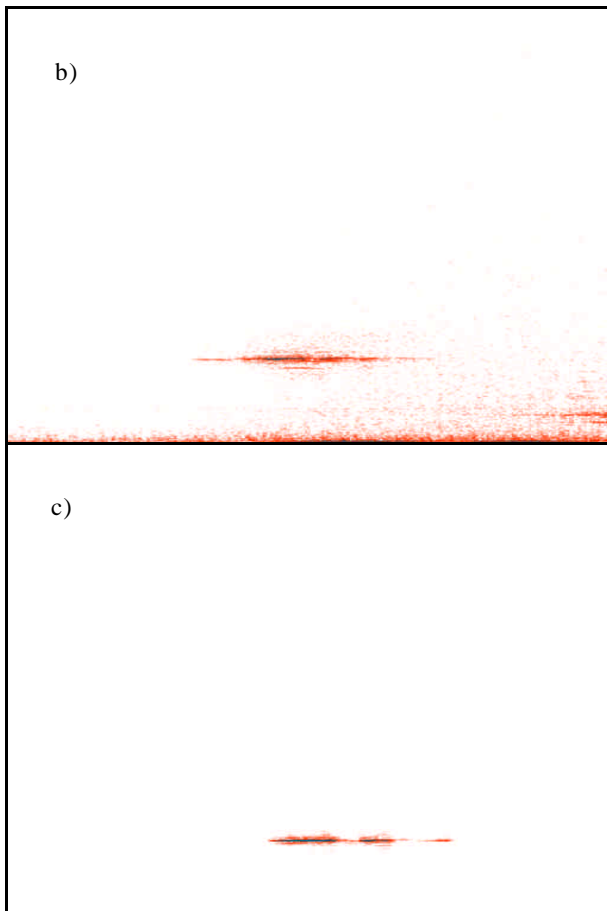


Figure 6. a) spectrogram of the original sound (digitalized from original tape); b) spectrogram of the restoration by Mirage laboratory; c) spectrogram of a commercial restoration.

## 7. CONCLUSIONS

In electro-acoustic music timbre has a capital importance just because every work is realized starting from the timbre research in order to create new sonorities and in order to let the sound object inspire the composer. Every sound is thought, experimented and elaborated; it is the sound which gives sense to the whole opera. Not having exact data about the sounds used in the various operas the illustrated approach appears as the most logical method of analysis. The attempt and the produced examples show how it is possible to identify and reproduce fairly successfully some categories of sounds. Although the various attempts it is still difficult to describe and synthesize many kinds of sounds due to their own complexity, the difficulty of recognizing their source and for the reason that it is not known if the sounds were wanted or effects of particular conditions or errors generated by the machines used for the synthesis. Once created the first database the difficulty will be to determine the nature of the different sounds.

This analysis, made to build the analysis standard and the database itself, is also useful to develop a more conscious and correct audio restoration.

The example given in fig. 6 represents only one of various problems related with audio restoration of such operas but it can form an opinion about the problems that an audio restorer must think about the nature of the sound to restore, the methods of creation of such sounds.

The audio restoration process at the moment generally follows two different types of methods: restoration using a signal model, based on modern signal processing techniques and restoration using a model of sound reproduction. In this context we want to explore the theorization and definition of a new paradigm for restoration: the one that uses a process model, by means of recovery-reconstruction, through the analysis of compositional processes and of the software that was used (to sound synthesis, aid for composing, etc.).

In the specific case of restoration of electronic music, the use of historical-musicological information about the use of timbre as fundamental parameter will be a necessary methodological assumption due to the impossibility to differently derive the sound models of the recordings. It is necessary that restoration be guided by all the pieces of information external to the signal that can be found in the history of the composing of electronic music and in that of the related musical technology. In this sense the study of timbre finalized to the audio restoration of electronic music on tape is part of the analysis of contemporary music.

## 8. REFERENCES

- [1] Stockhausen, K., "Musica elettronica e musica strumentale", in *La musica elettronica*, Henri Pousseur ed., 45-51, Feltrinelli, Milano, 1976.
- [2] Bayle, F., *Musique acousmatique. Propositions... Positions*, Parigi, Buchet/Chastel, 1993.
- [3] Schaeffer, P., *Traité des objets musicaux*, Parigi, Editions du Seuil, 1996.
- [4] Emmerson, S., "The Relation of Language to Materials", in *The Language of Electro-acoustic Music*, Emmerson S. ed., Londra, Macmillan, 1987.
- [5] McAdams, S., Spectral Fusion and the Creation of Auditory Images, in *Music, Mind and Brain*, Clynes ed., New York, Plenum Press, 1983.
- [6] Delalande, F., "En l'absence de partition, le cas singulier de la musique électroacoustique", *Analyse musicale*, 3: 197-708, 1986.
- [7] Camilleri, L., *Metodologie e concetti analitici nello studio di musiche elettroacustiche*, Rivista Italiana di Musicologia, 28(1):131-174, 1993.
- [8] Giomi, F. and Ligabue, M., "An Aesthetic – Cognitive Approach to the description of the Sound Object: from the Definition or a Phonematic Paradigm to the Individuation of



Analytic-Computational Strategies”, *Acta Semiotica Fennica*, Tarasti ed., Helsinki, University of Helsinki, 346-366, 1995.

- [9] Montecchi, G., “Continuo di Bruno Maderna”, in *I quaderni della Civica scuola di Musica* 21/22:43-57, 1992.
- [10] Anastasia, A., *Il ritratto di Erasmo. Per un’analisi storico-critica delle opere radiofoniche di Bruno Maderna*, Thesis, Dept. of Literature and Philosophy, University of Udine, 2000.
- [11] Lietti, A., “Gli impianti tecnici dello Studio di Fonologia musicale di Radio Milano”, *Elettronica* 3:116-122, 1956.

# Abstract musical timbre and physical modeling

Giovanni De Poli and Davide Rocchesso

June 21, 2002

## Abstract

As a result of the progress in information technologies, algorithms for sound generation and transformation are now ubiquitous in multimedia systems, even though their performance and quality is rarely satisfactory. For the specific needs of music production and multimedia art, sound models are needed which are versatile, responsive to user's expectations, and having high audio quality. Moreover, for human-machine interaction model flexibility is a major issue. We will review some of the most important computational models that are being used in musical sound production, and we will see that models based on the physics of actual or virtual objects can meet most of the requirements, thus allowing the user to rely on high-level descriptions of the sounding entities.

## 1 Introduction

In our everyday experience, musical sounds are increasingly listened to by means of loudspeakers. On the one hand, it is desirable to achieve a faithful reproduction of the sound of acoustic instruments in high-quality auditoria. On the other hand, the possibilities offered by digital technologies should be exploited to approach sound-related phenomena in a creative way. Both of these needs call for mathematical and computational models of sound generation and processing.

The timbre produced by acoustic musical instruments is caused by the physical vibration of a certain resonating structure. This vibration can be described by signals that correspond to the time-evolution of the acoustic pressure associated to it. The fact that the sound can be characterized by a set of signals suggests quite naturally that some computing equipment could be successfully employed for generating timbres, for either the imitation of acoustic instruments or the creation of new sounds with novel timbral properties.

The focus of this article is on computational models of timbre generation, especially on those models that are directly based on physical descriptions of generation process. The general framework of sound modeling is explained in sections 2 and 3. We will divide modeling paradigms into signal models and physics-based models. They will be illustrated in Section 5 and 6. Several techniques are presented for modeling sound sources and general, linear and nonlinear acoustic systems.

## 2 Computational models as abstract timbre models

In order to generate, manipulate, and think about timbre, it is useful to organize our intuitive sound abstractions into sound objects, in the same way as abstract categories are needed for defining visual objects. The first extensive investigation and systematization of sound objects from a perceptual viewpoint was done by Pierre Shaeffer in the fifties [1]. Nowadays, a common terminology is available for describing sound objects both from a phenomenological or a referential viewpoint, and for describing collections of such objects (i.e. *soundscapes*) [2], [3], [4].

For effective generation and manipulation of the timbre it is necessary to define models for sound synthesis, processing, and composition. Identifying models, either visual or acoustic, is equivalent to making

high-level constructive interpretations, built up from the zero level (i.e. pixels or sound samples). It is important for the model to be associated with a semantic interpretation, in such a way that an intuitive action on model parameters becomes possible. A sound generation model is implemented by means of sound synthesis and processing techniques. A wide variety of sound synthesis algorithms is currently available either commercially or in the literature. Each one of them exhibits some peculiar characteristics that could make it preferable to others, depending on goals and needs. Technological progress has made enormous steps forward in the past few years as far as the computational power that can be made available at low cost is concerned. At the same time, sound synthesis methods have become more and more computationally efficient and the user interface has become friendlier and friendlier. As a consequence, musicians can nowadays access a wide collection of synthesis techniques (all available at low cost in their full functionality), and concentrate on their timbral properties.

Each sound synthesis algorithm can be thought of as a computational model for the sound itself. Though this observation may seem quite obvious, its meaning for sound synthesis is not so straightforward. As a matter of fact, modeling sounds is much more than just generating them, as a computational model can be used for representing and generating a whole class of sounds, depending on the choice of control parameters. The idea of associating a class of sounds to a digital sound model is in complete accordance with the way we tend to classify natural musical instruments according to their sound generation mechanism. For example, strings and woodwinds are normally seen as timbral classes of acoustic instruments characterized by their sound generation mechanism. It should be clear that the degree of compactness of a class of sounds is determined, on one hand, by the sensitivity of the digital model to parameter variations and, on the other hand, the amount of control that is necessary to obtain a certain desired sound. As an extreme example we may think of a situation in which a musician is required to generate sounds sample by sample, while the task of the computing equipment is just that of playing the samples. In this case the control signal is represented by the sound itself, therefore the class of sounds that can be produced is unlimited but the instrument is impossible for a musician to control and play. An opposite extremal situation is that in which the synthesis technique is actually the model of an acoustic musical instrument. In this case the class of sounds that can be produced is much more limited (it is characteristic of the mechanism that is being modeled by the algorithm), but the degree of difficulty involved in generating the control parameters is quite modest, as it corresponds to physical parameters that have an intuitive counterpart in the experience of the musician. An interesting conclusion that could be already drawn in the light of what we stated above is that the generality of the class of sounds associated to a sound synthesis algorithm is somehow in contrast with the “playability” of the algorithm itself. One should remember that the “playability” is of crucial importance for the success of a specific sound synthesis algorithm as, in order for a sound synthesis algorithm to be suitable for musical purposes, the musician needs an intuitive and easy access to its control parameters during both the sound design process and the performance. Such requirements often represents the reason why a certain synthesis technique is preferred to others. From a mathematical viewpoint, the musical use of sound models opens some interesting issues: description of a class of models that are suitable for the representation of musically-relevant acoustic phenomena; description of efficient and versatile algorithms that realize the models; mapping between meaningful acoustic and musical parameters and numerical parameters of the models; analysis of sound signals that produces estimates of model parameters and control signals; approximation and simplification of the models based on the perceptual relevance of their features; generalization of computational structures and models in order to enhance versatility.

### **3 Sound Modeling**

In the music sound domain, we define *generative models* as those models which give computational form to abstract objects, thus representing a sound generation mechanism.

Generative models can represent the dynamics of real or virtual generating objects (*physics-based models*), or they can represent the physical quantities as they arrive to human senses (*signal models*) [5] (see figure ??). In our terminology, signal models are models of signals as they are emitted from loudspeakers or arrive to the ears. The connection with human perception is better understood when considering the evaluation criteria of the generative models. The evaluation of a signal model should be done according to certain perceptual cues. On the contrary, physics-based models are better evaluated according to the physical behaviors involved in the sound production process.

In classic sound synthesis, signal models dominated the scene, due to the availability of very efficient and widely applicable algorithms (e.g. frequency modulation). Moreover, signal models allow to design sounds as objects *per se* without having to rely on actual pieces of material which act as a sound source. However, many people are becoming convinced of the fact that physics-based models are closer to the users/designers' needs of interacting with sound objects. The semantic power of these models seems to make them preferable for this purpose. The computational complexity of physically-based algorithms is becoming affordable with nowadays technology, even for real-time applications. We keep in mind that the advantage we gain in model expressivity comes to the expense of the flexibility of several general-purpose signal models. For this reason, signal models keep being the model of choice in many applications, especially for music composition.

In the perspective of a multisensorial unification under common models, physics-based models offer an evident advantage over signal models. In fact, the mechanisms of perception for sight and hearing are very different, and a unification at this level looks difficult. Even though analogies based on perception are possible, an authentic sensorial coherence seems to be ensured only by physics-based models. The interaction among various perceptions can be an essential feature if we want to maximize the amount of information conveyed to the spectator/actor. The unification of visual and aural cues is more properly done at the level of abstractions, where the cultural and experience aspects become fundamental. Thus, building models closer to the abstract object, as it is conceived by the designer, is a fundamental step in the direction of this unification.

## 4 Classic Signal Models

Here we will briefly overview the most important signal models for musical sounds. A more extensive presentation can be found in several tutorial articles and books on sound synthesis techniques [6], [7], [8], [9], [10], [11]. Instead, section 5 will cover the most relevant paradigms in physically-based sound modeling.

### 4.1 Spectral models

Since the human ear acts as a particular spectrum analyser, a first class of synthesis models aims at modeling and generating sound spectra. The Short Time Fourier Transform and other time-frequency representations provide powerful sound analysis tools for computing the time-varying spectrum of a given sound.

#### 4.1.1 Sinusoidal model

When we analyze a pitched sound, we find that its spectral energy is mainly concentrated at a few discrete (slowly time-varying) frequencies  $f_i$ . These frequency lines correspond to different sinusoidal components called partials. If the sound is almost periodic, the frequencies of partials are approximately multiple of the fundamental frequency  $f_0$ , ie.  $f_i(t) \simeq i f_0(t)$ . The amplitude  $a_i$  of each partial is not constant and its time-variation is critical for timbre characterization. If there is a good degree of correlation among the frequency and amplitude variations of different partials, these are perceived as fused to give a unique sound with its timbre identity.

The sinusoidal model assumes that the sound can be modeled as a sum of sinusoidal oscillators whose amplitude  $a_i$  and frequency  $f_i$  are slowly time-varying

$$s_s(t) = \sum_i a_i(t) \cos[\phi_i(t)] , \quad (1)$$

$$\phi_i(t) = \int_0^t 2\pi f_i(\tau) d\tau + \phi_i(0) , \quad (2)$$

or, digitally,

$$s_s(n) = \sum_i a_i(n) \cos[\phi_i(n)] , \quad (3)$$

$$\phi_i(n) = 2\pi f_i(n)T_s + \phi_i(n-1) , \quad (4)$$

where  $T_s$  is the sampling period. Equations (1) and (2) are a generalization of the Fourier theorem, that states that a periodic sound of frequency  $f_0$  can be decomposed as a sum of harmonically related sinusoids  $s_s(t) = \sum_i a_i \cos(2\pi i f_0 t + \phi_i)$ .

This model is also capable of reproducing aperiodic and inharmonic sounds, as long as their spectral energy is concentrated near discrete frequencies (spectral lines). In computer music this model is called *additive synthesis* and is widely used in music composition. Notice that the idea behind this method is not new. As a matter of fact, additive synthesis has been used for centuries in some traditional instruments such as organs. Organ pipes, in fact, produce relatively simple sounds that, combined together, contribute to the richer spectrum of some registers. Particularly rich registers are created by using many pipes of different pitch at the same time. Moreover this method, developed for simulating natural sounds, has become the “metaphorical” foundation of a compositional methodology based on the expansion of the time scale and the reinterpretation of the spectrum in harmonic structures.

#### 4.1.2 Random noise models

The spread part of the spectrum is perceived as random noise. The basic noise generation algorithm is the congruential method

$$s_n = [a s_{n-1} + b] \bmod M . \quad (5)$$

With a suitable choice of the coefficients  $a$  and  $b$  it produces pseudorandom sequences with flat spectral density magnitude (white noise). Different spectral shapes can be obtained using white noise as input to a filter.

#### 4.1.3 Filters

Some sources can be modeled as an exciter, characterized by a spectrally rich signal, and a resonator, described by a linear system, connected in a feed-forward relationship. An example is the voice, where the periodic pulses or random fluctuations produced by the vocal folds are filtered by the vocal tract, that shapes the spectral envelope. The vowel quality and the voice color greatly depends on the resonance regions of the filter, called formants.

If the system is linear and time-invariant, it can be described by the filter  $H(z) = B(z)/A(z)$  that can be computed by a difference equation

$$s_f(n) = \sum_i b_i u(n-i) - \sum_k a_k s_f(n-k) . \quad (6)$$

where  $a_k$  e  $b_i$  are the filter coefficients and  $u(n)$  e  $s_f(n)$  are input and output signals. The model is also represented by the convolution of the source  $u(n)$  with the impulse response of the filter

$$s_f(n) = (u * h)(n) \triangleq \sum_k h(n - k)u(k) . \quad (7)$$

Digital signal processing theory gives us the tools to design the filter structure and to estimate the filter coefficients in order to obtain a desired frequency response. This model combines the spectral fine structure (spectral lines, broadband or narrowband noise, etc.) of the input with the spectral envelope shaping properties of the filter:  $S_f(f) = U(f)H(f)$ . Therefore, it is possible to control and modify separately the pitch from the formant structure of a speech sound. In computer music this model is called *subtractive synthesis*. If the filter is static, the temporal features of the input signal are maintained. If, conversely, the filter coefficients are varied, the frequency response changes. As a consequence, the output will be a combination of temporal variations of the input and of the filter (*cross-synthesis*).

If we make some simplifying hypothesis about the input, it is possible to estimate both the parameters of the source and the filter of a given sound. The most common procedure is linear predictive coding (LPC) which assumes that the source is either a periodic impulse train or white noise, and that the filter is all pole (i.e., no zeros) [12]. LPC is widely used for speech synthesis and modification.

A special case is when the filter features a long delay as in

$$s_f(n) = \beta u(n) - \alpha s_f(n - N_p) . \quad (8)$$

This is a comb type filter featuring frequency resonances multiple of a fundamental  $f_p = F_s/N_p$ , where  $F_s = 1/T_s$  is the sampling rate. If initial values are set for the whole delay line, for example random values, all the frequency components that do not coincide with resonance frequencies are progressively filtered out until a harmonic sound is left. If there is attenuation ( $\alpha < 1$ ) the sound will have a decreasing envelope. Substituting  $\alpha$  and/or  $\beta$  with filters, the sound decay time will depend on frequency. For example if  $\alpha$  is smaller at higher frequencies, the upper harmonics will decay faster than the lower ones. We can thus obtain simple sound simulations of the plucked strings [13], [14], where the delay line serves to establish oscillations. This method is suitable to model sounds produced by a brief excitation of a resonator, where the latter establishes the periodicity, and the interaction between exciter and resonator can be assumed to be feedforward. This method is called *long-term prediction* or *Karplus-Strong synthesis*. More general musical oscillators will be discussed in sect. 6.

## 4.2 Time domain models

When the sound characteristics are rapidly varying, as during attacks or non stationary sounds, spectral models tend to presents artifacts, due to low time-frequency resolution or to the increase of the amount of data used in the representation. To overcome these difficulties, time domain models were proposed. A first class, called *sampling* or *wavetable*, stores the waveforms of musical sounds or sound fragments in a database. During synthesis, a waveform is selected and reproduced with simple modifications, such as looping of the periodic part, or sample interpolation for pitch shifting. The same idea is used for simple oscillators, that repeats a waveform stored in a table (*table-lookup oscillator*).

### 4.2.1 Granular models

More creative is the *granular synthesis* model. The basic idea is that a sound can be considered as a sequence, possibly with overlaps, of elementary and short acoustic elements called *grains*. Additive synthesis starts from the idea of dividing the sound in the frequency domain into a number of simpler elements (sinusoidal). Granular synthesis, instead, starts from the idea of dividing the sound in the time domain into a

sequence of short elements called “grains”. The parameters of this technique are the waveform of the grain  $g_k(\cdot)$ , its temporal location  $l_k$  and amplitude  $a_k$

$$s_g(n) = \sum_k a_k g_k(n - l_k) . \quad (9)$$

A complex and dynamic acoustic event can be constructed starting from a large quantity of grains. The features of the grains and their temporal locations determine the sound timbre. We can see it as being similar to cinema, where a rapid sequence of static images gives the impression of objects in movement. The initial idea of granular synthesis dates back to Gabor [15], while in music it arises from early experiences of tape electronic music. The choice of parameters can be via various criteria, at the base of which, for each one, there is an interpretation model of the sound. In general, granular synthesis is not a single synthesis model but a way of realizing many different models using waveforms that are locally defined. The choice of the interpretation model implies operational processes that may affect the sonic material in various ways.

The most important and classic type of granular synthesis (*asynchronous granular synthesis*) distributes grains irregularly on the time-frequency plane in form of clouds [16]. The grain waveform is

$$g_k(i) = w_d(i) \cos(2\pi f_k T_s i) , \quad (10)$$

where  $w_d(i)$  is a window of length  $d$  samples, that controls the time span and the spectral bandwidth around  $f_k$ . For example, randomly scattered grains within a mask, which delimits a particular frequency/amplitude/time region, result in a sound cloud or musical texture that varies over time. The density of the grains within the mask can be controlled. As a result, articulated sounds can be modeled and, wherever there is no interest in controlling the microstructure exactly, problems involving the detailed control of the temporal characteristics of the grains can be avoided. Another peculiarity of granular synthesis is that it eases the design of sound events as parts of a larger temporal architecture. For composers, this means a unification of compositional metaphors on different scales and, as a consequence, the control over a time continuum ranging from the milliseconds to the tens of seconds. There are psychoacoustic effects that can be easily experimented by using this algorithm, for example crumbling effects and waveform fusions, which have the corresponding counterpart in the effects of separation and fusion of tones.

### 4.3 Hybrid models

Different models can be combined in order to have a more flexible and effective sound generation. One approach is *Spectral Modeling Synthesis* (SMS) [17] that considers sounds as composed by a sinusoidal part  $s_s(t)$  (see Eq. 1), corresponding to the main system modes of vibration, and a residual  $r(t)$ , modeled as the convolution of white noise with a time-varying frequency shaping filter (see Eq. 7)

$$s_{sr}(t) = s_s(t) + r(t) . \quad (11)$$

The residual comprises the energy produced in the excitation mechanism which is not transformed into stationary vibrations, plus any other energy contribution that is not sinusoidal in nature. By using the short time Fourier transform and a peak detection algorithm, it is possible to separate the two parts at the analysis stage, and to estimate the time varying parameters of these models. The main advantage of this model is that it is quite robust to sound transformations that are musically relevant, such as time stretching, pitch shifting, and spectral morphing.

In the SMS model, transients and rapid signal variations are not well represented. Verma et al. [18] proposed an extension of SMS that includes a third component due to transients. Their method is called Sinusoids+Transients+Noise (S+T+N) and is expressed by

$$s_{STN}(t) = s_s(t) + s_g(t) + r(t) , \quad (12)$$

where  $s_g(t)$  is a granular term representing the signal transients. This term is automatically extracted from the SMS residual using the Discrete Cosine Transform, followed by a second SMS analysis in the frequency domain.

#### 4.4 Abstract models: frequency modulation

Another class of sound synthesis algorithms is neither derived from physical mechanisms of sound production, nor from any sound analysis techniques. These are algorithms derived from the mathematical properties of a formula. The most important of these algorithms is the so called synthesis by *Frequency Modulation* (FM) [19]. The technique works as an instantaneous modulation of the phase or frequency of a sinusoidal carrier according to the behavior of another signal (modulator), which is usually sinusoidal. The basic scheme can be expressed as follows:

$$s(t) = \sin[2\pi f_c t + I \sin(2\pi f_m t)] = \sum_{k=-\infty}^{\infty} J_k(I) \sin[2\pi(f_c + k f_m)t] \quad (13)$$

where  $J_k(I)$  is the Bessel function of order  $k$ . The resulting spectrum presents lines at frequencies  $|f_c \pm k f_m|$ . The ratio  $f_c/f_m$  determines the spectral content of sounds, and is directly linked to some important features, like the absence of even components, or the inharmonicity.

The parameter  $I$  (modulation Index) controls the spectral bandwidth around  $f_c$ , and is usually associated with a time curve (the so called envelope), in such a way that time evolution of the spectrum is similar to that of traditional instruments. For instance, a high value of the modulation index determines a wide frequency bandwidth, as it is found during the attack of typical instrumental sounds. On the other hand, the gradual decrease of the modulation index determines a natural shrinking of the frequency bandwidth during the decay phase. From the basic scheme, other variants can be derived, such as parallel modulators and feedback modulation. So far, however, no general algorithm has been found for deriving the parameters of an FM model from the analysis of a given sound, and no intuitive interpretation can be given to the parameter choice, as this synthesis technique does not evoke any previous musical experience of the performer. The main qualities of FM, i.e. great timbre dynamics with just a few parameters and a low computational cost, are progressively losing importance within modern digital systems. Other synthesis techniques, though more expensive, can be controlled in a more natural and intuitive fashion. The FM synthesis, however, still preserves the attractiveness of its own peculiar timbre space and, although it is not particularly suitable for the simulation of natural sounds, it offers a wide range of original synthetic sounds that are of considerable interest for computer musicians.

## 5 Physics-based Models

In the family of physics-based models we put all the algorithms generating sounds as a side effect of a more general process of simulation of a physical phenomenon. Physics-based models can be classified according to the way of representing, simulating and discretizing the physical reality. Hence, we can talk about cellular, finite-difference, and waveguide models, thus intending that these categories are not disjoint but, in some cases, they represent different viewpoints on the same computational mechanism. Moreover, physics-based models have not necessarily to be based on the physics of the real world, but they can, more generally, gain inspiration from it; in this case we will talk about pseudo-physical models.

In this chapter, the approach to physically-based synthesis is carried on with particular reference to real-time applications, therefore the time complexity of algorithms plays a key role. We can summarize the general objective of the presentation saying that we want to obtain models for large families of sounding objects, and these models have to provide a satisfactory representation of the acoustic behavior with the minimum computational effort.



## 5.1 Functional Blocks

In real objects we can often outline functionally distinct parts, and express the overall behavior of the system as the interaction of these parts. Outlining functional blocks helps the task of modeling, because for each block a different representation strategy can be chosen. In addition, the range of parameters can be better specified in isolated blocks, and the gain in semantic clearness is evident. Our analysis stems from musical instruments, and this is justified by the fact that the same generative mechanisms can be found in many other physical objects. In fact, we find it difficult to think about a physical process producing sound and having no analogy in some musical instrument. For instance, friction can be found in bowed string instruments, striking in percussion instruments, air turbulences in jet-driven instruments, etc. . Generally speaking, we can think of musical instruments as a specialization of natural dynamics for artistic purposes. Musical instruments are important for the whole area of sonification in multimedia environments because they constitute a testbed where the various simulation techniques can easily show their merits and pitfalls.

The first level of conceptual decomposition that we can devise for musical instruments is represented by the interaction scheme of figure 1, where two functional blocks are outlined: a resonator and an exciter. The resonator sustains and controls the oscillation, and is related with sound attributes like pitch and spectral envelope. The exciter is the place where energy is injected into the instrument, and it strongly affects the attack transient of sound, which is fundamental for timbre identification. The interaction of exciter and resonator is the main source of richness and variety of nuances that can be obtained from a musical instruments. When translating the conceptual decomposition into a model, two dynamic systems are found [20]: the excitation block, which is strongly non-linear, and the resonator, supposed to be linear to a great extent. The player controls the performance by means of inputs to the two blocks. The interaction can be “feedforward”, when the exciter doesn’t receive any information from the resonator, or “feedback”, when the two blocks exert a mutual information exchange. In this conceptual scheme, the radiating element (bell, resonating body, etc.) is implicitly enclosed within the resonator. In a clarinet, for instance, we have a feedback structure where the reed is the exciter and the bore with its bell acts as a resonator. The player exert exciting actions such as controlling the mouth pressure and the embouchure, as well as modulating actions such as changing the bore effective length by opening and closing the holes. In a plucked string instrument, such as a guitar, the excitation is provided by plucking the string, the resonator is given by the strings and the body, and modulating actions take the form of fingering. The interaction is only weakly feedback, so that a feedforward scheme can be adopted as a good approximation: the excitation imposes the initial conditions and the resonator is then left free to vibrate.

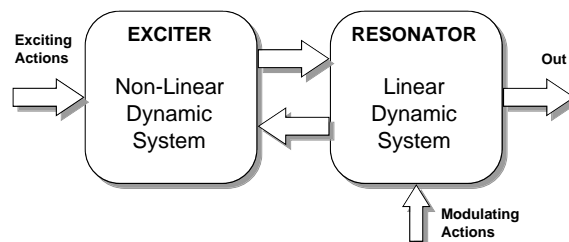


Figure 1: Exciter-Resonator Interaction Scheme

In practical physical modeling the block decomposition can be extended to finer levels of detail, as both the exciter and the resonator can be further decomposed into simpler functional components, e.g. the holes and the bell of a clarinet as a refinement of the resonator. At each stage of model decomposition, we are faced with the choice of expanding the blocks further (white-box modeling), or just considering the input-output behavior of the basic components (black-box modeling). In particular, it is very tempting to model just the input-output behavior of linear blocks, because in this case the problem reduces to filter design. However,

such an approach provides structures whose parameters are difficult to interpret and, therefore, to control. In any case, when the decomposition of an instrument into blocks corresponds to a similar decomposition in digital structures, a premium in efficiency and versatility is likely to be obtained. In fact, we can focus on functionally distinct parts and try to obtain the best results from each before coupling them together [21].

In digital implementations, in between the two blocks exciter and resonator, a third block is often found. This is an interaction block and it can convert the variables used in the exciter to the variables used in the resonator, or avoid possible anomalies introduced by the discretization process. The idea is to have a sort of adaptor for connecting different blocks in a modular way. This adaptor might also serve to compensate the simplifications introduced by the modeling process. To this end, a residual signal might be introduced in this block in order to improve the sound realism. The limits of a detailed physical simulation are also found when we try to model the behavior of a complex linear vibrating structure, such as a soundboard; in such cases it can be useful to record its impulse response and include it in the excitation signal as it is provided to a feedforward interaction scheme. Such a method is called *commuted synthesis*, since it makes use of commutativity of linear, time-invariant blocks [22], [23].

It is interesting to notice that the integration of sampled noises or impulse responses into physical models is analogous to texture mapping in computer graphics [24]. In both cases the realism of a synthetic scene is increased by insertion of snapshots of textures (either visual or aural) taken from actual objects and projected onto the model.

## 5.2 Cellular models

A possible approach to simulation of complex dynamical systems is their decomposition into a multitude of interacting particles. The dynamics of each of these particles are discretized and quantized in some way to produce a finite-state automaton (a cell), suitable for implementation on a processing element of a parallel computer. The discrete dynamical system consisting of a regular lattice of elementary cells is called a cellular automaton [25], [26]. The state of any cell is updated by a transition rule which is applied to the previous-step state of its neighborhood. When the cellular automaton comes from the discretization of a homogeneous and isotropic medium it is natural to assume functional homogeneity and isotropy, i.e. all the cells behave according to the same rules and are connected to all their immediate neighbors in the same way [25]. If the cellular automaton has to be representative of a physical system, the state of cells must be characterized by values of selected physical parameters, e.g. displacement, velocity, force.

Several approaches to physically-based sound modeling can be recast in terms of cellular automata, the most notable being the CORDIS-ANIMA system introduced by Cadoz and his co-workers [27], [28], [29], who came up with cells as discrete-time models of small mass-spring-damper systems, with the possible introduction of nonlinearities. The main goal of the CORDIS-ANIMA project was to achieve high degrees of modularity and parallelism, and to provide a unified formalism for rigid and flexible bodies. The technique is very expensive for an accurate sequential simulation of wide vibrating objects, but is probably the only effective way in the case of a multiplicity of micro-objects (e.g. sand grains) or for very irregular media, since it allows an embedding of the material characteristics (viscosity, etc.). An example of CORDIS-ANIMA network discretizing a membrane is shown in figure ??, where we have surrounded by triangles the equal cells which provide output variables depending on the internal state and on input variables from neighboring cells. Even though the CORDIS-ANIMA system uses heterogeneous elements such as matter points or visco-elastic links, a network can be restated in terms of a cellular automaton showing functional homogeneity and isotropy.

A cellular automaton is inherently parallel, and its implementation on a parallel computer shows excellent scalability. Moreover, in the case of the multiplicity of micro-objects, it has shown good effectiveness for joint production of audio and video simulations [30]. It might be possible to show that a two-dimensional cellular automaton can implement the model of a membrane as it is expressed by a waveguide mesh. How-

ever, as we will see in sections 5.3 and 5.4, when the system to be modeled is the medium where waves propagate, the natural approach is to start from the wave equation and to discretize it or its solutions. In the fields of finite-difference methods or waveguide modeling, theoretical tools do exist for assessing the correctness of these discretizations. On the other hand, only qualitative criteria seem to be applicable to cellular automata in their general formulation.

### 5.3 Finite-difference models

When modeling vibrations of real-world objects, it can be useful to consider them as rigid bodies connected by lumped, idealized elements (e.g. dashpots, springs, geometric constraints, etc.) or, alternatively, to treat them as flexible bodies where forces and matter are distributed over a continuous space (e.g. a string, a membrane, etc.). In the two cases the physical behavior can be represented by ordinary or partial differential equations, whose form can be learned from physics textbooks [31] and whose coefficient values can be obtained from physicists' investigations or from direct measurements. These differential equations often give only a crude approximation of reality, as the objects being modeled are just too complicated. Moreover, as we try to solve the equations by numerical means, a further amount of approximation is added to the simulated behavior, so that the final result can be quite far from the real behavior.

One of the most popular ways of solving differential equations is finite differencing, where a grid is constructed in the spatial and time variables, and derivatives are replaced by linear combinations of the values on this grid. Two are the main problems to be faced when designing a finite-difference scheme for a partial differential equation: numerical losses and numerical dispersion. There is a standard technique [32], [33] for evaluating the performance of a finite-difference scheme in contrasting these problems: the von Neumann analysis. It can be quickly explained on the simple case of the ideal string (or the ideal acoustic tube), whose wave equation is [34]

$$\frac{\partial^2 p(x, t)}{\partial t^2} = c^2 \frac{\partial^2 p(x, t)}{\partial x^2}, \quad (14)$$

where  $c$  is the wave velocity of propagation,  $t$  and  $x$  are the time and space variables, and  $p$  is the string displacement (or acoustic pressure). By replacing the second derivatives by central second-order differences, the explicit updating scheme for the  $i$ -th spatial sample of displacement (or pressure) is:

$$p(i, n + 1) = 2 \left( 1 - \frac{c^2 \Delta t^2}{\Delta x^2} \right) p(i, n) - p(i, n - 1) + \frac{c^2 \Delta t^2}{\Delta x^2} [p(i + 1, n) + p(i - 1, n)], \quad (15)$$

where  $\Delta t$  and  $\Delta x$  are the time and space grid steps. The von Neumann analysis assumes that the equation parameters are locally constant and checks the time evolution of a spatial Fourier transform of (15). In this way a spectral amplification factor is found whose deviations from unit magnitude and linear phase give respectively the numerical loss (or amplification) and dispersion errors. For the scheme (15) it can be shown that a unit-magnitude amplification factor is ensured as long as the Courant-Friedrichs-Lewy condition [32]

$$\frac{c \Delta t}{\Delta x} \leq 1 \quad (16)$$

is satisfied, and that no numerical dispersion is found if equality applies in (16). A first consequence of (16) is that only strings having length which is an integer number of  $c \Delta t$  are exactly simulated. Moreover, when the string deviates from ideality and higher spatial derivatives appear (physical dispersion), the simulation becomes always approximate. In these cases, the resort to implicit schemes can allow the tuning of the discrete algorithm to the amount of physical dispersion, in such a way that as many partials as possible are reproduced in the band of interest [35].

It is worth noting that if  $c$  in equation (14) is a function of time and space, the finite difference method retains its validity because it is based on a local (in time and space) discretization of the wave equation. Another advantage of finite differencing over other modeling techniques is that the medium is accessible at all the points of the time-space grid, thus maximizing the possibilities of interaction with other objects.

When the objects being simulated are rigid bodies, they can be described by ordinary differential equations

$$\begin{aligned}\dot{\mathbf{y}}(t) &= \mathbf{F}[\mathbf{y}(t), \mathbf{u}(t), t], \\ \mathbf{y}(0) &= \mathbf{y}_0,\end{aligned}\tag{17}$$

being  $\mathbf{u}(t)$  the vector describing the set of input signals and  $\mathbf{y}_0$  initial conditions. Numerical analysis developed a plethora of techniques for their integration [32] transforming Eq.(17) into difference equations

$$\mathbf{y}(n+1) = \mathbf{F}_d[\mathbf{y}(n), \mathbf{u}(n), n].\tag{18}$$

However, attention must be paid to stability issues and to the correct reproduction of important physical attributes. These issues are strongly dependent on the numerical integration technique and on the sampling rate which are to be used. In most of the cases, there is no better method than trying several techniques and comparing the results, but the task is often facilitated by the fact that the strong nonlinearities are lumped. For example, in [36], the dynamics of a clarinet reed is discretized by using a fourth-order Runge-Kutta method, Euler differencing, and bilinear transformation [37]. The Runge-Kutta method turns out to be unstable for low sampling rates, while Euler differencing shows a poor reproduction of the characteristic resonance of the reed, due to numerical losses. For that specific case, the best choice seems to be the bilinear transform, which corresponds to a trapezoidal integration of the differential equations, possibly with some warping of the frequency axis [38] for adjusting the resonance central frequency. The discretization by impulse invariance [37] is also a reliable tool when aliasing can be neglected, and its performance is often preferable to bilinear transformation in acoustic modeling because it is free of frequency warping and artificial damping. Other discretization methods have recently been compared using the clarinet reed as a testbed [39]. As a result, it seems that a technique that employs polynomial interpolation of the input signals [40] gives the best reproduction of the reed resonance at an affordable cost. This latter technique can be interpreted as an extension of the impulse invariance that includes some antialias filtering.

Further studies are needed to establish the most suitable discretization techniques for the many kinds of lumped dynamics, with special attention to be paid to the looped connection of lumped non-linear elements with memoryless nonlinearities and distributed resonators. Several techniques from signal processing and numerical analysis are yet to be experimented, while some general methodologies are just being proposed. In this respect, section 5.4 will show how, switching to a wave-variable representation of the physical quantities, it is possible to apply the paradigms of Wave Digital filters and Waveguide networks to the lumped and distributed elements respectively.

## 5.4 Wave models

When discretizing physical systems a key role is played by the efficiency and accuracy of the discretization technique. Namely, we would like to be able to simulate simple vibrating structures and exciters with no artifacts (e.g. aliasing, or non-computable dependencies) and with low computational complexity. Due to its good properties with respect to these two criteria, one of the most popular ways of approaching physical modeling of acoustic systems makes use of wave variables instead of absolute physical quantities. Given the dual physical variables  $p$  and  $u$  (let us call them pressure and velocity), the pressure waves are defined as

$$\begin{aligned}p^+ &= (p + Z_0 u)/2, \\ p^- &= (p - Z_0 u)/2,\end{aligned}\tag{19}$$

where  $Z_0$  is an arbitrary reference impedance.

When wave variables are adopted in the digital domain for representing lumped components this approach is called Wave Digital Filtering [41] It is possible to show that a lumped component having impedance  $Z(z) = P(z)/U(z)$  can be represented in pressure waves by

$$R(z) = \frac{P^-(z)}{P^+(z)} = \frac{Z(z) - Z_0}{Z(z) + Z_0}. \quad (20)$$

The reference impedance  $Z_0$  is chosen in such a way that there is at least one delay element in any signal path connecting  $p^+$  with  $p^-$ . The complete wave network is derived by applying the Kirchhoff principles [42] to junctions of components derived by the previous steps (wave-scattering formulation of the network) and to abrupt changes of the characteristic impedance.

On the other hand, when the components to be modeled are distributed wave-propagating media, Digital Waveguide Networks [43] can be used to simulate them. In these models the physical variables are decomposed into their respective wave variables and their propagation is simulated by means of delay lines. Low-pass and all-pass filters are added to simulate dissipative and dispersive effects in the medium.

As opposed to finite differencing, which discretize the wave equation (see eqs. (14) and (15)), waveguide models come from discretization of the solution of the wave equation. The solution to the one-dimensional wave equation (14) was found by D'Alembert in 1747 in terms of traveling waves:

$$p(x, t) = p^+(t - x/c) + p^-(t + x/c). \quad (21)$$

Eq. (21) shows that the physical quantity  $p$  (e.g. string displacement or acoustic pressure) can be expressed as the sum of two wave quantities traveling in opposite directions. In waveguide models waves are sampled in space and time in such a way that equality holds in (16). If propagation along a one-dimensional medium, such as a cylinder, is ideal, i.e. linear, non-dissipative and non-dispersive, wave propagation is represented in the discrete-time domain by a couple of digital delay lines (Fig. 2), which propagates wave variables, as defined in (19) with  $Z_0$  characteristic impedance of the medium. As a slight generalization, it can be seen that the wave equation in a cone is identical to the wave equation in a cylinder (eq. (14), except that  $p(x, t)$  is replaced by  $x p(x, t)$  where  $x$  is the radial position along the cone axis. Thus, the solution is a superposition of left- and right-going traveling wave components, scaled by  $1/x$  and can still be implemented by a couple of delay lines.

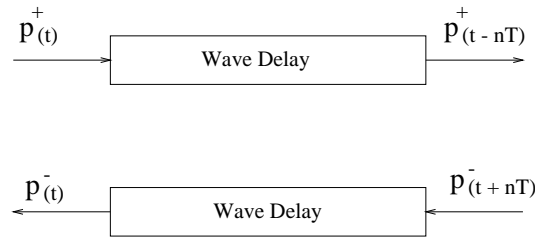


Figure 2: Wave propagation in an ideal (i.e. linear, non-dissipative and non-dispersive) medium can be represented, in the discrete-time domain, by a couple of digital delay lines.

Let us consider deviations from ideal propagation due to losses and dispersion in the resonator. Usually, these linear effects are lumped and simulated with a few filters which are cascaded with the delay lines. Losses due to terminations, internal frictions, etc., give rise to gentle low pass filters, whose parameters can be identified from measurements [23]. Wave dispersion, which is often due to medium stiffness, is simulated by means of allpass filters whose effect is to produce a frequency-dependent propagation velocity [44].

In order to increase the computational efficiency, delay lines and filters should be lumped into as few processing blocks as possible. However, when considering the interaction with an exciter or signal pick-up

from certain points of the resonator, the process of commuting and lumping linear blocks must be done with care. If the excitation is a velocity signal injected into a string, it will produce two velocity waves outgoing from the excitation point, and therefore at least two delay lines will be needed to represent propagation. The process of commuting and lumping must maintain the semantics at the observation points, while in the other points of the structure it is not necessary to have a strict correspondence with the physical reality.

Another aspect that we like to mention is that of simulating fractional delays. This is necessary when modeling musical instruments, since the proper tuning usually requires a space discretization much finer than dictated by the sample rate. More generally, fractional delay lengths are needed whenever time-varying acoustic objects (such as a string which is varying its length) are being modeled by digital waveguide networks. For this purpose, allpass filters or Lagrange interpolators of various orders can be used [45], the former suffering from phase distortion in high frequency, the latter suffering from both phase and amplitude distortion. However, low-order filters of both families can be used satisfactorily in most practical cases. In other cases, the problem of designing a tuning filter is superseded by the more general problem of modeling wave dispersion [46]. When the physical medium is changing its internal properties during vibration (e.g. a string exhibiting tension modulation), we should use a digital delay line that allows a continuous variation of the spatial sampling rate [47], or integrate the effects of the internal changes along the whole resonator length so that it becomes possible to treat them in a lumped fashion as length modulations [48].

So far, we have talked about one-dimensional resonators, but many musical instruments (e.g. percussions) and most of the real-world objects are subject to deformation along several dimensions. The algorithms presented so far can be adapted to the case of multidimensional propagation of waves, even though new problems of efficiency and accuracy arise. All the models grow in computational complexity with the increase of dimensionality, and for any of them, the choice of the right discretization grid is critical. For example, a rectangular waveguide mesh can be effective for simulating a vibrating flexible membrane, but the simulation of wave propagation turns out to be exact only along the diagonals of the mesh, while elsewhere it is affected by a dispersive phenomenon due to the fact that we are simulating circular waves by portions of plane waves. Waveguide meshes are shown to be equivalent to special kinds of finite-difference schemes [49], so that the von Neumann analysis can be used to evaluate the numerical properties of the algorithms. Special attention has to be paid to the dependence of the dispersion factor on frequency, direction, and mesh geometry, because this influences the distribution of resonances, and therefore affects the tone color and intonation. For membrane simulation, one of the most accurate yet efficient meshes is the triangular mesh [50], [51]. Interpolated waveguide meshes have recently been introduced for improving the accuracy while using simple geometries [52]. For three-dimensional wave propagation, the tetrahedral mesh is very attractive because its junctions can be implemented without any multiplication [53]. As compared to more conventional numerical techniques, such as finite difference schemes, the waveguide meshes offer the advantage that all the signals in the discrete-time system have a physical meaning, so that it is relatively straightforward to augment the model with some extra load (e.g. the air load for a membrane) or with nonlinearities.

## 6 Non linear musical oscillators

Nonlinearities assume a great importance in acoustic systems, especially where a wave-propagation medium is excited. A good model for these nonlinear mechanisms is essential for timbral quality, and is the real kernel of a physical model, which could otherwise be reduced to linear postprocessing of an excitation signal. Since the area where the excitation takes place is usually small, it makes sense to use lumped models for the excitation nonlinearity. In some cases, physical measurements provide a representation of the relation among some physical variables involved in excitation, and this relation can be directly implemented in the simulation. For example, for a simplified bowed string the transversal velocity as a function of force can be

found in the literature [54] for different values of bow pressure and velocity (which are control parameters). The instantaneous non-linear function can be approximated analytically or sampled and stored in a lookup table, which is in general multidimensional [55].

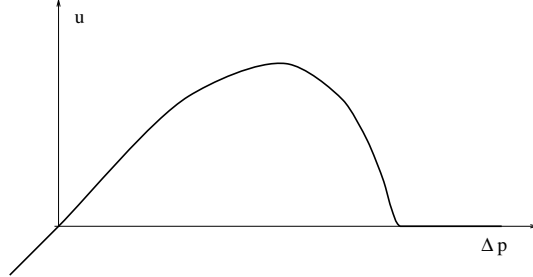


Figure 3: Nonlinear relation between pressure difference  $\Delta p$  (inside pressure - mouth pressure) and particle velocity  $u$  of the air entering into a clarinet.

As an example, let us consider a rough model of clarinet. The non-linear block representing the reed can simply be an instantaneous non-linear map (Fig. 3) relating the particle velocity  $u_r$  to the pressure difference  $\Delta p$ :

$$u_r = F_r(\Delta p) = F_r(p_m - p) , \quad (22)$$

where  $p_m$  is the player's mouth pressure and  $p$  is the pressure inside the bore at the excitation point [54]. Let us assume that there is, right after the exciter, a tract of constant-section tube having the characteristic impedance  $Z_0$ , and terminated by the radiation impedance  $Z(s)$  of the bell. In order to simplify the analysis, we can see the tube as a lossless transmission line for the dual quantities pressure and flow, so that the D'Alembert decomposition (21) does hold. This use of instantaneous nonlinearities gives rise to a general scheme of nonlinear oscillator, depicted in Fig. 4, which is composed of an instantaneous map  $y = G(x, u)$ , possibly dependent on input parameters or signals  $u$ , a delay-line section, which determines the periodicity, and a linear filter  $R(z)$  (see Eq. (7)) which can be tuned to give the desired spectral dynamics. In the case of the simplified clarinet  $x \doteq p^-$ ,  $y \doteq p^+$  and  $u \doteq p_m$ . If the filter  $R(z)$  is reduced to a constant  $r$ , and the input signal  $u$  is constant, the system evolution is described by the iterated map

$$y(n) = G[ry(n - m), u] = F[y(n - m)] \quad (23)$$

where  $m$  is the total length of the delay lines. This formulation permits us to introduce qualitative reasoning about the conditions for establishing oscillations and the periodic, multiperiodic, or chaotic nature of the oscillations themselves [56], [57]. Notice that Eq. (23) is the non linear generalization of Eq. (8).

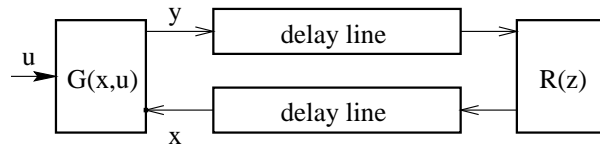


Figure 4: A computational model for non linear oscillators, useful for musical sounds.

In order to achieve a satisfactory audio quality of physics-based models, it is often necessary to use structures which are far more complex than that of the simplified clarinet. A first improvement over instantaneous non-linear excitation is considering the dynamics of the exciter: this implies the introduction of a state (i.e., memory) inside the non-linear block. When physicists study the behavior of musical instruments, they often use dynamic models of the exciter. A non-trivial task is the translation of these models into efficient computational schemes for real-time sound synthesis. A general structure has been found for good

simulations of wide instrumental families [20], [58]. In figure 5, this structure is schematically depicted. The block NL is a non-linear instantaneous function of several variables, while the block L is a linear dynamic system enclosing the exciter memory. The computational model is described by

$$\begin{aligned} \mathbf{y}(n) &= \mathbf{F}_{\text{NL}}[\mathbf{x}(n), \mathbf{u}(n), \mathbf{u}_{\text{E}}(n)] \\ \mathbf{x}(n+1) &= \mathbf{F}_{\text{L}}[\mathbf{x}(n), \mathbf{u}(n), \mathbf{u}_{\text{E}}(n), \mathbf{y}(n)] \end{aligned}$$

where  $\mathbf{x}$  is the linear system state,  $\mathbf{u}$  and  $\mathbf{u}_{\text{E}}$  are respectively the exciting actions and the signals coming from the resonator,  $\mathbf{y}$  is the output. Studies and simulations have shown that reeds, air jets, bows and percussions, can all be represented by this scheme to a certain extent.

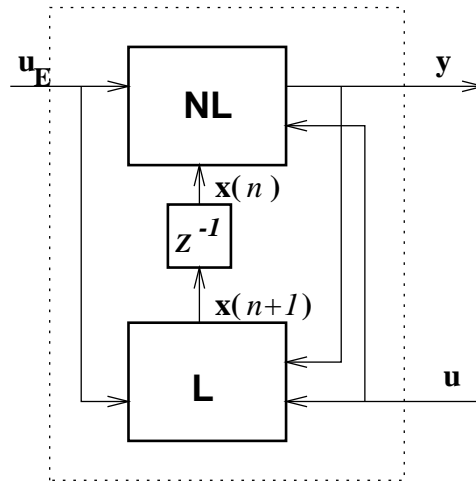


Figure 5: Scheme for a Dynamic Exciter

If a lumped modeling of dynamics is used, it has to be represented by discretization of some differential equations. To this end, one approach is to apply a finite difference scheme to these equations and introduce the instantaneous nonlinearity in the resulting signal flowgraph [59], [60]. This procedure must be tackled carefully because it is easy to come up with delay-free (noncomputable) loops: in some cases these inconsistencies can be overcome by introducing fictitious delay elements, but at the sampling rates that we can afford it is likely that these delays are not acceptable, so that other techniques are needed. A notable example is found in the piano hammer-string interaction [61], whose model can be applied to several percussive sound sources. The computable discretization scheme of the non-linear hammer is obtained from the straightforward finite difference approximation of the dynamics, from which a term dependent on previous and known terms is separated from an instantaneous and unknown term. Finally, the instantaneous nonlinearity is recast into new variables in such a way that the delay-free loop is canceled [62]. Fig. 6 compares the force signals obtained by the hammer-string model with and without the correct elimination of delay-free loops. It can be seen that the simple introduction of a decoupling unit delay in the delay-free loop leads to instabilities.

Another approach to the representation of the exciter dynamics is to resort to a model based on lumped linear or non-linear circuit components. The circuit components can be translated into the discrete-time domain using the wave digital filter method briefly explained in sec. 5.4 [41]. This method has recently been extended to cover non-linear elements without [63] or with [64] memory, i.e., resistances or reactances.

When the need is not for a special-case model, but for a model which is adaptable to several sound sources, or when it is desirable to tune the model from sampled sounds, other representations have to be chosen for the non-linear exciter. It has been proposed to express one-dimensional memoryless nonlinearities as a polynomial function whose coefficients can be identified by Kalman filtering [65], or adaptively by the LMS algorithm [66].



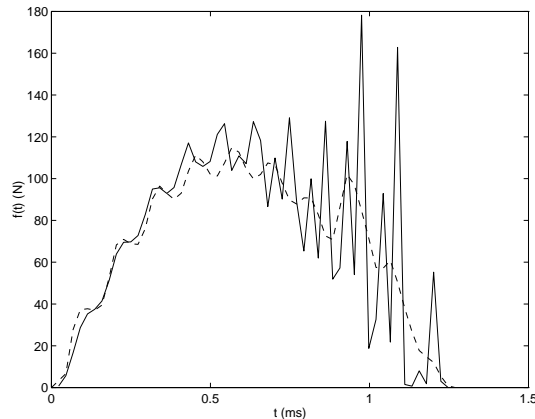


Figure 6: Force in hammer-string interaction computed adding a non physical delay (continuous line) and solving the noncomputability problem (dashed line).

Alternatively, some general-purpose approximation network might be used as a generalized nonlinearity. For example, a Radial-Basis-Function network [67] built with a small number of gaussian kernels has shown its effectiveness in representing the severe nonlinearities found in wind and string instruments [68]. The parameters of these networks have to be identified by some global optimization technique, such as the genetic algorithm, applied to solve a spectral or waveform matching problem.

Nonlinearities might occur in the resonator as well, especially when it is driven into large vibrations. These nonlinearities are usually mild, so that simple saturation characteristics introduced somewhere in the resonator work just fine. However, for some sound sources the resonator nonlinearity is more critical, and special techniques must be devised in order to ensure that energy is not introduced or lost improperly. For instance, a time-variant one-pole allpass filter has been proposed for reproducing the kind of nonlinearity found in the bridge of the sitar [69]. Similar nonlinearities are found distributed in two-dimensional resonators, such as plates, and it is not yet clear how to simulate them by means of a small number of filters, properly placed in the computational structure.

## 7 Conclusions

We have presented some of the main modeling approaches to musical sounds, most of them based on a mathematical description of sound sources, musical instruments, or enclosed spaces. Synthesis by physical models, begun as a method for studying the physical behavior of traditional musical instruments, is now one of the most useful methods of sound generation for music. Well-established models exist for several instrumental families and non-musical sound sources. For some other sources, effective and efficient models are still to be invented. Moreover, there are several structural and operational problems that need to be solved before physical modeling approaches the versatility and usability that are required by interactive and multimodal computer systems of the future. This is especially true when considering the problem of making the algorithms suitable for control purposes. The design of aids that facilitate an effective real-time control of the synthetic instruments should be considered as a task closely related to sound modeling and design. There should be an easy interpretation of the relationships among parameters, and physically-described sound models often lend themselves to this purpose. However, a higher level of abstraction is often desirable so that control can be exerted in terms of musical gestures.

Other interesting directions for future research are found in the abstraction of general computational structures, thus leveling up the idiosyncrasies of models of specific mechanical systems. These general

structures will be useful to generate new classes of sounds and to relate them to existing classes that have thorough mathematical and acoustical descriptions. Another aspect that is emerging as crucial to achieve high sound quality is the perceptual evaluation of physical models, especially as far as the approximations introduced in model formulation and computational realization are concerned.

To conclude, we think that the joint research efforts of mathematicians, engineers, and physicists, together with active experimentation by musicians and psychologists, will enrich the expressive capabilities of virtual musical instruments and will give rise to new forms of human-computer communication via non-speech sounds. This scenario is a natural extension of the old tradition of cooperation and mutual intersection between science and music.

## References

- [1] P. Schaeffer, *Traité des Objets Musicaux*. Paris, France: Éditions du Seuil, 1966.
- [2] J.-C. Risset, "An Introductory Catalog of Computer-Synthesized Sounds," tech. rep., Bell Laboratories, Murray Hill, N.J., 1969.
- [3] B. Truax, *Handbook for Acoustic Ecology*. Vancouver: A.R.C. Publications, 1978.
- [4] S. McAdams, "Music: a science of mind?," *Contemporary Music Review*, vol. 2, no. 1, pp. 1–61, 1987.
- [5] J. O. Smith, "Viewpoints on the History of Digital Synthesis," in *Proc.. Int. Computer Music Conf.*, (Montreal, Canada), pp. 1–10, ICMA, 1991.
- [6] J. A. Moorer, "Signal processing aspects of computer music," *Proceedings of the IEEE*, vol. 65, no. 8, pp. 1108–37, 1977.
- [7] G. Haus, *Music Processing*. Madison, WI: A-R Editions, 1993.
- [8] C. Roads, S. T. Pope, A. Piccialli, and G. De Poli, *Musical Signal Processing*. Lisse: Swets & Zeitlinger, 1997.
- [9] G. De Poli, A. Piccialli, and C. Roads, *Representations of Musical Signals*. Cambridge, Mass.: MIT Press, 1991.
- [10] C. Roads, *The Computer Music Tutorial*. Cambridge, Mass.: MIT Press, 1996.
- [11] G. De Poli, "A tutorial on digital sound synthesis techniques," *Computer Music Journal*, vol. 7, no. 4, pp. 8–26, 1983. Reprinted in *The Music Machine*, C. Roads Ed., 429–47, MIT Press 1991.
- [12] A. H. Gray and J. D. Markel, "A Normalized Digital Filter Structure," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 23, pp. 268–277, June 1975.
- [13] K. Karplus and A. Strong, "Digital Synthesis of Plucked String and Drum Timbres," *Computer Music J.*, vol. 7, no. 2, pp. 43–55, 1983.
- [14] D. A. Jaffe and J. O. Smith, "Extensions of the Karplus-Strong Plucked String Algorithm," *Computer Music J.*, vol. 7, no. 2, pp. 56–69, 1983.
- [15] D. Gabor, "Acoustical Quanta and the Theory of Hearing," *Nature*, vol. 159, pp. 591–594, May 1947.
- [16] C. Roads, "Asynchronous granular synthesis," in *Representations of Musical Signals* (G. De Poli, A. Piccialli, and C. Roads, eds.), pp. 143–186, Cambridge, MA.: MIT Press, 1991.

- [17] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing* (C. Roads, S. T. Pope, A. Piccialli, and G. De Poli, eds.), pp. 91–122, Swets & Zeitlinger, 1997.
- [18] T. S. Verma, S. N. Levine, and T. H. Y. Meng, "Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals," in *Proc.. Int. Computer Music Conf.*, (Thessaloniki, Greece), pp. 164–167, ICMA, Sept. 1997.
- [19] J. M. Chowning, "The Synthesis of Complex Audio Spectra by Means of Frequency Modulation," *J. Audio Eng. Soc.*, vol. 21, no. 7, pp. 526–534, 1973.
- [20] G. Borin, G. De Poli, and A. Sarti, *Sound Synthesis by Dynamic Systems Interaction*, vol. Readings in Computer-Generated Music, pp. 139–160. IEEE Computer Society Press, 1992. D. Baggi, editor.
- [21] G. Borin, G. De Poli, and A. Sarti, "Algorithms and Structures for Synthesis using Physical Models," *Computer Music J.*, vol. 16, pp. 30–42, Winter 1992.
- [22] J. O. Smith, "Efficient synthesis of stringed musical instruments," in *Proc. 1993 International Computer Music Conference*, (Tokyo, Japan), pp. 64–71, 1993.
- [23] V. Välimäki, J. Huopaniemi, M. Karjalainen, and Z. Jánosy, "Physical Modeling of Plucked String Instruments with Application to Real-Time Sound Synthesis," *J. Audio Eng. Soc.*, vol. 44, no. 5, pp. 331–353, 1996.
- [24] J. F. Blinn and M. E. Newell, "Texture and reflection in computer generated images," *Communications of the ACM*, vol. 19, no. 10, pp. 542–547, 1976.
- [25] J. von Neumann, *Theory of Self-Reproducing Automata*. Urbana, IL: Univ. of Illinois Press, 1966.
- [26] S. Wolfram, "Computational theory of cellular automata," *Comm. in Math. Physics*, vol. 96, pp. 15–57, 1984.
- [27] J. L. Florens and C. Cadoz, "The physical model: Modeling and simulating the instrumental universe," in *Representations of Musical Signals* (G. De Poli, A. Piccialli, and C. Roads, eds.), pp. 227–268, Cambridge, MA.: MIT Press, 1991.
- [28] C. Cadoz, A. Luciani, and J.-L. Florens, "CORDIS-ANIMA: A Modeling and Simulation System for Sound Synthesis - The General Formalism," *Computer Music J.*, vol. 17, pp. 19–29, Spring 1993.
- [29] E. Incerti and C. Cadoz, "Topology, geometry, matter of vibrating structures simulated with cordis-anima. sound synthesis methods," in *Proc. Int. Computer Music Conf.*, (Banff, Canada), pp. 96–103, ICMA, 1995.
- [30] C. Cadoz, A. Luciani, and J.-L. Florens, "Physical Models for Music and Animated Image. The use of CORDIS-ANIMA in "ESQUISSES" a Music Film by ACROE," in *Proc.. Int. Computer Music Conf.*, (Aarhus, Denmark), pp. 11–18, ICMA, Sept. 1994.
- [31] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*. New York: Springer-Verlag, 1991.
- [32] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*. Cambridge: Cambridge University Press, 1988.
- [33] J. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*. Pacific Grove, CA: Wadsworth & Brooks, 1989.

- [34] P. M. Morse, *Vibration and Sound*. New York: American Institute of Physics for the Acoustical Society of America, 1991. 1st ed. 1936, 2nd ed. 1948.
- [35] A. Chaigne, “On the Use of Finite Differences for Musical Synthesis. Application to Plucked Stringed Instruments,” *J. Acoustique*, vol. 5, pp. 181–211, 1992.
- [36] B. Gazengel, J. Gilbert, and N. Amir, “Time Domain Simulation of Single Reed Wind Instrument. From the Measured Input Impedance to the Synthesis Signal. Where are the Traps?,” *Acta Acustica*, vol. 3, pp. 445–472, Oct. 1995.
- [37] A. V. Oppenheim and R. W. Schaffer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1989.
- [38] J. A. Moorer, “The Manifold Joys of Conformal Mapping: Applications to Digital Filtering in the Studio,” *J. Audio Eng. Soc.*, vol. 31, no. 11, pp. 826–840, 1983.
- [39] F. Avanzini, “Higher order numerical methods in physical models of musical instruments. a study on the clarinet,” in *Proc. Diderot Forum on Mathematics and Music*, (Vienna, Austria), pp. 11–19, 1999.
- [40] C. Wan and A. M. Schneider, “Further improvement in digitizing continuous-time filters,” *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 533–542, 1997.
- [41] A. Fettweis, “Wave Digital Filters: Theory and Practice,” *Proc. IEEE*, vol. 74, pp. 270–327, Feb. 1986.
- [42] V. Belevitch, *Classical Network Theory*. San Francisco: Holden-Day, 1968.
- [43] J. O. Smith, “Physical modeling using digital waveguides,” *Computer Music J.*, vol. 16, pp. 74–91, Winter 1992.
- [44] D. Rocchesso and F. Scalcon, “Bandwidth of perceived inharmonicity for physical modeling of dispersive strings,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 597–601, Sept. 1999.
- [45] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, “Splitting the Unit Delay—Tools for Fractional Delay Filter Design,” *IEEE Signal Processing Mag.*, vol. 13, pp. 30–60, Jan 1996.
- [46] D. Rocchesso and F. Scalcon, “Accurate dispersion simulation for piano strings,” in *Proc. Nordic Acoustical Meeting (NAM-96)*, pp. 407–414, June 1996.
- [47] D. Rocchesso, “Fractionally-addressed delay lines,” *IEEE Transactions on Speech and Audio Processing*, 2000. Accepted for publication.
- [48] V. Välimäki, T. Tolonen, and M. Karjalainen, “Plucked-string synthesis algorithms with tension modulation nonlinearity,” in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, (Phoenix, Arizona), pp. 977–980, IEEE, March 1999.
- [49] S. A. Van Duyne and J. O. Smith, “The 2-d digital waveguide mesh,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (Mohonk, NY), IEEE, 1993.
- [50] F. Fontana and D. Rocchesso, “Physical modeling of membranes for percussion instruments,” *Acustica*, vol. 83, pp. 529–542, Jan. 1998. S. Hirzel Verlag.
- [51] F. Fontana and D. Rocchesso, “Signal-theoretic characterization of waveguide mesh geometries for models of two-dimensional wave propagation in elastic media,” *IEEE Trans. Speech and Audio Processing*, vol. 8, 2000. Accepted for publication.

- [52] L. Savioja and V. Välimäki, "Improved discrete-time modeling of multi-dimensional wave propagation using the interpolated digital waveguide mesh," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Munich*, pp. 459–462, Apr. 1997.
- [53] S. A. Van Duyne and J. O. Smith, "The tetrahedral digital waveguide mesh," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (Mohonk, NY), p. 9a.6, IEEE, Oct. 1995.
- [54] M. E. McIntyre, R. T. Schumacher, and J. Woodhouse, "On the oscillations of musical instruments," *J. Acoustical Soc. of America*, vol. 74, no. 5, pp. 1325–1345, 1983.
- [55] D. Rocchesso and F. Turra, "A Real-Time Clarinet Model on MARS Workstation," in *Proc. X Colloquium Mus. Inform.*, (Milano), pp. 210–213, AIMI, Dec. 1993.
- [56] X. Rodet, "Models of musical instruments from chua's circuit with time delay," *IEEE Trans. Circuits and Systems-II*, vol. 40, pp. 696–701, Oct. 1993.
- [57] X. Rodet, "Stability/instability of periodic solutions and chaos in physical models of musical instruments," in *Proc. Int. Computer Music Conf.*, (Aarhus, Denmark), pp. 386–393, ICMA, 1994.
- [58] D. Rocchesso, "Modelli Generalizzati di Strumenti Musicali per la Sintesi del Suono," *Rivista Italiana di Acustica*, vol. 17, no. 4, pp. 61–71, 1993. Paper invited by the panel of "Amedeo Giacomini" award, Associazione Italiana di Acustica.
- [59] A. Chaigne and A. Askenfelt, "Numerical Simulations of Piano Strings. I. A Physical Model for a Struck String using Finite Difference Methods," *J. Acoustical Soc. of America*, vol. 95, pp. 1112–1118, Feb 1994.
- [60] S. A. Van Duyne, J. R. Pierce, and J. O. Smith, "Traveling Wave Implementation of a Lossless Mode-Coupling Filter and the Wave Digital Hammer," in *Proc. Int. Computer Music Conf.*, (Aarhus, Denmark), pp. 411–418, ICMA, Sept. 1994.
- [61] G. Borin and G. De Poli, "A Hysteretic Hammer-String Interaction Model for Physical Model Synthesis," in *Proc. Nordic Acoustical Meeting (NAM-96)*, (Helsinki, Finland), pp. 399–406, June 1996.
- [62] G. Borin, G. D. Poli, and D. Rocchesso, "Elimination of delay-free loops in discrete-time models of nonlinear acoustic systems," *IEEE Transactions on Speech and Audio Processing*, vol. 8, Sept. 2000. Accepted for publication.
- [63] K. Meerkötter and R. Scholtz, "Digital simulation of nonlinear circuits by wave digital filter principles," in *Proc. IEEE Intl. Symp. on Circuits and Systems*, (Portland, OG), pp. 720–723, May 1989.
- [64] A. Sarti and G. De Poli, "Toward nonlinear wave digital filters," *IEEE Trans. Signal Processing*, vol. 47, pp. 1654–1668, June 1999.
- [65] P. R. Cook, "Non-linear periodic prediction for on-line identification of oscillator characteristics in woodwind instruments," in *Proc. Int. Computer Music Conf.*, (Montreal, Canada), pp. 157–160, ICMA, 1991.
- [66] G. P. Scavone, "Combined linear and non-linear periodic prediction in calibrating models of musical instruments to recordings," in *Proc. Int. Computer Music Conf.*, (Aarhus, Denmark), pp. 433–434, ICMA, 1994.
- [67] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 79, no. 9, pp. 1481–1497, 1990.

- [68] C. Drioli and D. Rocchesso, "A Generalized Musical-Tone Generator with Application to Sound Compression and Synthesis," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, (Munich, Germany), pp. 431–434, 1997.
- [69] J. R. Pierce and S. A. Van Duyne, "A Passive Nonlinear Digital Filter Design which Facilitates Physics-based Sound Synthesis of Highly Nonlinear Musical Instruments," *J. Acoustical Soc. of America*, vol. 101, pp. 1120–1126, Feb. 1997.
- [70] U. Zoelzer, D. Arfib, G. De Poli, P. Dutilleux, D. Rocchesso, and X. Serra, *Digital Audio Effects*. Chichester Sussex, UK: John Wiley and Sons, Ltd., 2001. To appear.
- [71] J. Blauert, *Spatial Hearing: the Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1983.
- [72] G. S. Kendall, "A 3-D Sound Primer: Directional Hearing and Stereo Reproduction," *Computer Music J.*, vol. 19, pp. 23–46, Winter 1995.
- [73] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. Boston, MA: Academic Press, 1994.
- [74] J. M. Chowning, "The Simulation of Moving Sound Sources," *J. Audio Eng. Soc.*, vol. 19, no. 1, pp. 2–6, 1971. Reprinted in the *Computer Music Journal*, June 1977.
- [75] F. R. Moore, *Elements of Computer Music*. Englewood Cliffs, N.J.: Prentice-Hall, 1990.
- [76] O. Warusfel, E. Kahle, and J. P. Jullien, "Relationships between Objective Measurements and Perceptual Interpretation: The Need for Considering Spatial Emission of Sound Sources," *J. Acoustical Soc. of America*, vol. 93, pp. 2281–2282, 1993.
- [77] R. Caussé, P. Dérogis, and O. Warusfel, "Radiation of Musical Instruments and Improvement of the Sound Diffusion Techniques for Synthesized, Recorded or Amplified Sounds (revisited)," in *Proc. Int. Computer Music Conf.*, (Banff, Canada), pp. 359–360, ICMA, 1995.
- [78] H. Kuttruff, *Room Acoustics*. Essex, England: Elsevier Science, 1991. Third Ed.; First Ed. 1973.
- [79] M. F. Cohen and J. R. Wallace, *Radiosity and Realistic Image Synthesis*. Cambridge, MA: Academic Press, 1993.
- [80] M. R. Schroeder, "Natural-Sounding Artificial Reverberation," *J. Audio Eng. Soc.*, vol. 10, pp. 219–233, July 1962.
- [81] M. R. Schroeder, "Computer Models for Concert Hall Acoustics," *American Journal of Physics*, vol. 41, pp. 461–471, 1973.
- [82] J. A. Moorer, "About this Reverberation Business," *Computer Music J.*, vol. 3, no. 2, pp. 13–18, 1979.
- [83] D. Rocchesso, "The Ball within the Box: a sound-processing metaphor," *Computer Music J.*, vol. 19, pp. 47–57, Winter 1995.
- [84] J. Borish, "Extension of the Image Model to Arbitrary Polyhedra," *J. Acoustical Soc. of America*, vol. 75, pp. 1827–1836, June 1984.
- [85] D. Rocchesso and J. O. Smith, "Circulant and Elliptic Feedback Delay Networks for Artificial Reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 51–63, Jan. 1997.
- [86] D. Rocchesso, "Maximally-Diffusive yet Efficient Feedback Delay Networks for Artificial Reverberation," *IEEE Signal Processing Letters*, vol. 4, pp. 252–255, Sept. 1997.

# Importance of phase in sound modeling of acoustic instruments

Tue Haste Andersen, Kristoffer Jensen  
Department of Computer Science, University of Copenhagen  
email: {haste,krist}@diku.dk

## Abstract

Phase information is left out in many additive synthesis applications, since it is sometimes believed to have a neglectable perceptual influence on the sound quality. In order to understand the validity of this assumption a number of experiments regarding the importance of phase are presented here. The sounds under examination are all recorded from acoustic instruments or singing voice, analyzed and synthesized using additive techniques, and evaluated using psycho-physic experiments. The experiments shows that phase is of crucial importance to localization of binaural recorded sounds, and thus is necessary to preserve the directional characteristics of the sounds. Regarding the monaural qualities of instruments sounds, the phase is also shown to be of great importance when compared to sounds synthesized with arbitrary chosen phase. The results suggest that a model of phase should be constructed for use in additive synthesis and effect applications.

## 1 Introduction

The current knowledge on perceptual qualities of sound mostly stems from psycho-acoustic experiments on humans, and medical experiments on animals. While the latter kind is extremely difficult and expensive to conduct, psycho-acoustic experiments are more accessible. Psycho-acoustic experiments involves humans as objects of measurement, and hence care must be taken in the design of the experiment to reduce the stimuli as much as possible so that only the desired effects are measured. For this reason the experiments found in the literature are often conducted using simple sound stimuli such as pure tones or filtered noise. The experiments are typically designed to reveal basic relations between sound and perception, thereby increasing the general understanding of phenomena related to perception of sound. The reduction of sound stimuli makes it difficult to gain understanding about complex sounds as a whole. Even though basic understanding of some phenomena exists, it is very difficult to conduct experiments and gain understanding about complex sounds with many partials, resonances, transients and so forth. In this chapter a number of psycho-acoustic experiments will be presented which documents the importance of one specific parameter in spectral sound models, namely the phase.

It is widely known that phase has a great importance in spatial perception, however no experiments involving complex sound stimuli such as musical sounds has been found. For this reasons a preliminary study was conducted [1] which has been extended and presented at the annual International Computer Music Conference [3] and in [2] as part of this study. Little is known about the importance of keeping the phase relations among partials of voiced sounds, and thus an experiment regarding perception of phase of monaural sounds has also been conducted. In section 2, a small overview

of experiments found in the literature will be given regarding spatial hearing in the horizontal plane, continued with an overview of previous experiments on monaural phase perception in section 3. Section 4 describes the binaural listening experiments regarding phase, and section 5 the monaural experiments.

## 2 Directional hearing and phase

Sound produced in the environment surrounding the listener has been distorted in amplitude, phase and spectral content when perceived. The distortion is caused by the human head and torso and the surrounding environment, and are used by the brain to decode spectral information. The two most prominent cues used to determine direction of sounds in the horizontal plane is governed by the *duplex theory*, developed by John Strutt better known as Lord Rayleigh [7]. According to this theory two primary cues are used - interaural time difference, most important for sounds of low frequency, and interaural level difference for high frequency sounds.

### 2.1 Interaural Time Difference

Interaural Time Difference (ITD), sometimes called interaural phase difference, is the difference in time between a sound arriving at the two ears. The difference is perceptible when the wavelength of the sinusoid is approximately larger than the length of the head [25], that is roughly,  $0.18\text{m}/344\frac{\text{m}}{\text{s}} = 0.5\text{ms}$ , which corresponds to sinusoids with frequencies below 2 kHz.

The perceived angle of incidence can be predicted as function of ITD by a model described in detail in [12]. The model describes the pressure on the surface of a rigid sphere

with a radius equivalent to the radius of an average human head. When only low frequency sounds are considered the model can be simplified to [12]:

$$\text{ITD}_p \approx \frac{3r}{c} \sin \Delta_{\text{inc}}, \quad (1)$$

where  $\text{ITD}_p$  is the predicted interaural time difference,  $r$  the radius of the head,  $c$  the speed of sound in air, and  $\Delta_{\text{inc}}$  the angle of incidence. A solution for the high frequency ITD can also be found, but is not as important as the ITD for low frequency sounds.

## 2.2 Interaural Level Difference

Interaural level difference (ILD) has little to do with phase preservation in signal models, but is together with ITD the most important cue for localization of sounds in the horizontal plane. The ILD is defined as [12]:

$$\text{ILD} = 20 \log p_L - 20 \log p_R \quad (2)$$

and is given in dB, where  $p_L$  and  $p_R$  is the pressure magnitudes at the left and right ear correspondingly. Like the ITD the ILD can be predicted using a model based on sound pressure on the surface of a rigid sphere, with radius equal to the average radius of a human head.

## 2.3 Other directional cues

Other cues than ILD and ITD play an important role in spatial perception; not only in the vertical plane, but also in the horizontal, where ILD or ITD's cause ambiguity. All these effects are related to distortion of the spectrum caused by head, body and especially pinna, where certain frequencies are attenuated, and others amplified, depending on the angle of incidence.

## 3 Perceptual qualities and phase

Experiments regarding phase and perception of other sound qualities than spatial information has also been conducted in the past by a number of researchers. These are often referred to as monaural since they are perceptible by listening with one ear. No complete overview of the literature has been found, as well as no throughout examination of the phenomena involved with perception of phase exists. In this section an overview of the results of importance to musical sound modeling is given.

In many of the experiments harmonic sounds are used, defined as *complex tones* by the sound pressure  $p$  as function of time,  $t$  [18]:

$$p(t) = \sum_{n=1}^N a_n \sin(2\pi nft + \phi_n). \quad (3)$$

Thus, the timbre of a complex tone might depend on the amplitude pattern  $a_1, a_2, \dots, a_N$  and the phase pattern  $\phi_1, \phi_2, \dots, \phi_N$  of the successive harmonics. The complex tones used in psycho-acoustic experiments are most often simple, in the sense that they are composed of few harmonics, and do not resemble most natural occurring sounds.

### 3.1 Pitch perception

Models of pitch perception found in the literature, e.g. Wightman [22] or Goldstein [9], often discard the phase of the frequency components. These models contradicts time domain models [20] where the pitch of a complex tone is given as function of the time interval between peaks in the waveform, in "some dominant region of the basilar membrane" [16]. To verify if the relative phase of harmonics of a complex tone is of importance to the perception of pitch, Moore conducted a number of experiments [16]. Here it was concluded that phase did in fact have an effect on perceived pitch in some cases. Most often however, it only affected the strength of the perceived pitch. This was later verified by Cariani and Delgutte [5].

### 3.2 Does phase affect timbre?

The first experiment concerning the perception of timbre of complex tone in relation to phase was conducted by von Helmholtz [19]. Using a special technique he was able to generate complex tones consisting of eight partials with fundamental frequencies of 120 and 240 Hz. The phase of each partial could be varied, and von Helmholtz was therefore able to perform experiments with the importance of phase to perception. From his experiments with different phase patterns of these two tones he concluded that [18]: "(1) the changes in timbre are not distinct enough to be observed after a few seconds required to alter the phases; anyhow these changes are too small to transform one vowel in another; (2) harmonics beyond the sixth to eighth give dissonances and beats, so it is not excluded that, for these higher harmonics, a phase effect does exist." These conclusions have often been interpreted as phase did not have any influence on timbre [18], even though experiments later showed that indeed it has an influence [14, 8, 18]. The answer to this question is of crucial importance to spectral sound models since up to one third of the information can be discarded if phase is not important to the perception of sound.

Plomp later conducted a number of experiments involving complex tones of ten harmonics, with equal spectral envelope but different phase shifts, producing four different waveforms as shown on figure 1. The most important findings in this work was:

- The maximum effect of phase on timbre is the difference between a complex tone in which the harmonics



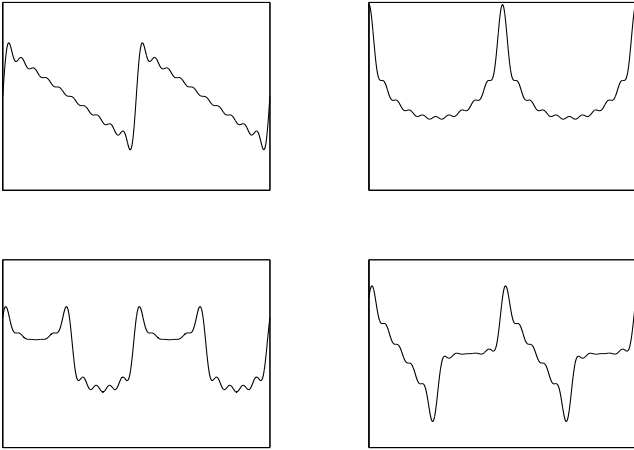


Figure 1: Waveforms used in experiments by Plomp to determine the importance of phase to complex tones. Reproduced after [18].

are in phase and one in which alternate harmonics differ in phase by  $90^\circ$ .

- The effect of lowering each successive harmonic by 2 dB is greater than the maximum phase effect described above.
- The effect of phase on timbre appears to be independent of the sound level and the spectrum.

Patterson [17] furthermore presented psycho-acoustic experiments involving so called altering phase waves or APH, which is complex tones with a flat spectrum of up to 31 harmonics where the phase difference  $D$  between each harmonic is constant. In the conducted experiment  $D$  was varied along with repetition rate, number of harmonics, and sound pressure level. The results showed that for lower repetition rates, the changes in  $D$  was better perceived, and furthermore that the changes was dependent on sound pressure level and frequency content.

These experiments all demonstrate that phase is perceptually important in many situations. However, it still remains a question to which extend phase is of importance to the modeling of natural occurring sounds using spectral sound models.

### 3.3 Other effects pertaining to monaural perception of phase

While there is no consensus on the importance of phase of complex tones and other periodic sounds, it is clear that for non-periodic sounds such as transients, the phase of the frequency components is of great importance to the perception of sounds.

Wakefield et.al. [21] conducted a study of the perception of transients using filtered noise, where a two-interval forced-choice adaptive psycho-physical procedure were used to find the just noticeable difference (JND) between a given sound and a copy of the sound, where the magnitude spectrum was smoothed and the phase spectrum held constant. The surprising result was that for the two sounds there were great differences in the JND level, depending on the phase pattern used. It was concluded that [21] “the effect for short duration signals is greater than what the (sparse) literature on the auditory perception of transients would suggest.”

Another effect pertaining to signal models where phase is left out or incorrectly modeled is the reverberance or phasiness [13]. This phenomena is most prominent in time or pitch scaled natural sounds, where a block based transformation system is used<sup>1</sup>.

## 4 Binaural experiments

To verify the importance of phase for natural occurring musical sounds, and to gain a better understanding of phase preservation in the additive model, a number of experiments with directional hearing in the horizontal plane of binaural recorded sounds were conducted.

### 4.1 Method

The system, which is also used here, consists of an analysis part, where amplitude, frequency and phase are extracted as time varying parameters, using the Short Time Fourier Transform (STFT). The sound is then synthesized by

$$s(t) = \sum_{n=0}^k a_n(t) \cos[\theta_n(t)], \quad (4)$$

where  $k$  is the number of partials,  $a_n(t)$  the time varying amplitude, and  $\theta_n(t)$  is the time varying phase. For binaural analysis/synthesis, there are of course two sets of parameters.

The different synthesis methods used in the experiments are:

1. Using amplitude, frequency and phase
2. Using only amplitude and frequency, leaving out the phase information
3. Using (1) with spectral envelope normalized in both channels

For synthesis with phase information,  $\theta_n(t)$  is calculated using cubic interpolation between frequency and phase values

<sup>1</sup>A demonstration of the phenomenon can be found at S.M. Sprenger’s webpages:  
<http://www.dspdimension.com/html/timepitch.html>

[15]. For synthesis without phase information  $\theta_n(t)$  is simply obtained by integration of the measured frequency values over time:

$$\theta_n(t) = \int_0^t \omega_n(\tau) d\tau. \quad (5)$$

$\theta_n(t)$ ,  $a_n(t)$ ,  $\omega_n(t)$  are estimated once for every period of the fundamental of the sounds. The maximum occurring ITD, according to (1) is:

$$\text{ITD}_{p\text{-max}} \approx \frac{3r}{c} \sin \Delta_{\text{inc}} = \frac{3 \cdot 0.09\text{m}}{344\text{m/s}} \sin \pi/2 = 0.78\text{ms}, \quad (6)$$

and since the fundamental frequency of the sounds used in the experiment is between 80 Hz and 659 Hz, corresponding to a period length of 12.5 ms and 1.52 ms, the ITD cues will in general not be present in  $a_n(t)$  or  $\omega_n(t)$ .

In previous formal listening tests [11] and in informal listening test performed in this study, the sound quality of the additive analysis/synthesis was found to be very good. The degradation of a large variety of musical sounds was previously found to be between *imperceptible* and *perceptible, but not annoying* [11], which also corresponds well with other studies [6].

For all the resynthesized sounds, each method retains the perceptually important features, such as the attack part, very well in monaural analysis/synthesis, even though there still is a slight difference in the waveforms between the original and resynthesized sounds with phase. The resynthesis with phase retains the non-instrumental part of the sound, such as noise, better than the method without phase information.

According to the duplex theory mentioned in section 2, the perceptual important cues for use in localization is expected to be less present in sounds synthesized without phase than with phase.

## 4.2 Experiments and results

To examine the importance of phase information and other localization cues in the sound, two psycho-acoustic experiments were conducted. The first was made as a preliminary study, whereas the second was carefully designed and verified using appropriate statistical tests. The variables under control in these experiments are: Reproduction method, sound source distance, instrument, angle of incidence and fundamental frequency.

The reproduction method is either using the original recorded sound, or a synthesized sound as described in section 4.1. In all the experiments an interval scale with eight different incidence angles was used, as shown in figure 2.

The different parameters were chosen so a reasonable diversity of sounds was available in the experiment, and to test if the same results can be obtained from these experiments, as has been concluded in previous studies of simple sound stimuli found in the literature.

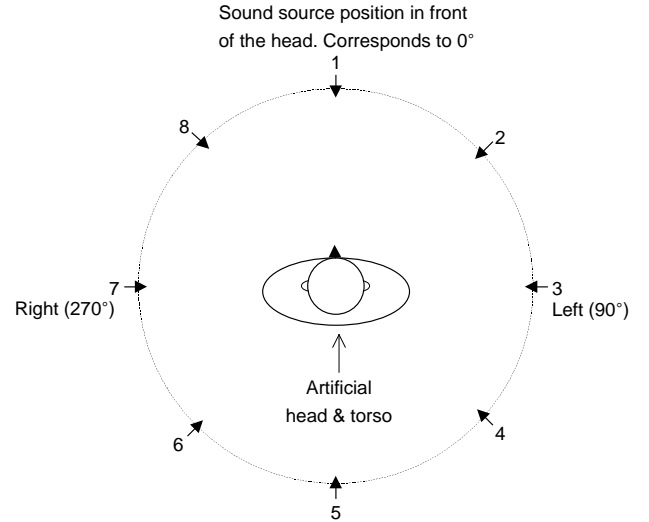


Figure 2: Setup during recording of sound sources. The artificial head and torso is placed in the middle while recordings are done at eight different angles of incidence.

### 4.2.1 Preliminary experiment

A preliminary experiment with seven subjects was first conducted and earlier reported in [1]. For the sound recordings an artificial head and torso was used, with Brel & Kjr type 4009 microphones attached to the inside of each ear. Sounds of a guitar were recorded at eight different angles, equally spaced around the head, at a distance of 0.5 meters as depicted in figure 2. The recordings were done in an office room using a DAT recorder. Two things were tested:

1. Absolute judgment: The subject listens to all sounds, both originals and resynthesized, in random order. For each sound the subject is asked to determine the direction of the sound.
2. Relative judgment: The subject listens to two sounds synthesized using the previously mentioned methods 1 and 2, played back one after the other in random order. The subject is asked which of them is easiest to determine the direction and distance of.

From the results shown on figure 3 of the first experiment where the subjects are asked to determine the absolute direction of the sounds, it is observed that the average error for original sounds and sounds synthesized with phase are almost equal. For sounds synthesized without phase, the error is substantially higher. The results clearly show that phase information is important when dealing with spatial information in binaural recorded sounds.

The notion of error is based on the difference in sound source position of the recordings, and the answer the subject

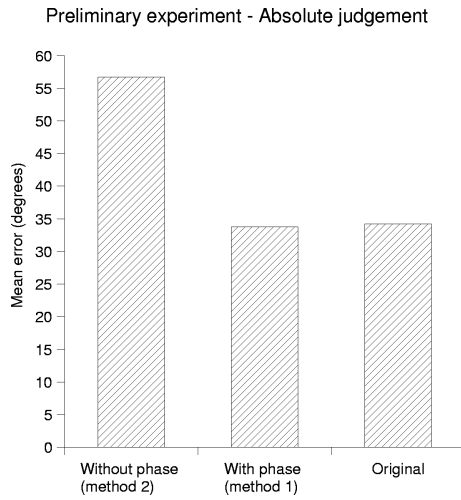


Figure 3: Average error in degrees of answers from preliminary experiment with 7 subjects, listening to 144 sounds each. The figure shows no difference in the subjects ability to localize.

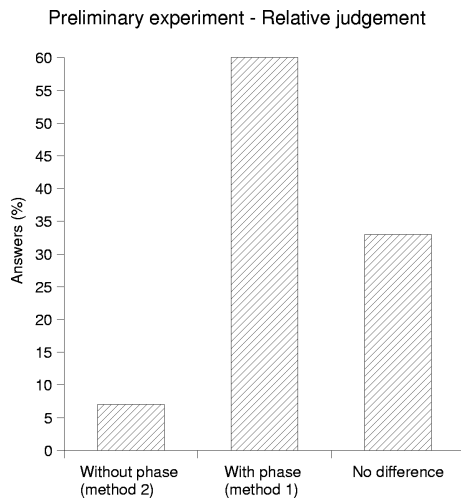


Figure 4: Answers from preliminary experiment with 7 subjects, listening to 48 sounds each. The figure shows which of the synthesis methods (1 or 2) sounds are easiest to locate in space.

gave, using the scale on figure 2. A number of corrections are made. The error is the sum of front/back error (weighted 1) and left/right error (weighted 2). Undetermined answer gives error 4.

Figure 4 shows the results from the second test with relative judgment. In 60% of the cases it is easier to determine the direction of the sounds when synthesized with phase information. For 33% of the sounds there is no difference, and for 7% it is easier to determine the direction when no phase information is used. Furthermore it is observed that for sounds coming directly in front or from the back of the listener, the sounds are more often judged to be undetermined regarding the spatial qualities. This test gives an idea about importance of phase information, but it is not clear what exactly is tested. Because sounds synthesized with and without phase information in some cases have small variations in timbre, they can be difficult to compare.

#### 4.2.2 Absolute judgment

In this experiment absolute judgment is used as in the previous experiment, but with a larger setup. The experiment was conducted as part of this study and reported in [3, 4, 2]. All sounds are recorded in an anechoic chamber to minimize reflections from the environment. Two instruments are used: singing voice and guitar. Each instrument is recorded at eight different angles equally spaced around an artificial head and torso, using two different distances, 0.5 meters, and 2.0 meters. In all, 160 sounds are recorded. The experimental setup includes 11 subjects listening to a set of the original recorded sounds and the corresponding synthesized sounds, played back over headphones (Beyer Dynamic DT 990). The singing voice is synthesized using method 1-3 described on page 4.1 and the guitar using method 1 and 2.

In this experiment all results are analyzed using Anova tests with multiple within-subject variables. Post-hoc tests were done using a Bonferroni test at 1% significance level.

**Reproduction method.** A significant influence is found in reproduction method with  $p < .001$ . Synthesis not including phase is found to be significantly different from the originals and synthesis including phase. Figure 5 shows the mean weighted error for the different synthesis types, for both guitar and singing voice.

The singing voice was furthermore tested using synthesis method 3: with phase information, and with the spectral envelope normalized in both channels. The difference between this synthesis method and synthesis with phase (method 2) is shown not to be significant.

No significant difference in performance was found between the original sounds and sounds synthesized with phase, which is surprising. Although the synthesized signal is close to the original in the stable part of the sound, sometimes tran-

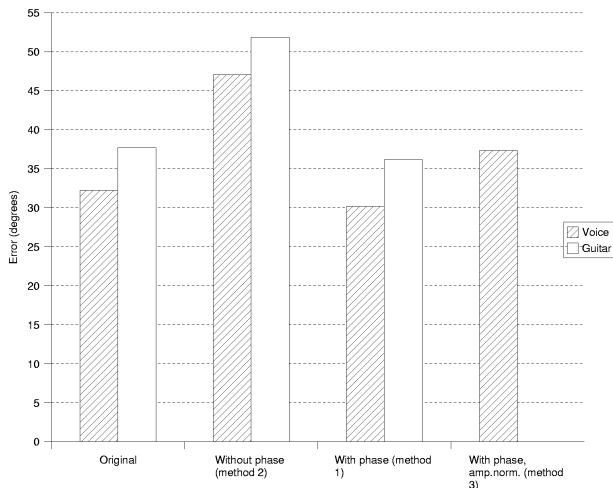


Figure 5: Mean error for the different types of reproduction, converted to degrees by multiplying by  $45^\circ$ .

sients are slightly modified, due to bad time resolution in the STFT analysis. Correctly reproduced transient signals, for instance the attack, should be easier to localize. These results indicate that this is not the case, although a more thorough examination is needed to determine the perceptual importance of transients in musical signals regarding directional cues.

**Distance to sound source.** Performance difference in the two recording distances is not observed. Because a small difference in mean sound level between the recorded sounds were present, no experiments was done regarding the test persons ability to perceive distance.

**Instrument type.** Figure 5 furthermore shows a difference in instrument type, with  $p < .04$ . On average the guitar received a higher error, compared to the voice. The finding can be explained by the fact that the guitar has a more diffuse sound source than the voice, which is better localized.

**Angle of incidence.** Angle of incidence is significant with  $p < .001$ . Figure 6 shows the mean error for all the subjects, as a function of angle of incidence and synthesis type. It is clear that it is difficult to judge sounds located at the front and back of the head. This corresponds to previous research of localization, where no correction of filtering by the subjects pinna and the headphones is done [23].

The spectral envelope normalization only gives a degradation in the sounds coming directly from the left or right of the listeners head ( $90^\circ$  or  $270^\circ$ ), whereas removing the phase gives a degradation in all angles of incidence, except the front/back ( $0^\circ$  or  $180^\circ$ ). This indicates that ILD plays a

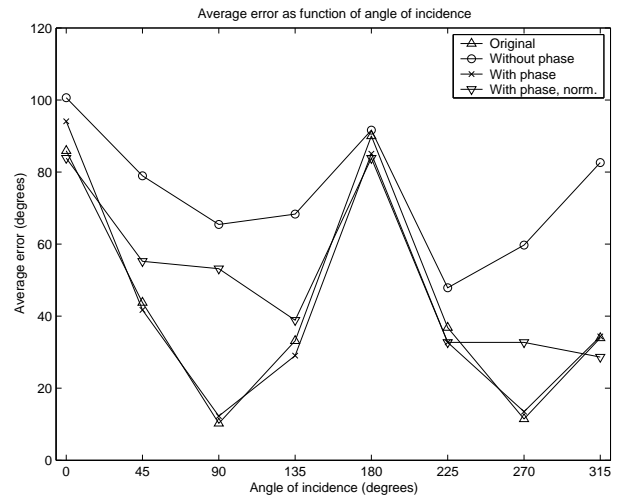


Figure 6: The mean error as a function of angle of incidence.  $0^\circ$  corresponds to a sound coming directly from the front of the listener.

significant role at these positions, where the recorded mean signal level difference is also relatively high.

**Fundamental frequency.** The experiments with guitar sounds have been done with recordings of tones with fundamental frequency in the range from 80 to 660 Hz. Figure 7 shows the average error of these different tones for the different reproduction methods. The observed difference is interesting because it shows that the localization ability improves as a function of fundamental frequency for the sounds synthesized without phase information. The interaction of fundamental frequency and reproduction method is significant with  $p < .03$ . It is difficult to conclude any general tendency from only four different fundamental frequencies, but it is interesting because it indicates that ILD or the spectral envelope is also used as a cue at low frequencies when no interaural phase difference is present. The decrease of error for the sounds resynthesized without phase may be explained by the fact that the ITD is more influent in the amplitude envelopes for the very high pitched sounds.

## 5 Monaural experiments and results

The purpose of this experiment is to determine if phase is important when synthesizing monophonic singing voice or other musical instruments. Additive analysis/synthesis is used to control phase as a parameter independent of the source sound, but the results may be applicable to other analysis/synthesis methods.

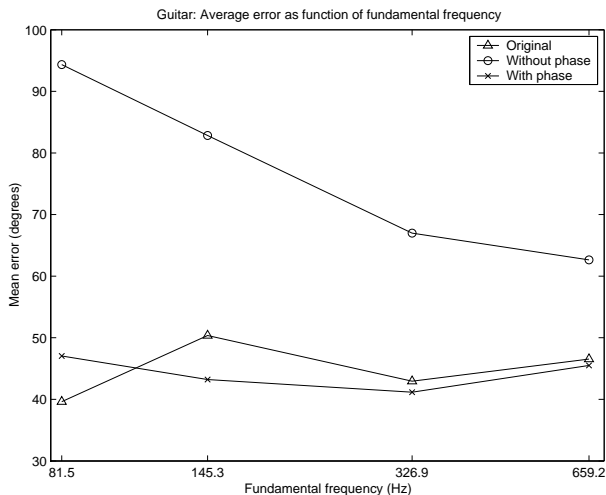


Figure 7: The mean error as a function of fundamental frequency of guitar tones for different reproduction methods.

## 5.1 Method

In the conducted experiments we use the double blind triple stimulus with hidden reference method [10], for assessing perceptual differences between original and synthesized sounds. This method was used in previous studies at DIKU to measure the sound quality of timbre models based on analyzed sounds without phase information [11]. In these experiments it was found that the sound quality was dependent on the fundamental frequency of the synthesized sounds. In general sounds with high fundamental frequency were rated to be closer to the original sound than sounds with low fundamental frequency. For this reason the sounds in this experiment was chosen to have a broad range of fundamental frequencies, to compare results with and without phase preservation at different pitches.

For each original sound, the subject hears first a reference sound, in this case the original recorded sound, followed by two sounds in random order, where one is the reference, and the other a synthesized sound. The synthesized sound is based on parameter estimation and synthesis with or without the estimated phase values, as described above. The subject is then asked to rate the two sounds relative to the reference on a scale from 1 to 5, where 1 is “Very annoying” and 5 is “Imperceptible.” One of the sounds must be given the score 5. The full scale is shown in table 1.

## 5.2 Experimental setup

All sounds are played back over Beyer Dynamic DT990 headphones, attached to a ProTools 882 sound interface. The sounds used in the experiment was piano, saxophone with and without vibrato, clarinet and singing voice, where each

Score	Impairment
5.0	Imperceptible
4.0	Perceptible, but not annoying
3.0	Slightly annoying
2.0	Annoying
1.0	Very annoying

Table 1: Scale used in listening experiment

instrument was present at five different pitches (except for the singing voice), and synthesized using the two above mentioned methods. In all eight subjects listened to 46 sets of sounds each.

## 5.3 Results

Statistical analysis of the results was done using Analysis of Variance with multiple within-subject variables. The results were evaluated according to a measure of degradation, which is the difference between the rating of the reference,  $S_r$  and the rating of the sound under experiment  $S_e$ :

$$d = S_e - S_r \quad (7)$$

The degradation measure takes into account the rating of the reference sound, to compensate for any erroneous rating by the subjects. In the conducted experiment no significant error in the rating of the reference sounds was found for any of the subjects.

### 5.3.1 Phase preservation

Figure 8 shows the mean degradation of each of the five instrument types, for synthesis with and without phase preservation. It is clear that for all instruments, synthesis without phase preservation is a substantial degradation of the sound quality compared to the sound synthesis with phase preservation. For the piano the error with phase synthesis is higher than the other sounds. This is likely to be caused by the artifacts introduced by the sinusoidal model, regardless of phase preservation, namely the errors introduced by the bad time resolution caused by the STFT, which makes it hard to model transients like the sound from the hammer of the piano. This was also one of the general comments from the subjects, that the attack of the piano, and the noise from the hammer were not well represented in any of the synthesized sounds. The main artifact of the instrument sounds synthesized without phase preservation, was that the low frequency modulations of the sound did not evolve naturally over time, according to comments made by the subjects. For the singing voice, the sounds sounded artificial, according to most of the subjects. This is interesting because it suggest that the phase is important during the whole sound, and not only in the transient part of sound.

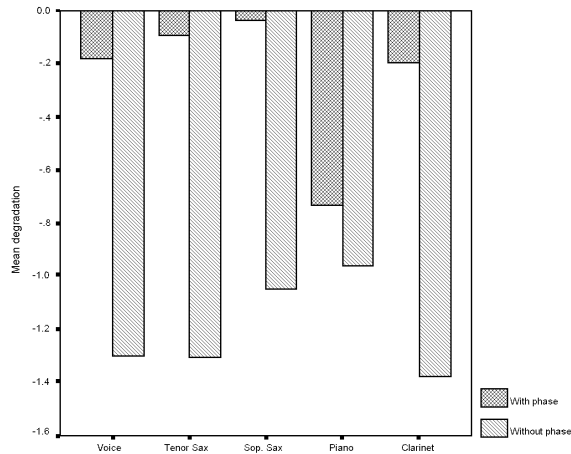


Figure 8: Mean degradation for the different types of reproduction, of each of the five instrument types.

### 5.3.2 Fundamental frequency

A correlation between degradation and fundamental frequency for any of the two synthesis methods was not observed. This may be explained by the fact that the sounds in the experiment are produced by different instruments, and thus not comparable. According to previous experiments [11] and literature [17], it was expected to see a larger degradation for sounds synthesized without phase preservation of low fundamental frequency than sounds with high fundamental frequency. However, using only harmonic sinusoids, or near harmonic sinusoids to represent a sound with high fundamental frequency may lead to a loss in the spectral contents between peak of high pitched sounds. This may be the reason why no fall in degradation for high pitched sounds is observed.

## 6 Discussion

A number of psycho-acoustic experiments involving the importance of phase in sound localization have been conducted. The results show that phase is an important parameter when performing additive analysis/synthesis of binaural recordings. If the phase is left out, the ability to perceive spatial qualities of the sounds is substantially degraded. For sounds in the examined frequency range the results indicate that there is no relation between fundamental frequency and the ability to perceive spatial qualities when phase information is used in synthesis. For the guitar sounds synthesized without phase, the error is substantially higher for sounds with low fundamental frequency, compared to sounds with high fundamental frequency. The phase is important in all incident positions, except front/back, whereas the spectral envelope is mainly in-

fluential in the lateral positions. Furthermore, indications are found that the attack part of the sound may not be particularly important for localization, as long as the interchannel time difference is preserved in the sustained part.

According to the conducted monaural experiments, it was shown that phase is of great importance when synthesizing sounds from music instruments and singing voice. Judging from the qualitative descriptions from the subjects under experiment, the phase is not only important in the transient part of the sound, where the block based sinusoidal analysis is clearly not well suited since the assumption about the signal under analysis is stationary, is violated, but is also an important parameter in the stable part of the sound. No correlation between fundamental frequency and mean degradation is found for any of the sounds synthesized without phase. This is contrary to previous experiments, but may be explained by bad overall synthesis in high pitched sounds, and the fact that too few samples of each instrument is present to observe a tendency.

From these experiments it is seen that phase should be preserved or modeled when synthesizing or processing sounds using additive techniques or other spectral based models.

## References

- [1] T. H. Andersen. Fasebevarelse ved rumlig opfattelse af lyde fra musikinstrumenter, October 2000. Student report. In danish.
- [2] T. H. Andersen. Phase models in real-time analysis/synthesis of voiced sounds, January 2002. Master thesis.
- [3] T. H. Andersen and K. Jensen. On the importance of phase information in additive analysis/synthesis of binaural sounds. In *Proceedings of the International Computer Music Conference, Havana, Cuba, 2001*.
- [4] T. H. Andersen and K. Jensen. Phase models in analysis/synthesis of voiced sounds. In *Proceedings of the DSAGM, 2001*.
- [5] P. A. Cariani and B. Delgutte. Neural correlates of the pitch of complex tones. II. pitch shift, pitch ambiguity, phase-invariance, pitch circularity, and the dominance region for pitch. *Journal of Neurophysiology*, 76(3), 1996.
- [6] A.S. Chaudhary. *Perceptual Scheduling in Real-time Music and Audio Applications*. PhD thesis, University of California, Berkeley, 2001.
- [7] R.O. Duda. 3-D Audio for HCI. <http://www-engr.sjsu.edu/~knapp/HCI/ROD3D/3D-home.htm>, Januray 2002.

- [8] J.L. Goldstein. Auditory spectral filtering and monaural phase perception. *Journal of the Acoustical Society of America*, 41:458–479, 1967.
- [9] J.L. Goldstein. An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, 54:1496–1516, 1973.
- [10] Methods for the subjective assessment of small impairments in audio systems, including multichannel sound systems. Technical report, International Telecommunication Union, Geneva, Switzerland, March 1994. ITU-R 8510, Recommendation.
- [11] K. Jensen. *Timbre Models of Musical Sounds*. Ph.D. dissertation, Department of Computer Science, University of Copenhagen, 1999. Report no. 99/7.
- [12] George F. Kuhn. Physical acoustics and measurements pertaining to directional hearing. In Yost and Gourevitch [24], pages 26–48.
- [13] J. Laroche and M. Dolson. Phase vocoder: About this phasiness business. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.
- [14] R.C. Mathes and R.L. Miller. Phase effects in monaural perception. *Journal of the Acoustical Society of America*, 19:780–797, 1947.
- [15] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal processing*, ASSP-34(4):744–754, August 1986.
- [16] B. C. J. Moore. Effects of relative phase of the components on the pitch of three-component complex tones. In E. F. Evans and J. P. Wilson, editors, *Psychophysics and physiology of hearing*. Academic Press, 1977.
- [17] R. D. Patterson. A pulse ribbon model of monaural phase perception. *J. Acoust. Soc. Am.*, 82(5):1560–1586, November 1987.
- [18] R. Plomp and H. J. M. Steeneken. Effect of phase on the timbre of complex tones. *J. Acoust. Soc. Am.*, 46:409–421, 1969.
- [19] T.D. Rossing. *The Science of Sound*. Addison-Wesley, 1990.
- [20] J.F. Schouten, R.J. Ritsma, and B.L. Cardozo. Pitch of the residue. *Journal of the Acoustical Society of America*, 34:1418–1424, 1962.
- [21] G.H. Wakefield, L.M. Heller, and L.H. Carney et. al. On the perception of transients: Applying psychophysical constraints to improve audio analysis and synthesis. In *Proceedings of the International Computer Music Conference*, pages 225–228, 2000.
- [22] F.L. Wightman. Pitch and stimulus fine structure. *Journal of the Acoustical Society of America*, 54:397–406, 1973.
- [23] L. Wightman, Diris J. Kistler, and Mark E. Perkins. A new approach to the study of human sound localization. In Yost and Gourevitch [24], pages 26–48.
- [24] W. A. Yost and G. Gourevitch, editors. *Directional Hearing*. Springer-Verlag, 1987.
- [25] W. A. Yost and E. R. Hafter. Lateralization. In Yost and Gourevitch [24], pages 49–84.

# A Hybrid Re-Synthesis Model for Hammer-Strings Interaction of Piano Tones

J. Bensa<sup>(1)</sup>, K. Jensen<sup>(2)</sup>, and R. Kronland-Martinet<sup>(1)</sup>

<sup>(1)</sup> LMA-CNRS, 31 chemin J.Aiguier, 13402 Marseille Cedex 20, France (e-mail: bensa, [kronland@lma.cnrs-mrs.fr](mailto:kronland@lma.cnrs-mrs.fr))

<sup>(2)</sup> Department of Computer Science, University of Copenhagen, Universitetsparken 1, 2100 Copenhagen, Denmark (e-mail: [krist@diku.dk](mailto:krist@diku.dk))

This paper presents some results obtained from a work collaboration between the LMA-S2M and DIKU- Musical Acoustics groups within the MOSART network. It consists in a preprint to be submitted in a journal, probably the *IEEE Trans. Speech and Audio Processing*.

## I. INTRODUCTION: CHOICE AND DESIGN OF THE MODEL

The design of a synthesis model is a crucial problem, which is strongly linked to the specificity of the sounds to be produced as well as the expected use of the model. This work has been made in the framework of the analysis-synthesis of musical sounds, meaning that we seek both at reconstructing a given piano sounds and at using the synthesis model in a musical context.

The perfect reconstruction of given sounds is a strong constraint. It necessitates the synthesis model to be designed so that the parameters can be extracted from the analysis of natural sounds. In addition, the playing of the synthesis model requires a good relationship between the physics of the instrument, the synthesis parameters and the generated sounds. Since the sound and the control levels mainly define the relationship between the “digital instrument” and the player, they constitute the most important aspects our piano model has to deal with.

The sound level is not only of importance for what the re-synthesis process of sounds is concerned. Actually, music based on the so-called “sound objects” -like electro-acoustic music or “musique concrète” [1] - looks for synthesis model allowing subtle and natural transformations of the sounds. The notion of natural transformation of sounds is relatively blurred. Here, we adopt the idea that it mainly consists in transforming sounds in a way that would correspond to a physical modification of the instrument. As a consequence, performing such sound transformations calls for the model to include physical descriptions of the instrument. Nevertheless, the physics of musical instruments is sometimes too complicated to be exhaustively taken into account. This is the case of the piano where hundreds of mechanical components are

connected together.

To take into account the necessary simplifications made in the physical description of the piano sounds, we have used hybrid models that are obtained by combining physical and signal synthesis models [2]. The physical model simulates the physical behavior of the instrument while the signal model tends at simulating the perceptual effect produced by the instrument. It allows in perfectly reconstructing a given sound and in manipulating it in both a physical and a perceptual way. Here, we have used a *physical model* to simulate the linear string vibration, and a physically informed *signal model* to simulate the non-linear interaction between the string and the hammer.

An important problem linked to hybrid models resides on the coupling of the physical and the signal models. Physical descriptions of the interaction between the string and the hammer have shown the complexity of the phenomenon. A huge part of the piano sound characteristics is due to this interaction [3]. Even though this observation is true from a physical point of view, this short interaction period is not in itself of great importance from a perceptual point of view. Actually, Schaeffer [1] shown that cutting the first milliseconds of a piano sound doesn’t alter the perception of the sound. We have carried out such an experiment and have concluded that, from a perceptual point of view, the string-hammer interaction is not audible in itself even though it undoubtedly plays an important role as *an initial condition for the string motion*. This is a substantial point justifying the dissociation of the string model and the source model in the design of our synthesis model. Thus, the resulting model consists in what is commonly called a “source-resonant” system (figure 1) [4]. The resonance has been modeled using a physically related model: the digital waveguide, while the source -the aim of which is to generate the initial condition for the string motion- has been modeled using a signal based non-linear model.

The advantages of such a hybrid model are numerous:

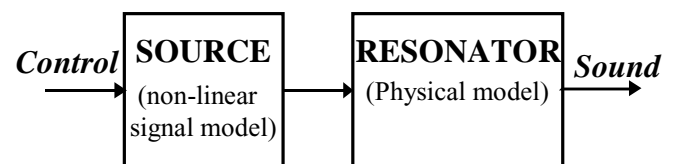


Fig. 1: Hybrid piano sound synthesis model



- it is simple enough so that the parameters can be estimated from the analysis of real sound,
- it takes into account the most relevant physical characteristics of the piano and permit the control with respect to the playing (the velocity of the hammer)
- it simulates the perceptual effect due to the non-linear behavior of the string-hammer interaction,
- it allows sounds transformation with both physical and perceptual approaches.

## II. THE RESONATOR MODEL

Several physical models of transverse wave propagation on a struck string have been published in the literature [3] [5] [6] [7]. The string is generally modeled using a one-dimensional wave equation. The specific features of the piano string that are of importance in the wave propagation (dispersion due to the stiffness of the string and frequency dependent losses) are further incorporated through several perturbation terms. To take into account the hammer-string interaction, this equation is then coupled to a non-linear force term, leading to a system of equations for which analytical solution cannot be exhibited.

Since the string vibration is only transmitted to the radiating soundboard at the bridge level, it is not worth to numerically calculate the entire spatial motion of the string. The digital waveguide [8] provides an efficient way of simulating the vibration at the bridge level of the string, while struck at a given location by the hammer. Moreover, the parameters of such a model can be estimated from the analysis of real sounds [9].

### A. The physics of vibrating strings.

We present here the main features of the physical modelization of piano strings. Those results will be very useful to understand the relationship between the parameters of our source/resonator model with the parameters of this physical model.

Consider the propagation of transverse waves in a stiff damped string governed by the motion equation [9]:

$$\frac{\partial^2 y(x,t)}{\partial t^2} = c_t^2 \frac{\partial^2 y(x,t)}{\partial x^2} - \kappa^2 \frac{\partial^2 y(x,t)}{\partial x^2} - 2b_1 \frac{\partial y(x,t)}{\partial t} + 2b_2 \frac{\partial^3 y(x,t)}{\partial x^2 \partial t} = P(x,t) \quad (1)$$

where  $y(x,t)$  is the transverse displacement,  $c_t$  the wave speed,  $\kappa$  the stiffness coefficient,  $b_1$  and  $b_2$  the loss parameters. When the force  $P$  is an impulse, this system can be solved and an analytical solution can be expressed as a sum of exponentially damped sinusoid:

$$y(x,t) = \sum_n a_n(x) e^{-\alpha_n t} e^{i\omega_n t} \quad \text{for } t \geq 0 \quad (2)$$

where  $a_n$  is the amplitude,  $\alpha_n$  is the damping coefficient and  $\omega_n$  the frequency of the  $n^{\text{th}}$  partial. Due to the stiffness, the waves are dispersed and the partial frequencies are not harmonic, given by [10]:

$$\omega_n = 2\pi n \omega_0 \sqrt{1 + Bn^2} \quad (3)$$

where  $\omega_0$  is the fundamental radial frequency and  $B$  is the inharmonicity coefficient ( $B = \kappa^2 \omega_0^2 / c^4$ ).

The losses are frequency dependant and expressed by [9]:

$$\alpha_n = -b_1 - b_2 \left( \frac{\pi^2}{2BL^2} (-1 + \sqrt{1 + 4B(\omega/\omega_0)^2}) \right) \quad (4)$$

where  $L$  is the length of the string.

Considering that the hammer strikes the string at a location  $x_0$  with an impulse force  $P$ , the amplitudes of the partials are related to the force  $P$  and string parameters by [6]:

$$a_n = \frac{2P}{\sqrt{\rho_L T}} \frac{\sin(n\pi x_0 / L)}{n\pi \sqrt{1 + n^2 B}} \quad (5)$$

where  $\rho_L$  is the linear mass and  $T$  the tension of the string.

The spectral content of the piano sound, as well as the majority of musical instruments sounds, is modified with respect to the dynamics. In the piano case, this non-linear behavior consists in an increasing of the brightness of the sound. This non-linearity is directly linked to the hammer-string contact. The stiffness of the hammer felt increases with the impact velocity. It behaves as a non-linear spring-mass oscillator [5][7]. In expression (5), the force term  $P$  depends on the characteristics of the hammer and will take into account the non-linear behavior whereas the other terms are related to the strings parameters.

We shall see in the next paragraph how the waveguide model parameters are related to the amplitude, damping coefficient and frequencies of each partial.

### B. Modeling using digital waveguides

#### 1) The single string case: the elementary waveguide

To model the wave propagation in a piano string, we use a waveguide model [8]. In the single string case, this elementary model consists of a unique loop system (figure 2) including:

- a delay line (a pure delay filter noted  $D$ ) simulating the duration the waves take to travel back and forth in the medium,
- a filter (noted  $F$ ) taking into account the dissipation and dispersion phenomena, together with the boundary

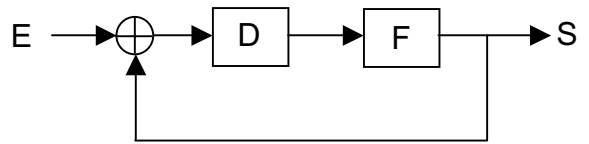


Fig. 2: the elementary digital waveguide

conditions. The modulus of  $F$  is then related to the damping of the partials and the phase to the inharmonicity in the string.

The output  $S$  of the model represents the vibrating signal measured at an extremity of the string (at the bridge level).

The input  $E$  corresponds to the frequency-dependent energy transferred to the string by the hammer. The transfer function of the elementary digital waveguide is given by:

$$T(\omega) = \frac{S(\omega)}{E(\omega)} = \frac{F(\omega)e^{-i\omega D}}{1 - F(\omega)e^{-i\omega D}} \quad (6)$$

The output of the digital waveguide driven by a delta function can consequently be expanded as a sum of exponentially damped sinusoids. It thus coincides with the solution of the motion equation of transverse waves in a stiff damped string for a source term given by a delta function force.

As shown in [9], the modulus and phase of  $F$  are related to the damping and the frequencies of the partials by the expressions:

$$\begin{aligned} |F(\omega_n)| &= e^{\alpha_n D} \\ \arg(F(\omega_n)) &= \omega_n D - 2n\pi \end{aligned} \quad (7)$$

## 2) The multiple strings case: the coupled waveguide

In the middle and the treble range of a real piano two or three strings are struck for the same note. The vibration produced by this coupled system is not the superposition of the vibrations produced by each string. It is the result of a complex coupling between the modes of vibration of these strings. This coupling leads to phenomena like beats and double decays on the amplitude of the partials, which constitute one of the most important features of the piano sound [11]. Beats are used by the professional to precisely tune the doublets or triplets of string since the beats periodicity is correlated to the frequency difference between the strings modes. In order to re-synthesize the vibration of several strings at the bridge level, we use coupled digital waveguides. Smith [12] proposed a coupling model with two elementary waveguides. He assumed that the two strings were coupled to the same termination, and that the losses were lumped to the bridge impedance. This technique leads to a simple model only necessitating one loss filter. Nevertheless, the decay times and the coupling of the modes are not independent, leading to unnatural synthetic sounds. Another approach proposed by Karjalainen [13] consists in coupling two digital waveguides through real gain amplifiers. In that case, the coupling is the same for each partial and the time behavior of the partials is similar. Here again, the generated signal sounds unnatural.

We have designed a model which is an extension of the Karjalainen's approach. It consists in separating the time behavior of the components by using complex-valued and frequency-dependant amplifiers (linear filters) to couple the waveguides (figure 3). This model accurately simulates the energy transfer between the strings as we will see in the next section.

Each string is modeled using an elementary waveguide (named  $G_1$ ,  $G_2$ ,  $G_3$ ). The coupled model is then obtained by connecting the output of each elementary waveguide to the input of the others through coupling filters. The coupling filters simulate the wave propagation along the bridge and are thus correlated to the distance between the strings. In the case of a doublet of strings, the two coupling filters are identical. In

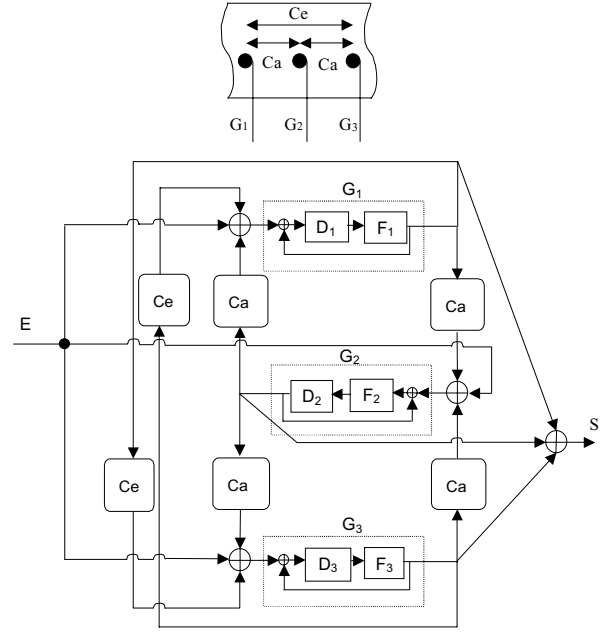


Fig. 3: Bottom: coupled digital model for three strings belonging to a same triplet and coupled through the bridge. Top: the corresponding physical system

the case of a triplet of strings, the coupling filters of adjacent strings (named  $C_a$ ) are equal but differ from the coupling filters to the extreme strings (named  $C_e$ ). The excitation signal is assumed to be the same for each elementary waveguide since we suppose the hammer to struck the strings in a similar way.

## C. Calibration of the waveguide models

This work takes place in the general analysis-synthesis framework, meaning that the objective is not only to be able to simulate sounds, but also to reconstruct a given sound. Therefore the calibration of the model must be done with a particular care. We first present the inverse problem allowing the calculation of the waveguide parameters from experimental data. Then, we describe the experiment and the measurements we have made in the case of one, two and three coupled strings. We finally show the validity and the accuracy of the analysis-synthesis process by comparing synthetic and original signals.

### 1) The inverse problem

We address here the estimation of the parameters of each elementary waveguide as well as the coupling filters from the analysis of a single signal (measured at the bridge level). For this purpose, we assume that in the case of three coupled strings the sound is composed of a sum of three exponentially decaying sinusoids (triplet). The measured signal can then be written as follow:

$$y(t) = \sum_n a_n e^{-\alpha_n t} e^{i\omega_n t} + b_n e^{-\beta_n t} e^{i\omega_{2n} t} + c_n e^{-\gamma_n t} e^{i\omega_{3n} t} \quad (8)$$

where  $a_n$ ,  $b_n$ ,  $c_n$  are the amplitudes,  $\alpha_n$ ,  $\beta_n$ ,  $\gamma_n$  the damping coefficients,  $\omega_{1n}$ ,  $\omega_{2n}$ ,  $\omega_{3n}$  the frequencies. In the piano case,

the coupled strings are closely tuned, leading to triplet of close frequency components.

The estimation method is a generalization of the one described in [14] for one and two strings. It can be summarized as follow: we start by isolating each triplet of the measured signal through band-pass filtering. We then extract from each triplet the three amplitudes, damping coefficients and frequencies of each partial using a parametric method. Further on, we identify the Fourier transform of equation (1), given by:

$$Y(\omega) = \sum_n \frac{a_n}{\alpha_n + i(\omega - \omega_{1n})} + \frac{b_n}{\beta_n + i(\omega - \omega_{2n})} + \frac{c_n}{\gamma_n + i(\omega - \omega_{3n})} \quad (9)$$

with the transfer function of the coupled waveguide. This identification leads to a linear system that admits an analytical solution in the case of one or two strings [14]. In the case of three coupled strings, the solution can only be found in a numerical way.

The process gives an estimation of the modulus and of the phase of each filter near the resonance peaks as a function of the amplitudes, damping coefficients and frequencies.

The resonant model being known, we finally extract the excitation signal using a deconvolution process with respect to the waveguide transfer function. Since the transfer function has been identified near the resonant peaks, the excitation is also estimated at discrete frequencies values corresponding to the partial frequencies. This excitation corresponds to the signal that has to be injected into the resonator to re-synthesize the actual sound.

## 2) Validation of the resonant model by an experiment

### a) Experimental setup

We have designed an experimental setup allowing the measurement of the vibration of one, two or three strings struck by a hammer for different velocities. On the top of a massive concrete support, we have attached a piece of a bridge taken from a real piano. On the other extremity of the structure, we have attached an agraf on a hard wood support. The strings are then tightened between the bridge and the agraf making our setup close to the conditions of a real piano. One, two or three strings are struck with a hammer linked to an electronically piloted key. By imposing different voltages to the system, one can control the hammer velocity in a reasonably reproducible way. The precise velocity is measured immediately after escapement using a photonic sensor pointing at the head of the hammer. The vibration at the bridge level is measured by an accelerometer. We have collected acceleration signals corresponding to hammer velocities varying between  $0.8 \text{ m}\cdot\text{s}^{-1}$  and  $5.7 \text{ m}\cdot\text{s}^{-1}$ .

### b) Filters estimation

From the signals collected on the experimental setup, a set of data has been extracted. For each hammer velocity, the waveguide filters and the corresponding excitation signals have been estimated using the techniques described above.

Figure 4 shows the modulus of the filter responses  $F$  for the

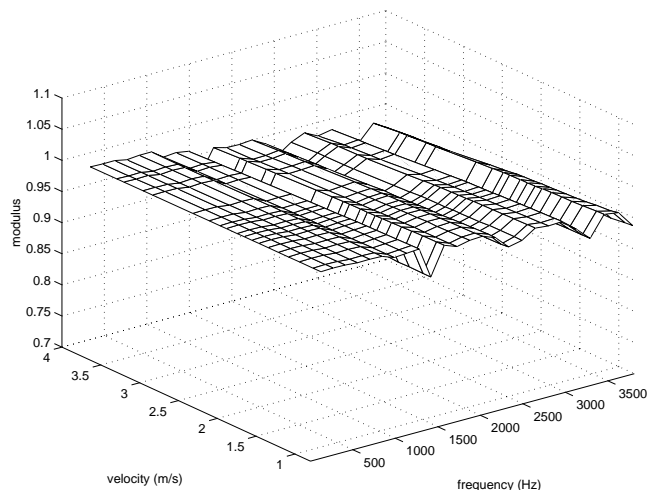


Figure 4 : amplitude of the filter  $F$  as a function of the frequency and of the hammer velocity

first twenty five partials in the case of tones produced by a single string. Here the hammer velocity varies from  $0.7 \text{ m}\cdot\text{s}^{-1}$  to  $4 \text{ m}\cdot\text{s}^{-1}$ .

One can notice that the modulus of the waveguide filters with respect to the hammer velocity are similar. This result validates our approach based on a source-resonator separation. Actually, the resonator represents the string that is unchanged in the experiment. Nevertheless, one can notice a slight decrease of the filter modulus as a function of the hammer velocity for high frequency partials. This non-linear behavior is not directly linked to the hammer string contact. It is mainly due to the non-linear phenomena involved in the wave propagation. The larger the amplitude of the motion, the larger the internal losses.

The filter modulus slowly decreases from a value close to 1. Since the higher partials are more damped than the lower ones, the amplitude of the filter decreases while the frequency increases. The value of the filter modulus (close to 1) suggests that the losses are weak. This is true in the piano string case and even more obvious on this experimental setup, since the lack of a soundboard limits the acoustic field radiation. More losses are expected in the real piano case.

Let's now consider the multiple strings case. From a physical point of view, the behavior of the filters  $F_1$ ,  $F_2$ ,  $F_3$  (which characterize the intrinsic losses) of the coupled digital waveguides should be similar since the strings are supposed identical. One can see this global behavior in figure 5, even though some artifacts pollute the drawing. These artifacts are mainly due to the poor signal/noise ratio at high frequency (above 2000 Hz) and low velocity. Nevertheless, this does not alter the synthetic sound since the corresponding partials are weak and of short duration.

Figure 6 shows the phase of the same filters. The phase is of great importance since it is related to the group delay of the signal and consequently directly linked to the frequency of the

partials. Piano sounds are not periodic (meaning that the

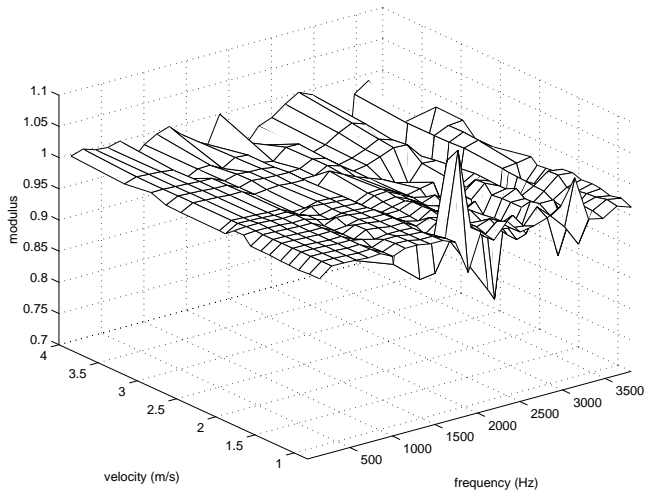


Figure 5 : amplitude of the filter F2 as a function of the frequency and of the hammer velocity

spectra is not harmonic), due to the bending stiffness of the steel wire. This leads to a non-linear behavior of the phase as a function of the frequency. One can notice that the phase is constant with the hammer velocity, since the frequencies of the partials are always the same (linearity of the wave propagation).

The coupling filters simulate the energy transfer between the strings and are frequency dependent. Figure 7 represent one of these coupling filters for different values of the hammer

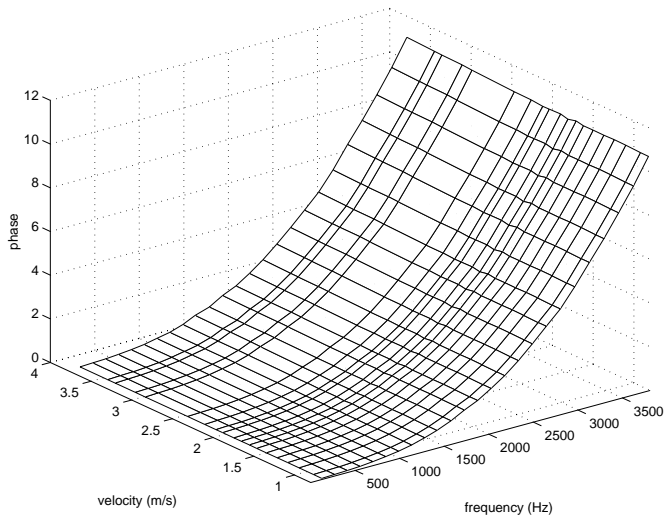


Fig. 6: phase of the filter F as a function of the frequency and of the hammer velocity

velocity. The amplitude is quite constant with respect to the hammer velocity (up to signal/noise ratio at high frequency and low velocity), showing that the coupling is independent of the amplitude of the vibration. The coupling seems to rise with the frequency, but a physical explanation of this phenomenon is out of the scope of this paper. The peaks at frequencies 700 Hz and 1300 Hz correspond to a maximum of the energy transfer and are related to the impedance of the experimental setup termination.

c) Accuracy of the re-synthesis

By injecting the excitations signal obtained by deconvolution into the waveguide model, one reproduce the signals measured on the experimental setup. From a perceptual

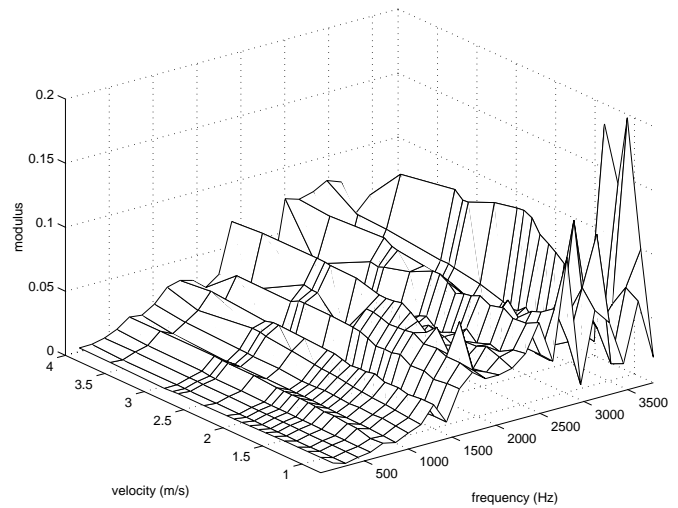


Figure 7 : amplitude of the filter Ca as a function of the frequency and of the hammer velocity

point of view, the resulting sound is undistinguishable from the original ones. Figure 8, 9 and 10 show the amplitude modulation laws of the first six partials of the original and the re-synthesized sound. The variations of the temporal envelope are generally well retained and for the coupled system, the beat phenomena are well reproduced. The slight differences that one can observe are not audible. They are due to fine physical phenomena that are not taken into account in our model. For example, let us consider the second and sixth partials of the original sound of the figure 8. We can see beats which show coupling phenomena on only one string. Indeed, the horizontal and vertical modes of vibration of the string are coupled through the bridge. This coupling was not taken into account in this study since the phenomenon doesn't alter the perceptual effect.

The accuracy of the re-synthesis validates a posteriori the

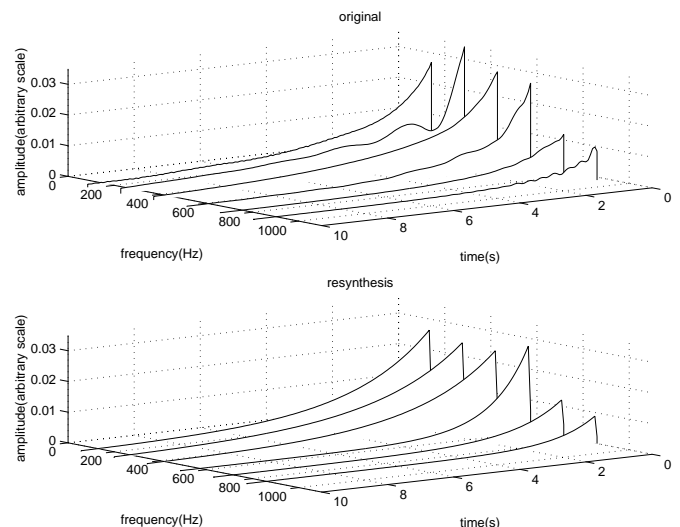


Fig. 8: amplitude modulation laws for the first sixth partials, one string

use of our model and the source-resonant approach.

### III. THE SOURCE MODEL

In this section, we examine in detail the excitation signals extracted by deconvolution with respect to the waveguide filter. In the previous chapter, we observed that the waveguide

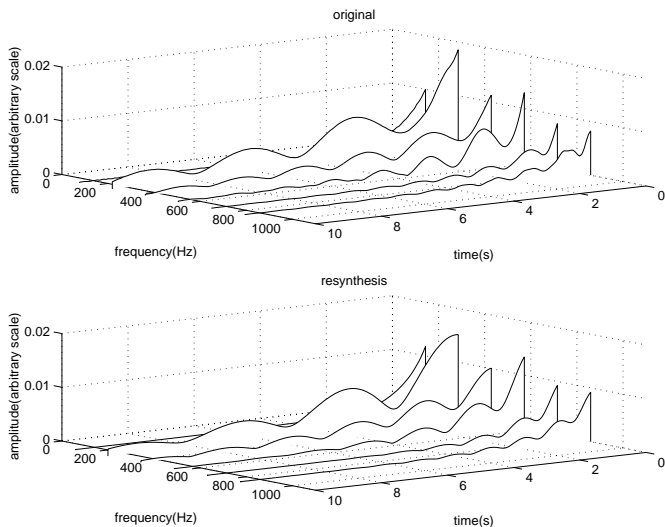


Fig. 9: amplitude modulation laws for the first sixth partials, two strings

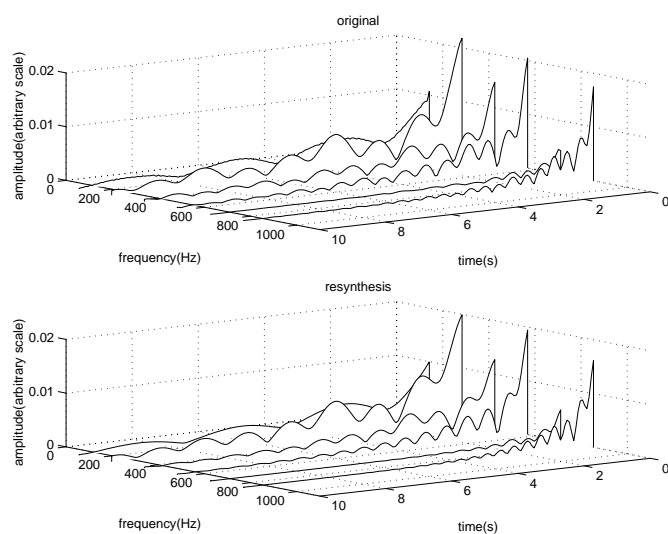


Fig. 10: amplitude modulation laws for the first sixth partials, three strings

filters are almost invariant with respect to the velocity. On the contrary, the excitation signal varies nonlinearly as a function of the velocity, thereby taking into account the timbre variations of the resulting piano sound. In order to have an accurate sound (re)-synthesis, the source model must be designed with care.

From the extracted excitation signals, we will study the source behavior as a function of the partial frequency and velocity (section IV.A). This will allow us to design the source model (section IV.B), which take into account these behaviors. In the next chapter, this model will be extended to take into account the note dependency, using data obtained from a real

piano. In this section, the model parameters are calculated using the excitation signal obtained from the experimental setup (described in III.C.2a), but we will show later that the data obtained from the real piano behave correspondingly.

#### A. Non-linear source behavior

Since the excitation is related to the hammer-string interaction, we expect it to be short and broadband. In addition, since the hammer-string interaction is non-linear, the source should behave non-linearly, basically giving more brightness to the high velocities. Finally, we expect the source spectrum to be modulated with respect to the hammer impact position. Some of the source characteristics have been shown previously [15] [16]. We shortly summarize here, the most important characteristics of the source behavior.

Figure 11 shows the excitation signals extracted from the measurement of the vibration of a single string struck by a hammer for three velocities corresponding to the pianissimo, mezzo-forte and fortissimo musical playing. The excitation duration is about 3.5 ms, which is in accordance with the duration of the hammer-string contact [17]. The excitation signals sound like an impact, with the higher velocity

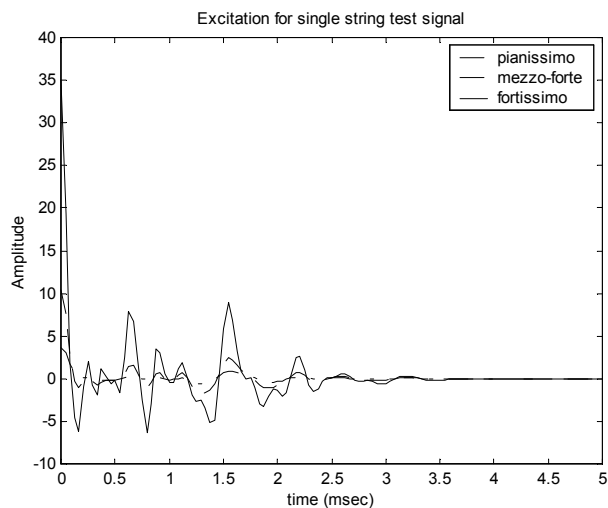


Fig. 11: Three excitation signals for the single string test signal.

excitation signals being brighter. This increase in brightness can be better visualized in the frequency domain.

The spectra of the excitation signals represented in figure 11 are shown in figure 12. They globally decrease at high frequency showing several irregularities among which appears a periodic modulation related to the location of the hammer impact on the string. The excitation corresponding to the fortissimo playing has more energy than the ones corresponding to mezzo forte and pianissimo, in particular for the higher partials. This corresponds to an increase of brightness with respect to the hammer velocity. In conclusion, the excitation signals can be seen as an invariant spectrum (which we call the static spectrum), shaped by a smooth

frequency response filter (which we call the spectral deviation), the characteristics of which depends on the hammer

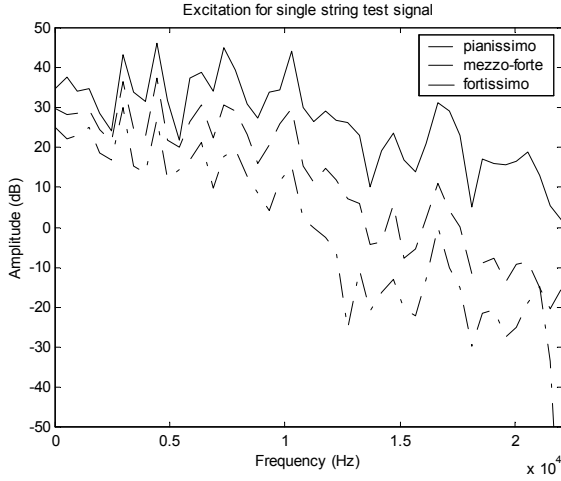


Fig. 12: Frequency domain plot of the excitation signals for the single string test signals

velocity.

### B. Design of a source signal model

The source model can be seen in figure 13. It has been elaborated using knowledge about the physical hammer-string interaction system, and by taking particular care that the model restitutes the perceptually important parameters well. The source model consists of the static spectrum and the spectral deviations. The static spectrum is a function of the hammer position and the string parameters. It is invariant with the velocity. The spectral deviation shapes the spectrum as a function of the velocity. The hammer position, string characteristics and the velocity dependency of the spectral

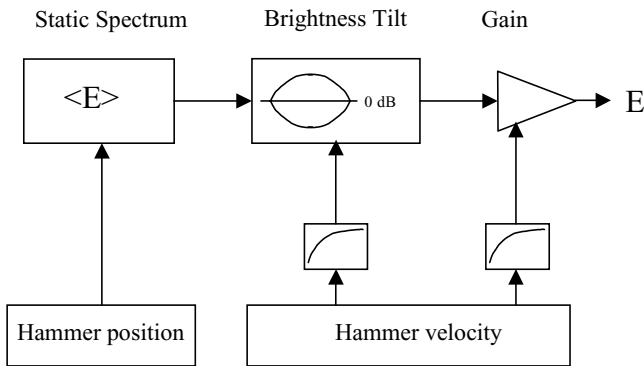


Fig. 13: The Source Model

deviation are subsequently a function of the note.

### C. Frequency dependency

This section describes the model of the source as a function of the partial frequencies.

#### 1) The static spectrum

We define the static spectrum as the part of the excitation that is invariant with respect to the hammer velocity. By considering the formula (equation (5)), we can deduce the analytical expression of the static spectrum:

$$E_{ss} = \frac{\sin(n\pi\alpha)}{\pi n\sqrt{1+n^2B}} \quad (10)$$

where  $B$  is the inharmonicity coefficient of the string [10],  $\alpha$  is the relative hammer impact position ( $\alpha=x_0/L$ ) and  $n$  is the partial number. The term  $1/\sqrt{\rho T}$  in equation (5) will be taken into account in the global gain. This expression gives an acceptable synthesis of the final piano sound (after spectral shaping by the spectral deviation), but in practice the excitation spectra contains more irregularities, which are also invariant with respect to the velocities. If we want an accurate resynthesis of an original sound, we have to include those irregularities in the static spectrum, by multiplying the theoretical static spectrum with an error term  $e$ . The final formula for the static spectrum is:

$$E_{ss} = \frac{\sin(n\pi\alpha)}{\pi n\sqrt{1+n^2B}} \cdot e \quad (11)$$

$e$  is the deviation from the theoretical string excitation spectrum, which is supposed to be independent of the hammer velocity (in practice, it can be a few dB difference for different

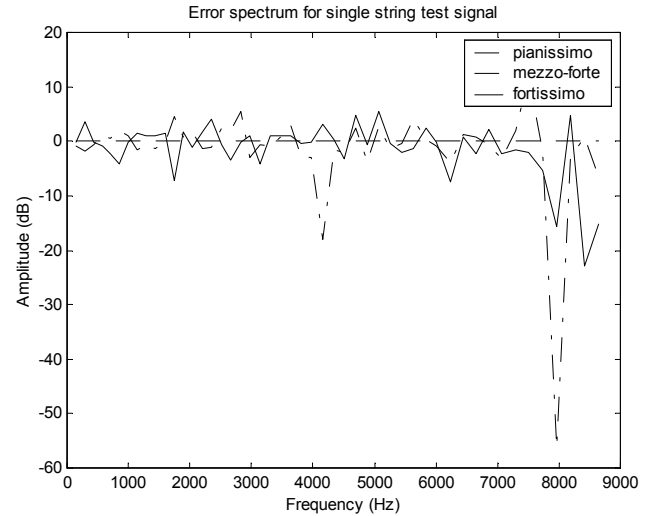


Fig. 14: Error spectrum for a single string.

velocities).  $e$  is obtained by comparing the original excitation signal with the excitation obtained by using the theoretical static spectrum.

## 2) The spectral deviation

The excitation spectrum is, however, not invariant with respect to the velocity, as we observed earlier. This is taken into account by the spectral deviation.

The spectral deviation is calculated by dividing the estimated excitation with the static spectrum:

$$d_v = \frac{E_v}{\langle E \rangle} \quad (12)$$

where  $E_v$  is the excitation signal for the velocity  $v$  obtained from the experimental data.

The spectral deviations for the single string experimental signals are shown in figure 15. The spectral deviation effectively strengthens the fortissimo, in particular the high partials. It can successfully be fitted to a second order exponential polynomial (as shown in figure 15):

$$\hat{d}_v = e^{c+bf+af^2} \quad (13)$$

In this model,  $c$  corresponds to the gain (since independent of the frequency),  $bf$  defines the spectral tilt, which is limited in the high frequencies by the term  $af^2$ . This modeling is sufficient both for synthesis, but also for resynthesis, in an analysis/synthesis situation.

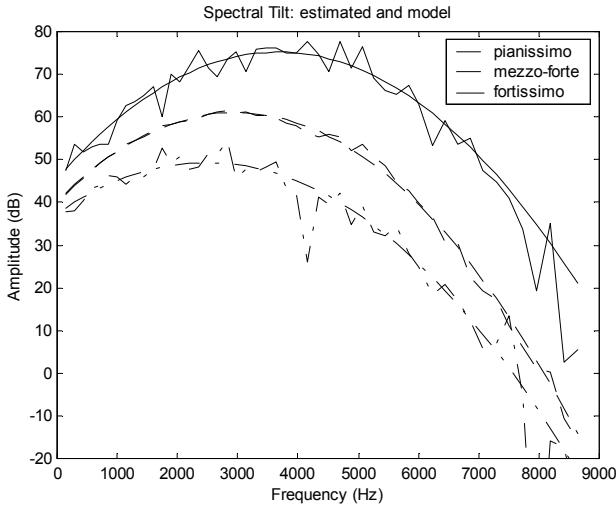


Fig.15: Spectral deviations for the single string signal. Original data and modeled data.

## D. Velocity dependency

In order to have a usable source model in a synthesis situation, it needs a model that both restitutes an acceptable sound quality, but also permits changing perceptually important attributes as a function of the hammer velocity.

The velocity dependent parameters of the source model ( $a$ ,  $b$  and  $c$  in equation (13)) are shown in figure 16.

The spectral tilt parameters ( $a$  and  $b$  in equation (15)) exhibit asymptotic behavior. This can, in part, be explained by the full compression of the felt at high velocities [5]. The gain parameter  $c$  exhibits less asymptotic behavior.  $a$ ,  $b$ , and  $c$  are modeled as a function of the hammer velocity by an asymptotic exponential with three parameters:

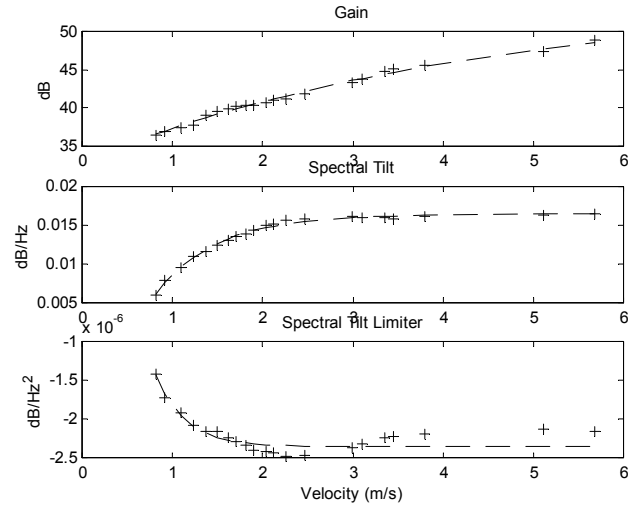


Fig.16: Gain, Spectral Tilt and Spectral Tilt Limiter for the single string. Original data (+), model data (solid).

$$A(v) = B_M - B_m \cdot e^{-B_v \cdot v} \quad (14)$$

where  $A(v)$  is the attribute to model (gain, spectral tilt, or spectral tilt limiter),  $B_M$  is the asymptotic value,  $B_m$  is the deviation from the asymptotic value at zero velocity, and  $B_v$  is the velocity exponential coefficient, governing how sensible the attribute is to a velocity change.

The parameters of the exponential model (equation (14)) are found using a non-linear curvefit [18]. The asymptotically modeled values of the gain, spectral tilt and spectral tilt limiter values are shown in figure 16 along with the original values.

## E. Resynthesis

The total source excitation can now be recreated by the equations (11)-(13), and the velocity model excitations can be recreated by additionally using the asymptotic velocity equation (14). The resulting excitation spectra are shown in figure 17. The deviations of the resulting excitations are perceptually insignificant.

The next chapter will show that the source model can be

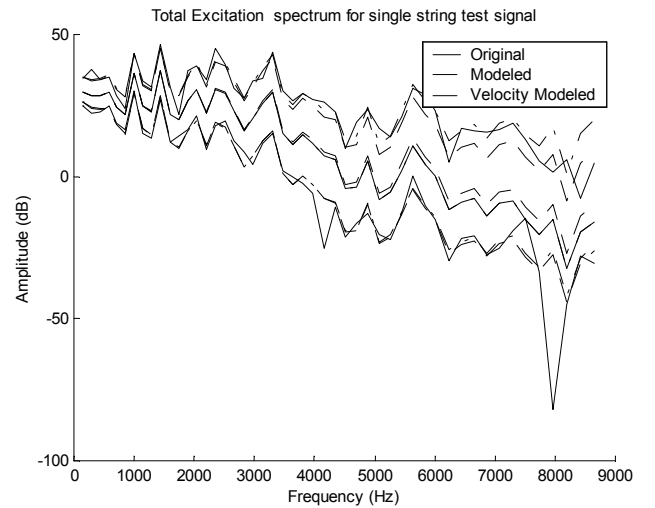


Fig.17: Recreated excitation spectra for the experimental data.

used for any fundamental frequency of the piano, and give a source model for both velocity and fundamental frequency.

#### IV. CONTROL OF THE HYBRID MODEL THROUGH MEASUREMENTS ON A REAL PIANO

##### A. Experimental setup

In order to take into account the note dependence of the resonator and source models, we have made a set of measurements on a real piano. Those measurements have been made on a Yamaha Disklavier C6 grand piano equipped with sensors. The vibrations of the strings were measured at the bridge level using an accelerometer whereas the hammer velocities were measured using a photonic sensor. Data were collected for several velocities and several notes. We have then used the estimation process described in III.C.1 for the previous experimental setup, and extracted for each note and each velocity the corresponding resonator and source.

##### B. Behavior of the velocity and note dependence of the resonator

As expected, the behavior of the resonator as a function of the hammer velocity is similar to the one described in III.C.2.b, for the signals measured on the previous experimental setup. The filters are quite similar with respect to the hammer velocity. Their modulus is close two one, but slightly weaker than previously, since it takes now into account the losses due to the acoustic field radiated by the soundboard.

On the contrary, the shape of the filters is modified as a function of the note. As we have seen, the filters are related to the physical features of the strings themselves which change for each note. In order to control our model in a relevant way, we have to take into account this behavior. figure 18 shows the modulus of the waveguide filter for several notes. We have not studied the behavior of each filter independently, in the case of doublet or triplet of strings. In order to compare this behavior with the one of the filters of the single string notes, we have calculated an average filter.

As shown in [9], the modulus and phase of those filters can be expressed in function of the physical parameters of the strings:

$$|F(\omega)| = \exp\left(-D\left[b_1 + \frac{b_2\pi^2\xi}{2BL^2}\right]\right) \quad (15)$$

$$\arg(F(\omega)) = \omega D - \pi\sqrt{\frac{\xi}{2B}}$$

$$\text{with } \xi = -1 + \sqrt{1 + 4B\omega^2 / \omega_0^2}$$

$L$  the length of the string has been measured on the piano,  $\omega_0$  is the radial fundamental frequency of the note,  $B$  is the inharmonicity coefficient estimated on the signal measured. By fitting the estimated filters using those relations, we have obtained the behavior of  $b_1$  and  $b_2$  as a function of the note. The corresponding curves are also plotted on figure 18.

The relations (15), relating the physical parameters to the waveguide parameters, allow the control of the resonator in a relevant physical way. We can either change the length of the strings, the inharmonicity, or the losses ... But in order to be in

accordance with the physical system, we have to take into account the inter-dependance of some of the parameters. For instance, the fundamental frequency is obviously related to the length of the string, as well as the tension or the linear mass. If we modify the length of the string, we also have to modify, for instance, the fundamental frequency, considering that the tension and the linear mass are unchanged. This aspect will be taken into account in the implementation of the model.

##### C. Behavior of the velocity and note dependence of the source

The source model parameters have been calculated for a

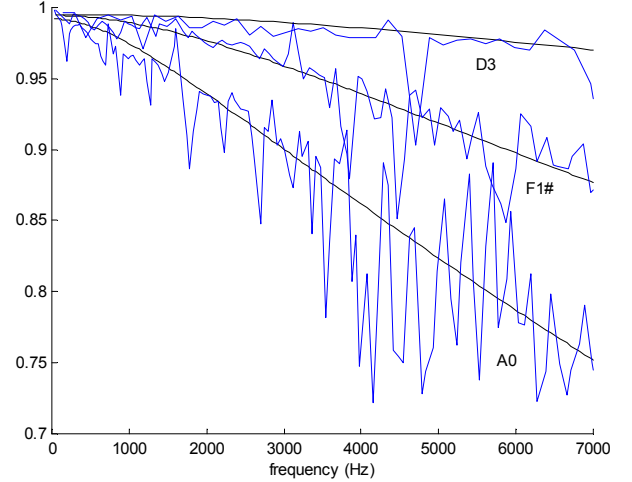


Fig.18: Modulus of the waveguide filters for notes A0, F1# and D3. Original (dotted line) and modeled filters.

subset of the data for the piano, namely the notes A<sub>0</sub>, F<sub>1</sub>, B<sub>1</sub>, G<sub>2</sub>, C<sub>3</sub>, G<sub>3</sub>, D<sub>4</sub>, E<sub>5</sub> and F<sub>6</sub>. Each note has approximately ten velocities, from around 0.4 m/s to approximately 3 m/s.

The source extracted from the piano signals behaves as the experimental data for all notes with respect to the hammer velocity. In particular, the spectral deviation parameters behave according to the model for all notes. The gain, spectral tilt and spectral tilt limiter all exhibit asymptotic behavior as a function of the velocity.

Since the source is dependent mainly on the characteristics of the hammer, which is different for each note, we do expect its behavior to change as well with respect to the note.

The velocity coefficient  $B_v$  for the gain  $c$  decreases with the note, however, stating that the high notes need more velocity range to obtain the higher dynamic range. The high pitch notes gain have more asymptotic behavior, i.e. the asymptotic values  $B_M$  are reached for a lower velocity, transforming into a higher velocity coefficient for the higher notes.

As shown in figure 19, the spectral tilt  $b$  asymptotic value  $B_M$  is positive for the low pitched notes and negative for the high pitched notes, reflecting that the low pitched notes excitations are less bright than the static spectrum, as opposed to the high partials. The low notes exhibits more exponential behavior, i.e. the low velocities for the low pitched notes are



lesser bright than the corresponding high pitched notes.

The spectral tilt limiter  $a$ , is positive for the high pitched notes and negative for the low pitched notes. In addition, the low velocities tends more in the same directions. The midrange notes does not have any noticeable 2<sup>nd</sup> order term, which may explain the lack of this term in previous related work [15][16].

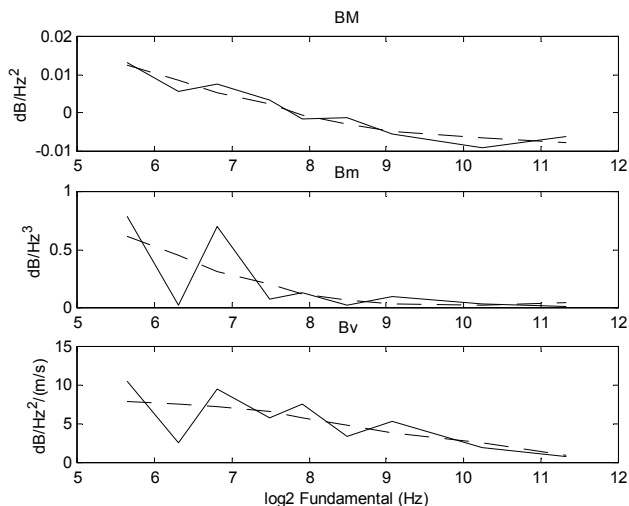


Fig.19: Asymptotic value  $B_M$ , deviation from the asymptotic value  $B_m$ , and velocity exponential coefficient  $B_v$  as a function of the note frequency for the spectral tilt coefficient  $b$ .

The spectral tilt limiter is thus effectively limiting the effect of the spectral tilt for the high partials.

One can finally notice that all those parameters are regular enough as a function of the note to be fitted by a second order polynomial as shown on figure 19. This approximation leads to a complete piano model allowing to control the source as a function of the velocity and the note.

## V. CONCLUSION

In this paper we have shown that a piano sound can be well reproduced using a hybrid model consisting of a resonant part and an excitation part. After an accurate calibration, the sounds obtained are perceptually close to the original ones for all notes and velocities. The resonator, which simulates the phenomena intervening in the strings themselves, is modeled using a digital waveguide model which is a very efficient way of simulating the wave propagation. It exhibits physical parameters such as the string tension, the inharmonicity coefficient, allowing a physically relevant control of the resonator. It also take into account the coupling effects, which are perceptually extremely relevant. The source is extracted using a deconvolution process and is modeled using a subtractive signal model. The source model consists of three parts (static spectrum, spectral tilt and energy) which are dependant on the velocities and the notes played. In order to reduce the number of parameters, a model of the note evolution is finally proposed leading to a fully playable piano model.

## REFERENCES

- [1] Schaeffer, P., *Traité des objet musicaux*, Edition du seuil, Paris, France, 1966, pp.219-220.
- [2] S. Ystad, "Sound Modeling Applied to Flute Sounds", in *Journal of Audio Engineering Society*, Vol. 48, No. 9, 2000, pp. 810-825.
- [3] A. Chaigne, A. Askenfelt, "Numerical Simulations of Struck Strings. {I}. {A} Physical Model for a Struck String Using Finite Difference Methods", *Journal of Acoustical Society of America*, 95(2), 1994, pp. 1112-1118.
- [4] J. Laroche and J. L. Meillier, "Multichannel Excitation/Filter Modeling of Percussive Sounds with Application to the Piano", *IEEE Trans. Speech and Audio Processing*, 2(2), 1994, pp.329-344.
- [5] X. Boutillon "Model for piano hammers: Experimental determination and digital simulation", *Journal of Acoustical Society of America*, 83(2), 1988, pp. 746-754.
- [6] C. Valette, C. Cuesta, "*Mécanique de la corde vibrante*". *Traité des nouvelles technologies*. Série Mécanique, Hermès, Ed., 1993.
- [7] D. E. Hall, "Piano string excitation. VI: Nonlinear modeling", *Journal of Acoustical Society of America*, 92(1), 1992, pp. 95-105.
- [8] J.O Smith III, "Physical Modeling Using Digital Waveguides", *Computer Music Journal*, 16(4), 1992, pp.74-87.
- [9] J. Bensa, S. Bilbao, R. Kronland-Martinnet, Julius O. Smith III, "The Simulation of Piano String Vibration: From Physical Models to Finite Difference Schemes and Digital Waveguides", to be published.
- [10] H. Fletcher and E. D. Blackham and R. Stratton, "Quality of Piano Tones", *Journal of Acoustical Society of America*, 34(6), 1961, pp. 749-761.
- [11] G. Weinreich. "Coupled piano strings", *Journal of Acoustical Society of America*, 62 (6), 1977, pp.1474-1484.
- [12] Julius O. Smith III, "Efficient Synthesis of Stringed Musical Instruments", *Proceedings of the ICMC*, Tokyo, Japan, 1993, pp. 64-71.
- [13] M. Karjalainen, V. Välimäki, T. Tolonen. "Plucked-string models: from the Karplus-Strong algorithm to digital waveguides and beyond", *Computer Music J.*, 22(3), 1998, pp. 17-32.
- [14] M. Aramaki, J. Bensa, L. Daudet, P. Guillemain, R. Kronland-Martinnet, "Resynthesis of coupled piano string vibrations based on physical modeling", *Journal of New Music Research*, 30(3), 2002, pp.213-226.
- [15] J. Bensa, K. Jensen, R. Kronland-Martinnet, S. Ystad, "Perceptual and Analytical Analysis of the effect of the Hammer Impact on the Piano Tones", *Proceedings of the International Computer Music Conference. I.C.M.A.*, 2000, pp. 58-61.
- [16] J. Bensa, F. Gibaudan, K. Jensen, R. Kronland-Martinnet, "Note and Hammer Velocity Dependence of a Piano String Model Based on Coupled Digital Waveguides", in *Proceedings of the ICMC*, Havana, Cuba, 2001.
- [17] A. Askenfelt "Five lectures on The Acoustics of the Piano" Publications issued by the Royal Swedish Academy of Music N°64. Stockholm: Kungl. Musikaliska Akademien, 1990.
- [18] J.J. More "The Levenberg-Marquardt Algorithm: Implementation and Theory", in *Numerical Analysis*, G.A. Watson, Springer-Verlag, Ed., 1977, pp. 105-116.
- [19] R. Kronland-Martinnet, Ph. Guillemain, S. Ystad "Modelling of Natural Sounds Using Time-Frequency and Wavelet Representations" in *Organised sound*, Cambridge University Press, Ed., Vol.2 n°3, 1997, pp.179-191.

# HARMBAL: A PROGRAM FOR CALCULATING STEADY-STATE SOLUTIONS TO NONLINEAR PHYSICAL MODELS OF SELF-SUSTAINED MUSICAL INSTRUMENTS BY THE HARMONIC BALANCE METHOD

*Snorre Farner*

Group S2M, Laboratoire de mécanique et d'acoustique, CNRS, Marseille, France

(exchanged from the Acoustics group at Department of Telecommunications, NTNU, Trondheim, Norway)

## 1. INTRODUCTION

A musician playing an acoustic instrument has many ways to control the quality of the sound of a single tone, i.e. pitch and timbre, some ways more subtle than others. In our search to understand and model the acoustical behaviour of the instrument, we try to quantify the player's means of control by parameters, like the blowing pressure or the firmness of the lips against the clarinet reed etc. The quality of the sound itself can be quantified by considering the peaks of the steady-state spectrum which is easily calculated from the real instrumental sound by a Fast Fourier Transform.

The *Harmonic Balance Method* is a method that is able to produce the peaks of such a spectrum given a set of physical equations describing the system. It works in the frequency domain and as it calculates the steady-state spectrum, only periodic solutions can be found, contrary to the time domain. Therefore, the result sounds static and boring. On the other hand, the great advantage is that it is very general and can be used on highly nonlinear systems, whether the nonlinearity is in the frequency or the time domain. Furthermore, it is rapid since it only calculates the a relatively small number of harmonics. To find a solution to a rather complex problem, it is possible to use a solution found for a simpler one and sufficiently gradually complexify the problem.

Now, models have been proposed and used for a large range of self-sustained instruments, but the problem is always to find solutions for these models with general conditions. The Harmonic Balance Method makes us able to study the model systematically and to learn how to control it. In addition, it efficiently provides verification for approximate analytical calculations like Variable Truncation Method [1].

This report presents the computer program *Harmbal* for using the Harmonic Balance Method to calculate the steady-state spectrum of a self-sustained musical instruments from a system of three equations: a nonlinear time-domain equation and a linear differential oscillation equation for the excitation, and the frequency response of the resonator. We will employ the clarinet as an example, but it is easy to change the program to use other equations and parameters to support other self-sustained instruments. In fact, the program should be sufficiently general to accommodate more fundamental changes, but this would require a good knowledge of the C programming language.

C was chosen as programming language because of the advantages of fast operation and good portability to most other operating systems. It will be distributed in September 2002 under the GNU General Public License (see [www.gnu.org/copyleft](http://www.gnu.org/copyleft)) as a free software with copyleft (a copyright that allows everyone to use,

change, and redistribute the program, but not to take copyright on anything based on this program, or to take away the copyleft freedom from it). With this everyone is encouraged to use the program and share their improvements and additions with everyone.

At last it should be mentioned that a similar program although less general was made for Matlab by S. Menigoz [2] in supervision of J. Gilbert [3]. This program was used in the early stages of programming and debugging of Harmbal and for verification of the results.

## 2. THE HARMONIC BALANCE METHOD

The Harmonic Balance Method is a numerical method to calculate the steady-state spectrum of periodic solutions of a nonlinear dynamic system. Although the method originally seems to have been designed for forced oscillations with a known fundamental frequency [4], it has been modified to self-sustained oscillations by adding the fundamental frequency as an unknown variable [3].

In short we search a solution  $\vec{x}$  containing the amplitudes of the  $N$  first harmonics as well as the constant component, i.e.  $N+1$  complex components, or  $2(N+1)$  real components which we organize as  $N+1$  real parts followed by  $N+1$  imaginary parts. The amplitude of the first harmonic ( $x_1 + ix_{N+2}$ ) can be made real by shifting the solution appropriately in the time domain, so that the imaginary part  $x_{N+2} = 0$ . The system may be written in the form

$$\vec{x}^\infty = \vec{F}(\vec{x}^\infty), \quad (1)$$

where  $\vec{x}^\infty$  is the solution. We thus search for the root of

$$\vec{G}(\vec{x}) = \frac{\vec{x} - \vec{F}(\vec{x})}{x_1}, \quad (2)$$

i.e.  $\vec{G} = 0$ . We have divided by the real and non-zero amplitude  $x_1$  of the first harmonic to avoid the trivial solution  $\vec{x} = 0$ .

By starting with a guessed value  $\vec{x}^i$ , we use the Newton-Raphson Method to calculate a step  $\Delta\vec{x}$  to a point  $\vec{x}^{i+1}$  closer to the solution and thus search the solution iteratively. As a change in the frequency will change  $\vec{G}$ , we take advantage of the fact that the imaginary part of the first harmonic,  $x_{N+2}$ , is zero and put the playing frequency  $f_1$  in its place.  $\Delta x_{N+2}$  is thus the change needed in the frequency to come closer to the solution.

The method is modified in the program with a *backtracking mechanism* to take a step shorter than  $\Delta\vec{x}$  if the new solution seems to be worse than the one we already have.

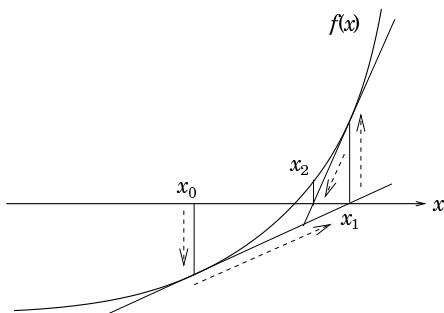


Figure 1: The iteration process of Newton's method

### 2.1. Newton-Raphson and backtracking

Starting with a one-dimensional problem, we want to find  $x$  for  $f(x) = 0$ . With Newton's method,

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}, \quad (3)$$

we calculate the slope  $f'(x_i)$  in our starting point  $x_i$  and follow it to where it crosses zero as a first approximation of the solution. This is repeated while increasing  $i$  till  $f(x_i) < \varepsilon$ , a small value, as illustrated in Figure 1.

In a multidimensional system, the problem is a vector problem: We search  $\vec{x}$  for which  $\vec{G}(\vec{x}) = 0$ . Newton's method is generalized to the Newton-Raphson Method, which can be written:

$$\vec{x}^{i+1} = \vec{x}^i - \left(\mathbf{J}_G^i\right)^{-1} \cdot \vec{G}(\vec{x}^i), \quad (4)$$

where  $\mathbf{J}_G^i = \nabla \vec{G}(\vec{x}^i)$  is the Jacobian of  $\vec{G}$  at  $\vec{x}^i$ . The step  $\Delta \vec{x} = \vec{x}^{i+1} - \vec{x}^i$  may be called the *Newton step* and follows the steepest tangent of  $\vec{G}(\vec{x}^i)$  to zero. Effectively, this is the multidimensional version of the well-known Newton's Method.

The Jacobian, a  $(2N+2) \times (2N+2)$  matrix, could be found analytically if the derivatives  $J_{ij} = \partial G_i / \partial x_j$  can be calculated analytically, but normally it is sufficient to use the first-order approximation

$$J_{ij} \simeq \frac{G_i(\vec{x} + \delta \vec{x}_j, f) - G_i(\vec{x}, f)}{\delta x_j}, \quad (5)$$

where  $\delta \vec{x}_j$  is zero except for the  $j$ th component which is the tiny perturbation  $\delta x$ . However, we exchanged  $x_{N+2}$  with the playing frequency  $f$ , so column number  $N+2$  of the Jacobian should be

$$J_{i,N+2} = \frac{\partial G_i}{\partial f} \simeq \frac{G_i(\vec{x}, f + \delta f) - G_i(\vec{x}, f)}{\delta f}. \quad (6)$$

Note that even though  $x_{N+2} = 0$ ,  $G_{N+2}(\vec{x}, f)$  is in general not.

Unfortunately, the tangent is not always sufficiently steep, and thus the Newton step brings us further away from the solution. Often this will result in  $|\vec{G}^{i+1}| > |\vec{G}^i|$ , in which case we apply the backtracking algorithm described in Numerical Recipes [5]. Acknowledging that the Newton step is the direction of the steepest descent, we simply take a shorter step  $\lambda \Delta \vec{x}$  in the same direction, where  $0 < \lambda \leq 1$ .

Now, a probably good value of  $\lambda$  can be found by defining the function

$$g(\lambda) = \frac{1}{2} |\vec{G}(\vec{x} + \lambda \Delta \vec{x})|^2 \quad (7)$$

with derivative

$$g'(\lambda) = \left( \nabla \vec{G} \cdot \vec{G} \right) \Big|_{\vec{x} + \lambda \Delta \vec{x}} \cdot \Delta \vec{x}. \quad (8)$$

We already have  $g(0) = \frac{1}{2} |\vec{G}(\vec{x})|^2$ ,  $g'(0) = -|\vec{G}(\vec{x})|^2$ , and  $g(1) = \frac{1}{2} |\vec{G}(\vec{x} + \Delta \vec{x})|^2$ , so we can model  $g(\lambda)$  as a quadratic function and find its minimum

$$\lambda = -\frac{\frac{1}{2} g'(0)}{g(1) - g(0) - g'(0)}. \quad (9)$$

It can be shown that  $\lambda$  should not exceed  $\frac{1}{2}$ , and we require  $\lambda \geq 0.1$  to avoid too short a step at this stage. If we still have  $|\vec{G}(\vec{x} + \lambda \Delta \vec{x})| > |\vec{G}(\vec{x})|$ , we model  $g(\lambda)$  as a cubic function, minimize it and keep  $\lambda$  to be between 10 and 50 % of the previously calculated  $\lambda$ . This calculation requires solving a system of two equations, so if also the new  $\lambda$  is not accepted because  $|\vec{G}|$  is still too large, we do not enhance to a fourth-order function but do instead another cubic calculation with the most recent values of  $\lambda$ . However, not many repetitions should be necessary before finding a better solution or realizing that this approach does not work.

### 2.2. Application on wind instruments

For a wind instrument the method takes the form shown in Figure 2 [3]. Capital letters denote quantities in Fourier space and small letters in the time domain, thus  $P(\omega) = \mathcal{F}(p(t))$  and  $p(t) = \mathcal{F}^{-1}(P(\omega))$ , where  $\mathcal{F}$  denotes the Fourier transform.  $p$  is the internal pressure in the mouthpiece and  $x$  the displacement of the mechanical excitor (the reed for a reed instrument, the lips for a brass instrument, for instance).  $u$  is the volume flow of air through the mouthpiece. For other types of instruments, the same variables may have different signification. To keep the equations as general as possible and with as few parameters as possible, we use consequently dimensionless quantities.

The excitation is modelled as a spring with mass and damping:

$$M\ddot{x} + R\dot{x} + Kx = p, \quad (10)$$

where  $M$ ,  $R$ , and  $K$  are dimensionless mass, damping, and spring constant.  $Kx$  includes possible constant terms on the right-hand side. Equation (10) is Fourier-transformed and calculated in the frequency domain. The volume flow, however, is nonlinear and must be calculated in the time domain:

$$u(t) = u(p, x). \quad (11)$$

A third equation describes the resonator (tube or string) and is most simply calculated in the frequency domain by means of its frequency response  $Z(\omega)$ :

$$P(\omega) = Z(\omega)U(\omega). \quad (12)$$

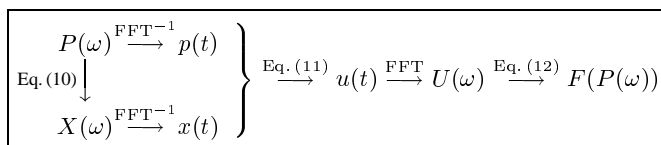


Figure 2: The iteration loop of the harmonic balance method for a clarinet

We discretize one period of the time-domain variables in  $N_t$  equidistant samples, i.e.  $u_n = u(n/N_t f)$ , implying that the sampling frequency  $N_t f$  adjusts to the current playing frequency  $f$  so that we always consider one period of oscillation while keeping  $N_t$  constant. Thus a discrete Fourier transform gives us the  $N_t/2$  first harmonics of the oscillation (including the constant component), each at a multiple of the playing frequency,  $f_k = kf$ . We limit our calculations to the first  $N_p$  harmonics, or partials, so that we have  $N_p+1$  components including the constant component. As for  $\vec{x}$  in start of Section 2, we separate the real and imaginary parts and represent  $P(\omega)$  with  $\vec{P}$  given by  $2(N_p+1)$  components.  $N_t$  must be at least  $2(N_p+1)$ , and it should be a power of two so that the Fast Fourier Transform (FFT) can be used.  $N_t$  should also be sufficiently large to avoid aliasing.

In Harmbal, Equation (10) is given by the three parameters  $M$ ,  $R$ , and  $K$ , Equation (11) is given by a C function that calculates  $u_n$  for all  $N_t$  time samples ( $0 \leq n < N_t$ ), and Equation (12) is given by a function that calculates the  $2(N_p+1)$  components of the impedance  $Z_k$ ,  $0 \leq k \leq N_p$ .

As Figure 2 implies, the calculations revolve around the pressure  $\vec{P}$ , and the Newton-Raphson Method is employed to find the root of the vector

$$\vec{G}(\vec{P}, f) = \frac{\vec{P} - \vec{F}(\vec{P}, f)}{P_1} \quad (13)$$

where  $P_{N+2} = f$ , the playing frequency.

### 2.3. The equations for a clarinet

The oscillation of the reed of a clarinet may be described using dimensional quantities by

$$\ddot{y} + g_r \dot{y} + \omega_r^2 y = \frac{1}{\mu_r} (p - p_m), \quad (14)$$

where  $y$  is the dynamic reed displacement,  $\mu_r$ ,  $g_r$ , and  $\omega_r$  are its mass per area, damping factor, and angular resonance frequency, and  $p$  and  $p_m$  are the dynamic pressure in the mouthpiece and the static pressure in the players mouth, respectively. In dimensionless form we define  $\tilde{p} = p/p_M$ ,  $\tilde{x} = y/H + \gamma/K$ ,  $\tilde{t} = \omega_t t$ , and  $\gamma = p_m/p_M$ , tilde ( $\tilde{\cdot}$ ) temporarily denoting dimensionless quantity, which give

$$K = \mu_r H \omega_r^2 / p_M, \quad (15)$$

$$M = K \omega_t^2 / \omega_r^2, \quad (16)$$

$$R = M g_r / \omega_t. \quad (17)$$

$\omega_t$  is the natural resonance frequency of the tube and  $H$  the reed opening, see Figure 3. When blowing too hard,  $p_m \geq p_M$ , i.e.  $\gamma \geq 1$ , the reed constantly blocks the opening, i.e.  $y = -H$ , so we get  $p = 0$  and can conclude that  $K = 1$ . We return to dimensionless quantities.

A usual basic model for the clarinet assumes a simple reed with no mass or damping, thus  $M = R = 0$  and  $K = 1$ . Equation (10) thus reduces to  $x = p$  and a blocking reed occurs for  $p \leq \gamma - 1$ .

Equation (11) can be written, for finite  $M$  and  $R$ :

$$u(p, x) = \zeta (1 + x - \gamma) \sqrt{|\gamma - p|} \text{sign}(\gamma - p) \quad (18)$$

as long as  $x > \gamma - 1$ ,  $u = 0$  otherwise.  $\zeta = Z_c w H \sqrt{2/\rho p_M}$  is a non-dimensional parameter for the mouthpiece construction,  $w$  being the width of the opening and  $\rho$  the density of the air.

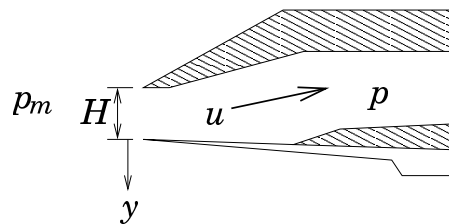


Figure 3: A schematic drawing of the mouthpiece

Finally, the input impedance of a cylindrical tube of length  $l$  with an open and a closed end can be written as  $Z(\omega) = i \tan(kl)$  (dimensionless). The wave number  $k = \omega/c + i\alpha$  where  $\alpha \simeq 1.3\nu\sqrt{\omega/\omega_t}$  is due to viscous losses.  $c$  is the speed of sound in air and its frequency dependency may be approximated by  $c = \omega c_0 / (\omega + \alpha c_0)$  to model dispersion.

## 3. PROGRAM LAYOUT AND USE

Harmbal consists of two parts, a core and a user part. The core part contains the harmonic-balance loop, the interface, numerics and so on. This is described in the README file following the program. However, the physical equations are put into the user part, which enables the user to formulate and add new equations for new purposes, as long as she conforms to the system of three equations as described in Section 2.

### 3.1. User's guide

Intentionally, calculations with Harmbal are performed from the command prompt, thus a command-line window must be opened and the command `harmbal` followed by options and arguments must be executed. The parameters as well as a guess values of the spectrum and playing frequency are given in a parameter file (and possibly changed by the command-line arguments). The program returns the results in another parameter file of the same format so that a parameter can be changed and the program be re-executed based on the new parameter file.

This low-level interaction between user and program was chosen to give a high degree of flexibility; Harmbal can be used directly in connection with other programs, such as Matlab, or by shell scripts to do batch jobs such as finding solutions for a range of a parameter or iteratively searching a solution for a complicated case by starting from a simple one. And even if there should be a bug in Harmbal that causes it to crash, nothing is lost and the user may modify whatever she thinks caused the error and re-run Harmbal.

Finally, experienced C programmers may compile the functions of Harmbal into other C programs performing related tasks (see however licence agreement).

#### The usage of Harmbal:

`harmbal [options] [frequency [all partials]]`

where brackets mean that the content is optional, *options* start with a minus sign (see below), *frequency* is obvious, and *partials* are given as  $\Re(P_0), \Re(P_1), \dots, \Re(P_{N_p}), \Im(P_0), \Im(P_1), \dots, \Im(P_{N_p})$ .

Possible options are:

- f *fn* input name *fn* of parameter file (default: `params.pmt`)
- o *fn* output name *fn* of parameter file (default: `pout.pmt`)
- p  $N_p$  change number of partials (excl. the constant component)
- t  $N_t$  change number of time samples (should be a power of 2 and  $\geq 2(N_p + 1)$ )
- c *p v* change a parameter *p* to the given value *v*
- h help information
- M calc. till min.  $|\vec{G}|$  is reached, not only till  $|\vec{G}| < \text{maxerr}$
- D direct Newton-Raphson, i.e. no backtracking

#### Some rules:

- If input file already is a solution, the result is merely copied to the output file. If Harmbal cannot find a solution, the output file is not made (nor touched if already existing).
- The command-line arguments take precedence to the values in the file, and if  $N_p$  is given, it is more important than the number of partials given in the file or even at the command line. Thus, there is no point in specifying at the same time  $N_p$  and the partials.
- The frequency must be given if the partials are given, and the partials must include the constant component, so for  $N_p$  partials, you should have  $2(N_p + 1)$  values plus the frequency.

#### The parameter file:

The parameter file is simply a text file that lists all necessary parameters and their values, one per line in the format *parameter value*. The last item in the file must be *P* which is given by *p* alone on a line immediately followed the  $2(N_p+1)$  partials, one on each line. Blank lines and lines starting with a hash (#) are ignored and not rewritten in the output parameter file.

There are some mandatory parameters which the program needs to run: *K*, *R*, and *M* (for Equation (10)), *nlmodel* and *impmodel* (nonlinear and impedance model numbers), *pmax* ( $p_M$ , a value of 1 for dimensionless), *denom* (which harmonic (default is 1, i.e.  $P_1$ ) to use in the denominator in Equation (13)), *maxitno* (maximum number of iterations), *maxerr* (maximum  $|\vec{G}|$  and  $|f^i - f^{i-1}|/f^i$ ), *freq* (initial playing frequency  $f^0$ ), *resfreq* (the lowest resonance frequency of the resonator  $\omega_t/2\pi$ ), and of course  $N_p$  and  $N_t$ . Both model numbers are the sum of the *instrument number* multiplied by 100 and the *model number*. For the nonlinear model (model 2) of the clarinet (instrument 1), for example, a line should read `nlmodel 102`.

Other parameters depend on the nonlinear and impedance models used. The use of the equations in Section 2.3 requires *nu* (damping factor) and *disper* (dispersion flag; 0=none, 1 dispersion; with intermediate values allowed if needed to converge) for the tube, and *zeta* ( $\zeta$ ) and *gamma* ( $\gamma$ ) for the nonlinear equation.

#### Examples of use:

```
harmbal 100 runs Harmbal on default parameter file,
  params.pmt, adjusting the initial frequency to 100 Hz,
  and return the solution, if found, in pout.pmt.

harmbal -f clar1.pmt -o clar2.pmt -c gamma
...0.5 -t 64 100 0 0.1 0 0 runs Harmbal on an
earlier parameter file clar1.pmt, changes gamma to
```

0.5,  $N_t$  to 64, the frequency to 100 Hz and the initial  $\vec{P}$  to have one harmonic with amplitude 0.1, makes the appropriate iterations to find a solution, and finally writes the result to `clar2.pmt`.

```
harmbal -f pout.pmt -c gamma 0.4 reads
pout.pmt, changes gamma, runs Harmbal, and writes to
the same file. If no solution was found, pout.pmt is not
changed. This is particularly practical when a gradual
change of a parameter is necessary, or to collect solutions
for a range of a parameter.
```

### 3.2. Adding instruments or functions

To add a nonlinear equation *u* or an impedance *Z*, it is necessary to go into the source code of the program and do some simple programming. However, this part of the code is separated from the core of the program to facilitate the change for unexperienced users of C. This is described in the following sections. Section 3.3 lists some pitfalls.

After whatever little change of the program, it must be recompiled. This is done in unix/linux systems by writing `make` in the directory of the program. If there are errors, these are often of the nature covered in the Section 3.3.

#### 3.2.1. Adding a new instrument

Assume that you want to add saxophone functions to the program. Then you should make a new file `saxophone.c` with a header file `saxophone.h` (see below). But first, tell about its existence by opening the file named `instr.h` and adding a line `#include "saxophone.h"` to the list close to the end of the file:

```
#ifdef INSTR
#include "clarinet.h"
#include "saxophone.h"
#endif
```

Then decide a number that is not already taken, say 2, for your new instrument package and add a line `#define SAXOPHONE 2` at the end of the file. Within the program, you should always refer to this instrument by the constant `SAXOPHONE`, not the number, in case it is necessary to change the number.

In the file `instr.c`, find the function `general_resonator()` and a new case in the `switch` statement, just before the default statement. This is where the instrument is chosen by the parameter `impmodel` (or `nlmodel` for the nonlinear model):

```
case SAXOPHONE:
  reson = saxophone_resonator
    (impmodel % INSTRFAC,paramlist);
  break;
```

Repeat this in the function `general_nonlin()`.

Now, it is time to make the file `saxophone.c` by copying the standard file `stdinstr.c`. Substitute `clarinet` with `saxophone` and update `saxophone_resonator()` or `saxophone_nonlin()` every time a new function is added, see next sections. Make also a file `saxophone.h` (from `stdinst.h`) containing the declaration of all the new functions you add.

In makefile you should add `saxophone.o` to the variable `INSTOBS`, which you find in the preamble of the file. This tells of the existence of the new file to the compiler.

### 3.2.2. Adding a new resonator

A new function should start with the name of the instrument file to avoid double-use of a name. The framework of an impedance function is as follows:

```
double *saxophone_nameimp(int N, double freq,
                          double *params)
{
    int i; //partial counter, 0=constant comp
    double nu, resfreq, disper; //parameter names
    double *Z; // output impedance array
    complex Za, Zb; // other temporary complex..
    double L, a; // ...and real variables

    resfreq = params[0]; // resonance freq. of
                        // full tube length
    nu = params[1]; // attenuation
    disper = params[2]; // dispersion flag
                        // (can be fractional)

    /* calculate Z = ?? */
    Z = allocvec(2*N);
    w1 = 2.0*PI*freq;
    for(i=1; i<N; i++){
        // calculations, dimensionless and in..
        Zb = ..?; // ..complex form
        Z[k] = Zb.re; // convert to real array
        Z[k+N] = Zb.im;
    }
    Z[0] = Z[N] = 0; //special case for freq=0
    return Z;
}
```

The function declaration, simply the first line followed by a semi-colon,

```
double *saxophone_nameimp(int N, double freq,
                          double *params);
```

must be added to `saxophone.h` for the main program to recognize it.

Close to the top of the file `saxophone.c` there should be a function `saxophone_resonator()`. The new resonator must be added to the switch clause, just before default. Here the function to be used is chosen by the parameter `impmodel`. We look at the already existing case 1 in `clarinet.c`:

```
case 1: // tanh(j(k+da)l + al);
        // d = dispersion flag
        np = 3;
        params = getparams(paramlist,
                            stringlist(np, "resfreq", "nu", "disper"),
                            YES);
        reson = initresonator(clarinet_tubeimp1,
                              params, np);
        break;
```

When you have made a new function, all the parameters that it needs should be listed with quotes in the line starting with `params`, and `np` in the line above should be changed to the number of parameters. It is important that `resfreq` is the first! In the line starting with `reson`, substitute the name of your function. To write the impedance function as a comment after `case` makes it easier to find the right case. The case number is the model number and should thus be changed accordingly.

At last, all new variables should be added to the parameter file used for the problem. Missing parameters will produce an error message. You should use dimensionless parameters as much as possible.

### 3.2.3. Adding a new nonlinear function

Adding a new nonlinear function is done in the same way as adding a new resonator. The framework is as follows:

```
double *saxophone_ul(double *x, double *p,
                    int Nt, double sampfreq,
                    double *params, int np)
{
    int n;
    double *u; // array to be returned
    double A, B; // parameters
    double deriv; // temporary variables

    A = params[0]; // get parameter from params
    B = params[1];

    /* calculate dimensionless u=Ax+Bp+dx/dt */
    u = allocvec(Nt); // allocate memory to u
    nold = Nt-1;
    for(n=0; n<Nt; n++){
        deriv = (x[n]-x[nold])*sampfreq;
        // the time derivative of x at n
        u[n] = A*x[n] + B*p[n] + deriv;
        // arbitrary example function!
        nold = n;
    }
    return u;
}
```

Similarly to the impedance, the new function must be added to `clarinet_nonlin()` as a case:

```
case 1:
    func = initfunc(saxophone_ul, 0);
    np = 2; // number of parameters
    valuelist = getparams(paramlist,
                          stringlist(np, "A", "B"), YES);
    break;
```

Remember that `np` should be updated to the number of parameters given in the line following and that the number after `case` must be changed. The new function name is substituted in the line starting with `func`.

### 3.3. Pitfalls

- Since we use dimensionless quantities, the impedance should be divided by the characteristic impedance, for instance  $Z_c = \rho c / A_c$ ,  $A_c$  being the characteristic cross section of the tube.

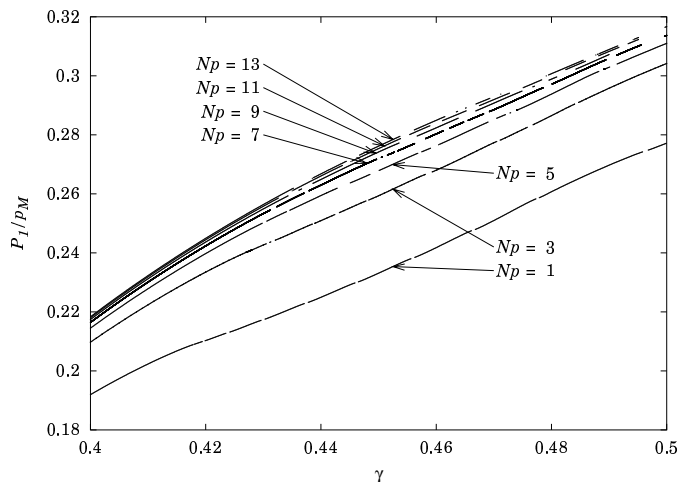


Figure 4: Solution holes in the curve of  $P_1$  versus  $\gamma$  for different  $N_p$  and  $\zeta = 0.5$ ,  $\nu = 10^{-3}$ , and  $N_t = 128$ . For even  $N_p$  we get the same as for  $N_p - 1$ .

This is easy to forget when using more than one impedance function at the same time.

- Arrays in C starts with the 0th element, so for  $N_t$  elements, the elements of  $u$  are  $u[0], u[1], \dots, u[Nt-1]$ .
- A forgotten semi-colon after a statement or mismatched braces ( $\{, \}$ ) may cause errors somewhere else.
- All variable must be declared (floating point variable as `double`, integer as `int`, and complex as `complex`).
- Calculations with complex variables must be performed with the functions given in `calc.h`, even adding or changing sign (since `complex` is a `struct`), e.g.  $ae^{iz+b}$  may be written `RCmul(a, Cexp(Cadd(ICmul(1, z), Complex(b, 0))))`, where  $a$  and  $b$  are `double` and  $z$  is `complex`. To access to the real and imaginary parts, add `.re` and `.im`, respectively.

#### 4. SOLUTION HOLES AND DIGITAL SAMPLING

Before the backtracking algorithm was employed, it was impossible to find a solution at special combinations of the parameters. In particular, when changing  $\gamma$  while keeping all other parameters constant, at certain values of  $\gamma$  and some of its neighbourhood, we call them *holes*, no convergence was obtained, even though the solutions at lower and higher  $\gamma$  seemed to go continuously through this hole, as shown in Figure 4. The Newton-Raphson method seemed to stop converging by switching between two values of  $\vec{P}$  and  $f$  or it just started to diverge. (The curve is traced by a script (`hbmap`) which runs `Harmbal` for a range of a parameter, in this case  $\gamma$  from 0.5 as low as possible in steps of  $10^{-4}$ .)

By setting  $N_p = 1$ , we have a one-dimensional problem since only  $P_1$  has a non-zero value and thus  $G_1$  is the only contributor to  $|\vec{G}|$ . With another program, `tracpar`, written in C, a graph of  $G_1$  for varying  $P_1$  around the solution  $G_1 = 0$  could be made, as shown in Figure 5 for several values of  $\gamma$  around a hole.

We see that the curve of  $G_1(P_1)$  experience soft jumps at rather regular distances, and at the center of the hole, i.e. for  $\gamma \simeq 0.4196$ , the jump is centered at the intersection of the  $x$ -axis. The

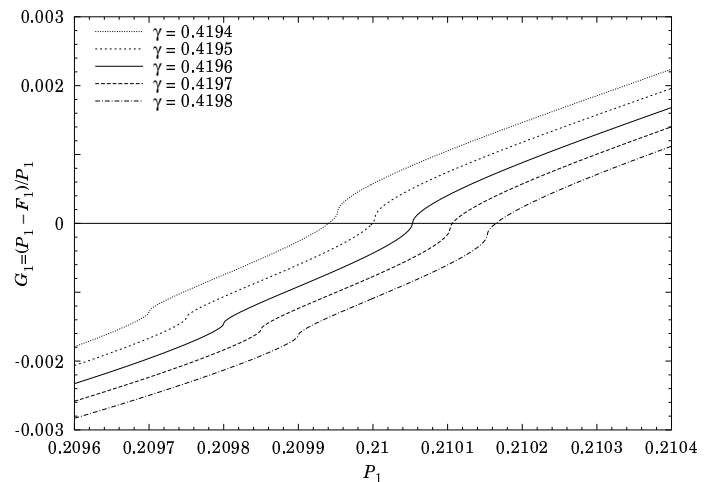


Figure 5:  $G_1$  as  $P_1$  varies around the solution  $G_1 = 0$  for various  $\gamma$  around a hole at  $\gamma \simeq 0.4196$ .  $N_t = 128$  and  $N_p = 1$ .

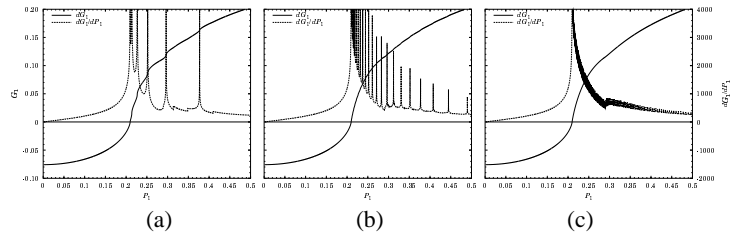


Figure 6: The effect of sampling rate on the smoothness of  $G_1(P_1)$ : (a)  $N_t = 32$ , (b) 128, and (c) 1024. The derivative  $G_1/P_1$  quantifies the roughness.

jump thus forms a school example of a situation where Newton's method does not converge because the Newton step  $\Delta P_1$  brings us alternatingly to the one and the other side of the solution, but not closer.

To solve the problem we ask ourselves why these jumps occur. The answer is the digital sampling of a continuous signal. If we increase the sampling rate, i.e. increase  $N_t$ , the jumps become smaller but occur more frequently, as shown for  $N_t = 32, 128$ , and 1024 in Figure 6a-c. The derivative  $\partial G_1/\partial P_1$  is added to quantify the size of the jumps. As  $N_t \rightarrow \infty$ , the system becomes continuous and the curve smooth. However, to increase  $N_t$  will only help convergence locally as the jump is likely to move away. Globally, there will probably be a greater number of holes, but they will cover a smaller range, in this case, of  $\gamma$ . Furthermore, to be able to use the Fast Fourier Transform,  $N_t$  should be a power of 2, so increasing the sampling rate will be costly in terms of computing time.

Instead, to cope with the problem, we simply shorten the Newton step sufficiently, and the backtracking algorithm described in Section 2.1 elegantly estimates how much. After adding the backtracking mechanism, the problem of convergence failure is almost completely vanished.

#### 5. FREQUENCY INSTABILITY

Another problem experienced, is that if the playing frequency  $f$  is badly guessed, it starts wandering and the system diverges rapidly. By plotting  $G_1$  for a range of  $f$ , see Figure 7, we see why: The

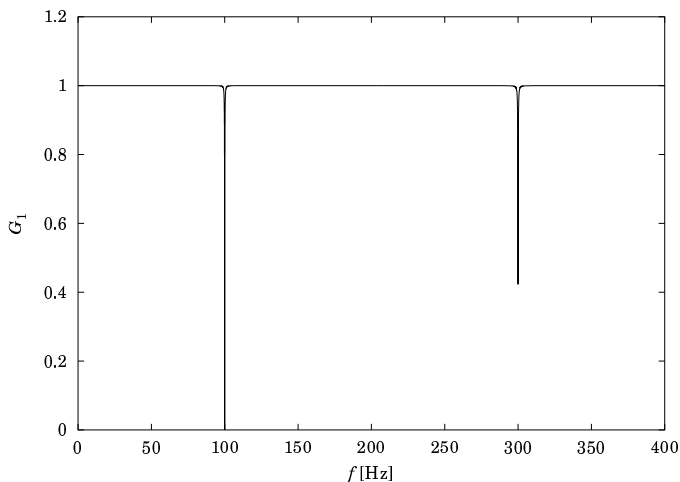


Figure 7: The variation of  $G_1$  for changes of the playing frequency around the solution for  $\omega_t/2\pi = 100$  Hz

slope of  $G_1(f)$  is zero (because  $\hat{P}$  is small and thus  $G_1 \simeq P_1/P_1$ ) except very close to the solution. Thus the Newton-Raphson Method cannot find a solution. For a simple system as the one shown in the figure where the well corresponds to a resonance peak of the tube, it is not difficult to guess, but it quickly becomes a challenge when adding a mass to the reed, dispersion in the tube, or interaction between to resonators.

In Harmbal the problem is temporarily mended to some extent by the possibility of changing the dispersion, for instance, continuously, so that the playing frequency can be followed quasi-continuously from a known solution without dispersion.

## 6. CONCLUSIONS

The program Harmbal is ready and free to be used by all of you and is in fact already in use by several researchers.

The problem of convergence failure has been examined and its origin was found to be the digital sampling of the signal. As this cannot be avoided, it was solved by implementing the backtracking mechanism.

The program has been compared to give the same solutions as an earlier program for Matlab which was made for a specific model of a clarinet and has been used several years already. Apart from being much faster than this earlier program, Harmbal is more general, meaning that it is simple to add new models and that it may be run automatically (by a script or other program) to produce ranges of solutions in a simple fashion. Thus, in principle, all researchers working on self-sustained musical instruments should be able to benefit from the presented program.

Two problems with the method are currently interesting to study: The method is very sensitive to the guessed playing frequency, and it does not tell whether a solution is physically stable or not.

## Acknowledgements

The European Union through the MOSART project is acknowledged for financial support, and apart from general help from my

colleagues at Laboratoire de mécanique et d'acoustique (LMA) at CNRS in Marseille, France, I would like to thank Claudia Fritz at IRCAM in Paris, for thorough testing and valuable feedback since she at a very early stage started to use the program, as well as Christophe Vergez at CNRS-LMA, Jean Kergomard at CNRS-LMA, and Joël Gilbert at Laboratoire d'acoustique de l'université du Maine (LAUM) in Le Mans for fruitful discussions during the work.

## 7. REFERENCES

- [1] Jean Kergomard, Sébastien Ollivier, and Joël Gilbert, "Calculation of the spectrum of self-sustained oscillators using a variable truncation method: Application to cylindrical reed instruments," *Acta Acoustica*, vol. 86, no. 4, pp. 685–703, 2000.
- [2] Sébastien Menigoz, "Oscillations non-linéaires des instruments à vent à anche simple: étude par équilibrage harmonique," July 1998, (French).
- [3] J. Gilbert, J. Kergomard, and E. Ngoya, "Calculation of the steady-state oscillations of a clarinet using the harmonic balance technique," *J. Acoust. Soc. Am.*, vol. 86, no. 1, pp. 35–41, 1989.
- [4] M. S. Nakhla and J. Vlach, "A piecewise harmonic balance technique for determination of periodic response of nonlinear systems," *IEEE Trans. Circuit Theory*, vol. 23, no. 2, pp. 85–91, 1976.
- [5] William H. Press et al., *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, 2 edition, 1992.





**IHP Network HPRN-CT-2000-00115 MOSART**  
**Music Orchestration System in Algorithmic Research and Technology**

**MOSART Task 3:**

**Control and Virtualisation of Instruments**

**Edited and compiled by**

**Jan Tro**

**Deliverable d23:**

Report on significant progress in experiments on **Control and virtualisation of instruments.**

**Table Of Content**

<b>Control and Virtualisation of Musical Instruments.</b> Jan Tro	Page 124
<b>A classification of controllers for parameters highlighting</b> Gabriele Boschi	Page 128
<b>Musical Instruments Control and Expression</b> Kristoffer Jensen	Page 134
<b>From Sounds to Music: Different Approaches to Event Piloted Instruments</b> P. Gobin, R. Kronland-Martinet, G-A. Lagesse, T. Voinier, S. Ystad	Page 143
<b>New Method for the Directional Representation of Musical Instruments in Auralizations</b> Felipe Otondo, Jens Holger Rindel, {	Page 169
<b>Directional representation of a clarinet in a room</b> Felipe Otondo and Jens Holger Rindel,	Page 173

# Control and Virtualisation of Musical Instruments.

Jan Tro ([tro@tele.ntnu.no](mailto:tro@tele.ntnu.no))

Norwegian University of Science and Technology

## 1 Introduction

The main goal of this report is to document significant progress on the Task 3 topic: **Control and Virtualisation of Instruments.**

Concerning Task 2 (Musical Sound Understanding, Timbre Modeling) and Task 3 we expected the research method would take perceptual evidence and experimentation into consideration, as well as mathematical descriptions of basic physical sound excitation phenomena and sound control parameters, in the attempt to look for alternative descriptors of sound quality and sound control. The close relationship between these two topics (Task 2 and 3) has made it necessary to include references and articles performed in the framework of both tasks, even though control and virtualisation aspects will be emphasised here.

## 2 Activities

This activity review will fall in two categories:

- **Networking:** The training of young researchers,
- and
- **research** accomplishments and progress of the network project.

Concerning networking young researchers directly involved in the research projects reported here is mentioned.

Regarding the presentation of the Task 3 progress research methods and strategy will be discussed.

Concerning *control aspects* performer controlled musical parameters will be focused. This includes the description of standard instruments as musical control devices, the development of new controllers for contemporary music and systematic measurements of control ability (performology).

*Virtualisation of instruments* will focus on the simulation of acoustical instruments (mainly physical modeling of sound sources) and the comparison with acoustical instruments.

Some of the main experiment reports are attached as

Appendices.

### 2.1 Networking

Contribution to the networking activity with young researchers or scientific reports related to Task 3 concerns the following institutions:

Norwegian University of Science and Technology (NTNU), Trondheim,  
Ørsted DTU, Acoustic Technology, Copenhagen,  
Centro Nazionale Universitario di Calcolo Elettronico (CNUCE), Pisa,  
Centre Nationale de la Recherche Scientifique (CNRS-LMA), Marseilles,  
Pompeu Fabre University (UPF), Barcelona.

In addition several other MOSART engaged institutions have stated interest in this task.

A complete activity report of young researchers is in process. Networking authors are mentioned in chapter 3.

### 2.2 Research

While musicians possess a craftsmanship in handling an instrument, it is not clear how a computational representation of actual or virtual instrument expressive features should be organised. Likewise it is not clear, how transitions of tones should be synthesised, when representations of single tones and instrument features are combined in a musical phrase. The state of art in digital instruments is still concerned with samples of single tones and isolated sounds. The combination of such, according to a certain style of composition or playing, is an entirely open problem within the synthesis of musical sound. The physical modeling of instruments and computational synthesis of musical plausible sound from these models have become a scientific study in its own right.

The analysis by synthesis approach has been a prevailing technique in these tasks. Concerning the broader scope of modeling of phrases and of complete instruments, several research strategies seem attractive. However, basic knowledge of excitation

mechanisms, control parameters and sound radiation needs still to be controlled, enhanced and purified.

On this background the development of sound recording databases for further detailed sound and control analyses seems to be relevant.

The following section includes scientific reports and references on performance control parameters, recordings in anechoic and reverberant environments, and event piloted instruments.

### 3 Selected scientific reports

#### 3.1 A Classification of Controllers For Parameters Highlighting.

Gabriele Boschi (MOSART short term employee at NTNU – Acoustics, spring 2002) discusses music controllers as part of one of the following three families:

- 1) Controllers that are actually **Acoustical instruments**, that is traditional instruments.
- 2) **Electronic instruments** quite close to the related traditional instrument, but also supporting new ways to communicate musical expressions.
- 3) **Electronic controllers** which use or are based on completely different concepts than the ones used in traditional instruments.

The report includes a comprehensive discussion on control parameters.

Full text is attached in Appendix I.

#### 3.2 Musical Instruments Control and Expression

This paper, written by Kristoffer Jensen (DIKU, Copenhagen), presents a study on the control of musical instruments and the expressive model, a generic signal model which models the perceptive effect of expressive changes of most musical sounds. The expressive model is based on a model of the hammer-string interaction of the piano, and on the observations of the control and expressions of some typical acoustic musical instruments.

The paper is divided into two parts: a first part with theoretical investigation of the control of musical instruments, and a second part, which details the expressive model.

Full text is attached in Appendix II.

#### 3.3 From Sounds to Music: Different Approaches to Event Piloted Instruments

This paper, written by Pascal Gobin, Richard Kronland-Martinet, Guy-André Lagesse, Thierry Voinier and Sølvi Ystad, addresses three different strategies to map real-time synthesis of sounds and controller events. The design of the corresponding interfaces takes into account both the artistic goals and the expressive capabilities of these new instruments. Common to all these cases, as to

traditional instruments, is the fact that their specificity influence the music which is written for them. This means that the composition already starts with the construction of the interface.

As a first approach, synthesis models are piloted by completely new interfaces, leading to "sound sculpting machines". An example of sound transformations using the Radio Baton illustrates this concept.

The second approach consists in making interfaces that are adapted to the gestures already acquired by the performers. Two examples are treated in this case: the extension of a traditional instrument and the design of interfaces for disabled performers.

The third approach uses external events such as natural phenomena to influence a synthesis model. The *Cosmophone*, which associates sound events to the flux of cosmic rays, illustrates this concept

Full text is attached in Appendix III.

#### 3.4 Experiments on instrument directivity.

##### 3.4.1 Directional patterns and recordings of musical instruments in auralization

This paper, written by Felipe Otondo (MOSART employee at Ørsted DTU), Jens Holger Rindel and Claus Lyng Christensen, outlines the developing ideas of the investigation started the 1<sup>st</sup> of September 2001. The problem of the spatial representation of sound sources that vary their directional pattern in time in auralizations is introduced. Musical instruments are used as a reference for the discussion of the traditional representations with assumed fixed directional characteristics. A new method for representation of the spatial sound contributions in time is proposed using multiple-channel recordings and virtual sources in the simulations. Further developments and applications of the solution are outlined.

Full text is attached in Appendix IV.

##### 3.4.2 New method for the Directional Representation of Musical Instruments in Auralizations

The paper is written by Felipe Otondo (MOSART employee at Ørsted DTU) and Jens Holger Rindel.

The issue of the representation of sound sources that vary their directional pattern in time in auralizations is introduced. Musical instruments are used as a reference for the discussion of the traditional representations with assumed fixed directional characteristics. A new method for the representation of the spatial sound contributions in time is proposed using multiple-channel recordings and various virtual sources in room auralizations. Possible developments of the proposed recording/reproduction method are described.

Full text in Appendix V.

In addition to this paper a multi-channel 24 bit version of the described recordings are available. These recordings have been done with multiple microphone setup surrounding the source as described in the paper.

These anechoic recordings have been done at the Technical University of Denmark(DTU).

The samples in this CD are at 16 bit quantisation and 44.1 kHz.

Most of the instruments have recorded in the whole compass. Some instruments have been recorded chromatically (Bb clarinet, Bass clarinet, Spanish guitar, tuba & violin), the rest include of the instruments scales and isolated tones.

The melodies recorded are short simple melodies intended for tests or comparisons.

### **3.4.3 Directional representation of a clarinet in a room**

This article, written by Felipe Otondo (MOSART employee at Ørsted-DTU) and Jens Holger Rindel, presents a study of the directional characteristics of a clarinet in the context of a real performance. Anechoic measurements of the directivity of a Bb clarinet have been done in the horizontal and vertical planes for isolated tones. Results are discussed comparing the particular directivity of tones and the averaged directivity over the whole range of the instrument. Room acoustic simulations with the measured and averaged directivities have been carried out in a concert hall as an example of a more realistic application. Further developments will consider measurements with other instruments as well as auralizations and tests with an alternative sound radiation representation.

Full text in Appendix VI.

## **3.5 Other projects**

### **Gesture capture techniques.**

A project on control, performed by Declan Murphy (MOSART employee at DIKU), consist of video gesture capture techniques to be applied to real-time control of computer/electronic instruments (eg PCM by cART or the Timbre Engine by DIKU) and to manipulation of musical structures for performance, composition and analyses.

Gesture capture techniques include arm tracking, baton tracking and hand posture tracking. Each involve two cameras and are prototyped on the EyesWeb platform by DIST. Work began at DIKU Copenhagen, continued at cART CNR Pisa, then DIST Genoa, and now at DAIMI Århus. It will continue at DAIMI/DIKU/Esbjerg.

Relevant publications: [1][2][3]

### **The X-Tiles**

Another project related to the MOSART control activity at DAIMI, Århus, performed by Declan Murphy (MOSART employee at DIKU) is the so-called "X-Tiles". It involves an Irish step dancer performing on a specially prepared surface. Traditional Irish (folk) step dancing involves intensely percussive footwork, which is recorded both via audio and with MIDI pressure sensors. The MIDI information (which includes spacial location) is used for higher-level control and manipulation of the amplified audio from the feet. The audio is picked up by surrounding microphones, which directly feed a surround sound system.

The control in this case would be in the form of a combination of the filling in of the basic rhythms and the overlaid spacio-temporal location of the steps, as a part of an interactive composition.

### **Votion**

A project performed by Declan Murphy (MOSART employee at DIKU) is called "Votion" and was a collaboration between DIKU, the Danish Design School, and DTU (all in Copenhagen). It was to involve using pointing gestures and spoken words to navigate a VR concept environment with a musical soundscape.

### **Recording of repeted performances.**

At NTNU a series of repeted performances have been recorded. The database includes lot of MIDI recordings. In addition the MOSART employee Gabriele Boschi, a top-class flute player, performed 40 recordings of the flute part of the opening movement of Francois Devienne's Concerto no. 7 for flute and piano during the summer 2002. The performances have been recorded digitally in different acoustical environments and will be made available for further research.

## **4 Conclusion and future plans.**

Copared to the high goals stated in the former MOSART agreement the scientific activity and documentation seems to comply with the scheduled plans concerning Task 3. The networking activity, however, has not been completely successful, at least for some of the network nodes where the lack of young research candidates has been a problem. An increase in future research employees may still give a positive outcome for the coming final Task 3 status.

The next stage of Task 3 will include comprehensive analyses of the established sound databases, extended music control parametre models, comparison of instrument recordings and computer simulations (physical modeling), and external presentation of industrialisable technological results and ideas.

## 5 References

The following publications are highly relevant to the Task 3. Some references may be reviewed in other MOSART tasks as well.

- [1] Declan Murphy, "Extracting Arm Gestures for VR using EyesWeb", Proc. Workshop on Current Research Directions in Computer Music, Barcelona, Nov 2001.
- [2] Declan Murphy, "Building a Hand Posture Recognition System: A Bottom Up Approach", Tech. Report, cART CNR Pisa, Mar 2002.
- [3] Declan Murphy, "An Improved Edge Detection and Ranking Technique", Proc. Danish Conf. on Pattern Recognition and Image Analysis, Aug 2002.
- [4] J. Bensa, K. Jensen, R. Kronland-Martinet, and S. Ystad. Perceptual and analytical analysis of the effect of the hammer impact on the piano tones. In Proceedings of the ICMC, Berlin, Germany, 2000.
- [5] K. Jensen and G. Marentakis. Hybrid perception. Papers from the 1<sup>st</sup> Seminar on Auditory Models, Lyngby, Denmark, 2001
- [6] J. Bensa, F. Gibaudan, K. Jensen, and R. Kronland-Martinet. Note and hammer velocity dependence of a piano string model based on coupled digital waveguides. In Proceedings of the ICMC, Havana, Cuba, 2001.
- [7] G. Marentakis and K. Jensen. Hybrid synthesizer: Progress report. In Workshop on current research directions in computer music, Barcelona, Spain, 2001.
- [8] K. Jensen and Murphy D. Segmenting melodies into notes. In Proceedings of the DSAGM, Copenhagen, Denmark, 2001.
- [9] K. Jensen. The timbre model. In Workshop on current research directions in computer music, Barcelona, Spain, 2001.
- [10] D. Murphy, G. Marentakis, T. H. Andersen, and K. Jensen. Scalable spectral reflections in conic sections. In Proceedings of the Digital Audio Effects Workshop, Limerick, Ireland, 2001.
- [11] Jordà, S. 'Afasia: the Ultimate Homeric One-man-multimedia-band'. Proceedings of New Interfaces for Musical Expression. Dublin, Ireland 2002.
- [12] Jordà, S. Barbosa, A. 'Computer Supported Cooperative Music: Overview of research work and projects at the Audiovisual Institute - UPF'. Proceedings of MOSART Workshop on Current Research Directions in Computer Music. Barcelona, 2001.
- [13] Jordà, S. Wüst, O. 'FMOL: A System for Collaborative Music Composition over the Web' Proceedings of Web Based Collaboration DEXA 2001. Munich, Germany, 2001.
- [14] Jordà, S. 'New Musical Interfaces and New Music-making Paradigms'. Proceedings of New Interfaces for Musical Expression. CHI 2001. Seattle, 2001.
- [15] Jordà, S. 'Improvising with Computers: A Personal Survey (1989-2001)'. Proceedings of International Computer Music Conference 2001. Havana, Cuba, 2001.
- [16] Wüst, O. Jordà, S. 'Architectural Overview of a System for Collaborative Music Composition over the Web'. Proceedings of International Computer Music Conference 2001. Havana, Cuba, 2001.
- [17] Tro, J. "Aspects of Control and Perception". Proc. DAFx00, Verona, Dec. 7-9, 2000, pp. 171-176.
- [18] Tro, J. "Measurements, Control and Precision in Music Performances". in "Recent Trends in Basic Psychophysics and Their Application to Acoustics", 5D.14.03. 17<sup>th</sup> International Congress on Acoustics, Rome, Sept. 2-7, 2001.
- [19] Waadeland, C.H. "Rhythmic Movements and Moveable Rhythms". Dr. Thesis, NTNU, Dept. of Musicology, Trondheim, 2000.
- [20] Handegard, H. "Modulation Effects in Guitar Tones". M.Sc. Thesis, NTNU-Acoustics, Dec. 1999.
- [21] Ystad, S. "Sound Modeling Using a Combination of Physical and Signal Models". Dr. Thesis, Universite de la Mediterranee - Aix-Marseille II and NTNU - Trondheim, 1998.
- [22] Smevik, T. "Analysis and Simulation of a French Horn". M.Sc. Thesis, NTNU - Fac. of Physics, Informatics and Mathematics, Dec. 2000.
- [23] Haugen, S.H. "Acoustic Properties of the Piano Soundboard". M.Sc. Thesis, NTNU-Acoustics, Dec. 1998.
- [24] Morset, L.H. "An Investigation of Vibrational and Acoustical Properties of the Violin Using TV Holography". 137<sup>th</sup> ASA Meeting, paper 3AMU-12, Berlin, 1999.
- [25] Svensson, U.P. et.al. "Effects of Wall Reflections on the Sound Radiation from a Kettledrum: A Numerical Study". Proc. ISMA, Leavenworth, June 26-July 1, 1998, pp. 371-376.
- [26] Torvmark, K.H. "Presentation and Evaluation of Timbral Microstructures". M.Sc. Thesis, NTNU-Acoustics, Dec. 1999.
- [27] Kristiansen, U., Støfringsdal, B., Svensson, P. & Tro, J. "Performance control and virtualization of acoustical sound fields related to musical instruments. Proceedings of Workshop on Current Research Directions in Computer Music. Barcelona, Nov. 15-17, 2001.

# A CLASSIFICATION OF CONTROLLERS FOR PARAMETERS HIGHLIGHTING

Gabriele Boschi, NTNU – Acoustics.

We can consider a music controller as part of one of the following three big families:

- 1) Controllers that are actually **Acoustical instruments**, that is traditional instruments.
- 2) **Electronic instruments** quite close to the related traditional instrument, but also supporting new ways to communicate musical expressions.
- 3) **Electronic controllers** which use or are based on completely different concepts than the ones used in traditional instruments.

Family two can be further divided into two sub-groups, in order to make the classification more clear:

- 2-A: Electronic Instruments as **extension** of a specific traditional instrument;
- 2-B: Electronic Instruments based on Traditional Instruments but with **different shape, material etc.**

Also for the third family some sub-division can be considered. In fact, electronic controllers are usually referred to belong to one of the two big groups: the free gesture and all the other ones (often called tactile or with tangible reference). Some nice considerations can be found in the following paragraph, extracted from "Tangible Music Interfaces Using Passive Magnetic Tags", Paradiso et Al, Massachusetts Institute of Technology [1].

“Although noncontact gesture sensors, generally based around capacitive sensing or computer vision and optical approaches, make very expressive musical controllers, they suffer from a lack of a tactile interface (leading to a deficit in precise, virtuosic input) and often the inability to reliably identify and distinguish different objects or body parts (limiting variation in control). The combination of free-gesture sensing with a tangible reference has the potential of producing a very expressive electronic musical interface that also encompasses a degree of tactile precision and versatility.”

Of course free gesture devices are part of the third group, since there are no traditional free gesture instruments.

The **parameters list** at the end of each group in the following classification is referred to the parameters that can be controlled by professional players with that particular instrument or instruments.

## **FAMILY 1: Acoustical Instruments**

Acoustical instruments, that is instruments that do not require electricity to be played, have been widely analysed in the past years and decades. In this paper we focus our attention on the kind of sound parameters control they allow.

**Wind Instruments:** In this kind of instruments all the parameters related to the air pressure, air direction, lip tension etc. are continuous, while the ones directly linked to the hands' positions (usually a finite number) are discrete.

**Bowed string Instruments:** String instruments have continuous control over all parameters, since both the excitation and the pitch-timbre controls are continuous.

**Plucked Instruments:** Instruments like the guitar, the harp, the clavichord fall into this category. Here the pitch normally has a finite number of values, while all the other parameters are usually continuous.

**Percussion instruments:** The number of available pitches is here fixed, but modulation and timbre variations are continuous controlled.

**Pipe/Electric Organ; Piano:** All these instruments have a keyboard which limit the available pitches to a discrete number. All the other parameters, such as timbre and Intensity, are usually continuous controlled.

*Parameters:*      **Pitch, Dynamic features, Spectral content, Articulation,  
Transition between notes, phrasing, rhythm.**

## **GROUP 2-A: Electronic Instruments as extension of a specific traditional instrument**

The instruments in this group are based on traditional instruments but some kind of sensors or transducers have been added in order to increase the available musical expression.

**Electric Guitar:** The electric guitar can be considered a new controller, or more precisely a new instrument, close to the Classic Guitar. It has been invented to fulfill the request of a certain amount of output power. From this point the fact that the output is electrical has made possible a dizzying array of sounds produced by electrically and

electronically modifying this electrical output. Besides the volume and tone controls on the guitar and on the amplifier, a variety of outboard devices are used to obtain custom sounds and effects.

*Parameters:* **All the parameters related to the following kind of sound processing: overdrive, distortion, compression, pitch shifting, flanging, phasing, chorus, delay, echo, reverberation etc.**

**Hypercello:** Hypercello, made by Yo-Yo Ma, Tod Machover, and Neil Gershenfeld, is a musical instrument, like any other, that makes sounds in response to the player's actions. It looks almost like a traditional cello, but the body has been fitted with a range of new sensors developed for the project to measure the player's actions. These then go into a real-time computing environment that calibrates the data, parses it to find features, and implements high level rules for how gestures control electronically-produced sounds. It can be played like an ordinary cello, but can also do more. For example, bowing ponticello (near the bridge), can open up entirely new sonic palettes rather than just sounding brighter. Or gestures can launch phrases or control algorithm parameters instead of just producing single notes.

*Parameters:* **The sensed parameters are used to trigger and modify synthesized sounds that accompanies the acoustic cello. The high level rules stated above should be considered as part of the mapping strategies.**

[<http://www.media.mit.edu/physics/yoyo.html>]

[<http://web.media.mit.edu/~joep/SpectrumWeb/captions/Cello.html>]

**Electric Violins:** These kind of instruments have most of the playing characteristics and nuances that one would expect from any traditional violin, but let the player extend the expressiveness in some way. In the MIDI-violin (Jensen Musical Instruments) it is possible, with a pitch-to-MIDI converter, to control sound processing and synthesis in a similar way to the Hypercello. Yamaha Silent Violin is equipped with some kind of signal processing, with which the sound of the violin is processed in order to have the rich acoustic environment typical of a concert hall without any external processing equipment.

*Parameters:* **In the MIDI-violin the mapping plays a big role. On the other hand a specially designed chip by Yamaha digitally enhances the Silent Electric Violin's sound on-violin with rich reverb. A switch on the underside of the instrument lets you select from Large Hall, Medium Hall or Room reverb types.**

*Other Parameters:* **The typical parameters of acoustical string instruments.**

[<http://www.halcyon.com/jensmus/violinop.htm>]

[[http://www.yamaha.co.jp/english/product/silent\\_v/features.html](http://www.yamaha.co.jp/english/product/silent_v/features.html)]

### **GROUP 2-B: Electronic Instruments based on Traditional Instruments but with different shape, material etc.**

This group comprises all the new instruments, close to the traditional ones, but either for the shape, for the material, or the way they work they can be different from the ones in group 2-A.

**Digital pianos, keyboards, electronic organs:** These are well known instruments, where big efforts were and are still made in order to make them play as close as possible to the related traditional instrument. Additive synthesis has been widely used but also PCM and sound samples are nowadays standards for these kind of devices, which also incorporate a big number of features not present in the traditional instruments (for example drum and bass synthesizers).

*Parameters:* **Pitch, Simple Timbre and Global Volume are present in all keyboards. Professional keyboards are equipped with more and more sound parameters and sound processing units. Starting from the dynamic features (hammer touch, aftertouch) it is possible to have Spectral content setup, Reverberation, Chorus, Multi effects etc.**

[<http://www.rolandus.com/products/>]

**Wind Controller Yamaha WX7:** The Yamaha WX7 is a woodwind-oriented MIDI wind controller. The fingerings are based upon saxophone fingerings, with a few major differences such as the 5 octave keys for the left thumb (2



down, 3 up, and a neutral position) which makes for 6 octaves. There are also alternate fingerings such that one can cover 7 full octaves.

*Parameters:* **The instrument sends MIDI note-on messages, with appropriate velocities, via a breath-pressure sensor. Breath pressure can be mapped to instrument volume (Control Change #7) for those synthesizers which don't listen to breath pressure (CC#2) via a DIP switch located on the back, above the left thumb. The other DIP switches include Aftertouch mapping for CC#2, Eb/Bb/8va key changing, key hold mode (same-note or parallel), breath response curve, and loose/normal lip bend mode. There are four trim pots on the instrument making it possible to tailor the response of the WX7 to your personal preference.**

[<http://www.kbspace.com/wx7/9>]

**Electric drums:** The first widely-marketed drum interface was the Moog 1130 Drum Controller. This device, introduced in 1973, employed an impact-sensing resistor in the drumhead and gave audiences their first exposure to synthesized drums in the concerts of progressive rock bands. Simmons SDX drumpads introduced the concept of "zoning", where hits of varying intensity in different areas of a single pad could trigger different sonic and MIDI events. Nowadays, although Simmons are long-vanished, nearly every musical instrument manufacturer makes electronic percussion interfaces. For instance Korg with Wavedrum supersedes the limited information in simple trigger detection by employing the actual audio signal received by transducers on the drumhead as excitation for the synthesis engine (various synthesis and processing algorithms are implemented, such as physical modelling), enabling a very natural and responsive percussion interface.

*Parameters:* **Sound samples triggering is the typical feature, even present in low cost controllers. Wavedrum, a quite advanced device, also allows the player to select/design his/her own drum sound, and because it uses a drumhead with microphones for its source waveforms, it accurately captures the nuances of various drumming techniques, including brushing and so on.**

[<http://web.media.mit.edu/~joep/SpectrumWeb/SpectrumX.html>]

[[http://www.sospubs.co.uk/sos/1994\\_articles/nov94/korgwavedrum.html](http://www.sospubs.co.uk/sos/1994_articles/nov94/korgwavedrum.html)]

**Voice:** For obvious reasons it is quite difficult to abstract the voice mechanism away from the sonic output, as was pursued in the guitar and wind controllers. For this reason all voice-driven electronic sound comes from processing the audio signals picked up by microphones. For example, Will Oliver, at the MIT Media Laboratory, has taken this approach in the Brain Opera's "Singing Tree", a realtime device that breaks the singing voice into 10 different dynamic parameters, which are then used to control an ensemble of MIDI instruments that "resynthesize" the character of the singing voice, but with entirely different sound sources.

*Parameters:* **In Singing Tree the extracted parameters from voice were found in the changing of a vowel to the next one, and also the amount of pitch deviation, the speed with which it deviates, the acceleration and zero points, and the periodicity of these deviations. It was found that this held enough information for creating the musical experience. The Singing Tree mapped the vocal parameters and their deviations using dynamic set assignment, probability, and random number generation. Mapping will be discussed in the next group.**

[[http://feynman.stanford.edu/people/Oliver\\_www/DIS/DIS2.html](http://feynman.stanford.edu/people/Oliver_www/DIS/DIS2.html)]

### **FAMILY 3: Electronic controllers far away from traditional musical instruments**

The list of controllers would here be very long, since very few worldwide standard controllers have been at present defined. When there is more than one device quite similar to others we here include the most interesting or the most known ones.

**Theremin:** The theremin, perhaps the first free gesture device in history, was invented in 1919 by a Russian physicist named Lev Termen (his name was later changed to Leon Theremin). Besides looking like no other instrument, the theremin is unique in that it is played without being touched. Two antennas protrude from the theremin - one controlling pitch, and the other controlling volume. As a hand approaches the vertical antenna, the pitch gets higher. Approaching the horizontal antenna makes the volume softer. Because there is no physical contact with the instrument, playing the theremin requires precise skill and perfect pitch.

*Parameters:*     **Independent control of Pitch and Volume of a single wave form.**

[<http://www.thereminworld.com>]

**Theremin extensions:** The basics of the Theremin have been taken again into account from many researchers worldwide, developing new and more powerful devices. For instance the sensor chair, developed at MIT, is a device that measures the body (hand and foot) positions and motions of a seated occupant, still using electromagnetic fields.

*Parameters:*     **No general parameters, see mapping.**

[<http://web.media.mit.edu/~joep/TTT.BO/chair.html>]

**Batons:** An interesting interface that began life in the 1980's as a percussion controller was computer music pioneer Max Mathews "Daton" (predecessor of the "Radio Drum"), where a sensitive plate responded to the location and force of a strike. The strike location was determined by measuring differential force with 4 pressure sensors at the corner plate supports. Many other researchers have explored related optical interfaces. For instance the Light Baton by Bertini and Carosi at the IEI-CNR in Pisa and the IR baton of Morita and colleagues at Waseda University both use a CCD camera and a frame-grabber to track the 2D motion of a light source at the tip of a wand in real time.

*Parameters:*     **Usually batons are used for "conducting" music, that is to tell the score player the speed and the expressive nuances during the performance.  
If more sensors are present (for instance accelerometers) it is possible to have an highly expressive control over electronic music; the performer can "conduct" the music at a high level, or descend into a "virtuoso" mode, actually controlling the details of particular sounds.**

[<http://www.newmusicbox.org/third-person/oct99/batons.html>]

In the free gesture devices world, almost all kind of sensors have been used in order to track the movement of the human body or of parts of it. Some of the widely used kind of sensors are considered in the following.

**Ultrasound:** The EMS Soundbeam, a wellknown ultrasound tracking system, is a distance-to-voltage-to-MIDI device which converts physical movements into sound by using information derived from interruptions of a stream of ultrasonic pulses. Also "Sound=Space" dance installation by Rolf Gelhaar uses ultrasounds to detect human motion.

*Parameters:*     **Besides the mapping issue, described later, Soundbeam features a 30 Factory-preset Pitch Sequences of 64 notes or 3-note chords, 70 User-definable Pitch Sequences of Up to 64 notes or 3-note chords - played in from a MIDI keyboard and recorded and 128 Soundbeam Controller Setups – i.e. 128 complete sets of all parameter settings (for each of the 4 Sensors and 8 switch inputs).**

[<http://www.soundbeam.co.uk/>]

[<http://www.newmusicbox.org/third-person/oct99/noncontact.html>]

**Infrared:** Infrared proximity sensors, most merely responding to the amplitude of the reflected illumination, are being used in many modern musical applications. For instance Soundstairs triggers musical notes as people walk up and down a stairway, obscuring or reflecting IR beams directed above the stair surfaces. Commercial musical interface products have appeared along these lines, such as the "Dimension Beam" from Interactive Light (providing a MIDI output indicating the distance from the IR sensor to the reflecting hand), and the simpler "Synth-A-Beams" MIDI controller, which produces a corresponding MIDI event whenever any of eight visible lightbeams are interrupted. One of the most expressive devices in this class is the "Twin Towers", developed by Leonello Tarabella and Graziano Bertini at the CNUCE-CNR in Pisa. This consists of a pair of optical sensor assemblies (one for each hand), each containing an IR emitter surrounded by 4 IR receivers. When a hand is placed above one of these "Towers", it is IR-illuminated and detected by the 4 receivers. Since the relative balance between receiver signals varies as a function of hand inclination, both range and 2-axis tilt are determined. The net effect is similar to a Theremin, but with more degrees of sonic expression arising from the extra response to the hand's attitude. Airsynth from Alesis is one of the latest infrared commercial devices available, where the hand position is tracked with a specific controller named XYZ™ 3 dimensional infrared controller.

*Parameters:*     **All the mentioned devices except Airsynth relies on mapping for**

**generating music. Airsynth is instead a stand alone instruments with 100 presets which comprises staccato, percussive, legato continuous pads, drum sounds and sounds that emulate things in nature.**

[[http://www.sospubs.co.uk/sos/1997\\_articles/jun97/dimensionbeam.html](http://www.sospubs.co.uk/sos/1997_articles/jun97/dimensionbeam.html)]

[<http://www.cnuce.pi.cnr.it/tarabella/Gesture.html>]

[<http://www.alesis.com/products/airsynth/>]

**Computer Vision Techniques:** Although they involve considerably more processor overhead and are generally still affected by lighting changes and clutter in compare to other kind of approaches, computer vision techniques are becoming increasingly common in free gesture musical interfaces and installations.

For over a decade now, many researchers have been devising vision systems for musical performance, and high increases in available processing capability have continued to improve their reliability and speed of response, while enabling recognition of more specific and detailed features. As the cost of the required computing equipment drops, vision systems become price-competitive, as their only "sensor" is usually a commercial video camera.

A straightforward example of this is the Imaginary Piano by CNUCE's Tarabella, which consists of a real-time image-analysis of video-captured system: here, the interaction "tools" are the mere bare hands of a pianist. The pianist is sitting as usual on a piano chair and has in front nothing but the camera few meters away pointed on his hands. There exist an imaginary line at the height where usually the keyboard lays: when a finger, or a hand, crosses that line downward, a specific message (actually a NoteOn MIDI message) is issued; "where" the line is crossed states the key number, "how fast" the line is crossed, states the velocity. This application should be considered an original performance rather than an original instrument due to the inaccuracy of gesture when striking the right key.

A package called BigEye, written by Tom DeMeyer and his colleagues at STEIM, tracks with any kind of video capturing system multiple regions of specified color ranges (ideally corresponding, for instance, to pieces of the performers' clothing or costumes). The output from BigEye (a MIDI or other type of data stream) is determined in a scripting environment, where sensitive regions can be defined, and different responses are specified as a function of the object state (position, velocity, etc.).

EyesWeb, developed at Laboratorio di Informatica Musicale (InfoMus) in Genoa by Antonio Camurri et Al., is an open software platform for the development of real-time music and multimedia applications. EyesWeb includes a hardware and software platform to support the user in the development and experimenting of computational models of expressive content communication and of gesture mapping strategies, and also in fast development and experiment cycles of interactive performance setups.

*Parameters:* **In the imaginary piano the available parameters are the note number (pitch) and the velocity (note value), which can be used for directly controlling the piano performance, but also for triggering and controlling other sound synthesis.**

**The latter methods clearly arise mapping issues, and also are the typical approaches used in the two other controllers.**

[<http://www.cnuce.pi.cnr.it/tarabella/Gesture.html#Impiano>]

[<http://www.steim.nl/bigeye.html>]

[[http://musart.dist.unige.it/sito\\_inglese/research/r\\_current/eyesweb.html](http://musart.dist.unige.it/sito_inglese/research/r_current/eyesweb.html)]

The following controllers fall into the sub-category of **wearable devices**, that is in these instruments sensors are affixed to the body or to the clothing of the performer.

**Percussive sensors:** The first experimental controllers (for instance electro-acoustic clothing by Benoit Maubrey) mostly used these sort of transducers placed on the performer's body to detect his or her actions.

On the other hand, starting with the early 90s, these devices were usually equipped with continuous sensors and RF units for wireless links.

**Biosignals:** Some devices have been built to detect biological signals and convert them into MIDI data stream. For instance Biomuse, produced by Biocontrol Systems, is able to acquire signals from heart, brain, eye movement etc, and convert them into MIDI messages.

**Gloves:** These kind of controllers have been widely used for music performances. One example, composed by Tod Machover at the MIT Media Lab, is "Bug Mudra", where an Exos "Dexterous Hand Master" was worn by the conductor, who had complete dynamic control over the audio mix and synthesis parameters through finger positions. Another commercial example is the Dataglove, designed by Tom Zimmerman and produced by VPL Research. Dataglove is equipped with optical flex sensor and is able to track 10 finger joints (lower two of each finger, two for thumb) and six DOF of the hand's position and orientation (magnetic sensor on back of glove).

**Others:** Musical jackets have been built at the MIT Media Lab, with a touch-sensitive MIDI keyboard embroidered directly into the fabric using conductive thread. The DIEM Digital Dance System is an interface designed especially for interactive dance. The dancer wears up to 14 bending sensors that measure the angles of the dancer's limbs. The bending sensors are connected to a small wireless transmitter worn by the dancer on a belt. Data is transmitted to a receiver unit which sends standard MIDI controller values for each sensor.

*Parameters:* **All the above mentioned wearable controllers are not complete, or more precisely stand alone instruments, since they rely on mapping for controlling sound synthesis.**

[<http://www.i-a-s.de/IAS/Maubrey/Maubrey.html>]

[<http://www.biocontrol.com/>]

[<http://www.geocities.com/mellott124/glove1.htm>]

[<http://www.cs.nps.navy.mil/people/faculty/capps/4473/projects/smithml/vplDataGlove.htm>]

[<http://web.media.mit.edu/~joep/SpectrumWeb/captions/Jackets.html>]

[<http://www.daimi.au.dk/~diem/digitaldance.html>]

### **Notes on mapping and sound parameters control:**

While the instruments in family one and two are clearly defined and outlined, when facing to the third group the instrument definition becomes less clear. In fact great part of these devices are mostly controllers, and instead of having sound outputs they have electronic signals carrying some information on the behaviour of the performer. The relationship between the values in this data stream and all the sound parameters is part of a wide aspect in computer music, the *mapping*.

The purpose of this paper is not to give a complete view on mapping, so only very few aspects will be taken into account: the hyper-instrument, the multi-level mapping and the instrument's playability.

A musical instrument is usually a compact tool which allows a performer to play music. If we consider any device in the third family, we can see that in order to play it must be connected to a synthesis appliance; in this case the instrument is the whole system while the device can be considered only a controller.

Since the behaviour of the whole instrument is mapping-dependent, that is the instrument relies on mapping to define the resulting sound from the performer's actions, changing the mapping change the instrument itself. This behaviour gives the life to the hyper-instrument definition.

In a hyper-instrument almost any kind of sound parameters can be controlled with opportune mapping strategies. The maximum number of parameters is usually related to how many degrees of freedom has the controller. If this number is big enough it is possible to gain control over all the parameters present in instruments in groups one and two, and also have something more, since the sound synthesis can be very complex and powerful.

Instead of listing all the parameters, interesting for computer music composers, it is possible to divide them in two parts: the low and the high level parameters.

Low level parameters directly affect the sound synthesis and thus the sound produced. This can be related to the notes, timbre, the single-note intensity and so on. On the other hand high level parameters affect the high level aspects of music, that is the global signs in scores, such as dynamics, articulation, phrasing, rhythm etc. The values coming from the controller are usually linked either to low or high level parameters.

Since writing a piece of music with third group controllers means to devise mapping strategies, it is usually up to the composer to take mapping into account. While carrying out this task it is important that he or she considers the *playability* of the instrument. In fact certain kind of mapping (usually complex ones) will make the instrument quite hard to play and the audience may experience a not too clear relationship between cause (the performer actions) and the effect (the sound characteristics). On the other hand simple mappings can increase playability but also may be obvious for the public. The composer and the performer should find a trade off between the two modes.

### **References:**

Links on each paragraph have been checked on 24-5-2002.

A very detailed article about controllers by Joe Paradiso:

<http://web.media.mit.edu/~joep/SpectrumWeb/SpectrumX.html>

# Musical Instruments Control and Expression

Kristoffer Jensen

Music Informatics Laboratory  
Department of Datalogy, University of Copenhagen  
Universitetsparken 1, 2100 Copenhagen Ø, Denmark  
[krist@diku.dk](mailto:krist@diku.dk), <http://www.diku.dk/~krist>

## Abstract

This paper presents a study on the control of musical instruments and the expressive model, a generic signal model which models the perceptive effect of expressive changes of most musical sounds. The expressive model is based on a model of the hammer-string interaction of the piano, and on the observations of the control and expressions of some typical acoustic musical instruments. The paper is divided into two parts: a first part with theoretical investigation of the control of musical instruments, and a second part, which details the expressive model.

## 1 Introduction

Acoustical instruments can be divided into two classes, envelope-based instruments (which are only given energy once), and continuous-control (which are continuously given energy) instruments. Some instruments, such as the bowed string instruments, permit the execution of both techniques, envelope-based by plucking the string, or releasing the bow, and continuous-control by stroking the bow on the strings. The sound of the musical instrument can be qualified by the timbre (or the identity) of the sound and the expressions caused by gestures. Expressions associated with musical instruments are well defined by common musical terms, such as note, intensity, tempo or style.

Section 2 presents the sound parameters, which are generally controlled in musical instruments and section 3 presents the control structure of some typical acoustic instruments. Most instruments only permit one class of control, which can further be divided into several subclasses. These classes, or different control mechanisms will be outlined in section 4, for some typical acoustic instruments, along with the perceptive outcome of this control. Based on this analysis, a model of continuous control of electronic music instruments is outlined, which permits the control of a large number of parameters on an instrument with few sensors. The continuous controls of electronic instruments are a possibility today when the instruments are becoming realistic enough, and especially considering that some of the new synthesis algorithms, i.e. physical modeling, are appropriate for controlling many physically related aspects of the sound.

A novel set of expression additions designed to be used with the timbre model [26], but general enough to be used with most synthesis algorithms

is presented in section 5. In related topics [9] deals with the representation of continuous music signals, [15] adds rules for duration and sound level in computer performance. [30] discusses the shortcomings, and [36] proposes a replacement of the commonly used musical interface standard MIDI.

## 2 The control parameters of musical instruments

In this section, a few important acoustic attributes are detailed. These attributes can be used in the parameter control of section 4, or in the expressive model in section 5. For an overview of the normal discrimination ability of some of these parameters, see [20].

### 2.1 The loudness

One important control is the loudness control. The loudness control can be executed in two manners, limited control on an envelope, or continuous control, as in the bowed string instruments. The piano sound, for instance, has an envelope with the typical attack, decay, and release behavior. The instrumentalist has two parameters that controls the time-varying loudness; the velocity, which controls the overall envelope, and the duration of each key, which controls the length of the decay period.

### 2.2 The pitch

The pitch has several different control techniques, the discrete pitch control, as in the flute, the contained pitch control, i.e. the pitch can be varied in a limited range, vibrato, as in the guitar, and finally there's the continuous pitch control, as in the violin.

## 2.3 The timbre

Timbre is defined as the difference between two sounds having the same pitch, loudness and duration in [1]. For an overview of timbre research, see [26]. The timbre generally applies only to quasi-harmonic sounds, and it is related to the amount of energy in the different harmonics, or frequency components, and the temporal evolution of this energy. Some instruments, such as the violin, has control over many timbre dimensions, others have a fixed timbre.

## 2.4 The noise

Noise is broadband components of the sound. Some instruments permit the addition of a noise component on the sound. This can be breathing noise, bowing noise, tapping noise, etc. Also, playing at the limit of a resonant mode can increase the noise component.

## 2.5 The inharmonicity

Inharmonicity is created when the sound contains spectral components whose frequencies are not multiples of the fundamental (i.e. not harmonics). The bowed string instruments have control over the inharmonicity. The inharmonicity increases for instance in the violin, when bowing in an angle to the string, which increases the longitudinal vibration. See [13], p. 133 for a mathematical study of this phenomenon. In addition, string instruments and decaying wind instruments have stretched or compressed harmonic frequencies which adds inharmonicity to the sound, and some wind instruments adds inharmonicity (or roughness) when tuned to the wrong note before playing.

## 2.6 The localization

Most acoustic instruments permits to be moved and pointed at different directions. Moving the instrument changes the amplitude and the timbre of the sound, which is interpreted by the listener, who thus understands the location of the instrument.

# 3 Acoustical instruments control

In this section, the main control structures of some typical musical instruments are detailed. This information has been found by extensive interviews with professional and semi-professional musicians, and by literature studies. More information about the different instruments can be found in, for instance, [2], [5], [13], [25].

## 3.1 The violin

The violin is a bowed string instrument, which share the control characteristics with most of the

other instruments in this family. The bowed instruments include the violin, the viola, the cello, and the contrabass. In [21], the control of the violin is investigated, and a set of continuous control parameters necessary to fully exploit this instrument is proposed. The intent is to estimate the number of parameters necessary to control a good instrument, rather than create a protocol for controlling a synthetic musical instrument.

The violin is a plucked, or a bowed instrument. In the first case, the amplitude follows a predefined envelope, whereas in the latter case, the amplitude is continually controlled by the bow. Likewise, the pitch and different timbre dimensions can be analyzed as being a function of the movement of the bow, and the fingers of the left hand.

The violin is tuned G3, D4, A4 and E5. The range of each string is about 2 octaves.

The violin has the following parts, which are of interest to us: strings, fingerboard and bridge. The bow is constituted of bow hair, and the hair length is about 65 cm. The sound is produced when the bow is touching the string. The string often continues to vibrate after the bow no longer touches it, thus producing an envelope-based sound.

The amplitude of the sound is dependent on the speed of the bow, the force (pressure) of the bow, and the position of the bow on the string. The faster the bow, the louder the sound, and the closer to the bridge of the string the louder the sound. The amplitude of the violin is roughly proportional to the force of the bow. This force has a maximum and minimum value, which are functions of the position of the bow.

The onset of the sound also changes with the force of the bow. With less force, the higher harmonics develop faster than the fundamental, whereas when the force increases, the fundamental develops as fast as the harmonics.

## 3.2 The trumpet

The trumpet is a lip-driven brass instrument, which produces sounds by buzzing the lips in the mouthpiece. This buzzing is produced when the blowing pressure forces the lips apart. The lips can vibrate only at certain frequencies, defined by the resonance of the horn, which in turn depends on the form, and the length of the horn. The trumpet can thus only be played at the modes of the instrument. The upper modes of the trumpet can be placed so as to correspond to musical intervals, but the lower modes lie too far apart to be useful in the diatonic scale. The gap is filled by changing the length of the cylindrical part of the horn. The trumpet has three valves, which lowers the tone one, two and three semitones. The valves are not tuned exact to the semitone, so pressing one and two semitone valves at the same time gives a slightly higher frequency than pressing the three-semitone valve. This permits to adjust the tones. The three-

semitone valve also has a variable extension that permits to lower the tone an additional halfnote. The range is about 2 1/2 octaves, but skilled players can reach even higher tones, by better control of the lips. Some tones can be reached either by changing the valves, or the lip form, but this is rarely used, the rule being to have as few valves open as possible.

The tone ceases when the airflow stops, either by controlling the lungs, or by stopping the airflow with the tongue. The tongue is also used to get a faster attack, and to repeat the tones. The most common tongue forms are the 'k' and the 't' forms, the 't' form being slightly sharper. The blowing force decides the amplitude and the timbre, more high frequency energy gets added with the blowing force. The form of the lips also influences the timbre, but it is very hard to change this without changing either the blowing pressure, or the frequency. Vibrato is introduced by hardening and softening the lips. A noise component, typical for the instrument, is added when playing at the border of the resonance. It is common to change the timbre on the trumpet by introducing an object in front of the mouth of the trumpet. This can be either the hand, or different kind of dampers.

### 3.3 The saxophone

The saxophone was developed by Adolph Sax 150 years ago. It is a woodwind reed instrument, and includes soprano, alto, tenor and baritone models.

The sound is created when the tongue is pressing the reed against the mouthpiece, thereby allowing it to vibrate when blowing air through it. The pitch is changed by opening and closing holes, with the help of an elaborate system. The instrument covers 2 1/2 octave, but higher frequencies can be reached by tuning the mouth to a higher mode. Some pitches can be reached by several different combinations, to facilitate the quick transition from one tone to another tone.

The sound radiates from the open mouth of the instrument, and from the open holes. The combination of radiation gives a stable sound at about 1 meter from the end of the saxophone. To get an even tone that covers the whole range of the instrument, it is necessary to change the blowing technique for each tone. The blowing strength, the blow direction (up/down), the stiffness of the lips, the tongue, and the size of the cavity in the mouth can influence the stability of the sound. When the mouth is tuned to the pitch, the attack is facilitated, and the tone appears strong and stable. It is, of course, necessary to tune the mouth prior to starting the tone. The strength of the air blow

decides the amplitude and the timbre of the sound. The upper harmonics gets more energy when the blowing force increases. The direction of the blowing jet is changed by moving the jaw back and forth. This changes the timbre of the sound. Moving the jaw back give a softer tone, moving the jaw forth yields a more percussive sound. The size of the cavity in the mouth decides the stability of the tone. Deeper tones demand a larger cavity. Vibrato is produced by lowering the lower jaw and lip. This lowers the pitch of the tone slightly. The sound also softens slightly when lowering the jaw. Other parameters includes sitting/standing position, and, of course, the direction of the sound. When all the relevant parameters are tuned up, the saxophone changes pitch. Higher tones can thus be reached. The octave is the most common choice, but the fifth over the octave can also be reached. It is also common to introduce a noise in the sound by playing at the limit of the resonance. This noise can also be produced by "singing" into the mouthpiece.

### 3.4 Conclusion

The musical instruments can be controlled by the fingers, hands, mouth and body in several different perceptual dimensions, including the amplitude, pitch, timbre, noise component, inharmonicity and space. Many of these controls are continuous.

It is important to notice that the different parameters of an acoustic instrument cannot be set outside the limits of the instrument. This means, for instance, that whatever you do to change the timbre, the instrument is still recognizable. Some of these limits are not the product of any physical law, but the result of many years of practice to avoid the displeasing sounds.

## 4 The parameter control structure

In the different musical instruments, there are several ways of controlling the different parameters of the sound. Some of the parameters have no control, as the pitch in a bell, whereas the violin has full control of the pitch. Some control patterns can be singled out, such as, no control, NC, discrete control, DC, contained control (vibrato), CC, fixed envelope, FE, adaptive envelope AE (the length of the sustain period is controllable), limited continuous control, LC, and full continuous control FC. Table 1 lists the control of some typical instruments on the different parameters of the instrument.

Table 1. The control structure of acoustic instruments

Instrument	Pitch	Amplitude	Timbre	Inharmonicity	Localization
Bell	NC	FE	NC	NC	NC
Guitar	CC	AE	LC	NC	LC
Violin	FC	FC/AE	LC	LC	LC
Trumpet	CC	FC	LC	NC	FC
Flute	DC	FC	LC	NC	FC
Voice	FC	FC	FC	NC	FC

#### 4.1 Ideas about the limitations of acoustic instruments

An acoustic instrument generates sound whose parameters are generally limited to a certain range. There is, for instance, a pitch range, a loudness range and a timbre range. In each parameters range, the parameter can be fixed, follow a slope, an envelope, or other template modalities. These templates guarantees the instruments identity, and enable the musician to concentrate on one, or a few parameters while the other parameters are controlled by the template modality.

#### 4.2 Discussion of a new control structure

In acoustic instruments, the musician can place less emphasis on one or more of the parameters, and still be sure that they obey the expectations, while instead concentrating on another parameter. For instance, when prolonging one note, emphasis might be put on the stable amplitude, leaving the pitch and timbre fixed, but when playing a succession of fast notes, emphasis will be put on the correct pitch. In an electronic instrument, it is important to define the range of each parameter, avoiding values which displaces the identity of the instrument, but allowing a wide enough range so the musician is able to express himself through the instrument. Furthermore, it is necessary to define some typical functions for each parameter, for instance, envelope, fixed value, low frequency oscillating, or a sampled envelope.

When the physical interface, as a musical keyboard, doesn't have enough manipulators, the musician can then remove one parameter from the sensor, leaving it on its trail, and assigning another parameter to the sensor, controlling another aspect of the sound. The musician can thus switch the control from parameter to parameter, by ensuring the correct behavior of the other, non-controlled, parameters, which follows a defined vector.

In most synthesizers, each control output is controlling one unique sound parameter, as illustrated in figure 1.

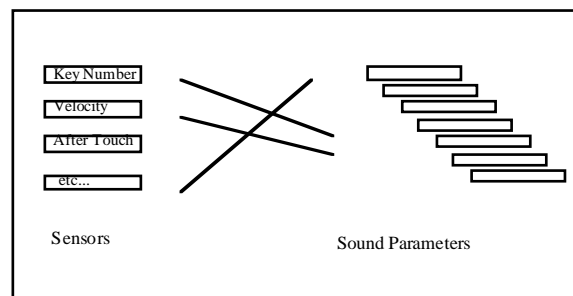


Figure 1. Typical sound synthesis mapping.

But in practice, one, or more control parameters can control any number of sound parameters.

In a dynamic situation as depicted in figure 2, where the control matrix is dynamically changed, what to do with a sound parameter without a sensor? A few rules have been extracted from the observation of the control mechanisms of some typical instruments, and the

**Rule 1:** Make sure the instrument stays recognizable.

**Rule 2:** Make sure the sensor rupture is unnoticed by the listener.

The first rule can be obeyed by using templates of envelope, vibrato, portamento slopes, etc, and pattern matching schemes to recognize the templates. In this way, the closest sound parameter template will be used in case no sensor is connected.

In order to obey the second rule, the following scheme is suggested. Use intelligent system to extract envelope, vibrato and slope, and loop vibrato on the extracted envelope and slope. In this way, the sound parameter is continuing on the trail it initialized when connected to a sensor. This demands, obviously, that a sensor was connected in an earlier stage.

In conclusion, there are generally more parameters to control than independent sensors. A control



mode is proposed, which permits to overrule this restriction.

The control mechanisms of some acoustic instruments have been studied in the previous section, and some typical controls have been shown. The acoustic instruments contain many more control possibilities than any electronic instrument. An approach to this problem is made

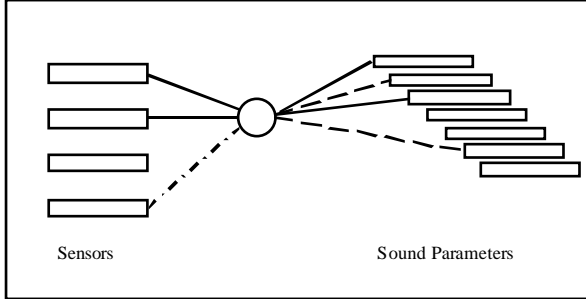


Figure 2. Adaptive sound synthesis mapping.

by proposing a control mode that will permit the musician to control more parameters with fewer sensors. Musicians and performance experiments should be involved in further research with the proposed model.

## 5 Expressive additions to the Timbre Model

The timbre model [25] has proven its value in a number of applications, including the analysis of different expression styles. This analysis will here serve as the basis for the inclusion of a number of parameters, which govern the behavior of the timbre model attributes when used in a synthesis context. The expressions are treated in an analysis/synthesis context, and adapted for real-time synthesis [28], where possible.

The expressive modes introduced to the timbre model include variants, pitch, intensity, vibrato and tremolo, and other expressions, such as legato/staccato.

The attempt to find expressive parameters that can be included in the timbre model is an initial study. Both additional literature studies and field test using the timbre engine [28] must be performed in order to assert the validity of the expression model. The music performance studies [16] gives a lot of information, which can also be gathered from research dealing with acoustics [2], [5] or physics [13] of musical instruments. In addition, appropriate gestures must be associated with the expression parameters [24], [37], [38], [18]. The vibrato problem [10] is an important topic, stating, for instance, whether to add cycles (vibrato) or stretch (glissando) a time-scaled continuous expression parameter. This is not considered a problem in the timbre model, because of the

division into clean envelopes, where only the sustain part is scaled, and periodic and non-periodic irregularities, which are not scaled.

### 5.1 Variants

In this work, the sound of a musical instrument is divided into an identity part, and an expression part, where the identity is the neutral sound, or what is common in all the expression styles of the instrument, and the expression is the change/additions to the identity part introduced by the performer. The expression can be seen as the acoustic outcome of the gesture manipulation of the musical instrument. This division, however, is not simple in an analysis/synthesis situation, since the sound must be played to exist, and it thereby, by definition, always contains an expression part. One attempt at finding the identity is by introducing the notion of variants, which is assumed to be the equivalent of the sounds coming from several executions of the same expressive style.

The variants are calculated by introduced an ideal curve to the different timbre attributes. The deviation from this ideal curve is then assumed to be stochastic, with a given distribution, and for each new execution, a new instance of the deviation is created, giving the sound a clearly altered timbre. Some of the timbre attributes have ideal curves corresponding to some perceptual or physical reality, such as the brightness creation function [25] for the spectral envelope,

$$a_k = a_0 \left( \frac{B}{B-1} \right)^{-k}, \quad (1)$$

where  $a_k$  is the amplitude of the partial  $k$ ,  $a_0$  is the fundamental amplitude, and  $B$  is the estimated brightness [4], see equation (5), and the equation for the ideal stiff string for the frequencies [12],

$$f_k = kf_0 \sqrt{1 + \beta k^2}, \quad (2)$$

where  $\beta$  is the inharmonicity coefficient. Studies of the discrimination of inharmonicity can be found in, for instance, [19] and [32].

Most timbre attributes, however, are fitted with a simple exponential curve,

$$c_k = v_0 \cdot e^{v_1 k}, \quad (3)$$

where  $v_0$  is the fundamental value and  $v_1$  is the exponential coefficient. This curve can model both almost linear curves with small  $v_1$ , but also exponential behaviors.

The parameters of the curves are found by minimizing the lms error using the Levenberg-Marquardt algorithm [31], except for the spectral envelope curve, which is created from the estimated brightness [4],

$$B = \left( \sum_{k=1}^N k a_k \right) / \sum_{k=1}^N a_k \quad (4)$$

The deviations from the ideal timbre attributes parameters are now calculated as,

$$d_k = c_k - \hat{c}_k, \quad (5)$$

where  $c_k$  are the estimated timbre attribute parameters (amplitude, frequency, or other parameter), and  $\hat{c}_k$  are the ideal parameters and  $d_k$  is the deviation (assumed to be white gaussian noise).

The deviation  $d_k$  between the clean exponential curve and the estimated attributes is assumed to be related to the execution of the sound, and the error can, if modeled properly, introduce new executions of the same sound, i.e. of the same instrument, player and style, in the same environment.

Although the clean exponential curves generally fit well with the estimated parameters, there can be discrepancies caused by bad parameter estimation, correlated deviations between attributes, or inadequate modeling. These discrepancies generally do not make artifacts in the original timbre attribute generated sounds, but they sometimes make the variant sounds too different. One way of minimizing this phenomenon is by using weighted curve-fits, which does remove some of the discrepancies. However, since the heavily different variants may be desired, the variants influence is scaled,

$$\tilde{d}_k = \hat{c}_k + v \cdot \hat{d}_k + (1-v) \cdot d_k, \quad (6)$$

where a variant scaling ( $v$ ) of zero gives the original timbre attributes, and a scaling of one gives entirely new timbre attribute deviations ( $\hat{d}_k$ ) for each execution. The total deviations can additionally be weighted [28], permitting more, or less, deviations from the identity of the sound.

## 5.2 Pitch

The modeling of the pitch evolution of the timbre attributes is important when executions for different notes are not available. This is the case, for instance, when creating new, or altered, timbre model parameters sets. Obviously, it's impossible to make pitch rules that encompasses all possible musical instruments, so instead, a simple interpolation scheme has been devised, which assures the proper modification of at least the important spectral envelope. In this scheme, all the timbre attributes are assumed to have the partial frequencies in the x axis, and the timbre attributes for the new pitch is found by interpolating between the neighboring values for each partial frequency. The values outside the original frequencies are found by using the extreme values. Although this scheme is rather simple, it has the double advantage of

handling the most important timbre attribute correctly, and assuring continuous variations in the resulting sound, as the pitch is altered. In addition, the new timbre attribute values should be interpolated after each pitch change, thereby allowing for subtle effects, caused by for instance resonances. When adding vibrato, the sound would have a continuously varying timbre, thereby adding more life to the execution.

The scheme does not handle sound level, however. In the work on generative rules for music performance, [15] suggest a rule, which raises the sound level 3 dB/octave. An initial study have shown that this effect is present in many musical instruments, although ranging from below 3 (violin) to around 3 (piano, clarinet) to 10 (flute), and 15 dB/octave (soprano). This effect is therefore parameterized (N dB/octave) and included into the expressive model.

## 5.3 Intensity

The intensity, or velocity, is another important expression attribute. In general, when increasing the velocity, blowing force, or bow velocity, etc., two things happen, the sound emitted gets louder, and it gets brighter. [6] showed that the increase in amplitude and brightness is asymptotic, i.e. the value changes less as velocity grows. In addition, it was shown that the change of brightness in the piano when changing the velocity of the hammer is governed by a straight line in the Hz/dB domain. Therefore, this is the model chosen for the intensity. The amplitude and spectral tilts (slope of the straight line) have an exponential form,

$$val = (v_M - v_m \cdot e^{-\beta \cdot v}), \quad (7)$$

where  $val$  can be either amplitude or spectral tilt [6], and  $v_M$  and  $v_m$  defines the maximum and minimum values,  $\beta$  the sensibility and  $v$  is the velocity.

The values of  $v_M$ ,  $v_m$  and  $\beta$  can be determined from the instrument, if enough velocity executions are available [6], or it can be user defined. In particular, the upper spectral tilt slope is a useful expression parameter, since it defines how bright the sound becomes, when increasing the velocity to a maximum. This model is also consistent with the analysis of piano executions performed in [25], which showed that the change of velocity (of the piano hammer) only affected the spectral envelope, except for an as yet unexplained change in the shimmer and jitter correlations.

## 5.4 Duration

The duration is of course also a very important expression parameter. Since the timbre model

encompasses both percussive and sustained sounds, a general strategy for modifying the length of the sound is necessary. This strategy could be found by following the clean curve of the sustain/decay part of each partial, with the given curve form,

$$\tilde{a}_k(t) = a_0 + (a_T - a_0) \frac{e^{t/T} - 1}{c - 1}, \quad (8)$$

where  $c$  is the curve form coefficient,  $T$  is the segment duration, and  $a_0$  and  $a_T$  are the start and end values. Unfortunately, the curve form is sometimes fitted to part of the release segment, or sometimes only part of the sustain/decay segment is used, therefore the curve form is error prone. Instead, the decay is assumed to be logarithmic, and modeled as a straight line in the dB domain,

$$\hat{a}^{dB}(t) = \hat{a}_0^{dB} + bt, \quad (9)$$

where  $\hat{a}_0$  and  $b$  are found by fitting to the known split-point amplitude/time values. If the execution is lasting (played a long time), and the second split-point has a higher amplitude than the first one ( $b > 0$ ), then clipping will eventually occur. The perceptual effect of this has been judged to be interesting enough to not prevent this happening. Instead, the amplitude of each partial is limited to a value at a user-defined percentage above the maximum value of the partial. Obviously, if this limit is set to 100%, then no crescendo effect is possible.

[15] tentatively suggests a modification to the attack and decay in durational contrasts. This is an interesting inclusion, but this effect has not been found [22] in relation to this work (these durational contrasts were not part of the sound material under test), and it is not included in the model.

## 5.5 Vibrato and Tremolo

The vibrato and tremolo are important expression parameters with specific values defined by the musical context in which the execution is produced. Therefore, a generative rule for the addition of vibrato and tremolo is not desirable in this work. Some vibrato or tremolo effects are, however, part of the identity of the sound, and these effects should not be user-controlled, but inherent in the sound. In particular, this is the case for the amplitude modulation caused by the beating of different modes or strings in, for instance, the piano [39].

The vibrato and tremolo are generally defined by three parameters, the rate (speed), the strength and the initial delay. [16] reviews several vibrato studies, and reports the vibrato rate of professional singers to be between 5.5 to 8 Hz, with a strength between one-tenth of a full tone to one full tone, averaging 0.5 to 0.6 of a whole-tone step.

[33] models the vibrato with the sum of a number of sinusoids with time-varying amplitudes and phases. The phase of the vibrato/tremolo is necessary, if the resulting sound should be perceptually identical to the original one. Care must be taken to model the resonances correctly when adding vibrato [29]. In addition, the perceived pitch of tones with vibrato is also an important research field [8].

In order to assert whether a sound contains a vibrato or tremolo expression, or whether it contains periodic vibrations in its identity, two things can be examined. First, if the partials are heavily correlated, secondly, if the rate and strength values are correlated, then the chances of it containing expressive vibrato/tremolo is great. If neither of the two cases occur, periodicity is assumed to be part of the identity of the sound, and not controlled by the performer. If expression periodicity is found, it is removed from the sound, and only added back, if and when the performer is signifying it.

## 5.6 Other expressions

The expressions can be any kind of acoustic change resulting from manipulation of the music instrument.

The other expression parameters used in classical music include mainly styles (legato/staccato, for instance) and tempi. Since some of these expressions are controlled continuously by the performer, they are not easily integrated into the timbre model. In particular, no good gesture capture device has been available to perform tests. In addition, not much timbre attribute changes have been found when analyzing executions of different styles [25]. Therefore, the conclusion must be that the styles are mainly a matter of duration, which is easily controlled in this model.

Another important expression is the transition [35]. Since the transition is the time-varying amplitude, fundamental frequency, and timbre, it should not be too difficult to create timbre attribute sets with appropriate values for different transition.

Finally, another possible expression is the generic timbre navigation [40], [34]. In this case, the timbre is manipulated using sensors and various mapping strategies [24], [38].

## 6 Conclusion

The paper presents the study of the control of musical instruments, which outlines some basic control compartments of typical musical instruments and the resulting general control structure, including ideas about how to overcome the limited sensors problem in digital musical instruments.

In addition, the expression model, which enhances the use of a sound model in a musical context, is presented. The expression model models the behavior of the timbre attributes, when playing a different note, intensity, duration, or other style. It also introduces the vibrato and tremolo into the model. A major contribution is the notion of variants, in which the timbre is divided into an identity part, and an expression part. The variant is then a stochastic model of the deviation from the identity part of the sound, with a new instance at each new execution. Therefore, an interesting novelty is added to each new sound emitted, even if it's played with the same pitch, loudness and duration.

## References

- [1] American Standard Association. *Acoustical Terminology*, New York, 1960.
- [2] Backus, J. *The acoustical foundation of music*. John Murray Ltd. London, 1970.
- [3] Barrière, J-P (editor). *Le timbre, métaphore pour la composition*, C. Bourgois Editeur, IRCAM, 1991.
- [4] Beauchamp, J. *Synthesis by spectral amplitude and "Brightness" matching of analyzed musical instrument tones*. J. Acoust. Eng. Soc., Vol. 30, No. 6. 1982.
- [5] Benade, A. H. *Fundamentals of musical acoustics*. Dover publications inc. New York. 1990.
- [6] Bensa J., K. Jensen, R. Kronland-Martinet, S. Ystad. *Perceptual and Analytical Analysis of the effect of the Hammer Impact on the Piano Tones*, Proceedings of the ICMC, Berlin, Germany. 2000.
- [7] Bregman, A. S. *Auditory Scene Analysis*, The MIT Press, Massachusetts. 1990.
- [8] d'Alessandro, C., M. Castellengo. *The pitch of short-duration vibrato tones*, J. Acoust. Soc. Am, 95(3), pp. 1617-1630, March. 1994.
- [9] Desain, P. & Honing, H. On continuous Musical Control of Discrete Musical Objects. ICMC Proceedings 1993.
- [10] Desain, P., H. Honing. *Time functions function best as functions of multiple times*. Computer Music Journal, 16(2), 17-34, 1992.
- [11] Fitz, K., L. Haken, P. Christensen. *Transient Preservation Under Transformation In an Additive Sound*
- [12] Fletcher, H. *Normal vibrating modes of a stiff piano string*, J. Acoust. Soc. Am., Vol. 36, No. 1, 1964.
- [13] Fletcher, N. H., T. D. Rossing. *The physics of musical instruments*, Springer-Verlag. 1990.
- [14] Freedman, M. D. *Analysis of musical instrument tones*. J. Acoust. Soc. Am. Vol. 41, No. 4, 1967.
- [15] Friberg, A. *Generative rules for music performance: A formal description of a rule system*. Computer Music Journal, Vol. 15, No. 2, summer 1991.
- [16] Gabrielson, A. *The performance of music, in The psychology of music*, D. Deutsch (editor), pp. 501-602, AP press, San Diego, USA, 2<sup>nd</sup> edition, 1999.
- [17] Helmholtz, H. *On the Sensations of Tone*, reprinted in 1954 by Dover, New York, 1885.
- [18] Hunt A., M. Wanderley. *Interactive Systems and Instrument Design in Music Working Group*, <<http://www.notam.uio.no/icma/interactive/systems/wg.html>>, October 4, 2001.
- [19] Järvinen, H., V. Välimäki, M. Karjalainen. *Audibility of inharmonicity in string instrument sounds, and implications to digital sound synthesis*. ICMC Proceedings, Beijing, China, 359-362. 1999.
- [20] Jensen, K., G. Marentakis, *Hybrid Perception*, Papers from the 1<sup>st</sup> Seminar on Auditory Models, Lyngby, Denmark, 2001.
- [21] Jensen, K. 1996. *The Control Mechanism of the Violin*. Nam-96 Proceedings.
- [22] Jensen, K. *Envelope Model of Isolated Musical Sounds*, Proceedings of the DAFX, Trondheim, Norway, 1999.
- [23] Jensen, K. *Pitch Independent Prototyping of Musical Sounds*, Proceedings of the IEEE MMSP Denmark, 1999.
- [24] Jensen, K. *The Control of Musical Instruments*, Proceedings of the NAM. Helsinki, Finland. 1996.
- [25] Jensen, K. *Timbre Models of Musical Sounds*, PhD. Dissertation, DIKU Report 99/7, 1999.
- [26] Jensen, K., *The Timbre model*, Workshop on current research directions in computer music, Barcelona, Spain, 2001.
- [27] Marchand, S. *Musical sound effects in the sas model*, Proceedings of the 2nd COST G-6

- Workshop on Digital Audio Effects (DAFx99), NTNU, Trondheim, December 9-11, 1999
- [28] Marentakis, G., K. Jensen. *Timbre Engine: Progress Report*, Workshop on current research directions in computer music, Barcelona, Spain, 2001.
- [29] Mellody, M., G. H. Wakefield. A model distribution study of the violin vibrato. Proc ICMC, 1997.
- [30] Moore, F.R. The Dysfunction of MIDI. CMJ 12(1), 1988.
- [31] Moré, J. J. *The Levenberg-Marquardt algorithm: Implementation and theory*. Lecture notes in mathematics, Edited by G. A. Watson, Springer-Verlag, 1977.
- [32] Rocchesso, D., F. Scalcon. *Bandwidth of perceived inharmonicity for physical modeling of dispersive strings*. *IEEE Transactions on Speech and Audio Processing*, 7(5):597-601, September 1999.
- [33] Rossignal, S. *Segmentation et indexation des signaux sonores musicaux*, Thèse de doctorat de l'université Paris VI, Paris, France 2001.
- [34] Rován, J. B., M. M. Wanderley, S. Dubnov, Ph. Depalle. *Instrumental Gestural Mapping Strategies as Expressivity Determinants in Computer Music Performance*, Kansei- The Technology of Emotion Workshop - Genova - Italia, Oct. 3/4, 1997.
- [35] Strawn, J. *Orchestral Instruments: Analysis of performed transitions*, J. Audio. Eng. Soc., Vol. 34, No. 11, pp. 867-880. November 1986.
- [36] The ZIPI Music Interface Language, CMJ 18(4), 1994.
- [37] Wanderley M., M. Battier (eds). *Trends in Gestural Control of Music*. Ircam - Centre Pompidou, 2000.
- [38] Wanderley, M. *Interaction musician-instrument: Application au contrôle gestuel de la synthèse sonore*. Thèse de doctorat de l'université Paris VI, Paris, France 2001.
- [39] Weinreich, G. *Coupled piano strings*, J. Acoust. Proc. Am., Vol 62, No. 6, pp. 1474-1484, December 1977.
- [40] Wessel, D. *Timbre space as a musical control structure*, Computer Music Journal 3(29), pp.45-52, 1979.

*Paper submitted for publication to "Organised Sound" (Cambridge Univ. Press)*

## **From Sounds to Music: Different Approaches to Event Piloted Instruments**

Pascal Gobin<sup>(1)(2)</sup>, Richard Kronland-Martinet<sup>(3)</sup>, Guy-André Lagesse<sup>(2)</sup>,  
Thierry Voinier<sup>(3)</sup>, Sølvi Ystad<sup>(4)</sup>

(1) Conservatoire National de Région  
2, Place Carli F-13001 Marseille France [pgobin@wanadoo.fr](mailto:pgobin@wanadoo.fr)

(2) Association Les Pas Perdus  
10, Rue Sainte Victorine F-13003 Marseille France [lespasperdus@wanadoo.fr](mailto:lespasperdus@wanadoo.fr)

(3) C.N.R.S. (Centre National de la Recherche Scientifique)  
Laboratoire de Mécanique et d'Acoustique  
31, Chemin Joseph Aiguier F-13402 Marseille France [{kronland,voiniier}@lma.cnrs-mrs.fr](mailto:{kronland,voiniier}@lma.cnrs-mrs.fr)

(4) N.T.N.U. (Norges Teknisk-Naturvitenskapelige Universitet)  
Department of Telecommunications N-7034 Trondheim Norway

### **Abstract**

This paper addresses three different strategies to map real-time synthesis of sounds and controller events. The design of the corresponding interfaces takes into account both the artistic goals and the expressive capabilities of these new instruments. Common to all these cases, as to traditional instruments, is the fact that their specificity influence the music which is written for them. This means that the composition already starts with the construction of the interface. As a first approach, synthesis models are piloted by completely new interfaces, leading to "sound sculpting machines". An example of sound transformations using the Radio Baton illustrates this concept. The second approach consists in making interfaces that are adapted to the gestures already acquired by the performers. Two examples are treated in this case: the extension of a traditional instrument and the design of interfaces for disabled performers. The third approach uses external events such as natural phenomena to influence a synthesis model. The *Cosmophone*, which associates sound events to the flux of cosmic rays, illustrates this concept

### **1 Introduction**

Nowadays computers can make sounds and musical sequences evolve in real-time. Musical interpretation can thus make sense when disposing possibilities of expression adapted to such new digital instruments. Unlike with traditional instruments, an interface based on a digital device is not restrained by the mechanics of the instrument. These new interfaces naturally prepare for new gestures aiming at exploiting, as efficiently as possible, the possibilities offered by the computer. Nevertheless, these huge possibilities of sound piloting can sometimes slow down the creative processes if they are not clearly thought through. Even worse, these new technologies might make the technician more valuable than the composer if the technology is used to impress the audience ignoring the importance of creativity and musical intention.

This can be the case if the interfaces are made without reflections about the musical context in which it is to be used. In our opinion, the design of an interface is already part of the creative and the compositional processes.

In this paper we present four examples of the design of sound interfaces based on three different strategies: the piloting of synthesis parameters to perform intimate transformations on the sounds, the adaptation of interfaces to specific gestures, and the construction of

between scientists and musicians. It has mainly involved the group *S2M (Synthèse Sonore et Modélisation)* of the *Laboratoire de Mécanique et d'Acoustique*, the *Conservatoire National de Région* and the association *Les Pas Perdus* all of which are located in Marseille (France).

The first example is an attempt to control sound synthesis models by giving the musician a set of new tools to sculpt sounds. These new sound possibilities naturally ask for new interfaces. We found that the Radio Baton, initially designed by Max Mathews to conduct a musical sequence, can be used with great effect to pilot synthesis model parameters. Here the aim was to give the musician an intuitive and easy-to-use tool for improvised sound transformations.

The second application points at a quite different issue. Here the problem is to give the musician the possibility of expanding his or her own instrument. Actually, even though a large number of new interfaces have been made, the most common digital instruments use keyboard interfaces. Even though these interfaces offer a lot of possibilities, the musical interpretation can not be the same as the interpretation obtained when playing a wind instrument or a string instrument, since the instrumental play is closely related to the physics of the instrument. This means that for example the linear structure of a keyboard is not easily adapted to the playing of a trumpet, and that the information given by a keyboard is poor compared to the possibilities that a flute player has when playing a sound. Several MIDI controllers with structures close to traditional instruments like wind or string instruments (Pousset 1992, Machover, 1992) have been proposed. In most cases these controllers are quite limited, since the sensors generally don't give access to fine playing information (for example the lack of reed vibration information in the Yamaha WX7) and are not dedicated to a given synthesis model allowing for example natural sound transformations. Here we show how an interface based on a traditional flute equipped with sensors and a microphone would give the performer access to synthesized sounds through flute-like synthesis models, without modifying the traditional playing techniques obtained through years of practicing.

Another example of interfaces suited to special gestures addresses the design of controllers adapted to a set of unconventional motions. This is an artistic project, which has been going on for about five years, with the participation of four people with limited gestural faculties. The main features of this work lie in the fact that the technological realization is closely linked to the artistic problems and the heavy conditions of the handicap itself. In particular, rather than setting ourselves a target leading to a conventional situation (e.g. 'playing music'), and of imagining with the handicapped people the necessary technological methods with this target in view, we preferred to start from scratch and to adapt completely new tools for them.

As we already mentioned, the last group of interfaces connects sound models to natural phenomena. This is a degenerate case where the "motion" can not even be learnt by the "player" but has to be taken as it is. The last example, which illustrates this approach, has been made possible thanks to collaboration with Claude Vallée and David Calvet of the *Centre de Physique des Particules de Marseille*. In this case a device, which we have called the *Cosmophone* has been designed to make the flux and properties of cosmic rays directly perceptible within a three dimensional space. This is done by coupling a set of elementary particle detectors to an array of loudspeakers by a real time data acquisition system and a real time sound synthesis system. Even though the aim of the original installation was to make people aware of a physical phenomenon in an entertaining way, such a system can also be used in a musical context.

## **2 Sculpting the sounds using the Radio Baton**

The conception of sensors detecting movements is an important part of Max Mathews' current research work. One of his inventions, the Sequential Drum, consists of a surface equipped with sensors. When an impact is detected on this surface, the apparatus indicates the intensity and the position of the impact on the surface (Mathews and Abbot, 1980). A collaboration with Boie (Boie et al. 1989) led to the design of the Radio Drum which detects

Drum is in fact able to continuously detect the position of the extremities of the two drumsticks (emitters).

These first prototypes were connected to a computer containing an acquisition card piloted by a conductor program allowing a real-time control of the execution tempo of a partition already memorized by the computer (Mathews 1991a, 1991b 1997). The Radio Drum is also called the Radio Baton, since the drumsticks launching musical sequences can be used in the same way as a baton used by a conductor of an orchestra. Max Mathews designed the Radio Drum for two reasons:

- to make it possible to actively listen to a musical play by releasing musical sequences and thus give a personal interpretation of a musical work.
- to make it possible for a singer to control his or her own accompaniment.

The Radio Drum comprises two sticks (batons) and a receiver (a 55 x 45 x 7 cm parallelepiped) containing the electronic parts. In a simplified way, each drumstick can be considered as an antenna emitting radio frequency waves. A network with five receiving antennas is placed inside the drum. It measures the coordinates of the placement of the extremities of the drumsticks where the emitters are placed.

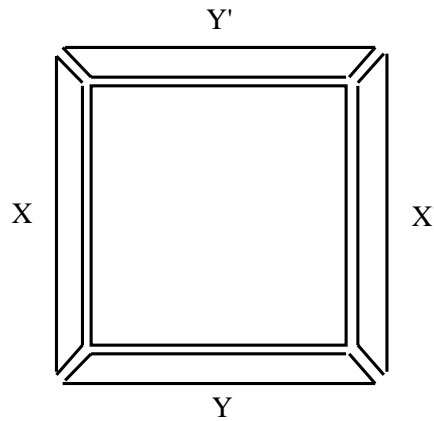


Figure 2.1

The intensity of the signal received by one of the antennas depends on its distance from the emitter: it gets stronger as the emitter gets closer to the antenna. To evaluate the position of the extremity of the baton as a function of the x-axis, it is sufficient to calculate the difference in intensity between the antennas x and x' (Figure 2.1). In the same way the antennas y and y' give the position of the emitter as a function of the y-axis. In order to get information about the height of the baton (the z-coordinate), it is sufficient to add the intensities received by the five antennas. Since each baton emits signals with different frequencies, it is relatively easy to discriminate between the signals sent by the two batons to each antenna.

The relation between the coordinates and the intensity of the received signals is not linear. A processing of these data is necessary so that the system gives information proportional to the coordinates of the batons. This operation is realized by a microprocessor making it possible to add other functions to the instrument. The latest version contains additional software either making it possible to transmit the coordinates of the batons (and information about the controllers) when requested from a computer, or to transmit the coordinates of one of the batons when it cuts a virtual plane parallel to the surface of the drum the same way as when one hits its surface. When the program detects the virtual plane, it calculates and transmits the velocity of the movement as a function of the z-axis. The height of the plane can of course be modified. This leads us back to the functioning of the Sequential Drum.

These working modes (transmission of the data when there is an external request or strike



microprocessor of the Radio Baton makes the implementation of this communication possible by the use of the MIDI protocol and of a program like MAX (Puckette and Zicarelli 1990) to communicate with the Radio Baton. Thus the control possibilities with instruments that can be programmed are almost unlimited.

Numerous musical applications of the Radio Baton have already been published (see e.g. (Boie et al., 1989), (Boulanger and Mathews, 1997), (Gershenfeld and Paradiso, 1997), (Jaffe and Schloss, 1994)) and most of them mainly exploit the drum-like structure of the device which is well adapted to the launching of sound events and to the conduction of their time evolution. We shall briefly describe how the Radio Baton can be diverted from its original aim and used to perform sound transformations with additive and physical synthesis techniques.

A presentation of the Radio Baton piloting sound synthesis models was given by our research group at an international colloquium on new expressions related to music organized by GMEM (*Groupe de Musique Expérimentale de Marseille*) in Marseille (Kronland-Martinet et al. 1999; Ystad, 1999). Intimate sound transformations using the Radio Baton were demonstrated. The instrument has also been used in a musical context with the goal of performing improvised sound transformations.

### **2.1 Radio Baton and additive resynthesis of sounds.**

The amplitude modulation laws of the sound components together with their frequency modulation laws give the parameters defining an additive synthesis model.

When connected to analysis techniques, additive synthesis methods give resynthesized sounds of high quality (Kronland-Martinet et al. 1997). However these models are difficult to manipulate because of the high number of parameters that intervenes. Mapping the Radio Baton to such a sound synthesis technique is closely linked to the compositional process since it limits the possibilities of sound transformations. We tried several mapping strategies, aiming at intimately control the sound by altering different parameters of the synthesis process. We eventually concluded by the fact that this particular interface is hard to seriously use for other purposes than the ones originally planned by the designer. Actually, the precision of the mapping cannot allow piloting subtle parameters. Nevertheless, the Radio Baton can be used to globally improvise with a sound transformer process. We demonstrated this fact by musically using the interface in a way rather close to a traditional instrument. For that, three parameters are used for manipulating the sound using the Radio Baton, namely the duration of the note, its fundamental frequency and its amplitude. In each case the manipulations can be done independently for each modulation law or globally on all the modulation laws. In our case the duration of the sound and the frequency manipulations are done in a global way. This corresponds to a simple acceleration or slowing down of a note when the duration is altered, and to a simple transposition when the frequency is altered. The amplitude modulation laws have been modified differently, giving the possibility of effectuating a filtering or equalization on the sound. In figure 2.2, the control possibilities of the Radio Baton are illustrated. The sound is generated when one of the sticks cuts a virtual plane the height of which is predefined by the user. The x-coordinate is related to the duration of the generated sound and the y-coordinate to the transposition factor. The second stick is used to control the note after it has been triggered (aftertouch) and uses the y coordinate to act on the frequency transposition (like for the first baton) and the z-coordinate to fix the slope  $\alpha$  of a straight equalization line. This slope is positive when the baton is over a predefined plane (0 point in figure 2.2) corresponding to an attenuation of low-frequency components and thus to a high-pass filtering. This is illustrated in figure 2.3. When the baton is below the zero point, the slope is negative, corresponding to a low-pass filtering. This allows a continuous modification of the brightness of the sound.

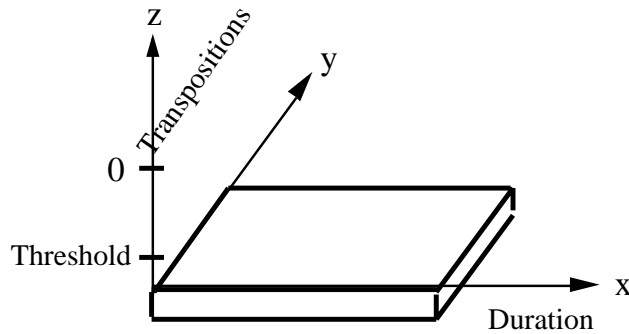


Figure 2.2

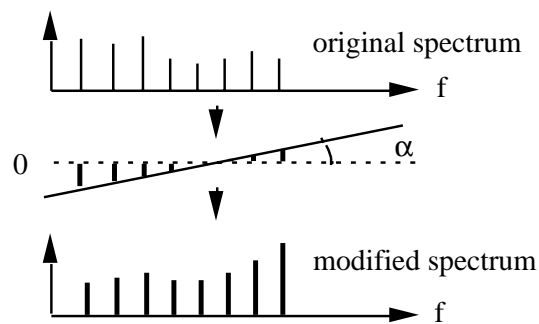


Figure 2.3

The second stick could have controlled a third parameter corresponding to the x-coordinate, but since playing the Radio Baton was difficult, with the possibilities already described, this parameter was not used. Even though the possibilities offered by such instruments are almost infinite, it is important to point out a crucial issue: the complexity of the playing. Actually the lack of absolute reference in the 3D space makes this kind of approach preferably suitable for improvisation purposes, without expecting to exactly reproduce a sequence. This assumption will, of course, be denied when a musician accepts spending thousand of hours learning this new instrument, despite the lack of already written music for it.

## 2.2 Radio Baton and physical modeling of sounds

In a way similar to the description in the previous section, we have used the Radio Baton to pilot a physical synthesis model. This application tends to prove that physical models are well adapted to sound transformations by providing a small amount of meaningful parameters making the mapping easier. Physical models describe the sound generation system using physical considerations. An interesting method consists in simulating the way the acoustical waves propagate in the instrument by using a looped system with a delay line and a linear filter (figure 2.4). Such a model is called the digital waveguide and it has been widely used for synthesis purposes (Smith 1992). The delay is proportional to the length of the resonator and thus to the fundamental frequency of the note played, while the filter takes into account the dissipation and the dispersion phenomena. Analysis-synthesis techniques adapted to the digital waveguide can be designed, allowing the resynthesis and the manipulation of a given natural sound (Kronland-Martinet et al. 1997; Ystad 2000). When the Radio Baton acts on a digital waveguide model the parameters to be modified have a physical significance.

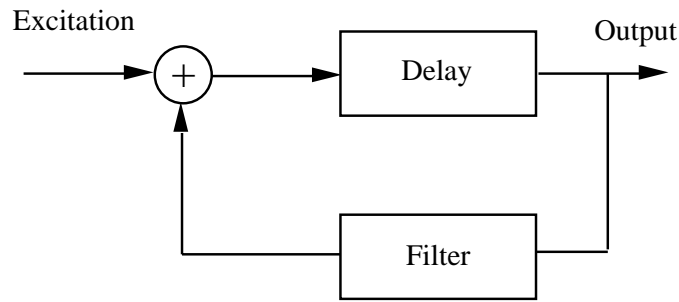


Figure 2.4

The parameters that are controlled by the sticks in this example only correspond to the delay and to the excitation of the model (the source), the filter being chosen once and for all. One of the batons acts on the choice of the excitation: each of the 4 corners of the receiver corresponds to a different excitation as shown in figure 2.5. In the middle of the plate, the source corresponds to a mixture of the four sources with weights depending on the distance from each corner. The second baton acts on the delay of the resonator (thus on the frequency of the note played) given by the y-coordinate. In addition it acts on the frequency of the excitation signal when a saw tooth source is used.

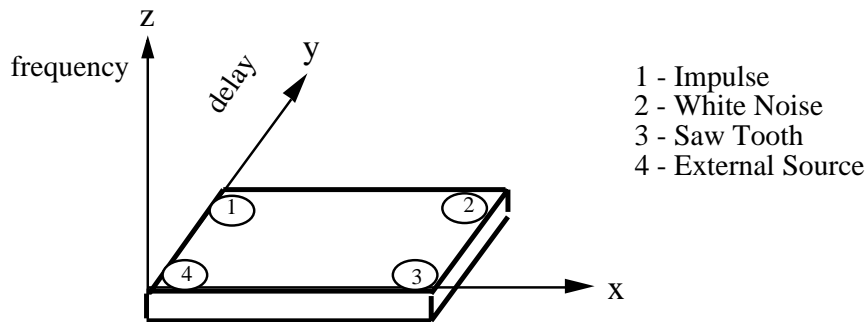


Figure 2.5

The Radio Baton is an interesting instrument offering numerous playing possibilities and we enjoyed the numerous brainstorming sessions we have had while designing some of them. Once again, the instrument has been used in musical contexts, but the close relationship between the synthesis process and the physics of real instruments made it frustrating for some musicians. Like any new instruments, playing the Radio Baton has to be learned, and although the instrument offers a great number of possibilities, it is difficult to pilot certain playing effects like for instance the vibrato played on a wind instrument. On the other hand, the Radio Baton is well adapted to control percussive sound models (Schloss, 1990), (Boulanger and Mathews, 1997). It is also well adapted to the construction of new sounds not directly related to real instruments, and we eventually realized that it was best suited to generate sounds produced by signal synthesis models such as, for example, the FM synthesis (Kronland-Martinet et al., 1999).

### 3 Extending the possibilities of a traditional instrument: the flute case

Even though most digital instruments do not sufficiently take into account the personal touch a musician wants to give when playing, musicians are interested in sound effects and transformations. Many composers dream of adding new possibilities to existing instruments. Playing non-realizable instruments like a gigantic flute or blowing into a string could be

interface, which is presented in this section is made using a traditional flute connected to a computer by magnetic sensors detecting the finger position and a microphone at the embouchure level detecting the pressure variations (Ystad and Voinier 2001). An earlier attempt to extend the possibilities of a traditional flute was made at IRCAM (Pousset, 1992). However, this instrument was mainly made to act as a MIDI controller. Other attempts in designing meta-instruments have been made and the reader can for example refer to (Bromwich, 1997), (Cadoz et al., 1990), (Jensen, 1996), (Pierrot and Terrier, 1997).

The interface we developed has been designed to pilot a so-called hybrid synthesis model, which has been made to resynthesize and transform flute sounds (Ystad 2000), meaning that the mapping and the musical perspective were already set at the design stage. The resonator of the instrument is modeled with a physical model simulating the propagation of the acoustical waves in a tube (digital waveguide model), and the source is modeled using a non-linear signal synthesis model. The reason why a signal model has been used in this case is related to the fact that the physical phenomena observed at the embouchure of a flute are not fully understood. Even though some models describing the interaction between the air jet and the labium have been proposed (Verge 1995), most of the physical parameters intervening are difficult to measure, and the resolution of the equations is generally not compatible with real-time implementations.

The hybrid synthesis model makes it possible to “imitate” the traditional instrument and in addition make very subtle modifications on the sound by changing the model’s parameters. Hereby one can obtain physically meaningful sound transformations by acting on different parts of the model to simulate effects like exaggerated vibratos or change the properties of the source and/or the resonator. It is important to underline that the aim of such an interface is not to replace the traditional instrument, but to expand its possibilities. This augmented instrument will give musicians the possibility of making use of already acquired playing techniques. The general problem, when proposing new instruments to musicians, is that they often are afraid of investing a lot of time to learn to play an instrument that may not be used in the future or for which no music is written. In this sense, this approach will hopefully be attractive to musicians.

The flute has been equipped with Hall effect sensors detecting the distance to magnets connected to each keypad. The sensors have been placed on an aluminum rail fastened to the flute, while the magnets are fastened to the support rods where the keypads are fixed so that they approach the sensors when the keypads are closed as illustrated in Figure 3.1 and Figure 3.2. The state of the keypads is related to the frequency of the note played. The speed at which the keypad is closed can be calculated by continuously measuring the distance between the sensor and the magnet. All these parameters make it possible to calculate the fundamental frequency, which is correlated to the delay line length of the resonator model.

The fundamental frequency is also used at the input of the sine generator feeding the non-linear function of the source model.

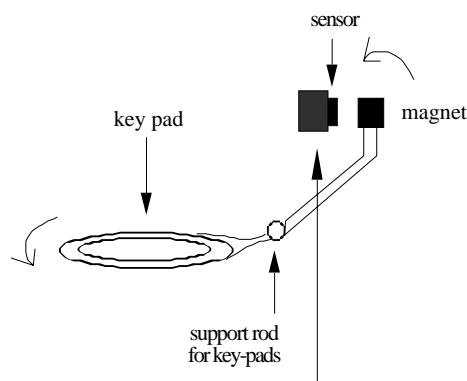


Figure 3.1

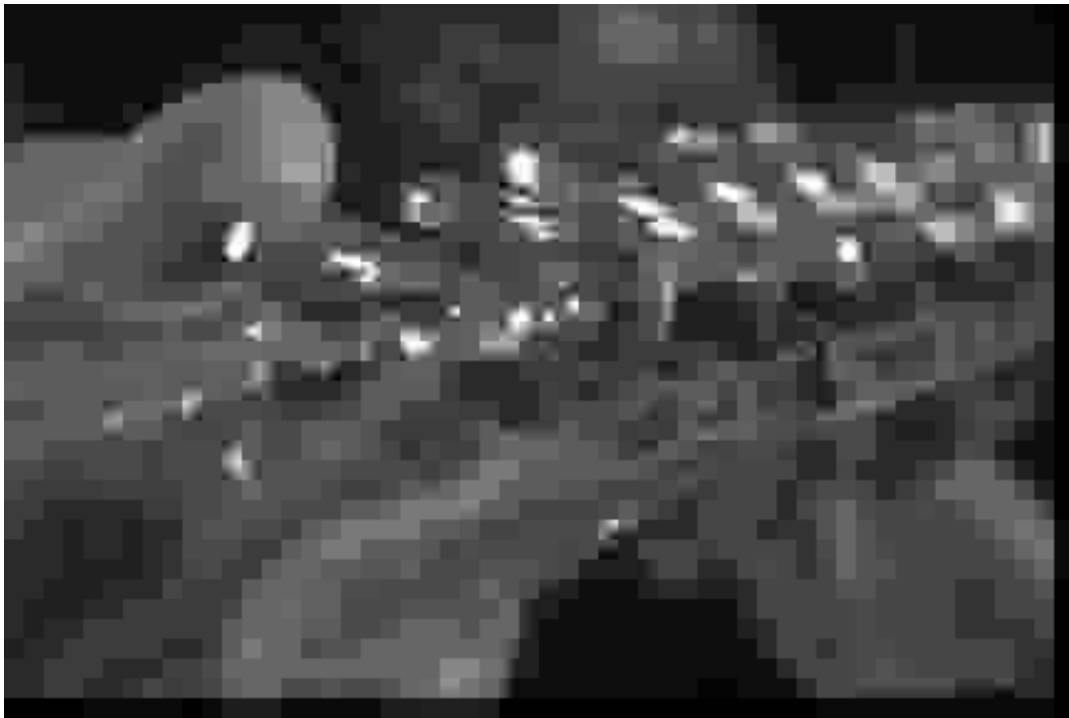
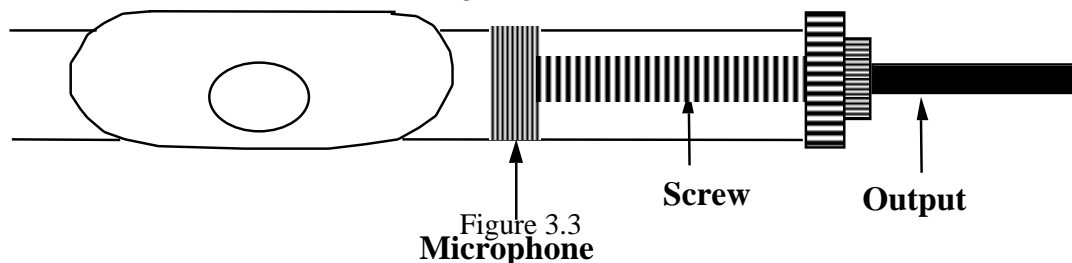


Figure 3.2

The pressure variations are detected by a microphone situated inside the flute at the cork position near the embouchure as seen in figure 3.3.



This system does not represent a complete way of characterizing the play, since for instance the angle at which the air jet hits the labium, or the position of the player's lips, are not taken into account. However, these important features influence the internal pressure which is measured by the microphone and which acts on the generated sound. The detection of the pressure variations makes it possible to estimate the vibrato, which will be added to the frequency information and used at the input of the sine generator. The logarithm of the pressure envelope will further be used as amplitude at the input of the sine generator. It will also be used to determine the level of the non-deterministic source signal that will be added to the deterministic source signal before the input of the resonator (figure 3.4). In some cases, when the state of the keypads does not change when the same note is played in different registers, the measurements of the pressure level at the embouchure allows the determination of the frequency of the note played.

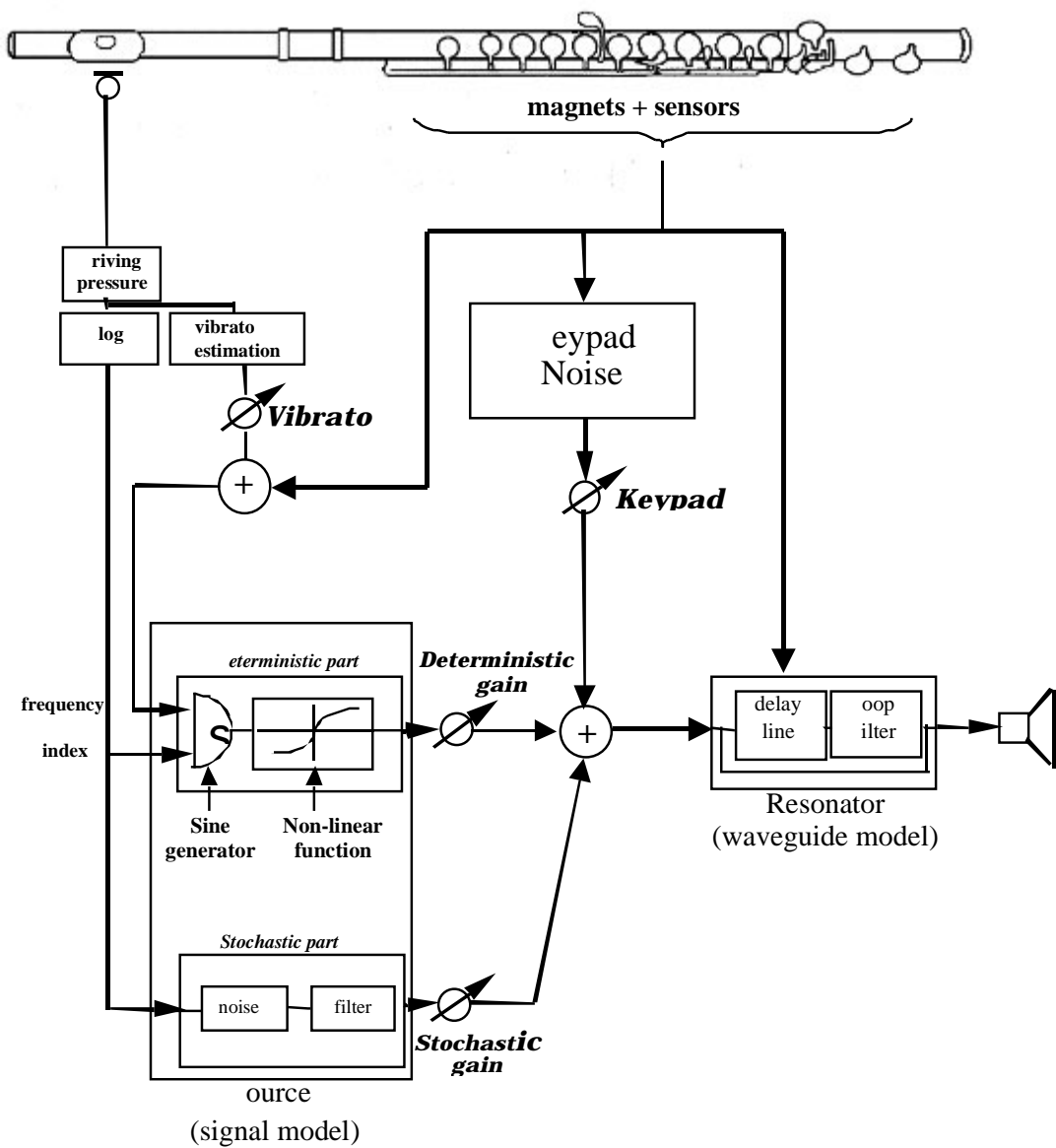


Figure 3.4

Although the synthesis model is the most important part of the flute interface, it should be mentioned that it also is MIDI compatible. Actually, the data from the sensors are processed to generate MIDI codes, which are sent to the real-time processor.

The MIDI flute interface and the synthesis model are both implemented using the MAX/MSP development environment (Puckette et al. 1990).

Since the flute interface generates MIDI codes to control the synthesis model, one can use these codes to pilot any kind of MIDI equipped instrument. In the same way, since the real-time processor is also considered as a MIDI instrument, assigning arbitrary frequency values for a given key state can change the tune of the flute. At this point, the musical possibilities are only limited by the imagination.

The parameters of the synthesis model can dramatically act on the sound itself. Since the digital flute model has been designed to respond to MIDI codes, one can act on the parameters of the model using MIDI controllers such as pedals, sliders, ... (Ystad and Voinier

of the flute while replacing the tube model by a string model, or inversely, by injecting a voice signal or a source of another instrument into the resonator model of the flute. The last alternative would also allow the generation of the noises made by the flautist while playing. Actions on the non-linear source model make it possible to modify the spectral content of the sound so that clarinet-like sounds or brass effects can be obtained. The model we implemented also comprises a keypad noise generator, which can be altered by modifying the corresponding table. This keypad noise could for instance be replaced by any percussive sound. Finally, by acting on the characteristics of the physical model simulating the tube of the instrument, physically unrealizable instruments like a gigantic flute can be simulated. All these manipulations show the advantage of a hybrid sound model associated to a meta-instrument, which enables the modification of a natural sound in the same way as synthetic sounds, while keeping the spirit of the traditional instrument.

#### **4 From gesture to sound: adapting technology to disabled people**

*Un Bon Moment* is an artistic project, which has been going on for about five years, with the participation of four people with limited gestural faculties. The main features of this work in continual development lies in the fact that the technological realization is closely linked to the artistic problems and the heavy conditions of the handicap itself. We did not wish to give ourselves a particular goal as to the final form, but by working patiently on the development and awareness, create visual, sound and technological methods that would shape the rather indefinable realization (neither show nor installation) that we have called *Un Bon Moment, A Sound and Vision Walk About*. For the Australian Aboriginals, the notion of *A Walk About* denotes a short informal holiday period, far from work, when they can wander through the bush, visit relations, or go back to native life.

We have worked on the principle that it is the framework or the *a priori* concerning the results, which create the handicap rather than the person himself (if one considers him as such). For example, it is obvious that traditional musical instruments are all made with reference to specific gestural possibilities (movements and space between the arms, hands and fingers, the speed and precision of these movements etc...). Even if the apprenticeship is long, an able-bodied person is able to play any traditional instrument. This is not the case for "someone with a handicap", who will neither necessarily adapt the characteristics of his body (size, space, force and precision...) to the instruments, nor in the music that has been composed with or for these instruments. Therefore we no longer consider handicap as disabling, but rather as new particular gestural possibilities in the creation of sound laying out, and, more generally, as particular conditions in the elaboration of the artistic side of things. It means that the purpose of this project goes far beyond the question of "giving handicapped people access" to music (which in the long run relies on the instrumental capacities of an able-bodied person), but rather of being aware of musical models linked to limited gestural possibilities and it tackles the question of musicality which does not depend on virtuosity.

##### **4.1 The question of time in *Un Bon Moment***

One of the initial observations made during this project, concerned a completely unusual relationship with time. It takes more time for handicapped than for able-bodied people to communicate (linked notably to difficulties in articulating), and therefore it takes more time to work together. The handicapped need some time between an idea and its' expression, or a decision (to accomplish a gesture for example) and the moment when the decision really takes shape.

Thus it prevents us from synchronizing the sound events in a fixed, predetermined order, and therefore from planning a musical score in its usual way. Finally, the notion of time is felt differently according to each person. It is related to the interior rhythm, underlined by music through beat, figure, gesture and movement. One can imagine that the notion of time for

at a time in an extreme state of tension, is very different to that for an able-bodied person. Thus it seemed essential to us to not only take into account these particular relationships with time, but also to bring them to the foreground.

- when creating the instruments : each instrument was made with the previous remarks in view, with the aim of enabling the musician to physically get involved ( taking into consideration the gestural particularities, movement, energy, muscular tension etc.) in the invention of sound processes, rather than simply triggering a set of sound events.

- In realizing the performances: we have worked to create situations and potential rather than the sound result itself (sound, visual, gestural possibilities for each musician, and possibilities of meetings and play for several persons).

We have also aimed at offering the audience the idea of a *walk about*, where there was no longer a question of coherence with the idea of shape, but rather of an awareness of events that were about to happen in a space of time that belonged to the spectator himself, as he could go in and out of the performance area as he pleased.

#### **4.2 The electronic instrument**

The notion of electronic instrument, which is presented here, is different from the one described for a specific purpose in section 3. The particularity of the electronic musical instrument lies in the absence of an acoustic mechanism system, which closely links a certain gesture to a certain sound event (the instrumental tone). A musician has to learn to play the instrument, by mastering a gestural repertoire that enables him to produce sounds with the instrument and to tackle its musical range. The problem of the electronic instrument depends to a certain extent on how much the different stages of monitoring of the sound phenomena are left in the hands of the user/creator (gestural abilities, sound synthesis device). Between gesture and sound, one does not necessarily manipulate masses in movement (hammer, air column, string...) but rather digital data. This concept has the effect of separating gesture and produced sound in their immediate (conventional) relationship, and we have to invent new relationships between gesture, produced sound and musical creation. Consequently:

- it allows new musical gestures outside the realism of conventional musical attitudes.  
- it allows the creation of other forms of musical script, linked to imprecision and approximation.

- it shows the need of a reflection on musical choices simultaneously with the creation of instrumental devices. In a certain way it shows that through the concept of an instrument, one has already started to compose the music with it.

- elsewhere, beyond the opposition between acousmatics and live instrumental music the playing of an electronic musical instrument have induced us to put forward a more fundamental aspect - more innovative – which lies in the fact that the composer-musician (and/or improviser) can intervene directly on the sounds (sound phenomena), precisely at the moment where they develop, last, and establish themselves. Therefore she or he can construct and develop an idea, a musical form, on the basis of the sensation of duration, and of the modes of articulation, which are the converse of a structure of logic and rhetoric. These playing techniques closely link together sound matter and musical composition, (both are generated at the same time), and evidently in a new way, they question the relationships between organization - the implementation of rules- and sound material.

In that way, this method of creating music implies that the work is looked on as a living moment, whereas the music created for traditional instruments is considered as a constructed object.

#### **4.3 The gestures and the music**

Obviously, the player's gesture is no longer just the one of a performer involved in the mere production of a piece in which the organization of its constitutive elements has been



thought in terms of organised objects but rather in terms of a sensitive sound experience set in time and space. The gesture, which consists in producing a sound, can therefore be considered as the origin of the shape which is afterwards gradually elaborated. A lot of work related to gestures and music has been recently published and we refer the reader for example to (Camurri et al., 1998), ( Coutaz and Crowley, 1995), ( Azarbayejani et al., 1996), (Sawada et al., 1997) for the state of the art in this field.

In *Un Bon Moment*, our main interest is precisely to use this gesture which produces and controls the audio-visual events as a source of creation. The gestural model is then free from any conventions (and particularly from musical ones). The main argument in *Un Bon Moment* is not to integrate these unexpected gestures into the framework of preexisting musical productions, but to use them in the implementation of specific musical means, as factors of artistic invention and reflection. In this context, impossibilities of determining time and synchronizing events become elements, which could involve situations of sensitive creation.

#### **4.4 The actual state of the project**

So far, two public performances of "Un Bon Moment" over a period of about one month have been carried out. Four performers, having few and particular gestural possibilities were running moving machines (called "Mobil Home"), each one being a sculpture based on a wheelchair, equipped with instrumental devices. One of them was intended to record, transform and project images, and the three others were more specifically intended for playing sounds. The three instrumental sound devices were conceived with the same idea: a gesture sensor, a module which transforms information given by the sensor into MIDI data, a sound generating module and a video projector which projects the whole or a part of the computer screen. The three gesture sensors are:

- A Headmouse System (Origin Instruments), controlling the mouse pointer on the computer screen, by means of an optical sensor following the movements of a disc stuck on the operator's forehead.
- A lever, large-sized ( and large range), as a joystick.
- A breath and lip-pressure sensor, which is part of a Yamaha WX7 wind controller.

These three gesture sensors communicate with the MAX software, either directly in the case of the WX7, or through an I-Cube (Infusion Systems) in the case of the lever, or with a MAX object (the PAD object) especially created for this work. For the moment, the sound generating modules for the three instruments, are supplied by the audio section of the MAX (MSP) software. The computer screen is projected either on a surface which has been integrated into the mobile instrumental device, or on an outside one. It enables the visualization of texts which are heard and treated in a sonorous way, of images which have been transformed by MIDI data picked up by the gesture sensor (Imagine Software), and of the redesigned mouse pointer (See below, Rafika Sahli-Kaddour's instrument).

#### **The PAD object**

This is an external rectangular graphical object. Like all the MAX objects, it can be moved on the patch and at the same time it keeps its intrinsic characteristics. It is resizable by a simple mouse click and slide. The PAD object analyzes the movements and actions of the pointer located in the screen zone that it covers, and sends back:

- The mouse's pointer position relative to the lower left hand corner (an offset and a multiplying factor can also be given),
- The speed and acceleration of the pointer's movement.
- A message if the pointer enters into the object or leaves it.
- A message if a click occurs when the pointer is inside the object.
- A message if the mouse button is held down when the pointer is inside the object and moves ( click and drag).

#### **An example : Rafika Sahli-Kaddour's instrument**

Rafika Sahli-Kaddour is one of the four performers. She has the ability and practice to drive a

enables her by the use of the PAD object to act upon three instrumental devices working differently.

- The first device projects the computer screen on a surface or a space where visual indicators have been set up: images, objects, etc... Each of them corresponds to invisible PAD objects. In this way, Rafika can direct the mouse pointer (redesigned as a red circle) towards real objects in space, move it on to the objects, produce and control the sound events according to three visual indication points. It is a kind of "non-linear score".
- The second device is fairly similar to the first, apart from the fact that a computer screen image is projected (it could be a film), and the objects that allow the control of sound events can correspond to certain zones of the image (or the film).
- The third device does not need screen projection. It uses the speed information and the horizontal or vertical movement of the pointer, given by the PAD object. The player controls sound transmissions by the head movements, and their speed. The movement repertoire calls upon, for memorization, more abstract notions of touch and intention.

#### **4.5 On-going development**

The instrument developed for Rafika Sahli-Kaddour is inspiring for future developments of more general instruments which could be used by anyone with limited mobility. Actually, the existing instrument is adapted to Rafika's gestural possibilities (she can only move her head). She can set off sound events, and control sound evolution, by moving the pointer on the screen thanks to a Headmouse System. We have observed that Rafika has gained a certain dexterity by using this instrument. However, outside the framework of *Un Bon Moment*, the use of the device is subject to the following constraints :

- the need to visualise the pointer on the screen or an external surface.
- the need to rewrite the program which determines the speed and movement of the mouse pointer at each time that the sound configuration is modified.
- the need of a gestural apprenticeship linked to a special device. In Rafika's case we have not encountered any problems at all, but it could be completely impossible to use for other disabled people.

Therefore, the idea has consisted in conceiving an instrument (opposite to a traditional acoustic instrument), which is neither a device that imposes a gestural form, nor a device adapted to specific morphological and motor abilities, but rather a device able to "learn" the player's gestures. This type of instrument could obviously be in another framework than the handicapped one. For the moment, we have set out three main stages for the realization of this project:

- The qualification of a few gestures (a maximum of ten or so) which correspond to musical aims.
- The creation of an "apprenticeship motor", which follows the player's gestures by a computer (detection of the pointer's movements and speed).
- The creation of an interface which enables complex gestures "learnt" by the computer to be linked to the previously created lexicon, and certain gestures or types of gestures, to be linked to sound events.







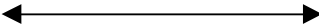
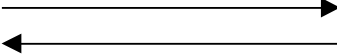



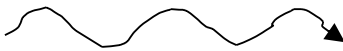
During the first stage, we use two MAX objects: the MouseState object and the PAD object described in section 4.5. They seem to be adapted to an "appraisal" of Rafika's gestures which are actually the only focal point of this study.

Actually, we are involved in a preliminary stage of definition:

- of a certain number of elementary attributes which enable us to define the gestures of the player. These gestures can then be expressed as a linear combination of the elementary attributes, which correspond to the eigenvectors in a mathematical sense.
- of a set of relevant elementary basis gestures in the context of a musical sound production.

In both cases it's obviously a question of defining prototypical reference points, while keeping in mind the complexity of reality, and the difficulty of translating "vague ideas", as when we

the implementation of the device is concerned, several solutions are considered (like neural networks, techniques of classification of pictures through contents) to link a set of quantitative data (in terms of coordinates, speed and acceleration) sent back by the computer sensors, to the whole set of qualitative features determined by the eigenvectors.

Gestures	Trajectory
Vibrating	
Trembling	
Throwing	
Put down	
Rubbing	
Doubt	
Swaying	
Following	
Sadness	
Question	
Greeting	
Sawtooth	

<b>Basis vectors:</b>
confined/extensive
soft/violent
periodical/in evolution
undulating/flat
direction

Figure 4.1

## 5 The "particles motion": The Cosmophone

### 5.1 The Cosmic Rays

Interstellar space is filled with a permanent flux of high-energy elementary particles called "cosmic rays". These particles have been created by violent phenomena somewhere in our galaxy, for example when an old massive star explodes into a supernova. The particles then stay confined in the galaxy for millions of years by the galactic magnetic fields before reaching our planet. When colliding against the earth's atmosphere, cosmic rays create showers of secondary particles. Though partly absorbed by the atmosphere, these showers induce a large variety of phenomena, which are measurable at the sea level.

The main phenomenon is a flux of muons, a kind of heavy electron absent from usual matter because of its short lifetime. Muons are produced at a large rate in cosmic showers. Thanks to their outstanding penetrating properties, they are able to reach the ground. At the sea level, their flux is about hundred muons per second per square meter.

Highly energetic cosmic rays produce bunches of muons, or multi-muons, having the same direction and a few meters apart from each other. The number of muons within a bunch is a function of the energy of the primary cosmic ray. Within an area of a hundred square meters, the rate of multi-muons bunches ranges from one per second (bunches of two or three muons) to one per minute (bunches of ten muons or more).

Muon interaction within the matter is another phenomena. When muons pass close to atomic nuclei, electromagnetic showers composed of electron-antielectron pairs are created. This phenomenon can be observed for example inside buildings with metallic structures, at a rate of about one per minute per ten square meters.

### 5.2 The Concept of Cosmophone

Human beings are insensitive to particles passing through their body. The Cosmophone is a device designed to make the flux and properties of cosmic rays directly perceptible within a three dimensional space. This is done by coupling a set of elementary particle detectors to an array of loudspeakers by a real time data acquisition system and a real time sound synthesis system. In that way, information received from the detectors triggers the emission of sounds, depending on the parameters of the detected particles. These parameters and the rate of occurrence of the different cosmic phenomena allow a large variety of sound effects to be produced. Because of the fluctuations in the occurrence of the phenomena, the set of detectors

phenomena to compose music, from Mozart with his "*Musical Dices Game*" to John Cage with his piece "*Reunion*", where player's moves on a specially equipped chessboard trigger sounds. Natural random-like phenomena have also been used, as in the piece entitled "*The Earth's Magnetic Field*" by Charles Dodge for which a computer translated fluctuations in the magnetic field of the Earth into music. Many other concepts for sound generation from such a source of random events can be (and are actually) explored, but we have chosen for this installation to keep the initial concept and further try to give the listeners the impression of being immersed in a particle rain.

### **5.3 Sound Generation and Spatialization in the Cosmophone**

According to the concept explained above, the synthesis system has to generate sounds when triggered by the particles detection system. To simulate a particle rain, in which listeners are immersed, we have grouped the loudspeakers in two arrays; one above the listeners (placed above a ceiling), and the other one below them (placed under a specially built floor). The arrays of loudspeakers are disposed so that the ears of the listeners (who are supposed to be standing up and moving inside the installation), are located approximately at an equal distance from the two groups. Both ceiling and floor are acoustically transparent, but the speakers are invisible to the listeners. A particle detector is placed close to each loudspeaker. When a particle first passes through a detector of the top group, then through a detector of the bottom group, a sound event is triggered. This sound event consists in a sound moving from the ceiling to the floor, "materializing" the trajectory of the particle. Because of the morphology of the human ears, our hearing system can accurately localize moving sources in a horizontal plane, but is far less accurate in the vertical plane. Initial experiments have shown us that the use of a panpot to distribute the signal energy between two loudspeakers was not sufficient to create the illusion of a vertically moving sound source. To improve the illusion of a vertical movement, we have used the Doppler effect, which is perceivable when there is a relative movement between an acoustic source and a listener. It then leads to a modification of the pitch of the perceived sound during time, as well as a modification of its amplitude. This effect is very common in every day life, and our hearing system is used to recognizing it. Chowning (Chowning-71) has shown that this effect is essential for the realism of moving source simulation. Using a Doppler effect simulation along with the energy panpot between the ceiling and the floor speakers greatly improve the illusion of a vertical movement of the sound source. But the departure and arrival points of the moving source in the space remain rather imprecise for listeners. To improve the localization, we have then added two short sounds as starting and ending cues. The first cue is emitted from the high loudspeaker at the beginning of the sound event; the second comes from the low loudspeaker, at the end of the event. Because they are chosen to be very precisely localizable, these two cues greatly improve the illusion, giving the impression of a sound crossing the ceiling, then hitting the floor.

We shall now describe a Cosmophone built for the *Cité des Sciences et de l'Industrie* in Paris, to be a part of an exposition area on particle physics: the "*Théâtre des Muons*". For this installation, two arrays of twelve speakers and detectors are in two concentric circles; the inner one comprises four speakers and detectors, the outer one the eight others. The outer circle is about five meters in diameter, which is enough space to allow several listeners to stand in the installation, as shown on figure 5.1, which represents a picture of the listening space.



Figure 5.1

#### **5.4 The Particles Detection System**

Each particle detector is composed of a slat of plastic scintillator associated to a photomultiplier. A particle passing through the scintillator triggers the emission of a photon inside the scintillator. The photon is then guided into the scintillator until it reaches the photomultiplier. An electrical impulse is then generated. This kind of detector is very

be sure to detect only particles generated by cosmic rays, a coincidence system is used. Such particles are first passing through a ceiling located detector, then through a floor located detector. Knowing that these particles travel at light speed, and knowing the distance between high and low detectors, the two events occurring in a given time (a few nanoseconds) is a signature of a cosmic induced particle. The coincidence triggers the readout of all the detector signals, which allows identifying the speakers in which the sound is to be produced. These signals are transmitted to a PC computer and processed to identify the kind of event detected, single particle, or bunch(es) of multi-muons. Depending on the energy of the cosmic rays and on their occurrence, it is possible to observe during a single event a combination of single muons and bunches of multi-muons. The program makes use of its knowledge of the geometrical arrangement of the detectors to recognize which phenomena were detected. Depending on the location of the installation, it is possible to observe a high rate of particle, which would generate a high rate of sound events and make the listeners rather confused. We can therefore decide to ignore some of these events to avoid overloading the listeners auditory systems. Finally, the information is sent to the sound synthesis system through a MIDI interface, using a custom protocol.

### **5.5 The Sound Synthesis System**

The sound synthesis system is built on an Apple Macintosh computer, running the MAX/MSP real time synthesis software. The computer is equipped with a MOTU (Mark Of The Unicorn) audio interface which is able to output simultaneously twenty-four channels of audio signals, which are sent to the amplifiers and loudspeakers. The synthesis program is composed of a set of modules or instruments; each of them is able to generate sounds associated with one event. The more powerful the computer, the more instruments that can be run at the same time, and the more events that can be played simultaneously. As mentioned above, different phenomena are detectable. In particular three cases were distinguished: a single muon reaching a pair of detectors (high then low); a "small bunch", where more than one pair of detectors hit simultaneously, but less than four, and a "big bunch", when at least four pairs are hit. We decided that the three cases would be illustrated by different sound sequences. When using a lot of detectors, the three events may occur simultaneously, which means that one instrument should be able to generate several sound sequences at the same time. A dynamic router has been implemented to allocate one sound event to one free instrument by passing the incoming MIDI data on to it. Therefore the instrument becomes busy during the sound generation process then returns to a free state. A synoptic of the sound synthesis system is shown in figure 5.2, and a synoptic of an instrument in figure 5.3.

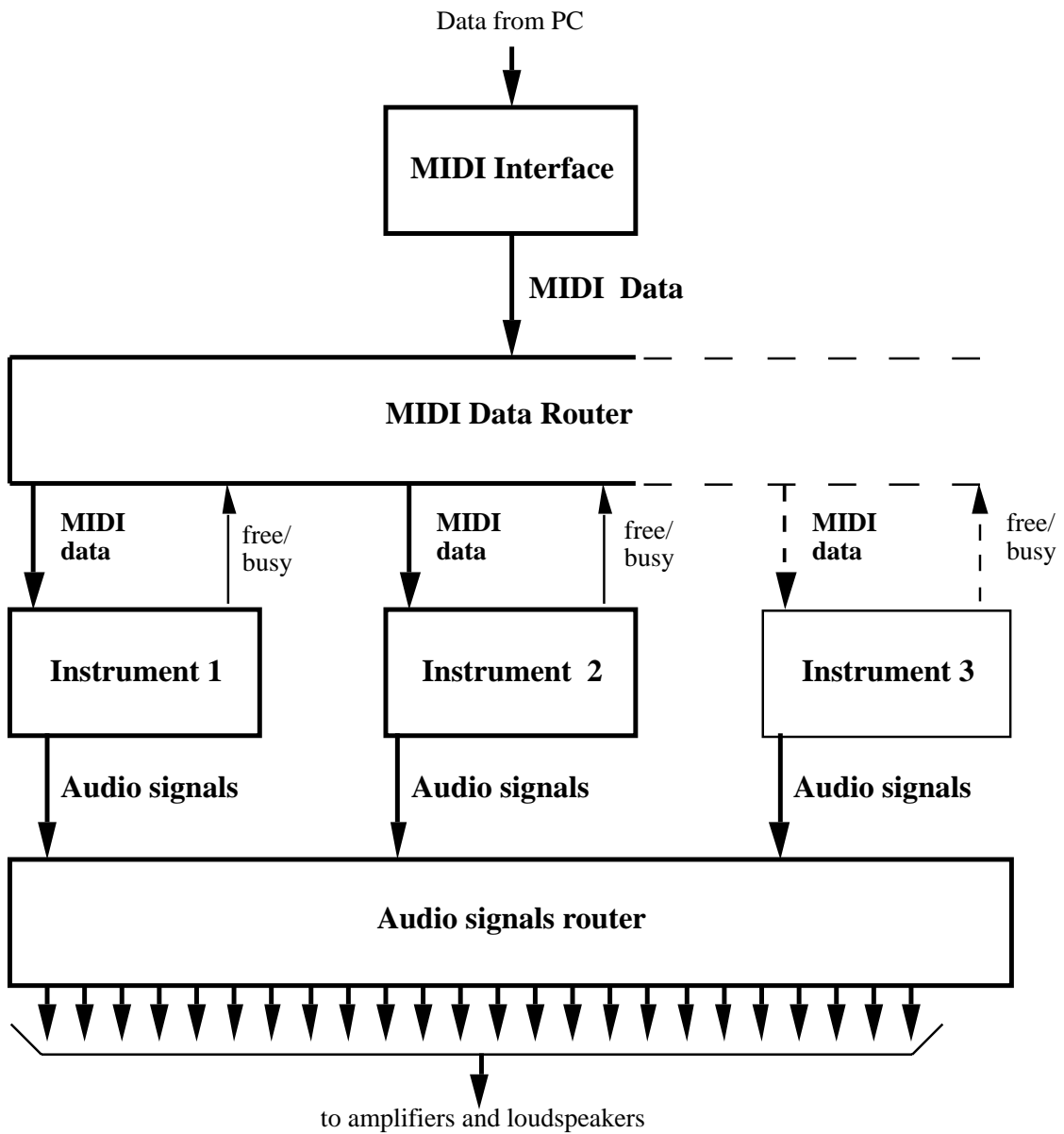


Figure 5.2



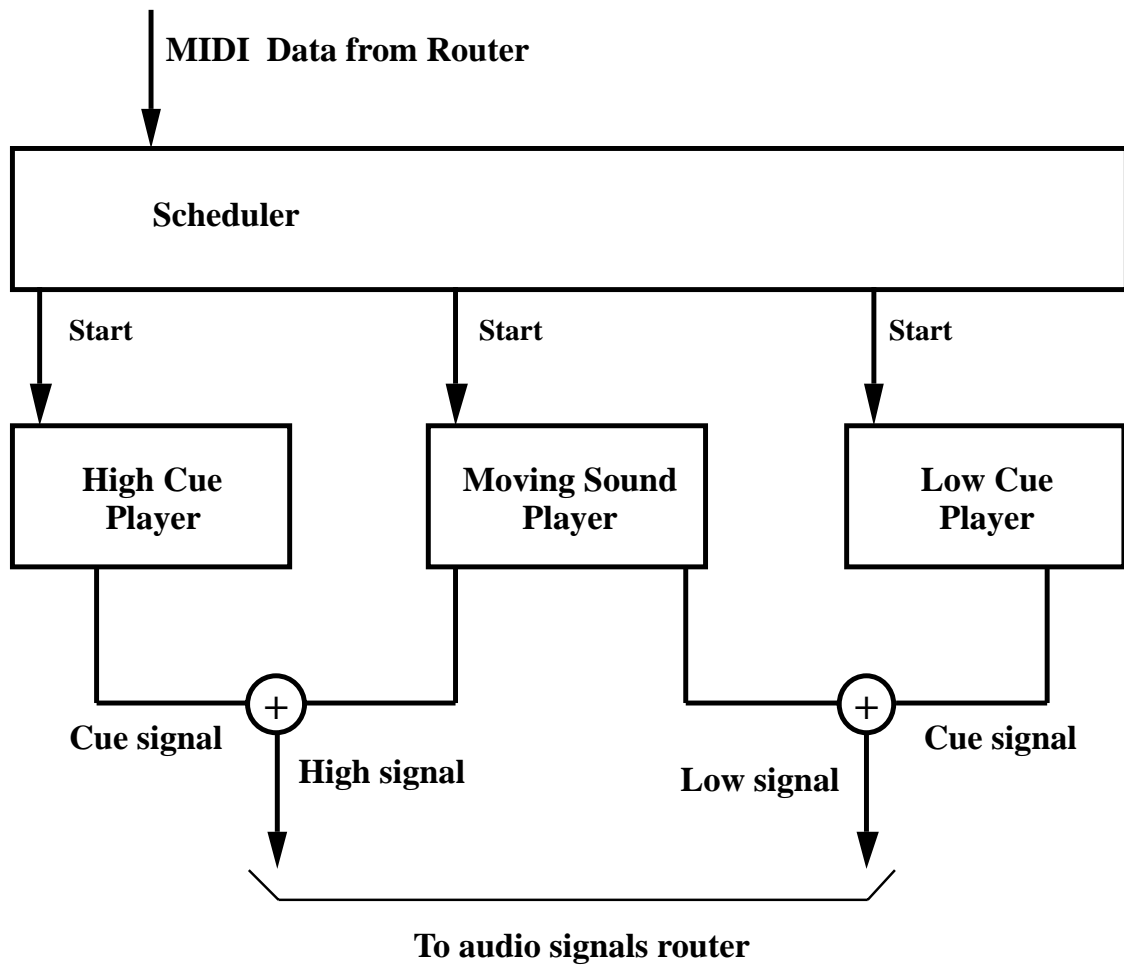


Figure 5.3

An instrument consists in a scheduler that trigger the three successive sounds: the cue in the ceiling located speaker, followed by the sound moving from ceiling to floor located speakers, and finally by the ending cue in the floor located speaker. Depending on the event's content, up to three sequences are played simultaneously. The scheduler manages the appropriate timing of the sequences, sending start messages to the appropriate sound players. Two kinds of sound players are used: the cue sound players and the moving sound player. The cue sound player is very simple: it reads a memory stored sound sample. As shown in figure 5.4, the moving sound player is a little bit more complicated, as it has to apply the Doppler frequency shift and amplitude modification. It makes use of a variable speed sound sample player and pre-computed tables for frequency and amplitude modifications. The final part of the instrument is an audio signal router that feeds the appropriate output channel, according to the information received for the event.

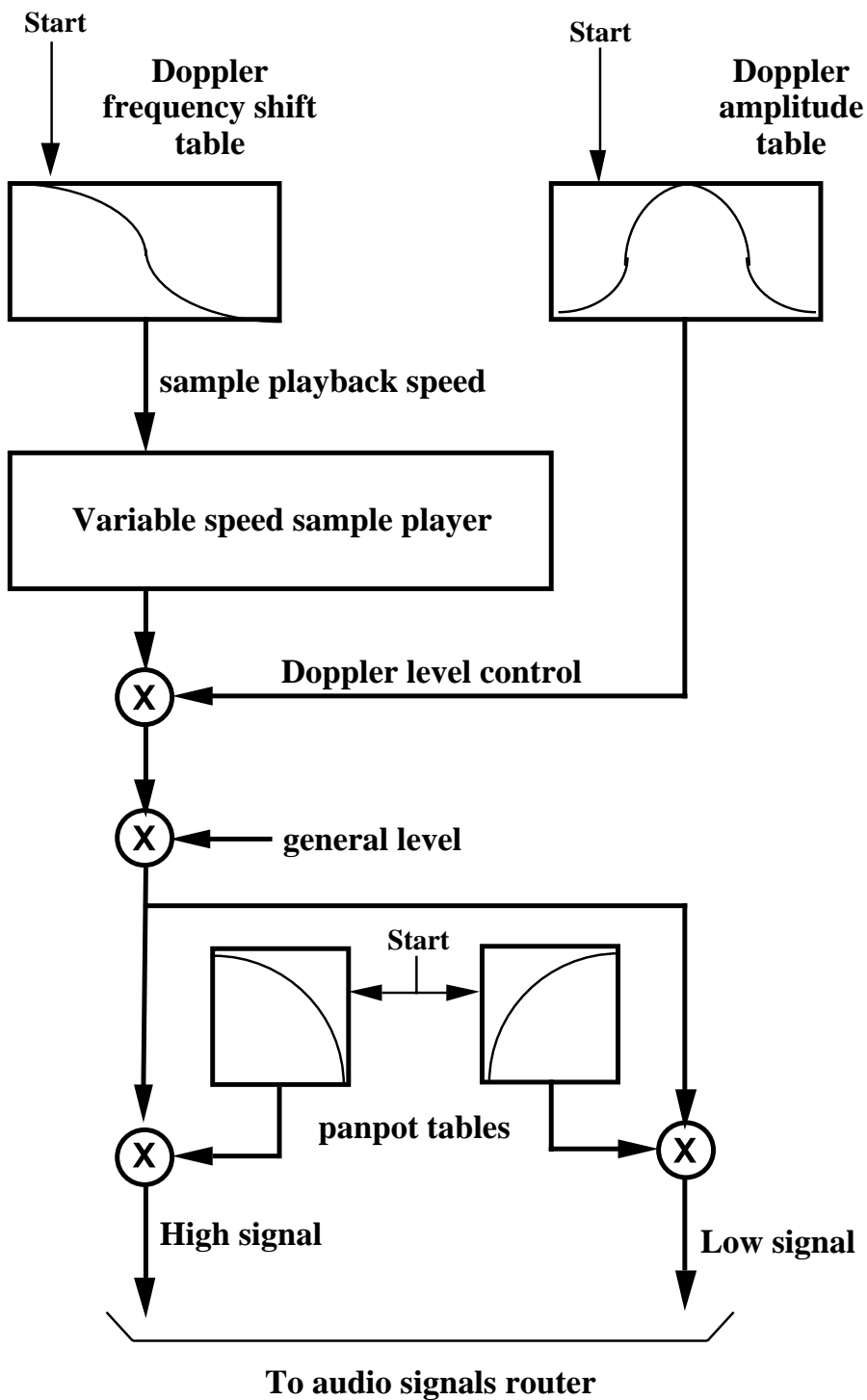


Figure 5.4

### 5.6 The Musical Potential

The final sound quality of this installation mainly depends on the nature of the sounds associated with the detected events. Because real particles travel at light speed and do not generate audible sounds, the real phenomenon could not be reproduced, but we can create with sounds its mental representation to the listener. The main difficulty consists in creating

of sounds were created by different synthesis techniques, and about ten moving sounds and ten cue sounds were selected and submitted to a panel of listeners for final judgement. We have tested many associations of moving and cue sounds to obtain the final result. The synthesis system allows fine-tuning of the Doppler effect parameters, which has appeared to be very useful for the final "tuning" of the sound impression. To help listeners concentrate on their auditive impressions when entering this installation, we have decided to diffuse some continuous background music through all the speakers. The installation has appeared to be a helpful tool for music spatialization. One of the authors has composed an original music score to be played in the *Cosmophone*, taking advantage of these possibilities. The installation described is open to the public in the *Cité des Sciences et de l'Industrie* in Paris, and a prototype composed of two arrays of four detectors and speakers is actually installed at *L.M.A.* in Marseille.

## Conclusion

The four developments that are presented in this article lead, by different approaches, to a conclusion where we discuss the musical intention behind computer systems used as musical instruments.

The creation of electronic instruments, which from our point of view is a part of the musical composition, inevitably puts forward questions about traditional relations between movements and sound production, thoughts and musical realizations.

The main difference between electronic and acoustic instruments seems to be the relation between the gestures and the sounds. What acoustic instruments are concerned, there is a one-to-one relation between these two aspects since a particular gesture generates a particular sound and since variations produced by muscular energy are closely related to variations in timbre, intensity and pitch. When it comes to electronic instruments these relations no longer exist, and everything has to be invented. This leads to a new way of considering the interface as an element of the musical composition.

Even though this article aims at describing different approaches of interface design, we always kept the relationship between music and events in mind. In the first example we used an already existing interface, the Radio Baton, to pilot different synthesis models. At an early stage we realized that this interface was restrained due to the fact that it, in the first place, was intended to conduct musical sequences or to be used as a drum-like instrument. Diverting this interface from its primary goal showed us the danger of rapidly falling into a "spectacular" use rather than a musical one. Nevertheless, its use for sound manipulation in an improvising context was found interesting.

The construction of meta-instruments, like the augmented flute, is a natural approach to give already skilled musicians access to new technologies. These interfaces mainly are of interest when they are related to synthesis models which correspond to the behavior of the instrument. This makes such controllers well adapted to physical synthesis modeling which naturally increases the musical possibilities of the instrument.

When constructing interfaces the gesture and the musical goal should be taken into account. Interfaces adapted to disabled people are a good example of this, since the motion possibilities are different for each of them, leading to a development of specific interfaces for each particular case. The interface has to be designed by carefully taking into account both the gesture possibilities and the musical objective.

The last example we develop in this article, the *Cosmophone*, represents the most extreme case, where there are no gestures, and where natural phenomena generate a three dimensional sensation of the trajectories of cosmic rays. In this case a specific set of detectors associated to an adequate mapping directly creates this objective.

These examples show that the musical dimension has to be part of the conception of an interface, and led us to the conclusion that: even though new technologies provide numerous

devices which can be used to pilot sound processors, a genuine musical interface should go past the technical stage to integrate the creative thought.

### **Acknowledgments**

Part of this work has been supported by the Norwegian Research Council.

The Cosmophone was designed from an original idea of C. Vallée (Centre de Physique des Particules de Marseille) who built the particle detector together with D. Calvet.

The authors thank D. Zicarelli for his valuable help during the development of the Cosmophone sound synthesis system, Mitsuko Aramaki for her precious help during the review of the paper and the referees for their helpful comments.

## **6 Bibliography and References**

Azarbayejani, A. Wren, C. Pentland, A. 1996. *Real-time 3-D tracking of the human body*, in Proceedings of the IMAGE'COM96, May 1996.

Battier M. 1999. *L'approche gestuelle dans l'histoire de la lutherie électronique. Etude d'un cas : le theremin*. in R. de Vivo and H. Genevois (eds) *Les nouveaux gestes de la musique*, Editions Parenthèses, Collection Eupalinos, ISBN 2-86364-616-8.

Boie, R. Mathews, M. Schloss, A. *The radio drum as a synthesizer controller*, in Proc. Int. Computer Music Conf. (ICMC'89), pp. 42-45, 1989.

Boie, R.A. Ruedisueli, L.W. and Wagner, E.R. 1989. *Gesture Sensing via Capacitive Moments*. Work Project N° 311401 AT&T Bell Laboratories.

Boulanger, R. Mathews, M. 1997. *The mathews' radio baton and improvisation modes*, in Proc. Int. Computer Music Conf. (ICMC'97), pp. 395-398.

Bromwich, M.-A. 1997. *The metabone: An interactive sensory control mechanism for virtuoso trombone*. in Proc. Int. Computer Music Conf. (ICMC'97), pp. 473-475.

Cadoz, C. Lisowski, L. Florens, J. L. 1990. *A modular feedback keyboard design*, Computer Music J., vol. 14, no. 2, pp. 47-51, 1990.

Camurri, A. Ricchetti, M. Di Stefano, M. Strocchio, A. 1998. *EyesWeb - toward gesture and affect recognition in dance/music interactive systems*, in Proc. Colloquio di Informatica Musicale CIM'98, AIMI.

Chowning, J. 1971. *The Simulation of Moving Sound Sources*. Journal of the Audio Engineering Society 19:1.

Coutaz, J. Crowley, J. 1995. *Interpreting human gesture with computer vision*, in Proc. Conf. on Human Factors in Computing Systems (CHI'95), 1995.

Gershenfeld, N. Paradiso, J. 1997. *Musical applications of electric field sensing*, Computer Music J., vol. 21, no. 2, pp. 69-89, 1997.

Jaffe, D.A., Schloss, W.A. 1994. *A Virtual Piano Concerto* in Proc. Int. Computer Music Conf. (ICMC'94), 1994.

Jensen, K. 1996. *The control mechanism of the violin*. in Proceedings of the Nordic Acoustic

- Kanamori, T. Katayose, H. Simura, S. Inokuchi, S. 1993. *Gesture sensor in virtual performer*, in Proc. Int. Computer Music Conf. (ICMC'93), pp. 127-129, 1993.
- Katayose, H. Kanamori, T. Simura, S. , and Inokuchi, S. 1994. *Demonstration of Gesture Sensors for the Shakuhachi*. In Proceedings of the 1994 International Computer Music Conference. San Francisco, International Computer Music Association, pp. 196-199.
- Keane D. and Gross, P. 1989. *The MIDI baton*. in Proc. Int. Computer Music Conf. (ICMC'89), pp. 151-154, 1989.
- Kronland-Martinet R., Guillemain Ph., Ystad S. 1997 “*Modelling of Natural Sounds Using Time-Frequency and Wavelet Representations*” Organised Ssound, Vol.2 n°3, pp.179-191, Cambridge University Press.
- Kronland-Martinet, R. Voinier, T. Guillemain, P. 1999. *Agir sur le son avec la baguette radio* in R. de Vivo and H. Genevois (eds) *Les nouveaux gestes de la musique*, Editions Parenthèses, Collection Eupalinos, ISBN 2-86364-616-8.
- Laubier, S. de 1999. *Le Méta-Instrument a-t-il un son? Emergence de lois ou de constantes dans le développement d'instruments virtuels*. In H. Genevois and R. de Vivo, (eds) *Les nouveaux gestes de la musique*. Marseille: Editions Parenthèses, pp. 151-156.
- Machover, T. 1992. *Hyperinstruments - a Composer's Approach to the Evolution of Intelligent Musical Instruments*. In L. Jacobson, ed. *Cyberarts: Exploring Arts and Technology*. San Francisco: MillerFreeman Inc., pp. 67-76.
- Mathews, M.V., Schloss, W.A. 1989. *The Radio Drum as a Synthetizer Controler* in Proc. Int. Computer Music Conf. (ICMC'89), 1989.
- Mathews, M.V. 1991a. *The Conductor Program and Mechanical Baton*. in M. Mathew and J. Pierce (eds) *Current Directions in Computer Music Research*. Cambridge, MA, MIT Press, 1991.
- Mathews, M.V. 1991b. *The Radio Baton and Conductor Program or: Pitch, the Most Important and Least Expressive Part of Music*. *Computer Music Journal* 15:4.
- Mathews, M.V. 1997. *Exécution en direct à l'âge de l'ordinateur*. in H. Dufourt and J.M. Fauquet (eds) *La musique depuis 1945 - matériau, esthétique, perception*. Mardaga, Brussels, 1997.
- Mathews, M.V. Abbot C. 1980. *The Sequential Drum*. *Computer Music Journal* 4:4.
- Moog, R. 1987. *Position and Force Sensors and their Application to Keyboards and Related Controllers*. In Proceedings of the AES 5th International Conference. New York, NY: Audio Engineering Society, pp. 179-181.
- Moog, R., and T. Rea. 1990. *Evolution of the Keyboard Interface: The Bosendorfer 290SE Recording Piano and the Moog Multiply-Touch-Sensitive Keyboards*. *Computer Music Journal*, 14(2):52-60.
- Moore, F. R. 1987. *The disfunctions of MIDI*. in Proc. Int. Computer Music Conf. (ICMC'87), pp. 256-262, 1987

- Mulder, A. 1995. *The I-Cube system: moving towards sensor technology for artists*. in Proceedings of the ISEA, 1995.
- Pierrot P., Terrier A. 1997. *Le violon MIDI*. tech. rep., IRCAM, 1997.
- Pousset, D. 1992. *La flûte-midi, l'histoire et quelques applications*. Mémoire de Maîtrise, 1992. Université Paris-Sorbonne.
- Puckette, M. Zicarelli, D. 1990. *Max - an Interactive Graphic Programming Environment*. Opcode Systems.
- Rovan, J., Wanderley, M. Dubnov, S. Depalle, P. 1997. *Instrumental Gestural Mapping Strategies as Expressivity Determinants in Computer Music Performance*. Kansei, The Technology of Emotion. Proceedings of the AIMI International Workshop, A. Camurri, ed. Genoa: Associazione di Informatica Musicale Italiana, October 3-4, 1997, pp. 68-73.
- Sawada, H. Onoe, N. Hashimoto, S. 1997. *Sounds in Hands - a Sound Modifier using Datagloves and Twiddle Interface*. In Proceedings of the 1997 International Computer Music Conference. San Francisco, International Computer Music Association, pp. 309-312.
- Schloss, A. 1990. *Recent Advances in the Coupling of the Langage MAX with the Mathews/Boie Radio Drum*. Proceedings of the International Computer Music Conference 1990.
- Smith J.O. 1992. *Physical Modeling Using Digital Waveguides*. Computer Music Journal 16:4.
- Smith, J. R. 1996. *Field mice: Extracting hand geometry from electric field measurements*, IBM Systems Journal, vol. 35, no. 3/4, pp. 587-608, 1996.
- Snell, J. 1983. *Sensors for Playing Computer Music with Expression*. In Proceedings of the 1983 International Computer Music Conference. San Francisco, International Computer Music Association.
- Verge, M.P. 1995. *Aeroacoustics of Confined Jets with Applications to the Physical Modeling of Recorder-like Instruments*. PhD thesis, Eindhoven University.
- Vergez, C. 2000. *Trompette et trompettiste: un système dynamique non linéaire analysé, modélisé et simulé dans un contexte musical*. Ph.D. thesis, Université de Paris VI.
- Wanderley, M. M. Battier, M. eds. 2000. *Trends in Gestural Control of Music*. Paris: Ircam - Centre Pompidou.
- Wanderley, M. M. Viollet, J.-P. Isart, F. Rodet, X. 2000. *On the Choice of Transducer Technologies for Specific Musical Functions*. In Proceedings of the 2000 International Computer Music Conference. San Francisco, CA: International Computer Music Association, pp. 244-247.
- Wright, M. Wessel D. Freed, A. 1997. *New musical control structures from standard gestural controllers*, in Proc. Int. Computer Music Conf. (ICMC'97), pp. 387-390.

Ystad, S. 1999. *De la facture informatique au jeu instrumental*. in R. de Vivo and H. Genevois (eds) *Les nouveaux gestes de la musique*, Editions Parenthèses, Collection Eupalinos, ISBN 2-86364-616-8.

Ystad, S. Voinier, T. 2000. *A Virtually-Real Flute*. Computer Music Journal (MIT Press), 25:2, pp 13-24, Summer 2001.

Ystad, S. 2000. *Sound Modeling Applied to Flute Sounds*, Journal of Audio Engineering Society, Vol. 48, No. 9, pp. 810-825, september 2000.

Yunik, M. Borys M. Swift G.W. 1985. *3A Digital Flute*. Computer Music Journal 9(2): 49-52.

# New Method for the Directional Representation of Musical Instruments in Auralizations

Felipe Otondo, Jens Holger Rindel

Ørsted DTU, Acoustic Technology, Technical University of Denmark

*email:* {fo,jhr}@oersted.dtu.dk

## Abstract

The issue of the representation of sound sources that vary their directional pattern in time in auralizations is introduced. Musical instruments are used as a reference for the discussion of the traditional representations with assumed fixed directional characteristics. A new method for the representation of the spatial sound contributions in time is proposed using multiple-channel recordings and various virtual sources in room auralizations. Possible developments of the proposed recording/reproduction method are described.

## 1 Introduction

The term “auralization” has been coined as an analogous term to visualization – it therefore names the process of rendering audible (imaginary) sound fields. Room auralizations have as main objective a simulation as accurate as possible of the binaural listening experience in a certain location within a modeled space (Kleiner, Dalenback, Svensson 1993), (Odeon 2002). An important factor to be taken into consideration in an auralization is the directional characteristics of the sound source. Musical instruments have a complex directivity pattern, which generates a particular acoustic behavior in a room. The aim of this investigation is to take a closer look at their directivity in the case of a real performance and to provide a better representation of this behavior in room auralizations.

## 2 Directional characteristics of musical instruments

The sound produced by musical instruments involves many different acoustic features that are related to intrinsic characteristics of the instrument. One of these features is the directional characteristic, or directivity, which is the way in which the sound of the instrument is radiated in different directions at different frequencies. The directivity of a musical instrument is affected by the different notes played on the instrument (Meyer 1978), the different performing intensities (Rossing 1990) and the different playing techniques. These changes are different for the different families of musical

instruments, due to the complexities of the musical instrument as a multi-resonating system (Fletcher, Rossing 1998). An example of the measured directional characteristics of four isolated notes played on a Spanish guitar at the 500 Hz octave can be seen in Figure 1. In this case the notes were played within two octaves by the performer trying to maintain the same intensity.

## 3 Musical instruments as sound sources in auralizations

When musical instruments are used as sound sources for auralizations, it is important to take into consideration their radiation characteristics in order to have a representation of the sound source in the room model. As already mentioned, musical instruments are sound sources that have a complex directivity which cannot be easily described in a real performance situation. If a fixed directivity pattern per octave were to be considered, such as the case of a loudspeaker, the result would be rather poor and inaccurate. The directivity changes in time would be ignored and the consequence would be a wrong directional pattern with the level at certain frequencies of the particular spectra either emphasized or diminished. Perceptually this will deteriorate the listening experience due to the added colourations. A more accurate representation that will contain the source directivity changes in time is therefore necessary.

## 4 Improvement of the spatial representation of sound

One could offer a better representation of the spatial sonic characteristics of a musical instrument in a room auralization, or of any source that changes its directivity in time, by taking into account the various samples of the sound field created by the source; they are to be used afterwards in the reproduction process. One method of achieving this is through simultaneous anechoic recordings of the musical instruments with microphones surrounding the source in order to capture the sound radiated in different directions. An example of an anechoic 4-track recording of a



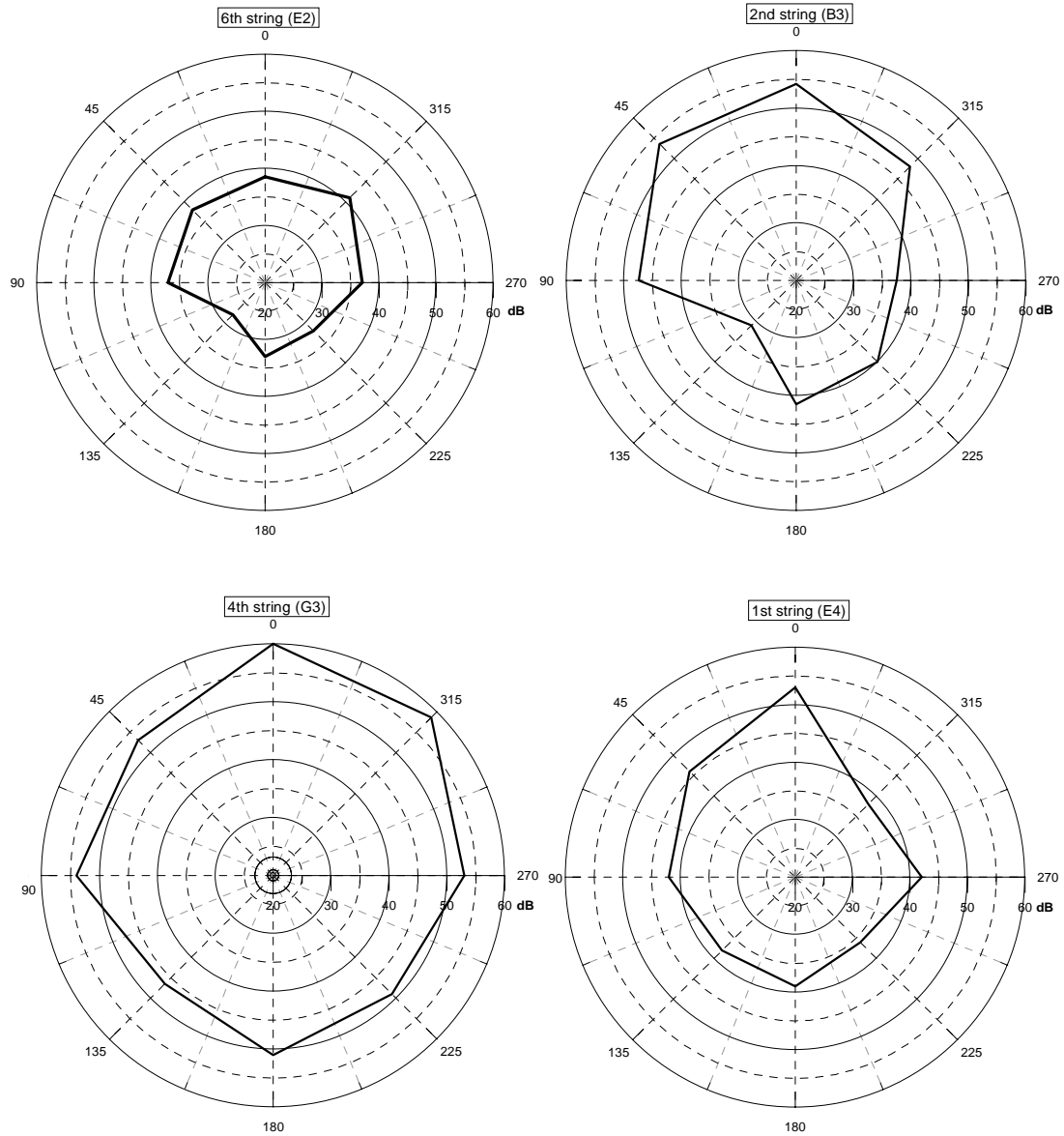


Figure 1. Polar diagram of the directivity of a Spanish guitar in the 500 Hz octave band for four isolated notes within two octaves. The notes were played by a performer trying to keep the same level of intensity. The position of the guitar was vertical, as in a normal classical music performance. The dynamic range is plotted from 20 to 60 dB.

musical instrument can be seen in Figure 2, where four microphones are located around the source. After making the multi-channel recording of the instrument, each of the particular recordings registered by the microphones is played by a particular virtual source in the auralization according to the original position in the recordings. This can be done easily in the simulation program by defining sources that have a neutral directivity pattern (omnidirectional) within a discrete span of radiation. In the case of a 4-track recording of figure 2, each source span would correspond to a quarter of a sphere. Figure 3 shows a room acoustic simulation for the example of figure 2, where an auralization considering four virtual sources has been done, each source with an omnidirectional characteristic within a span of a quarter of a sphere and radiating in the direction of 0, 90, 180 and 270 degrees. The new source (consisting of the four virtual sources together) will radiate in a distinctive way in each of the four directions following changes in level, movements, asymmetries and orientation of the original source that were recorded by the individual microphones.

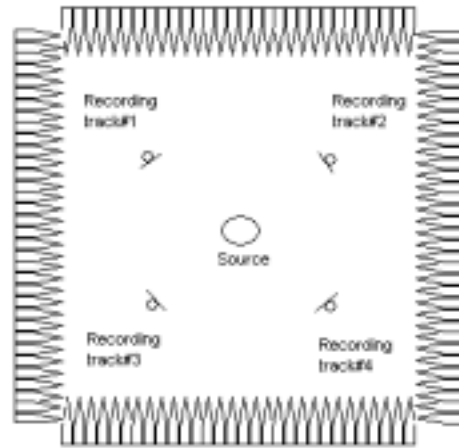


Figure 2. Setup for a 4-track anechoic recording of a source.

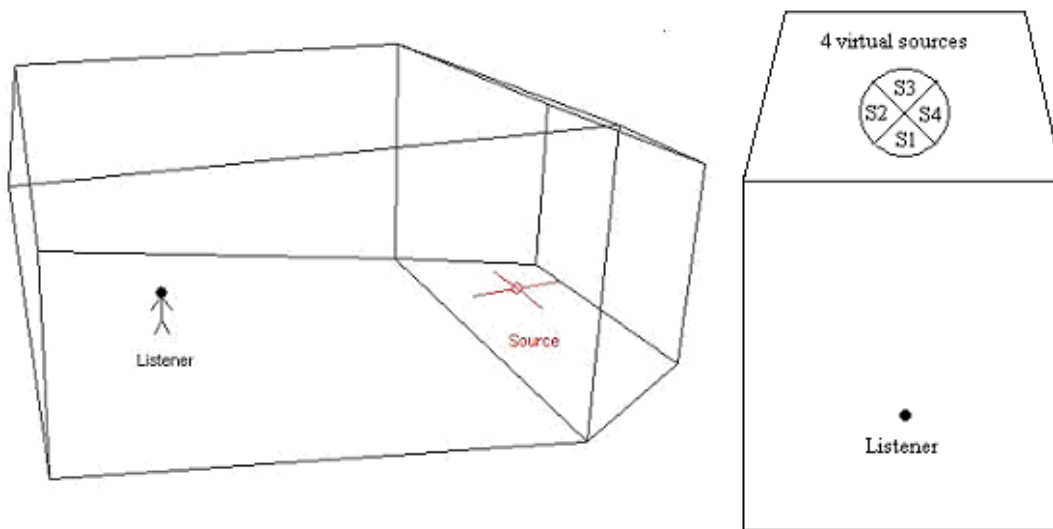


Figure 3. Room acoustic simulation with one listener and four virtual sources (left). View from above of the room showing the four virtual sources, each in a different direction with a discrete radiation pattern of a quarter of a sphere (right).

## 5 Concluding remarks

The directivity of musical instruments in a real performance situation in room auralizations is an issue that needs a better representation, as the one already existing only assumes a fixed directivity per octave band. The proposed method offers an alternative directivity representation without the need of any directional data of the instruments.

Further developments of the system will be aimed at an optimization of the recording setup, by considering the influence of the number and the position of microphones and their perceptual consequences in the room auralizations.

Binaural virtual reality systems for room models usually lack a sound source definition and have problems with the spatial representation of the source (Kleiner, Dalenback, Svensson 1993). The use of a different kind of representation of the source's directional characteristics in these systems, such as the one proposed by this work, can help to improve and make more reliable the spatial representation of a sound source in sound demonstrations avoiding sound colorations.

The directivity representation of sound sources in movement (like the real performance case of a saxophone player or an actor in movement) is not possible with the fixed directivity representations available nowadays. An implementation of this work could help to make more reliable representations of a live performance situation.

The use of headphones in virtual reality systems limits the possibilities of sound reproduction. This is one of the reasons for the lack of commercial success of such systems. The use of multi-channel loudspeaker reproduction systems for multi-track recordings could be a further development of this project, considering a crosstalk cancellation system or some other filtering technique to avoid destructive sound interference (Gardner 1998).

## 6 Acknowledgments

The work reported in this article has been financed by the European Community project MOSART (Music Orchestration Systems in Algorithmic Research and Technology) HPRN-CT-2000-00115.

## References

- Kleiner, M., Dalenback, B. I., and Svensson, P. 1993. "Auralization - An Overview." *Journal of the Audio Engineering Society*, 41(11):861-874.
- ODEON. 2002. "Odeon Room acoustics software." <http://www.dat.dtu.dk~odeon>
- Meyer, J. 1978. *Acoustics and the performance of music*. Verlag Das Musikinstrument, Frankfurt. pp. 75-102.
- Rossing, T. D. 1990. *The science of sound*. Second Edition, Addison-Wesley (chapter 11, pp. 208-209).
- Fletcher, H.N. and Rossing, T. D. 1998. *The physics of musical instruments*. Second-Edition, Springer-Verlag, New York. (Parts III, IV and V).
- Kleiner, M., Dalenback, B. I., and Svensson, P. 1993. "Audibility of Changes in Geometric Shape, Source Directivity, and Absorptive Treatment - Experiments in Auralization." *Journal of the Audio Engineering Society*, 41(11): 905-913.
- Gardner, W. 1998. *3D Audio using loudspeakers*. Kluwer Academy Publishers, Boston.

# DIRECTIONAL REPRESENTATION OF A CLARINET IN A ROOM

Felipe Otondo and Jens Holger Rindel

*Ørsted-DTU, Acoustic Technology, Technical University of Denmark*

## Abstract:

This article presents a study of the directional characteristics of a clarinet in the context of a real performance. Anechoic measurements of the directivity of a Bb clarinet have been done in the horizontal and vertical planes for isolated tones. Results are discussed comparing the particular directivity of tones and the averaged directivity over the whole range of the instrument. Room acoustic simulations with the measured and averaged directivities have been carried out in a concert hall as an example of a more realistic application. Further developments will consider measurements with other instruments as well as auralizations and tests with an alternative sound radiation representation.

**Keywords:** musical acoustics, clarinet, directivity, room acoustics simulations

## 1. INTRODUCTION AND GOALS

The directivity of musical instruments has been studied by several authors [1, 2, 3], Jürgen Meyer being probably the one who has contributed with more specific information about the radiation characteristics of musical instruments in a real performance situation [4]. The data on the directional characteristics of different classical music instruments provided by Meyer are mostly concerned with averages of the directivity over the whole performing range of the instruments. Very little information is included about the directivity of instruments for particular tones, even though, as shown by Meyer, the directivity of instruments changes dramatically over the performing range. Most of the available data on the directivity of musical instruments used nowadays for room acoustic simulations and auralizations consider the averaged directivities from Meyer's results. Very few attempts have been made to use a different directional representation that would include the directivity changes of the musical instruments within the performing range [5]. On the other hand, experiments using room acoustics auralizations have shown that the directional representation of sources in room acoustic simulations is important and changes in their directivity can affect the perceived sound in a room [6]. For these reasons there is a need to better understand how large the variations of the directivity of musical sources are in a real performance situation and how important these variations can be both acoustically and perceptually.

The first goal of this study has been to measure and compare the particular directivities of a clarinet for particular tones and the averaged directivity over the whole compass. This has been done in order to compare the traditional representations (averaged directivity) with a more realistic representation of a performance situation (directivities of particular tones). The second goal has been to use the measured and averaged directivities for room acoustic simulations in order to evaluate their possible influence on the perceived sound according to different room acoustical parameters.

## 2. DIRECTIVITY OF THE CLARINET

### 2.1 Choice of instrument

The idea of making directivity measurements was inspired by the goal of achieving a comparison between the particular directivities of a musical instrument and the averages of the directivity over the whole compass. The instrument chosen for the directivity measurements was a clarinet in Bb, mainly due to its sound radiation characteristics as well as its large register possibilities [7]. The directivity measurements of the clarinet were planned and made in a way that would allow a study of the directivity of the instrument in a performance situation, always with the representation of the source in computer room simulations in mind. For this reason, it was very important to make simultaneous measurements and have a setup that could be used for other purposes such as simultaneous recordings for room auralizations [5]. It was also important to make the measurements using the whole performing range of the instrument in order to have the average and the particular directivities available for comparison.

### 2.2 Directivity measurements of the clarinet

The directivity measurements of the clarinet were made using simultaneous recordings of 13 microphones in the anechoic chamber at 45° intervals, considering a distance of 1.5 meters from the source and measured in octaves from 125 to 8000 Hz. The measurements were made using a 24-bit quantisation and a sampling frequency of 44.1 kHz. Single tones were measured over the whole compass of the instrument (44 tones) with a similar musical intensity played by the performer. Figure 1 shows the performer playing the clarinet in the anechoic chamber during the measurements and Figure 2 shows the measuring setup in the horizontal and vertical planes.



Fig. 1. Clarinet player during the directivity measurements in the anechoic chamber.

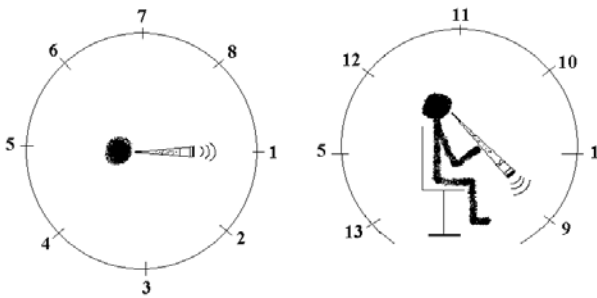


Fig. 2. Setup for the simultaneous directivity measurements with 13 microphones. The left part of the figure shows the setup in the horizontal plane and the right part shows it in the vertical plane. Microphones 1 and 5 appear in both planes.

### 2.3 Results

In order to compare particular directivities with the averaged directivity over the whole range, short samples of the sound of the tones were chosen and used as representative. Five different tones, with ascending pitches over the whole compass of the instrument, were considered for the comparisons with the calculated average. These tones with their respective fundamentals were: C4 (262 Hz), A4 (442 Hz), E5 (667 Hz), B5 (999 Hz) and E6 (1327 Hz). Figure 3 shows an example of polar plots comparisons between the averaged and the particular directivity for five ascending tones at 2000 Hz. The horizontal and vertical results were normalised to the level in the frontal microphone (microphone 1 in Figure 2) so as to correspond to the way the directivity of sources is represented in computer room simulations and in order to have a basis for comparison for later simulations. Figure 4 shows the graphs with the level differences between the particular tones directivity and the averaged directivity for

the octave bands from 500 to 4000 Hz. In this case the curves were normalised to the level in the front and a suitable range of frequencies was chosen for comparisons (500-4000 Hz). The graphs displayed in Figure 4 show the filtering process over the fundamental of the tones.

The results in the horizontal plane show that the directivity differences increase with the filtered frequency. For most of the curves at 500 and 1000 Hz the level differences are within a range of  $\pm 5$  dB, with some punctual exceptions where the difference can be up to 10 dB (C4, 270° at 1000 Hz and A4, 90°, 270° at 1000 Hz). At 2000 Hz, the directivity differences become greater within a range of  $\pm 10$  dB, with some exceptions where the differences can be up to almost 30 dB (E6, 315°). At 4000 Hz the directivity differences are clearly very big within a range of  $\pm 20$  dB.

In the vertical plane the differences become greater than in the horizontal plane and also less predictable within the same measured tone. The differences at 500 and 1000 Hz are within a range of  $\pm 15$  dB with some particular greater differences in some cases (C4 for 135° and 315° at 500 Hz; A4 for 90° at 500 Hz; E5 for 90° at 1000 Hz). At 2000 Hz the differences are in a range of  $\pm 10$  dB, while at 4000 Hz the differences are within a range of  $\pm 15$  dB.

### 2.4 Discussion

The comparisons between the measurements of the directivity of the clarinet and the average show fluctuating differences that clearly increase with frequency. When comparing the directivity differences in both planes it becomes clear that the differences are much greater in the vertical plane than in the horizontal plane. As shown by Meyer [4], in the vertical plane the clarinet becomes more and more directional towards the axis of the bell (315°) over 1000 Hz. This could make the particular directivities at higher frequencies have greater differences resulting in larger variations compared to the average. This could also mean that the changes at higher frequencies are more dramatic and that a different representation might imply severe consequences in the sound perceived in a room. To provide an example, one can consider tone E6 from Figure 3. In this case the instrument clearly becomes more directional in both planes. It can be guessed that in this case if the average representation were to be used in the auralization of a room instead of the particular directivity of E6, the instrument's directional characteristic would be transformed into something much more omnidirectional, raising the level in the front direction (0°) instead of the axis of the bell and also at the sides (45° and 315°) of the instrument. In an auralization these differences could be very noticeable if one considers a change in the direct sound of 6 dB and in the lateral reflections in almost 10 dB as seen in the figure.

Another interesting result of the comparisons is that they show that the pitch of the selected tones does not seem to affect the directivity differences in the horizontal plane. In the vertical plane an increase in differences can be noticed with the pitch of the tones. As stated before, the vertical plane seems to be much more unstable and dependent on the particular tones played.

### 3. THE SOUND OF THE CLARINET IN A ROOM

#### 3.1 Room acoustic simulations

As a way to get a more clear picture of how the directivity of a clarinet affects the sound in a room, computer room simulations were carried out with the measured directivities assuming the same power for the sources. The software used for the simulations was ODEON version 6.0 [8]. The purpose was to contrast the differences in the sound in the room using the traditional averaged representation of the directivity of the clarinet and the representation of particular ones. For this purpose the directivities of two of the five tones considered previously (C4 & A4) were used. The room simulations were carried out in a model of the concert hall ELMIA located in Jönköping, Sweden. The different directivities were simulated with the sources located in the normal position of the soloist on the stage, at a height of 1 meter from the floor pointing to the audience.

The first comparison was made between the grid response for the sound pressure level (SPL), calculated both with the averaged directivity and the directivity of the two tones respectively. Figure 5 shows the grid response for the SPL in the concert hall at 500, 1000, 2000 and 4000 Hz with the three directivities (average, C4, A4). As it can be seen in the figure, the simulations showed the SPL to be directly dependent on the directivity of the sources in most of the cases, while in some cases the directivity changes dramatically for the same tone filtered at two different octaves. In this case the directivity of the average maintains a good SPL homogeneity in the room compared with the tones for all the measured frequencies, except for 4000 Hz, where it is clearly worse. Another interesting fact shown in the figure is that the directivity affects the SPL on the stage, showing in some cases a great difference within a very short distance from the source.

A second comparison made with the simulation software concerned the clarity factor (C80). Figure 6 shows the grid response of the C80 in the room at the different frequencies, with the averaged directivity and the directivity of the tones. In this case the clarity also seems to be directly dependent on the directivity of the source and the differences seem to be more critical. When comparing the C80 with the averaged directivity and the ones of the tones, one can see that in some cases the C80 with the average directivity is clearly lower than the one of the tones (1000 and 4000 Hz). In the other cases it is similar or slightly higher. The C80 of the averaged directivity seems to be more symmetrical in the distribution in the room and also quite similar for 500, 1000 and 2000 Hz. One could say that the C80 with the averaged directivity is not necessarily much better than the one of the tones but more stable and predictable.

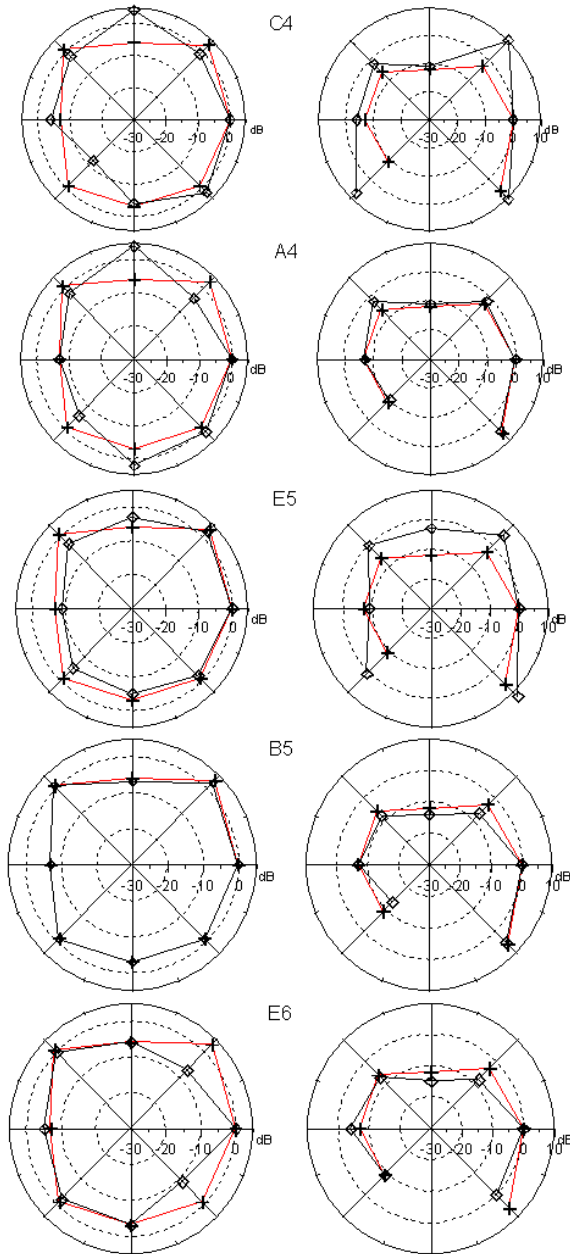


Fig. 3. Comparisons of the averaged and the particular directivities of the clarinet at 2000 Hz for different tones. Left: Horizontal Plane. Right: Vertical plane. The averaged directivity is plotted with crosses while the particular directivity is plotted with diamonds. The orientation is the same as in Figure 2 with the instrument pointing to the right.

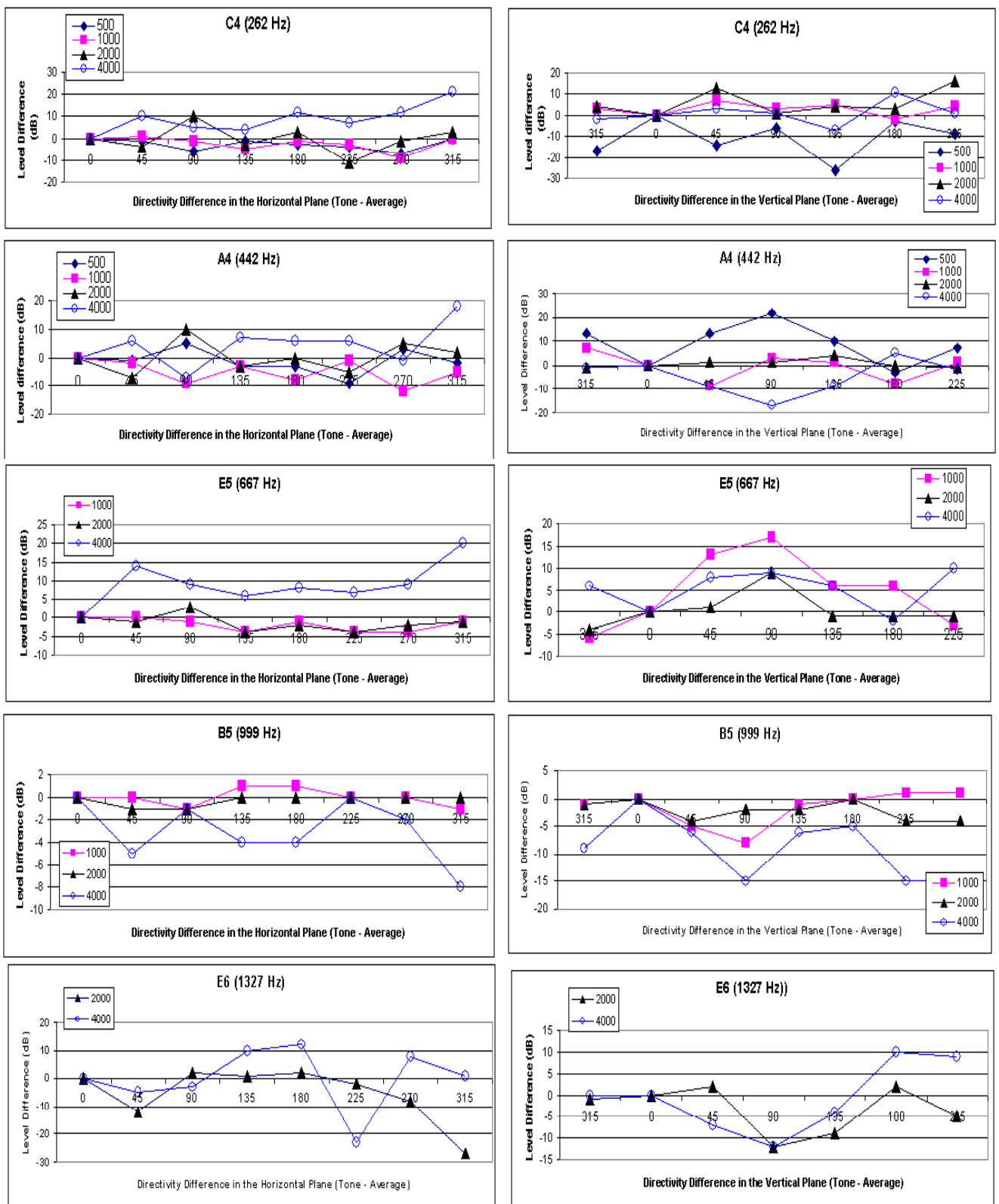


Fig. 4. Directivity difference in the horizontal and vertical planes for five ascending tones of the clarinet when compared with the average. The left column of figures corresponds to the horizontal axis while the right column corresponds to the vertical axis.

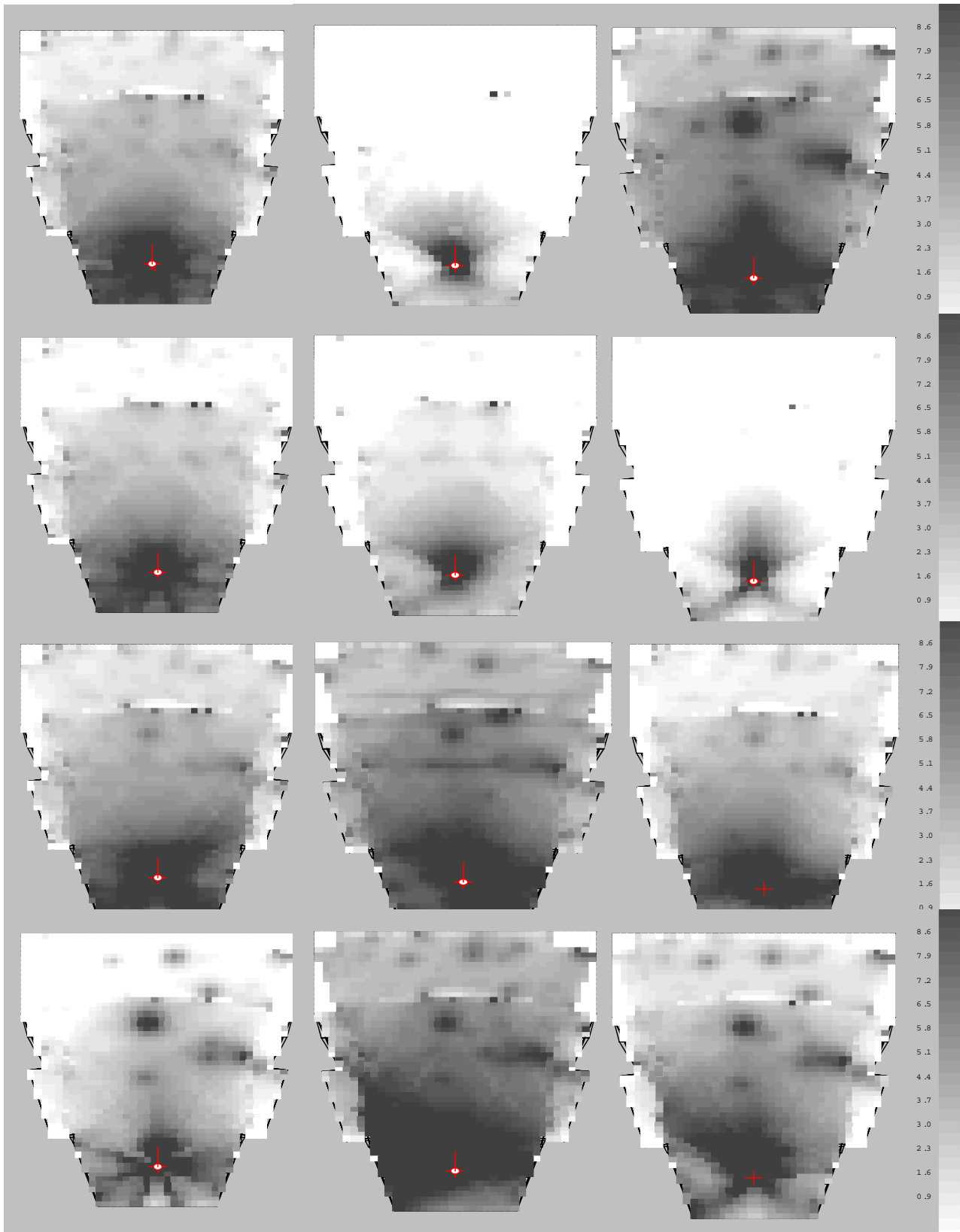


Figure 5. Grid-response of the sound pressure level (SPL) with three different directivities at 500 (top), 1000, 2000 and 4000 Hz (bottom) at the ELMIA concert hall. The figures on the left correspond to the averaged directivity, the figures at the center to the first tone's directivity (C4) and figures to the right to second tone's directivity (A4). The scale is relative and shown from 0 to 10 dB with white and black as the lowest and maximum values, respectively.



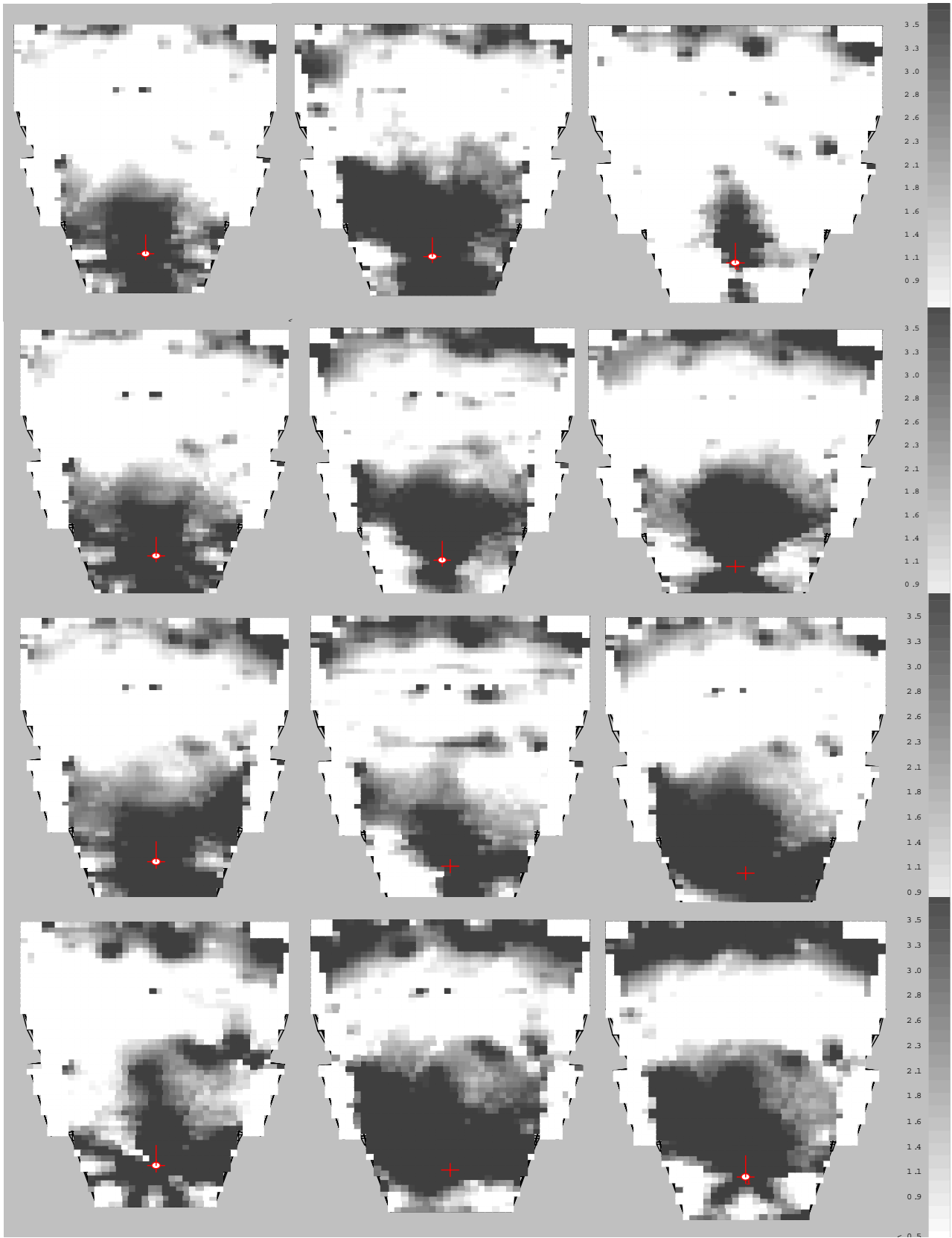


Figure 6. Grid-response of the clarity (C80) with three different directivities at 500 (top), 1000, 2000 and 4000 Hz (bottom) at the ELMIA concert hall. The figures on the left correspond to the averaged directivity, the figures at the center to the first tone's directivity (C4) and figures to the right to second tone's directivity (A4). The scale is shown from 0 to 4 dB with white and black as the lowest and maximum values, respectively.

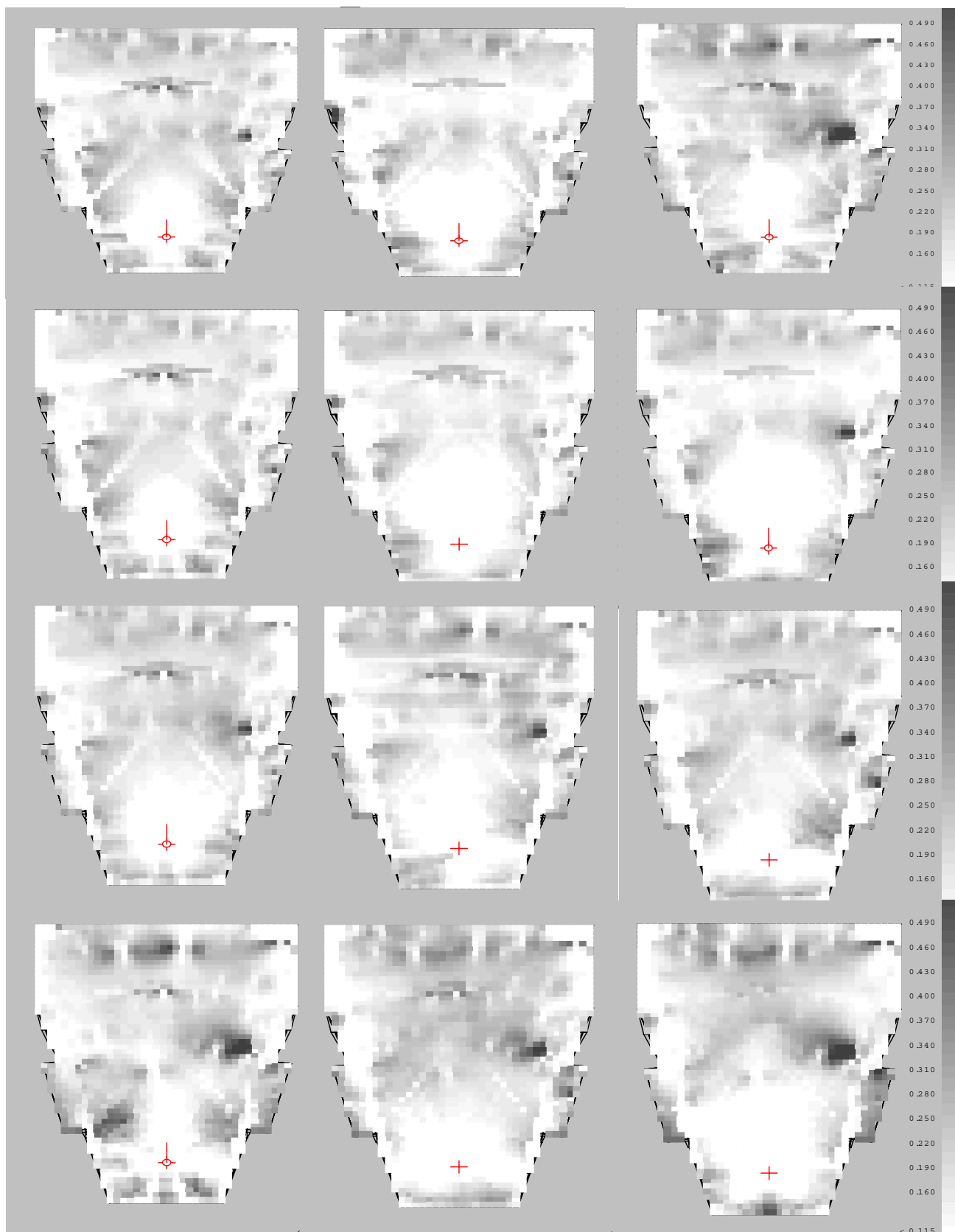


Figure 7. Grid response of the Lateral Energy Fraction (LF) with three different directivities at 500 (top), 1000, 2000 and 4000 Hz (bottom) at the ELMIA concert hall. The figures on the left correspond to the averaged directivity, the figures at the center to the first tone's directivity (C4) and figures to the left to second tone's directivity (A4). The scale is shown from 10% to 50% with white and black as the lowest and maximum values, respectively.

The last comparison made with room simulations concerned the Lateral Energy Fraction (LF). Figure 7 shows the LF grid response in the room at the different frequencies with the averaged directivity and the directivity of the tones. The simulations show in this case that the LF is also dependent on the directivity of the source. The LF distribution in the room with the averaged directivity showed to be more stable and symmetrical than the one for the tones in most of the cases.

### 3.2 Discussion

The room simulations with the measured and averaged directivities showed that the changes in the directional pattern of the source clearly affect the SPL, C80 and LF in the room. The symmetry of the directivity in the horizontal plane of the sources showed to affect the homogeneity of the simulated sound in the room in the horizontal plane, especially notorious for the C80 and LF. The changes of the directivity from one filtered frequency to the other were greater for the particular tones measured than for the averaged directivity. The importance of the symmetry and stability in the representation of a musical instrument in a room will need to be studied in detail in order to assert if the desired representation should resemble the directivity of particular tones, with differences in level and asymmetries in the sonic distribution in the room, or an averaged directivity which is more stable and less directional.

Further developments of the room acoustic simulations could consider comparisons of auralizations using the measured and averaged directivities used for the simulations. Other issues to be studied are the importance of the directivity representation of the source in the sound distribution on stage [9] and the importance of the changes of the directivity in the vertical plane in the representation.

### 4. CONCLUSIONS

The directivity measurements of particular tones of the clarinet show greater differences with the averaged directivity. These differences increased with frequency but not with the pitch of the tones measured and showed to be much larger in the vertical plane than in horizontal axis.

Room simulations with the averaged and particular directivities of the clarinet showed that the changes in the directional pattern of the source clearly affect the sound distribution in a room. The symmetry of the directivity in the horizontal plane of the sources showed to be important in the homogeneity of the simulated sound in the room in the horizontal plane. The averaged representation of the directivity implied in some cases a more even distribution of the acoustical factors than the one of the particular tones; in others it implied a more scattered distribution.

### AKNOWLEDGMENTS

The authors would like to thank Jørgen Rasmussen, Finn Jacobsen, Claus Lyng Christensen, Anders C. Gade, Nina Gade, Anne K. Snaslev and Christoffer Weitze for their help and support during this work. The work reported in this article has been financed by the European Community project MOSART (Music Orchestration Systems in Algorithmic Research and Technology) HPRN-CT-2000-00115.

### REFERENCES

1. T.H. Rossing, "The science of sound". Addison-Wesley. Second edition (1990).
2. N.H. Fletcher & T.D. Rossing, "The physics of musical instruments". Springer-Verlag. New York. Second edition (1997).
3. J. Štěpánek & Z. Otčenášek, "Sound Directivity Spectral Spaces of Violins". Proceedings of the International Symposium on Musical Acoustics, Perugia Italy (2001).
4. J. Meyer, "Acoustics and the performance of music". Verlag Das Musikinstrumenter. Frankfurt/Main (1978).
5. F. Otondo & J.H. Rindel, "New method for the representation of musical instruments in auralizations". paper submitted to the International Computer Music Conference, Gotemborg, Sweden. (2002).
6. B.-I. Dalenbäck., M.Kleiner & P. Svensson, "Audibility of Changes in Geometric Shape, Source Directivity, and Absorptive Treatment-Experiments in Auralization" .41, 905-913 (1993).
7. W. Piston, "Orchestration". W.W. Norton & Company. 164-169 (1955).
8. "The Odeon home page." <http://www.dat.dtu.dk/~odeon>
9. A.C. Gade, "Investigations of Musicians' Room Acoustic Conditions in Concert Halls. Part I: Methods and Experiments" *Acustica*. 69, 193-203 (1989).

**IHP Network HPRN-CT-2000-00115 MOSART**  
**Music Orchestration System in Algorithmic Research and**  
**Technology**

**MOSART Task 4:**

**Detection of Human Motion and Interactive**  
**Musical Performance.**

**Edited by Jens Arnsfang**

**Deliverable d24**

Evaluation Reports on experiments with Detection of Human Motion, Interfacing to Musical Devices and use for Musical Performance

**Table of Content**

<b>Detection of Human Motion and Interactive Musical Performance</b> Jens Arnsfang, Kristoffer Jensen, Declan Murphy	Page 182
<b>Gesture Recognition for Conductor and Dance Interpretation</b> Volker Krüger	Page 187
<b>A Smart Analog to MIDI Interface</b> Gabriele Boschi	Page 191
<b>Building A Hand Posture Recognition System from Multiple Video Images: A Bottom-Up Approach</b> Declan Murphy	Page 193
<b>The Votion Project</b> B. Stang, E. Tind, D. Murphy, J. Arnsfang, K. Jensen, A.-M. Bach Jensen, C. Beyer, M. Gugliemi	Page 227
<b>An Improved Edge Detection and Ranking Technique</b> Declan Murphy	Page 232
<b>Extracting Arm Gestures for VR using EyesWeb</b> Declan Murphy	Page 240

# Detection of Human Motion and Interactive Musical Performance

Jens Arnsfang, Kristoffer Jensen, Declan Murphy

## Introduction

When it comes to human-computer interaction in a musical performance situation, several modalities are involved. In this task in particular the interaction between vision and music is studied. One example is a conductor, controlling both human soloists and a virtual orchestra, using a computational vision interface. Another example is a virtual orchestra, following the movements and moods of a human dancer. See for example [1-6] for computer vision tools and application of such techniques within computer music. Other modalities than vision have been used and further are conceivable. This task contains open windows within Music Informatics towards both the research field of Human Computer Interaction (HCI) and towards contemporary performance art.

Interactive musical performance includes synchronization and control of patterns and structure of music. In order to be able to make sense of the rhythmic structure of a live performer the computer needs to be able to quantize the time durations in the performance to their score-bound note durations. This process is tightly intertwined with the ability to follow tempo changes and the detection of beat and/or meter in a real-time situation [8-9].

## Tasks steps

The task contains three major issues: (a) How to detect information in human body motion, such as dance and conducting, (b) how to link this information to control mechanisms in general and (c) how to synchronize such control structures to musical events.

Concerning (a): Detection of human motion is itself a complicated task, where electronic sensors, such as bend sensors [10] or magnetic sensors [11], is one methodology, and the search for vision based methods is another. The strive is concentrated upon a breakthrough on the possibilities, while comparing and incorporating these in the current state-of-the-art motion capture paradigms. The method aims towards taking elements from motion sequence analysis and from biomechanics, in order to build a first prototype of a body motion detector, capable of revealing time evolving parameters in human motion.

Concerning (b). The link and control structures (mapping) between the parameters output from the motion detection and the end goal, performance of music, will be studied in two ways. Taking the natural approach, that certain motions should be given classical meaning in conformance with conducting and dancing traditions, directly or context dependent; or taking the more innovative attitude, that chosen motions may be linked to certain new

music attributes, and given control over musical events. These events may be reactive or interactive, evolving over time, using algorithmic behavior of control structures.

Concerning (c). This task can be divided into two parts: One part is focused on finding hardware and software solutions capable of satisfying the demands of music performance (flexible, intuitive, easy to use, robust, compatible with previous releases over a reasonably long time). The (commercially) available software systems include Max/MSP, PD, EyesWeb, BigEye, jMax, the Very Nervous System. The other part is focused on identifying structure in the music that can be manipulated by the control structure identified in (b).

To some extent the methodology and the findings in this task will provide a basis for the task T6 of Computer Music Composition Tools, as far as insight into high level description of music is concerned.

## **Contributions**

Other contributions in this publications include *Gesture Recognition for Conductor and Dance Interpretation*, by Volker Krüger, which is concentrating on the issue (a) above, and contains a short survey of the state of art, and of further plans for reaching the goal of visual detection of human motion, *A Smart Analog to MIDI Interface*, by Gabriele Boschi who is concentrating on the issue (b) above, and contains a well carried through exercise in producing an actual interface, producing the technical foundation for carrying out the sub task required. Furthermore, see G. Boschi's contribution to Task 3. Other visual based contributions are the contributions by Declan Murphy, who is concentrating on the issues (a) and (c) above, and contains a major and very promising effort in this important and ultimately central goal of the entire task 4. The cross-field, large collaboration matrix, *Votion project* published here has shown some very promising aspects of the interaction between artistic and virtual reality collaborators, mixing audio and visual feedback with multi-modal sense interaction.

## **Practical examples of cross-field production by the partners**

Although being at the point where task 7 (dissemination) has just been initiated, several completely new applications, used in industrial or artistic production (although not yet firmly installed in cultural institutions) have been produced. Forerunners are found, not initiated in the MOSART context only, but produced by the partners, and likely to be upgraded by the end of the project, due to the outcome of task 4 and others. A few such are mentioned below.

### **The DIEM Digital Dance System**

The DIEM Digital Dance [12] system is an interface designed especially for interactive dance. The dancer wears a large number bending sensors that measure the angles of the dancer's limbs. The bending sensors are connected to a small radio transmitter worn by the dancer on a belt. Data is transmitted to a receiver unit that sends standard MIDI controller values for each sensor. The system can be used in any MIDI setup to control music or lighting in a dance performance. The system has been used by instrumentalists to control live computer processing in a concert performance. The DIEM Digital Dance System has been made commercially available on request by numerous performers, choreographers and composers. It is used in music productions on many locations with great success today.

### **The Genoa Opera House and the Salzburg Opera Festival**

The MOSART partner DIST has a long standing effort within the Genoa Opera House for building interactive tools for dance following, exactly the topic of this MOSART task 4. Visual detectors (cameras), light detectors (infrared), sound detectors (sonars) and others are combined in order to provide a tool for a performance situation. The parts in the DIST artistic production performances are humans, sensor technology, and robots performing together with the live dancers.

Public performances are often given. One major event was an opera at the Salzburg festival, where a camera was placed over the stage in order to capture the motion pattern and body language of the main character. Subsequently the timbre and musical expression of the human voice was slightly modified, according to the detected human motion. In this way the artist obtained a tool for further expressing moods and feelings.

While the public was aware of these processed sounds and music from the artist, they were however, by and large, unaware of the technology (the cameras and other sensors were discretely out of sight). We consider this to be a tribute to both artist and technology. The tools developed in the MOSART task 4 is planned to be used in similar cultural events. Declan Murphy and others in MOSART besides people at DIST are working and further developed techniques for movement analysis around the EyesWeb platform. Several papers have been published by Declan Murphy and from the DIST researchers concerning these issues. The EyesWeb web page (<http://www.eyesweb.org/>) is a good source for further information.

An event in London at Theatre Almeyda (composer Battistelli) was performed successfully by the staff of DEI-University of Padua using the EyesWeb platform. The researchers at DIST had a performance at the Opera House of Genoa, Teatro Carlo Felice, in collaboration with Alvisé Vidolin and composer Roberto Doati, based on EyesWeb (live electronics with real-time movement analysis and mapping). The InfoMus laboratory at DIST had joint work with KTH and DEI at Mestre (Venice) Atelier, including collaboration in a concert with two DJs.

## **The CNUCE Interactive Concerts on Virtual Instruments**

Being a partner with desire for, and furthermore capabilities in both electronic, computer science, music informatics, composing and artistic performance, the staff at CNUCE often perform in public themselves, not least team leader Leonello Tarabella. At these occasions, Leonello Tarabella carries a suitcase with the necessary equipment: A few sensors, light equipment, software and interfaces the size of a portable computer.

This kind of equipment has shown to be adequate not only for large performance halls, but also for minor stages. It is easily envisaged that such equipment can become every man's possession. By combining the existing tools and technology with the outcome of MOSART task 4, a foundation could be laid down for mass production of such every man's portable interactive music performance tool kit, and at the same time the professional traveling artists and composers tool kit. With the popularization of small cheap sensors in many fields, and small, cheap and capable portable computers, only knowledge about the current technology, as found in the MOSART network, is needed to enable the production of such kits.

Although already visible, the impact of MOSART onto such equipment, and onto such industrial productions and cultural performance situations, awaits the outcome of task 7.

## **Conclusion**

The MOSART task 4 is an important step in the network collaboration. This task is linked to several others, including those involved in sound control and parameterization, music pattern recognition, interactive music performance and composition. The success of the network collaboration is dependent on respecting the important goals of each task, stated in the contract technical annex. This has been done by structuring the work in the steps (a), (b) and (c), as reported above. The fact that the work in these steps continues to gain strength and relevance, and that cooperation within the consortium is evident, and that young researchers in the net actually commute among the partners, show that networking is indeed taking place within task 4 and other tasks. This is to the benefit of both science and the training of young researchers.

## **References**

- [1] Arnspang, Jens. *Motion Constraint Equations in Vision Calculus*. Elinkwijk Drukkerij, Utrecht. 1991.
- [2] Antonio Camurri, Paolo Coletta, Massimiliano Peri, Matteo Ricchetti, Andrea Ricci, Riccardo Trocca, and Gualtiero Volpe. A real-time platform for interactive dance/music systems. In Proc. Int. Computer Music Conf., Berlin, Germany, Aug2000. ICMA.
- [3] Claudia Nölker and Helge Ritter. Detection of fingertips in human hand movement sequences. In Wachsmuth and Fröhlich, editors. *Gesture and Sign Language in Human-Computer Interaction*. Proc. Int. Gesture Workshop, pages 209-218. ISBN: 3-540-64424-5.



- [4] Claudia Nölker and Helge Ritter. GREFIT: visual recognition of hand postures. volume 1739 of LNAI, pages 61-72, Gif-sur- Yvette, France, Mar 1999. Springer-Verlag. For video clips and examples of hand gesture applied to audio synthesis, see also <http://www.techfak.uni-bielefeld.de/~claudia/vishand.html>.
- [5] Moritz Störing and Erik Granum, 2001, “Con-straining a statistical skin color model to adapt to illumination changes”, Proc. Dansk Selskab for Automatisk Genkendelse af Mønstre, 30–31. Aug 2001, Copenhagen.
- [6] Camurri, A., KANSEI, The Technology of Emotion. AIMI International Workshop, Genoa 1997. DIST Press, Univ. of Genoa, 1997.
- [7] STEIM products > BigEye, <<http://www.steim.nl/bigeye.html>>, August 1, 2002.
- [8] Desain, Aarts, Cemgil, Kappen, van Thienen and Trilsbeek. Robust Time-quantization for Music, from Performance to Score. Proceedings of the 106th AES convention, Munich: Audio Engineering Society 1999.
- [9] Honing, H., From time to time: the representation of timing and tempo, Computer Music Journal, 25(3). Pages 50-61, Fall 2001.
- [10] Siegel, W., “The Challenges of Interactive Dance - an overview and case study”, Computer Music Journal, vol. 22, no. 4, 1998.
- [11] Ascension Products - Flock of Birds, <<http://www.ascension-tech.com/products/flockofbirds.php>>, August 1, 2002.
- [12] Diem Digital Dance System, <<http://www.daimi.au.dk/~diem/digitaldance.html>>, August 1, 2002.

# Gesture Recognition for Conductor and Dancer Interpretation

Volker Krüger

Our principle goal is to design, develop and evaluate a novel system using digital cameras for the detection and interpretation of human motions for interaction with musical performances, e.g. for the interpretations of conductors, the control of musical instruments or the automatic choreographic assistances for dancers. Such a system will provide 3-D measurements of human movements of the whole bodies, body parts and joints. The availability of these descriptions will allow us to describe, manipulate and interpret the human movements according to our goals.

General human activity recognition has been subject of much research in recent years. Excellent surveys of previous work in this area can be found in [Moeslund and Granum, 2001; Gavrilu, 1999]. Success without the use of explicit models, e.g., [Davis and Bobick, 1997] and with the use of such models, e.g., [Wren and Pentland, 1998] has been reported. Recent progress in gesture recognition covers a variety of different applications. Among these applications are

- Handgesture recognition for sign language[Wang *et al.*, 2002],
- Recognition of actions in, e.g., smart rooms and augmented desk interfaces[Oka *et al.*, 2002; Ren and Xu, 2002; Bretzner *et al.*, 2002; Duric *et al.*, 2002; Herda *et al.*, 2002],
- Recognition of individuals though their actions, using, e.g. gait-gestures[Collins *et al.*, 2002; Kale *et al.*, 2002].
- Detection of joints and body structure[Sullivan and Carlsson, 2002; Sidenbladh *et al.*, 2002; Bregler and Malik, 1997; Wren and Pentland, 1998]

Despite the recent advances there are still many hurdles in achieving reliable detection and estimation of whole body human motion. Some of the most challenging issues are due to the complexity and the variability of the appearance of the human body, the nonlinearity of a conducting motion, non-rigid nature of dancing movements, a lack of sufficient image cues about 3D body pose, including self-occlusion as well as the presence of other occluding dancers. A further challenge is the exploitation of multiple video streams. We have conducted research on single-camera and multi-camera based tracking of human body parts using a rigid body model and a particle filter. In this approach the body parts are approximated using simple geometric solids and the human tracking problem is posed as a state estimation problem. The state vector contains both the 3D geometry of the body parts and their motion parameters. We employ a branching particle method, the system of particles that mimics the conditional density of states[Crisan *et al.*, 1998] to converge to the target distribution. Shape filtering

viewed as a measurement process is also elegantly incorporated into the non-linear filtering framework, which contributes to the accurate computation of the particle weights. The four figures below are taken from a video sequence and show human body tracking results, the original image taken from the video (left) and the corresponding geometric solids, viewed from the frontal (center) and the side (right).

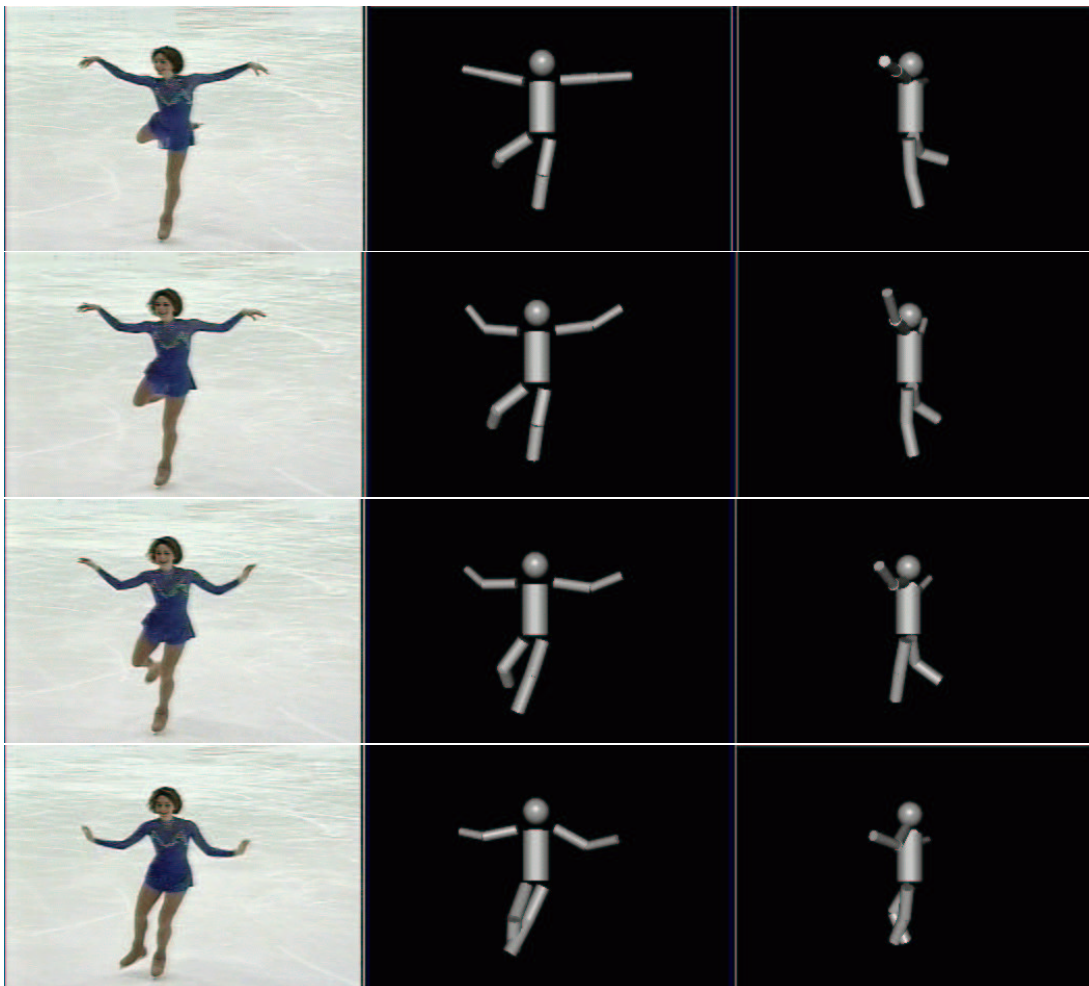


Figure 1: The figures below are human body tracking results, overlaid on the original image(left) and the corresponding geometric solids, viewed from the frontal (center) and the side (left).

Human body self-occlusion is a moajor cause of ambiguities in body part tracking using a single camera. Although by using particle filtering, the body part tracking ambiguities can be described by a multi-mode posterior distribution of body part motion parameters, these ambiguities cannot be removed using only a single camera. We have experimented with multiple, calibrated cameras within

our particle framework, however, the results are preliminary.

Once the 3D geometric parameters of the body parts are recovered, continuous tracking using these 3D description parameters can be pursued in a Bayesian framework similar to that in the classical 2D case. The challenge is to arrive at a sparse description of a moving human to add to the state vector.

## References

- [Bregler and Malik, 1997] Christoph Bregler and Jitendra Malik. Learning Appearance Based Models: Mixtures of Second Moment Experts. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 845. The MIT Press, 1997.
- [Bretzner *et al.*, 2002] Lars Bretzner, Ivan Laptev, and Tony Lindeberg. Hand Gesture Recognition using Multi-Scale Colour Features, Hierarchical Models and Particle Filtering. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pages 423–428, Washington, DC, USA, May 21-22, 2002.
- [Collins *et al.*, 2002] Robert Collins, Ralph Gross, and Jianbo Shi. Silhouette-based Human Identification from Body Shape and Gait. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pages 366–371, Washington, DC, USA, May 21-22, 2002.
- [Crisan *et al.*, 1998] Dan Crisan, Jessica Gaines, and Terry Lyons. Convergence of a Branching Particle Method to the Solution of the Zakai Equation. *SIAM Journal on Applied Mathematics*, 58(5):1568–1590, 1998.
- [Davis and Bobick, 1997] J. Davis and A. Bobick. The representation and recognition of action using temporal templates, 1997.
- [Duric *et al.*, 2002] Z. Duric, F. Li, and H. Wechsler. Recognition of Arm Movements. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pages 348–373, Washington, DC, USA, May 21-22, 2002.
- [Gavrila, 1999] D. M. Gavrila. The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.
- [Herda *et al.*, 2002] Lorna Herda, Raquel Urtasun, and Pascal Fua. An Automatic Method for Determining Quaternion Field Boundaries for Ball-and-Socket Joint Limits. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pages 95–100, Washington, DC, USA, May 21-22, 2002.
- [Kale *et al.*, 2002] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Krueger. Human Identification Using Gait. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Washington, DC, USA, May 21-22, 2002.

- [Moeslund and Granum, 2001] Thomas B. Moeslund and Erik Granum. A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding: CVIU*, 81(3):231–268, 2001.
- [Oka *et al.*, 2002] Kenji Oka, Yoichi Sato, and Hideki Koike. Real-time Tracking of Multiple Fingertips and Gesture Recognition for Augmented Desk Interface Systems. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pages 429–343, Washington, DC, USA, May 21-22, 2002.
- [Ren and Xu, 2002] Haibing Ren and Guangyou Xu. Human Action Recognition in Smart Classroom. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pages 417–422, Washington, DC, USA, May 21-22, 2002.
- [Sidenbladh *et al.*, 2002] Hedvig Sidenbladh, Michael Black, and Leonid Sigal. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In *Proc. European Conf. on Computer Vision*, number 2350 in LNCS, pages 784–800, Copenhagen, Denmark, June 27-31, 2002. Springer.
- [Sullivan and Carlsson, 2002] Josephine Sullivan and Stefan Carlsson. Recognizing and Tracking Human Action. In *Proc. European Conf. on Computer Vision*, number 2350 in LNCS, pages 629–644, Copenhagen, Denmark, June 27-31, 2002. Springer.
- [Wang *et al.*, 2002] Chunli Wang, Wen Gao, and Shiguang Shan. An Approach Based on Phonemes to Large Vocabulary Chinese Sign Language Recognition. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pages 411–416, Washington, DC, USA, May 21-22, 2002.
- [Wren and Pentland, 1998] C. Wren and A. Pentland. Dynamic Models of Human Motion. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pages 10–15, Nara, Japan, April 14-16, 1998.

## A SMART ANALOG-TO-MIDI/USB INTERFACE

**Gabriele Boschi** – CNUCE - Pisa

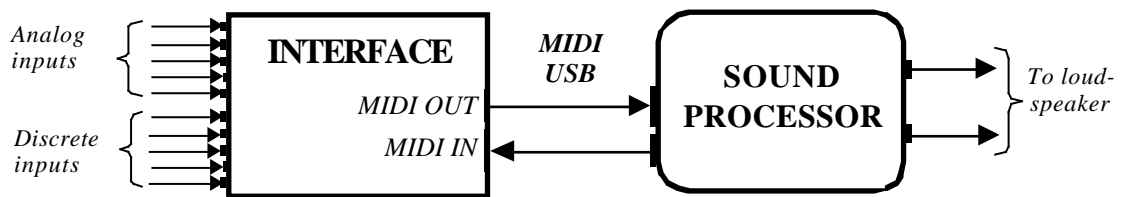
[gabrieleboschi@interfree.it](mailto:gabrieleboschi@interfree.it), [gabriele@daimi.au.dk](mailto:gabriele@daimi.au.dk)

**Brian Mayoh, Steffen Brandorff, Morten Breinbjerg** – DAIMI – Aarhus

**Wayne Siegel** – DIEM – Aarhus

Mosart funded, task 4 relevant

The first phase of the work comprises the development and the practical realization of an **interface for converting analog and discrete values into MIDI or USB data stream**, as shown in the following figure.



Gestures, movement and actions, captured by different kind of sensors (light, traction, temperature, pressure, touch, rotation etc.) are translated into data stream for a generic sound processor, in order to create interactive and creative sound spaces, where sensors can be placed in various places all around. The destination device, that in the first approach can be a Personal Computer, maps received data into sound synthesis parameters.

This interface is similar to some others available in commerce, but has the properties to be less complicated, easier to use and perhaps more appropriate for not too complex applications.

FPGA/PAL,  $\mu$ P and PIC implementations have been evaluated.

FPGAs or PALs offer low cost implementation, high speed, with the benefits of high-level features description using HDL, but at the other side do not provide needed data processing power and flexibility.

Microprocessors offer a high data processing power, but the related hardware and software is complex, and above all implies high developing time delay.

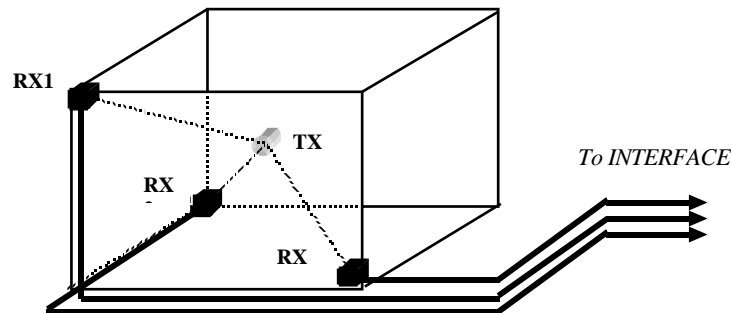
A *Microchip PIC* solution has instead been judged as a good choice. The ready to use analog and digital ports, the wide on board FLASH and EEPROM memories and the on board USART make the hardware not a critical point. Furthermore, software control allows quite powerful data processing, high reliability and high flexibility and adaptability to various users' requests.

At first, a MIDI ports are being used, for their easy implementation and easy ways of use, but since their own limitations in bandwidth and flexibility, USB I/O are to be considered.

The interface is equipped with a MIDI IN port, in order to obtain an easy and *on the fly* configuration of various parameters; in this way is so possible to set up, via MIDI, the output

data format, the thresholds on the analog channels, the inversion function, the different kind of linearization etc.

The second phase of the work regards the development of a **3D object position tracking system** using radio frequency signals. An RF transmitter (in the performer's hand) and more RF receivers (at the vertices of the room) will be used, the latter connected, with ad hoc circuits, to the above mentioned interface, as shown in the following figure.



This part, in detail still to be evaluated, will comprise study and practical experiences on electro-magnetic fields.

It is expected, at the end of the first and both jobs, a demonstration respectively with a real interactive sound space and with the 3D tracking system.

---

## Reference

Gabriele Boschi  
A smart Analog-to-MIDI / USB Interface  
DAIMI Technical Report,  
Computer Science Department, University of Aarhus, Denmark

Building A Hand Posture Recognition System  
From Multiple Video Images:  
A Bottom-Up Approach

Declan Murphy  
cART Lab, CNR, Pisa.  
`declan@diku.dk`

15 March 2002



## Abstract

This report presents an overview of work carried out during a visit to the cART lab of CNUCE, CNR, Pisa, under the MOSART TMR project. The task was to investigate the feasibility of developing a system for accurate real-time recognition of hand posture from multiple video images, and to set about developing such a system. This system is to be suitable for application to real-time control of computer and electronically generated music.

**Keywords:** Hand Posture, Gesture Capture, Computer Vision, Gestural Control, Musical Interface.

## Acknowledgments

I want to thank very much Leonello Tarabella, Graziano Bertini, Gabriele Boschi, and all the team in Pisa for their generous hospitality and help. My visit to Pisa was both enjoyable and productive, and I hope this comes across in this report to some extent.

I owe some further thanks to Claus B. Madsen and Mads Sørensen of Ålborg University for providing me with some useful pointers to related background work, and for the stimulating exchange of ideas.

I also wish to thank Antonio Camurri and all the team at the InfoMus lab, DIST, University of Genoa, for allowing me the flexibility to finish the constraints in §4.2 during my visit there. [This section was revised shortly after the original report.]

I would also like to express thanks to Jens Arnsparang and Kristoffer Jensen for initiating the MOSART project, and particularly for “adopting” me as their PhD student.

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Background . . . . .	5
1.1.1	Manual Dexterity and Expressivity . . . . .	5
1.1.2	The Handel System . . . . .	6
1.1.3	The Approach . . . . .	6
1.2	Literature Review . . . . .	7
<b>2</b>	<b>Platform</b>	<b>9</b>
2.1	Background . . . . .	9
2.2	Linux . . . . .	10
2.3	EyesWeb . . . . .	10
<b>3</b>	<b>Image Pre-Processing</b>	<b>12</b>
3.1	Canny Edge Detection . . . . .	12
3.1.1	How it Works . . . . .	13
3.1.2	The Y-Junction Effect . . . . .	13
3.2	Edge Ranking . . . . .	13
3.2.1	Perimeter Extraction . . . . .	14
3.2.2	Bridging Gaps . . . . .	15
3.2.3	Ranking . . . . .	16
3.3	Finger Recognition . . . . .	17
<b>4</b>	<b>Physical Model</b>	<b>18</b>
4.1	The Model . . . . .	18
4.2	Constraints . . . . .	19
4.3	The Palm . . . . .	20
<b>5</b>	<b>Fitting the View(s) to the Model</b>	<b>23</b>
5.1	Calibration . . . . .	23
5.2	The Geometry of Reconstructing the 3D Image . . . . .	24
5.2.1	The Trivial Case, A Point . . . . .	24

5.2.2	A Line Segment . . . . .	25
5.2.3	Two Views, Convex Polyhedra . . . . .	25
5.2.4	Fleshy Segments . . . . .	26
5.2.5	Articulation . . . . .	27
5.2.6	The Digital Calculus . . . . .	27
<b>6</b>	<b>Further Work and Conclusion</b>	<b>29</b>
6.1	Further Work . . . . .	29
6.1.1	Completion . . . . .	29
6.1.2	Animation . . . . .	29
6.1.3	Audio/Musical Parameters . . . . .	29
6.2	Conclusion . . . . .	29

# List of Figures

2.1	EyesWeb Patch using the Image Pre-Processing block. . . . .	11
3.1	Canny Edge Detection and its “Y-Junction” effect. . . . .	13
3.2	Edge Ranks . . . . .	14
3.3	Maintaining the “last outer pixel”. Each pixel of the perimeter is represented by a square, with a straight line segment from its centre to the centre of its last outer pixel, rotating anticlockwise. . . . .	15
3.4	Bridging the perimeter gaps. . . . .	16
3.5	A counter example of why rotation must be backwards, not onwards through the edge. . . . .	17
4.1	Transformations between representations of the hand. . . . .	19
4.2	The Physical Model of the Hand. . . . .	21
5.1	How eyes and cameras perceive 3D projected onto 2D. . . . .	23
5.2	The Trivial Case: a single point. . . . .	25
5.3	Twin View, Convex Polyhedra Fitting. . . . .	26
5.4	Thick Vector Fitting. . . . .	28
5.5	Articulated Segment Fitting. . . . .	28

# Chapter 1

## Introduction

This report outlines the status of the research carried out by the author during a visit to the Computer Art Lab (cART) of CNUCE, CNR, Pisa, under the MOSART Transfer and Mobility of young Researchers (TMR) EU research network project.

The object of the project was to investigate the feasibility of developing a system for accurate real-time recognition of hand posture<sup>1</sup> from multiple video images, and to set about developing such a system. This system is to be suitable for application to real-time control of computer and electronically generated music.

### 1.1 Background

#### 1.1.1 Manual Dexterity and Expressivity

Many evolutionists consider that the dexterity of the human hand (in particular, its opposable thumb) was the single most decisive factor contributing to our advantageous evolution amongst the primates. The hand is certainly the most articulate part of the human body when it comes to physical manipulation<sup>2</sup>, and it is well known in anatomy that the human hand has, for its size, many times more nerve endings than almost any other part of the body.

---

<sup>1</sup>The word *posture* is used in this report to refer to a particular (mutually relative) position of the hands and fingers, as opposed to the word *pose* which is more commonly used in its stead in the literature. In non-technical English, the word *pose* refers – not so much to the position itself – but to its affected impression on other people. This is not (yet) a concern at the low level of the system described in this report. Therefore, *posture* is preferred.

<sup>2</sup>The word *manipulation* comes from the Latin *manus*, meaning hand.

Next to facial gestures, manual gestures rank as perhaps the most important and surely the most common sort. In [8, ch. 2], McNeill presents a survey (collating earlier work of Morris and Kendon) of twenty of the most ubiquitous gestures worldwide: all of them involve hand movements, most exclusively so.

Practically all types of musical instruments (the voice being about the only exception) are controlled manually – for many, their expressive capacity is entirely by finger control.

Despite all the recent interest in gesture amongst the music informatics/technology field, there remains no satisfactory definition of what a gesture actually is [22].

However, it does seem that – whatever a gesture actually is – it can be completely determined (albeit at a lower level than is usually intended) by describing the posture of the relevant body parts over the time from the preparatory phase – through the stroke phase – to the retraction phase [7].

Thus it would seem that a hand posture recognition system should be a good starting point for investigation of gestural expression in control of music.

### **1.1.2 The Handel System**

This work started out by considering extending the Handel[19] system, which is based on matching real-time silhouette spectra to spectra of known hand postures. The extension would be from a single view of a single hand, to two hands using two (or more) cameras. However, it soon became apparent that there is just too much hidden detail (occlusion) in the way that the fingers overlap each other in most postures from most views. (Also the palm occludes the fingers when the hand is closed into a fist (again, from most views).)

### **1.1.3 The Approach**

Everybody else (in the literature survey, §1.2) has used “stab in the dark” or non-linear dynamical estimation methods (for determining hand posture from video images). The approach of this work has been to try an analytical, linear and projective geometrical, well-founded approach (at least until unfeasible complexity is run into, or time is run out of).

1. First of all, a good physical model of the hand is constructed, with a neat representation of its possible posture, orientation and position.

2. Next, as an image pre-processing stage, the outline of the hand (with its overlapping fingers) is extracted (for each frame of each camera view).
3. Then, by considering the inverse projection of the hand images, the camera views are reconciled with the model to give the required hand posture, orientation and position.

The reconciliation (3 above), is the most challenging stage to develop. The work began by considering the projection of a single point to a camera image, and generalises up to the full hand model (§5).

## 1.2 Literature Review

A good deal of the literature concerns recognition of (a rather limited number of) specific hand postures as with recognition of sign language or for control gestures. This type of system usually uses neural networks or hidden Markov models (HMM). However, as they provide no parametric representation of generic hand posture, they will not be considered further here.

The GREFIT[14] system developed by Nölker and Ritter is based on a technique they developed[13]: using a sequence of up to three Local Linear Mapping (LLM) neural networks per finger, operating on a feature list of Gabor and Gaussian filters applied to the (pre-processed) images. They are initially only concerned with locating the fingertips in two dimensions; subsequently they apply a Parametrised Self-Organising Map (PSOM) to estimate the finger joint angles. Their system is also relevant in that it has been applied to modulating audio parameters and formants of synthesised voice[14, URL].

Shimada et al. [16] report impressive results using only monocular silhouette images. Their approach basically consists of progressively reducing covariance ellipsoids generated by fitting progressively finer models.

Their initial, coarse, feature extraction consists of mid-wrist and finger tip location, with which they measure the closeness of matching of generated candidates. Various tactics are used to reduce the number of viable candidates, and Bayesian matching is used to reduce to some of the most likely candidates.

The finer scale fitting starts with a comprehensive feature list. It then uses the physical constraints of the hand model to introduce truncation (according to the corresponding constraint inequalities) into an Extended Kalman Filter (EKF). Finally, they repeat this procedure for the other most likely candidates with a post-fit-evaluation. (The rationale here is that, if the wrong

choice is made due to ambiguity (or inaccuracy), the system can subsequently correct itself to one of the other likely candidates.)

A different approach, which begins with projective geometry, is that proposed by Stenger et al. in [17]. Their model consists entirely of truncated quadrics (ellipsoids and cones, which have a neat, already understood, projective geometry). Then they use the Unscented Kalman filter (UKF, which they claim is superior both to the EKF and to random sampling (Monte Carlo) methods) to estimate the best-fit posture by iteratively projecting the modeled hand (after dealing with self-occlusion) corresponding to the (iterated) state onto the detected edges of the observed image.

Delamarre and Faugeras propose a method based on 3D reconstruction from stereo images<sup>3</sup>. They use a calibration algorithm first, to estimate the camera settings, and then another computer vision algorithm to reconstruct the 3D scene from the stereo images.

Their model is made of truncated cones and spheres for the fingers only (i.e. omitting the palm), and they begin their fitting algorithm by assuming that the initial posture is known. (They use a Kalman filter to predict the user’s posture between subsequent frames.) From there, they use the Iterative Closest Points (ICP) algorithm to simulate attractive forces between points on the model and points on the finger surfaces. To overcome the problem of the forces cancelling each other out as the model and reconstruction intersect, they shunt the intersecting reconstruction forces along the normal (a technique known as Maxwell’s demons).

For each iteration of convergence, they use a technique from robotics to translate the model according to the above forces. As explained by Schwertassek and Roberson, they simplify the computation of the acceleration of each component of the model by representing the system as a graph and then having three layers of forward and backward recursion.

Ueda et al. [20] use a rather neat technique<sup>4</sup> to construct a discrete 3D representation from multiple silhouettes using a tree representation. They also use a simulated force technique in their algorithm to fit the constructed “voxel” representation to a skeletal model.

---

<sup>3</sup>By stereo images, they mean either two cameras not far apart or one camera with a mirror image from not far away – both views being of the same foreground object(s).

<sup>4</sup>They refer to their technique as “building a voxel model by octree representation”.



# Chapter 2

## Platform

### 2.1 Background

It was considered desirable from the outset that the source code of the system should be in C++. This would readily facilitate its integration into existing cART systems (e.g. Pure C Music (PCM)[18]), the potential development environments (EyesWeb, Microsoft Windows, Linux) and the authors own performance tools project, as well as being an all-round suitable language for real-time processing and portability.

The hardware available for development was:

- a laptop computer (ACER *TravelMate 525TXV*),
- 2 USB WebCams (Logitech QuickCam Pro 3000).

It is desired that the eventual system should run in real-time. There is a problem here with the latency of the USB bus of about 200 ms from camera to computer. However a capture rate of up to 30 frames per second is possible, so that the hardware listed is good enough for development.

The future possibility of using high-end custom-built image-DSP hardware (currently under development at CNR, Pisa) for this system has arisen, thereby substantiating the feasibility of a more analytic approach working in real-time.

In any case, the final working system with source code will be downloadable by anonymous FTP from [11] under the GNU General Public License (GPL)<sup>1</sup>. In order to run on different platforms, there will be some low level camera driver API issues specific to the local operating system and hardware; otherwise it should only be a matter of recompiling on the local host machine.

---

<sup>1</sup>The GPL may be viewed/downloaded from <http://www.gnu.org/licenses/gpl.txt>

## 2.2 Linux

A Linux driver for these WebCams, [12], was not available when the project started, but became available soon afterwards. The operating system was installed (with a dual boot configuration). Reading frame information from several cameras at once is trivial once the platform is appropriately configured first. This configuration consists of having:

- a correct kernel release version:
  - recent enough for the driver to support the particular camera type,
  - a kernel that is not specific to any particular distribution,
- an appropriate kernel configuration:
  - having loadable module, video4linux and USB support compiled into the kernel,
  - having the USB and camera drivers compiled into the kernel as loadable modules,
- loading the PWC camera driver and PWCX<sup>2</sup> camera decompressor modules into the running kernel,
- loading the USB driver module *after* the above device drivers:
  - uhci or ohci, as appropriate to your needs.

The Advanced Linux Sound Architecture (ALSA) was chosen as the audio driver layer on top of Linux.

## 2.3 EyesWeb

EyesWeb[2], a software package developed by our MOSART partners at the DIST lab in Genoa, provides a very nice prototyping environment for video gesture applications (see figure 2.1). However, it only runs on recent versions of Microsoft Windows, which cannot (at the time of writing) recognise more than one video device at a time under the video4windows protocol used.

The latest release of EyesWeb (version 2.4.1) prefers the Windows Device Manager (WDM, another protocol) under which dual framegrabber input has been reported with various low level hacks – hardware multiplexing is

---

<sup>2</sup>The PWCX decompressor module does not come with the kernel source, but may be freely downloaded from [12].

the recommended approach by the EyesWeb team (hardly practical for USB WebCams).

The developer kit (SDK)<sup>3</sup> for the WebCams was downloaded from the Logitech developer web-site. This allows development of stand-alone programs using the Logitech WebCams. Reading information from two cameras is possible, but not at the same time: the restriction of accessing only a single video device at a time is at the level of the operating system. It is possible to switch alternately between cameras, but not in real-time.



Figure 2.1: EyesWeb Patch using the Image Pre-Processing block.

---

<sup>3</sup>The Logitech WebCam SDK is currently only available for recent releases of MS Windows. Their WebCam drivers are also available for recent versions of MacOS and Linux on IBM style PC and Apple PowerPC.

# Chapter 3

## Image Pre-Processing

Before trying to fit the image to the model, the image is pre-processed to simplify the task by extracting (as far as possible) only key information. A binary image tracing only the outline of the hand with the fingers is extracted from the 24-bit colour camera image. This takes place in three stages: (§3.1) edge detection, (§3.2) edge ranking, and (§3.3) finger detection.

### 3.1 Canny Edge Detection

One standard, or simple, approach to finding the perimeter of an object is to take its silhouette, which is very easy to compute if (as in this case) there is a single foreground object with a known or controlled background. However, the exact position of the silhouette border varies with the thresholding level used to compute it, and with both the source direction and the level of the ambient lighting.

A more accurate approach – almost immune to these weaknesses – is to use edge detection. As we may be dealing with low resolution images (to enhance real-time performance), such accuracy is desirable. Another significant advantage is the ability to make out some detail of the overlapping fingers instead of perceiving just a single blob.

Edge detection, in computer vision, is defined as the process of assigning a value to each pixel of an image in proportion to the likelihood that the pixel is on the boundary between two regions of different intensity values.

The Canny edge detection algorithm[3] is used – as implemented in the Intel Open Computer Vision Performance Library (OpenCV)[6]. This was chosen for its sharp results (in comparison to rival edge detection techniques), and for its efficient implementation (which was readily available).

### 3.1.1 How it Works

First the image is smoothed by Gaussian convolution; then a simple 2D first derivative operator is applied to the smoothed image to highlight regions of the image having high first spatial derivatives. Edges give rise to ridges in the gradient magnitude image. The algorithm then tracks along the top of these ridges and sets all pixels that are not on top of the ridge to zero, so as to give a thin line in the output.

Two thresholds limit the tracking:  $t_1$  and  $t_2$ , with  $t_1 < t_2$ . Tracking can only begin at a point on a ridge higher than  $t_2$ . Tracking then continues in both directions out from that point until the height of the ridge falls below  $t_1$ . This hysteresis helps to ensure that noisy edges are not broken up into multiple edge fragments.

### 3.1.2 The Y-Junction Effect

One problem with the basic Canny operator is that of so-called “Y-junctions”: places where three (or more) ridges meet in the gradient magnitude image. Such junctions occur where an edge is partially occluded by another object. The tracker will treat two of the ridges as a single line segment, and the third one as a line that approaches, but doesn’t quite connect to, the other two. See figure 3.1.



Figure 3.1: Canny Edge Detection and its “Y-Junction” effect.

## 3.2 Edge Ranking

A rank (low non-negative integer) is associated with each edge according to the following rule:

- 0 corresponds to edges on the perimeter,
- 1 corresponds to edges terminating on the perimeter which are not of rank 0,
- $n$  corresponds to edges terminating on an edge of rank  $n - 1$  which are not of rank  $< n$ .

See figure 3.2.

Later stages of processing have the option of having an image frame composed of all edges up to a specified rank, thereby selecting the appropriate level of detail.

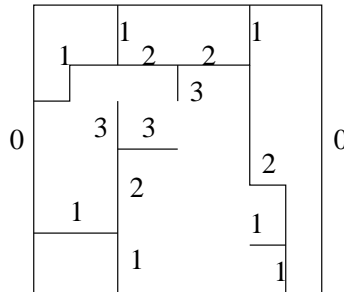


Figure 3.2: Edge Ranks

### 3.2.1 Perimeter Extraction

Taking the edge detection from above §3.1, the perimeter is in turn extracted. While this may sound trivial – as indeed it is to our human eyes – it is not so for the computer: there may be (and in general will be) some spurious noise outside of the hand, and the perimeter will not be continuous in general.

First the image is scanned, from an outer edge of the frame inwards, for any “blob” in such a manner that any enclosed shape will be approached from the outside (even if this shape is not connected<sup>1</sup>).

Next, the perimeter of the found shape is traced, paying careful attention to always keep to the *outside* of the shape while considering the next pixel along the perimeter. This is achieved by:

1. maintaining the concept of “the last outer pixel”, for each pixel as we trace along the perimeter,
2. rotating, from the last outer pixel to the next pixel of the shape, in a direction (clockwise or anti-clockwise) consistent with the angular polarity of the trace.

<sup>1</sup>The term *connected* is borrowed from topology: here it simply means that the shape in question can be entirely traced through adjacent pixels.

In other words, every time we locate another perimeter pixel, we record the last non-perimeter pixel (on the outside). In order to find the next perimeter pixel, we rotate about the immediate neighbouring pixels – starting with the next pixel after the last outer pixel, in a consistent direction of rotation. See figure 3.3.

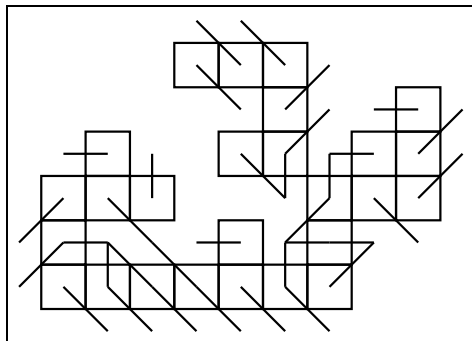


Figure 3.3: Maintaining the “last outer pixel”. Each pixel of the perimeter is represented by a square, with a straight line segment from its centre to the centre of its last outer pixel, rotating anticlockwise.

We are assuming here that edges outside of the hand (if any) have been removed or are disregarded in the processing. This may be achieved by such techniques as gauging the length, whether it lies close to other edges, etc., as described in, for example, [10]. This has not been a problem in the prototype.

### 3.2.2 Bridging Gaps

Now if we try to trace along the outside of the shape we have found, we will find that, very often, the trace only runs around a section of the desired hand perimeter because of gaps left by the Canny edge detection (§3.1.2).

Such gaps come in various forms, so that a general technique to bridge them all is required. Some examples taken from observation of real data appear in figure 3.4. The technique found to cover all cases, tested with real hand data, is as follows:

1. Starting with the first perimeter pixel found, continue tracing anti-clockwise.
  - (a) If trace describes a full circuit, then we are finished.
  - (b) Otherwise record the end pixel and the last outer pixel of the second<sup>2</sup> from end pixel.

---

<sup>2</sup>We use the last outer pixel of the second to end pixel because that of the end pixel

2. From the initial pixel, trace clockwise.
  - (a) Record the end pixel and the second<sup>2</sup> from end last outer pixel.
3. For all ends, try to join them.
  - (a) Rotate from outside to inside, scanning at a radius of a given size<sup>4</sup>.
  - (b) If another segment is located, join to the nearest pixel on it.
  - (c) If this new segment is already on the list, merge it (on the list) with the last one; otherwise, trace the extent of the new segment, updating its end data.
4. For any remaining segments (i.e. if all segments end in spurious diver-  
sions), try to join them.
  - (a) Back-trace along the segment away from the end, testing within the aperture from outside to in, and otherwise proceed as in 3.

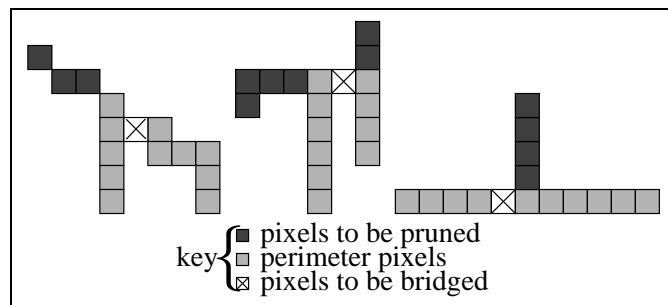


Figure 3.4: Bridging the perimeter gaps.

### 3.2.3 Ranking

If the desired rank is greater than zero, then all perimeter ranks are first assigned their zero rank. Then, using the above gap bridging technique (without heed to the concept of outside anymore), all adjoining edges are given rank one. This process is repeated until the desired rank is attained.

will have swung around into the inside. Recovery of the outside may be made by rotating back<sup>3</sup>(if the segments tend to be long) or by placing the last outer pixels in a two cell Last In First Out (LIFO) buffer (if the segments tend to be very short).

<sup>3</sup>In rotating backwards, the temptation to take the short cut of rotating on through the edge must be resisted! See figure 3.5.

<sup>4</sup>There is a certain radius (or aperture) size observed to be both large enough to bridge all Y-junctions and small enough to not interfere with other edges in the vicinity, for a fixed image resolution and a fixed distance from camera to hand.



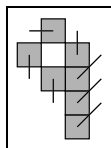


Figure 3.5: A counter example of why rotation must be backwards, not onwards through the edge.

### 3.3 Finger Recognition

Knowing the rough position/orientation of the fingers (e.g. from previous frames), it is not difficult to imagine a technique for selecting those edges which correspond to the perimeters of the fingers (excluding knuckles, skin creases, nails and other edges): particularly since we already have a technique for extracting perimeters and edges of a given rank.

The accurate outline that we have extracted so far is enough to begin reconstructing a representation of the hand in three dimensions – the outline of overlapping fingers can be considered at a later stage. As the fitting of the 3D observation representation to the model is the most challenging and problematic stage of the system, the choice was made to focus on that as soon as the feasibility of the earlier stages was assured.

# Chapter 4

## Physical Model

The model of the hand is based on that of [9] (although a completely different fitting algorithm is used), and is not much different from any other hand model used in previous literature except for its more precise formulation of the constraint between adjacent fingers, and of the non-interference of fingers<sup>1</sup>. Only those postures attainable by a single hand, uninfluenced by any other force than its own muscles, without straining itself, are considered. (It is possible to place the metacarpophlangeal joints out of alignment, for example, by tightly clenching a fist, by grasping a solid object, or by external force.)

### 4.1 The Model

The precise measurements are based on the authors own hands, and remain to be verified for generality with people of different gender and age. They do, however, correspond well to those models published in the literature.

The basic premise is that any hand posture can be represented by specifying the position and orientation of the palm, and the angles of the joints, and that we can transform between this representation and all others used and required by the system: see figure 4.1. To this end we have a skeletal model consisting of sixteen rigid bodies (three bones per finger plus the palm) joined by fifteen joints (knuckles) as shown in figure 4.2. Considered individually, all the proximal and distal interphlangeal joints move in one plane, and all five metacarpophlangeal joints move in two planes.

Thus the physical joint representation may be stored as a  $4 \times 5$  matrix  $\Theta \equiv \theta_{i,j}$ ,  $0 \leq i \leq 4$ ,  $0 \leq j \leq 3$ , where  $i, j$  correspond to the angles as per

---

<sup>1</sup>The term *finger* is often taken to exclude the thumb, but – for the purposes of this report – the hand has five fingers.

i	finger	j	movement
0	thumb	0	lateral metacarpophlangeal
1	index	1	axial metacarpophlangeal
:	:	2	axial proximal interphlangeal
4	little	3	axial distal interphlangeal

Table 4.1: The physical joint representation.

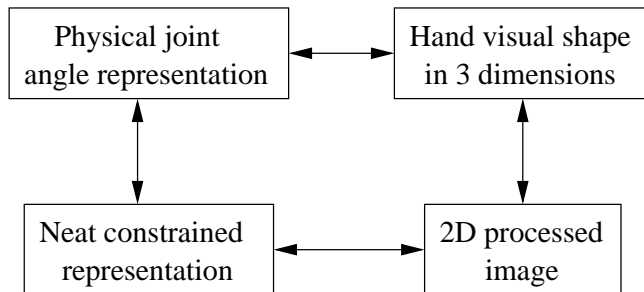


Figure 4.1: Transformations between representations of the hand.

table 4.1.

## 4.2 Constraints

[This section was revised after the original report date, during the author’s subsequent visit to the InfoMus lab, DIST, University of Genoa, during March–April 2002.]

However, there are various constraints on the joints when considered as a system.

First of all, there are limitations on how far the joints will flex (without being hindered by any other constraint). These  $\theta_{i,j}^{\min}$  and  $\theta_{i,j}^{\max}$  are recorded in table 4.2.

Then, for each finger  $i$ , the range of movement of  $\theta_{i,0}$  is constrained by  $\theta_{i,1}$ :

$$\theta_{i,0}^{\min} \left| \frac{\theta_{i,1} - \theta_{i,1}^{\max}}{\theta_{i,1}^{\max}} \right| \leq \theta_{i,0} \leq \theta_{i,0}^{\max} \left| \frac{\theta_{i,1} - \theta_{i,1}^{\max}}{\theta_{i,1}^{\max}} \right| \quad \forall i, 0 \leq i \leq 4. \quad (4.1)$$

Next,  $\theta_{i,2}$  and  $\theta_{i,3}$  are tightly coupled as follows:

$$\text{For } i = 0, \quad \frac{d\theta_{i,3}}{d\theta_{i,2}} \approx 4 \quad \text{independent of } \theta_{i,j}, \quad j \neq 2, 3. \quad (4.2)$$

$$\text{For } 1 \leq i \leq 4, \quad \frac{d\theta_{i,3}}{d\theta_{i,2}} \approx 2 \quad \text{independent of } \theta_{i,j}, \quad j \neq 2, 3. \quad (4.3)$$

However, these relations from (4.2) and (4.3) become inaccurate as  $\theta_{i,2}$  approaches its limits (in table 4.2) and the tendons become strained.

For all  $i$ ,  $0 \leq i \leq 4$ ,  $\theta_{i,3} = 0 \Rightarrow \theta_{i,2} = \hat{\theta}_{i,2}$ , and these  $\hat{\theta}_{i,2}$  are recorded in table 4.3.

For all  $i, i'$ ,  $i' = i \pm 1$ ,  $1 \leq i, i' \leq 4$ , we have  $|\theta_{i,0} - \theta_{i',0}| \leq \hat{\theta}_{|i,i'|,0}$ , for some constants  $\hat{\theta}_{|i,i'|,0}$ . (Of course,  $\hat{\theta}_{|i,i'|,0} = \hat{\theta}_{|i',i|,0}$ .) Similarly for  $\hat{\theta}_{|i,i'|,1}$ , except that in this case the finger motion is no longer co-planar – and in fact the constraint is not symmetric. The convention is that  $\theta_{i,1} - \theta_{i',1} \leq \hat{\theta}_{|i,i'|,1}$  when  $\theta_{i,1} \geq \theta_{i',1}$  (i.e. when finger  $i$  is more towards a closed fist than finger  $i'$ ). These measurements may be found in table 4.4.

The remaining constraints are due to the interference of the fingers with each other and with the palm. Before formalising this, it is convenient to introduce “fleshy segments”, c.f. §5.2.4.

### 4.3 The Palm

The palm is considered to consist of five straight line segments, rigidly fixed together at a common point. A reference frame for its orientation must be specified, which will serve to specify the orientation of the whole hand posture, and the relative angles (in three dimensions) of the palm segments must be specified in this frame (since the global representation of the position and orientation of the fingers depends on it).

The planes of axial movement of the joints 1–4 are fixed such that, as these fingers close, they remain coplanar with the metacarpophlangeal joint of the thumb.

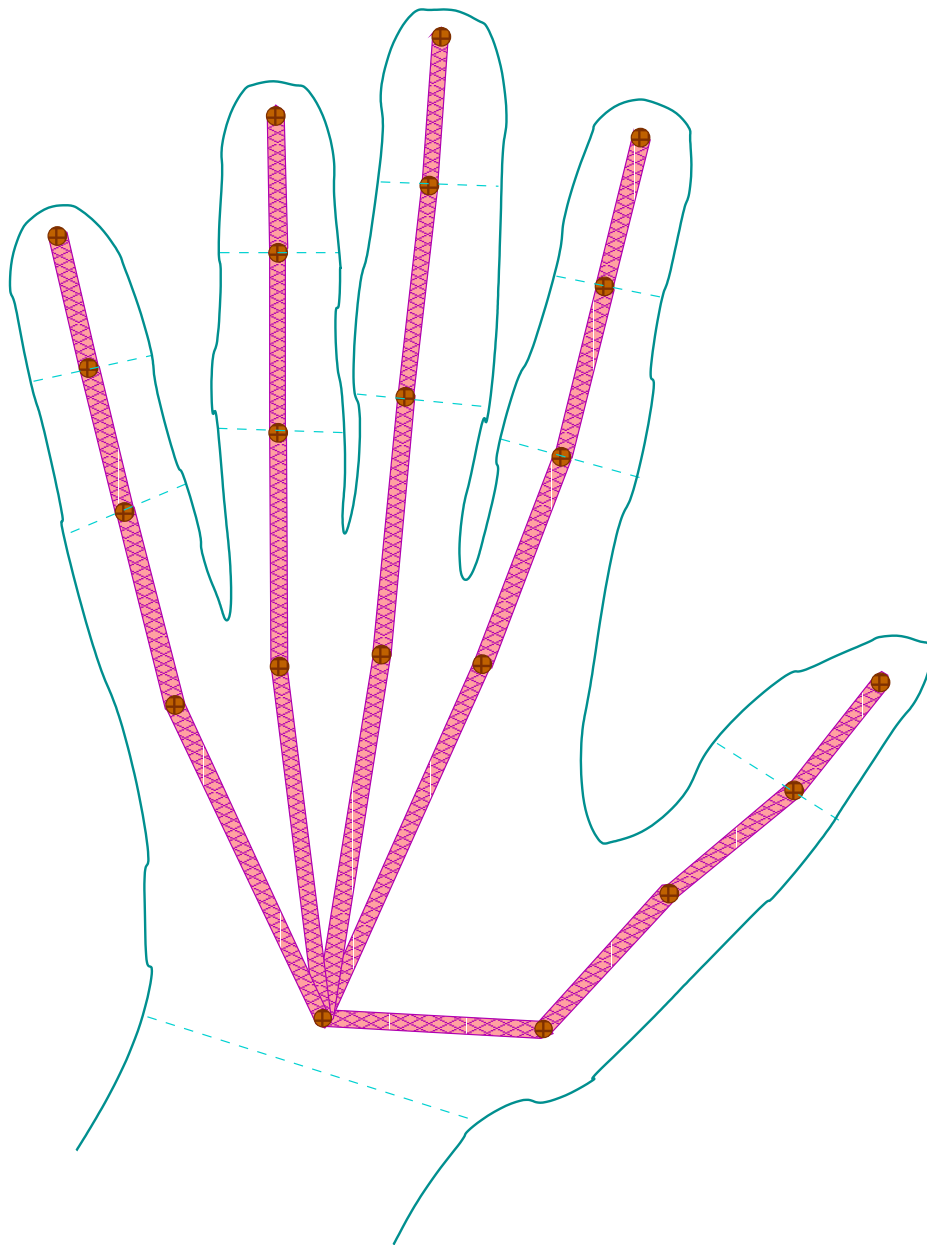


Figure 4.2: The Physical Model of the Hand.

$\theta_{i,j}^{\min}$	0	1	2	3
0	-0.288	-0.262	-0.026	-0.175
1	-0.105	-0.105	0.026	-0.044
2	-0.096	-0.148	0.052	-0.052
3	-0.157	-0.087	0.026	0.000
4	-0.183	-0.070	0.026	-0.017

minimum

$\theta_{i,j}^{\max}$	0	1	2	3
0	0.201	0.175	0.227	0.515
1	0.113	0.689	0.855	0.585
2	0.079	0.742	0.899	0.654
3	0.070	0.820	0.916	0.576
4	0.201	0.890	0.812	0.550

maximum

Table 4.2: The maximum and minimum values of  $\theta_{i,j}$  in radians, where 0 corresponds to straight ahead. For axial angles, positive values correspond to a closing fist. For lateral angles,  $\theta_{i,0}$ , positive angles correspond to towards the thumb (away from the palm in the case of the thumb), and 0 corresponds to the relaxed position. All angular measurements are considered accurate to  $\pm 0.010$ .

$i$	0	1	2	3	4
$\hat{\theta}_{i,2}$	0.079	0.166	0.262	0.332	0.070

Table 4.3:  $\hat{\theta}_{i,2} = \theta_{i,2} \Big|_{\theta_{i,3}=0}$

$i, i'$	$\hat{\theta}_{ i,i' ,0}$	$\hat{\theta}_{ i,i' ,1}$	$\hat{\theta}_{ i',i ,1}$
1, 2	0.166	0.436	0.524
2, 3	0.175	0.393	0.458
3, 4	0.201	0.480	0.419

Table 4.4:  $|\theta_{i,0} - \theta_{i',0}| \leq \hat{\theta}_{|i,i'|,0}$ ,  $\theta_{i,1} - \theta_{i',1} \leq \hat{\theta}_{|i,i'|,1}$ .

# Chapter 5

## Fitting the View(s) to the Model

It is generally maintained that one of the most important factors for the inspiration of the Renaissance was the new understanding of the perception of vision, which facilitated a fruitful cross-fertilisation between science and the arts. It is this very understanding of vision that is used here (with analogous hopes for fruitful cross-fertilisation).

Basically, the 3D image is projected onto a single point, and the intersection of the rays of projection with a plane form the image, as illustrated in figure 5.1.

### 5.1 Calibration

Before the location of the hand in 3D space, it is necessary for the program (model) to know the relative positions of the image planes and the vanishing points (as determined by the camera lenses) for each camera. This may be accomplished by an initial calibration step, after the cameras have been set in position and before hand tracking begins. A few simple measurements can be typed in via a user interface, corresponding to the relative positions of the cameras, or an interactive routine requiring the user to indicate the extent

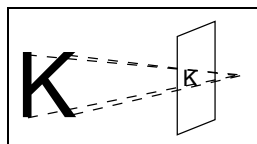


Figure 5.1: How eyes and cameras perceive 3D projected onto 2D.

of each view may be used.

Specifications specific to the camera will need to have been already determined by calibration software, in particular:

- the focal distance,
- the angle of vision, and
- the magnification factor.

Settings specific to the installation (essentially the relative positions and orientations of the cameras) will have to be established by a calibration routine before we can proceed.

## 5.2 The Geometry of Reconstructing the 3D Image

Under these sort of projections, metric distances are not preserved by linear transformations. We must instead generalise up, through similarity and affine geometry, to projective geometry in order to deal with invariant properties under these, projective, transformations. See [15], or [4] for further reading on projective geometry, and [5] for its applications to computer vision.

### 5.2.1 The Trivial Case, A Point

The most simple case to consider is that of a single point in two dimensions. It illustrates the inherent inaccuracy introduced by digital images.

Let  $F = \mathbb{R} \times \{y \in \mathbb{R} : y > w\} \cap \{(x, y) \in \mathbb{R}^2 : \frac{w}{b} < \frac{y}{x} < \frac{w}{a}\}$  be the camera view, for a fixed  $w \in \mathbb{R}$ , and let  $V = [a, b] \subset \mathbb{Z}$  be the image plane for fixed  $a, b \in \mathbb{Z}$ ,  $a < b$ . Define  $P : F \rightarrow V : (x, y) \mapsto \lfloor \frac{wx}{y} + \frac{1}{2} \rfloor$ , so that  $P : \mathbf{x} \mapsto P\mathbf{x}$  is the projection mapping from  $F$  to  $V$  with the origin as the vanishing point. (See figure 5.2 left.)

Now consider the the inverse function (illustrated in figure 5.2 right)

$$P^{-1} : V \rightarrow 2^F : v \mapsto \{\mathbf{x} \in F : P\mathbf{x} = v\},$$

$$P^{-1}v = \{(x, y) \in F : \frac{w}{v + \frac{1}{2}} \leq \frac{y}{x} \leq \frac{w}{v - \frac{1}{2}}\}.$$

Already the most simple case introduces some complexity which will snowball if we carry it through the next stages. Instead we will bear it in mind but out of the equations until we reintroduce it when we come to the fitting in §5.2.6.



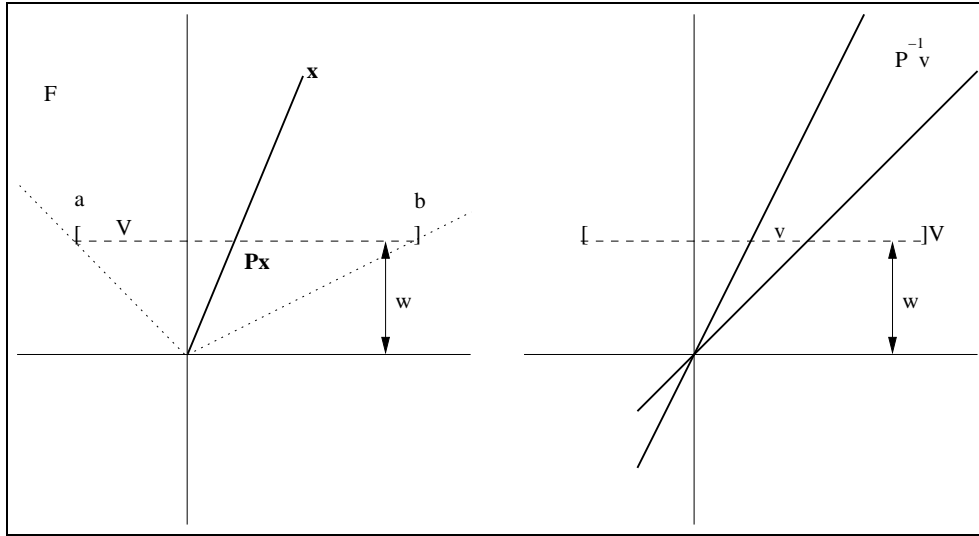


Figure 5.2: The Trivial Case: a single point.

### 5.2.2 A Line Segment

Existence proofs and formulae, for calculating maximum and minimum range along each ray where possible fit may lie, have been established. Also, a formula for calculating exact point(s) of fit on the opposite ray given one point on either ray, has been derived. These are omitted however, as they may not be necessary when multiple view information becomes available.

### 5.2.3 Two Views, Convex Polyhedra

Consider the intersection of the rays from the different views, as illustrated in figure 5.3. Exactly one of a set of co-linear ray intersections is projected. This is not sufficient to eliminate the others. Co-linear sets are not disjoint. There must be at least one vertex fit at an intersection common to co-linear sets (although this does not hold for curved shapes). Fit as follows, maintaining a data structure of how vertices fit:

1. Start offering the vertices at the “corners” for fitting.
2. Continue offering remaining vertices along opposite edges of the intersection quadrilateral (excluding points from a ray common with the corner) testing for fitting by distance.
3. If a fit is found, then continue offering remaining vertices along unused rays, using the opacity constraint.

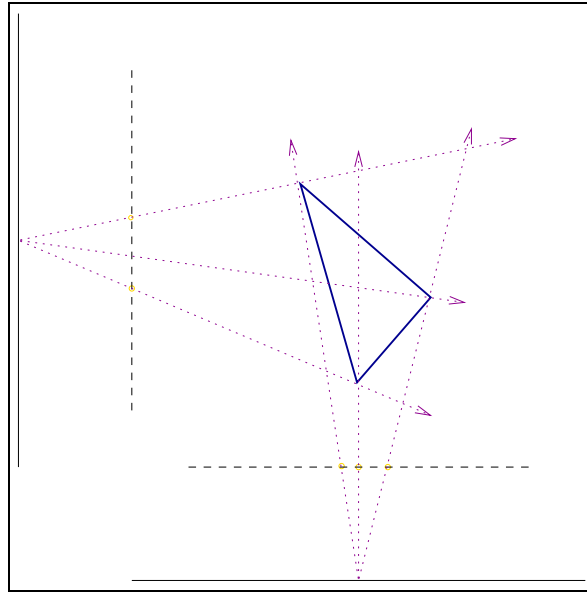


Figure 5.3: Twin View, Convex Polyhedra Fitting.

If more than one vertex lies on a co-linear set, then only the one nearest the camera (origin) is visible. Thus probabilities of finding fits in given areas can be ranked, and fit offering can proceed in this order.

#### 5.2.4 Fleshy Segments

With a bijection between the skeletal and 3D surface representations of the hand as motivation, we introduce the concept of a “thick geometric vector”. The hand can be represented as a balanced tree with the common fixed joint of the palm as root and each bone being represented as a tree node having:

- length,
- angle of attachment to next segment,  $\alpha : -\pi < \alpha < \pi$ ,
- thickness (constant  $r$  for the time being),

where  $\alpha = \pi$  or  $\alpha = \emptyset$  corresponds to a terminal segment.

Let  $K \subset \mathbb{R}^2$  be the line segment corresponding to a node. Then the required bijection (c.f. figure 4.1) is given by

$$f : K \rightarrow f(K) \subset \mathbb{R}^2 : \mathbf{k} \mapsto \{(x, y) \in \mathbb{R}^2 : \|(x, y) - \mathbf{k}\|_2 \leq r\}.$$

Determining the translation and rotation of the tree is the object. See figure 5.4. Consider the region about the end points of diameter  $r$ . Offer each vertex at the center of the tangent circle to both rays of radius  $r$ .

### 5.2.5 Articulation

Now we begin to get realistic by generalising up to allow articulation in the joints. Consider figure 5.5. There are two free variables, with complete occlusion of one joint. A function to calculate the intermediate knuckle location is needed for fitting.

Let  $S = \{s_1, s_2, s_3\}$  be the hand tree,  $d(s_i)$  be the length of  $s_i$ , and  $d(s_i, s_j)$  be the Euclidean distance from the start of  $s_i$  to the end of  $s_j$ . Let  $\alpha_i$  be the actual angular setting of  $(s_i, s_j)$  and  $f(i, j) = d(s_i, s_j)$ . We have

$$f(i, i + 1) = \sqrt{d^2(s_i) + 2d(s_i)d(s_{i+1}) \cos \alpha_{i+1} + d^2(s_{i+1})},$$

and so  $f(i, j)$  can be calculated iteratively.

### 5.2.6 The Digital Calculus

Further generalisations include three dimensions and including information from previous frames to track velocity of hand parts.

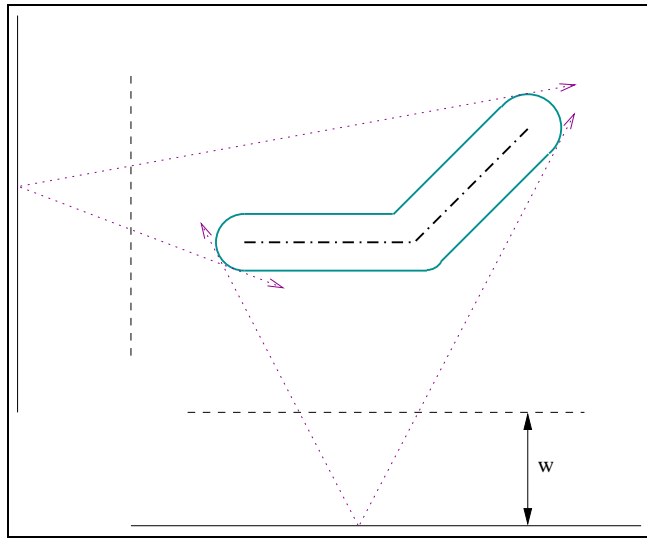


Figure 5.4: Thick Vector Fitting.

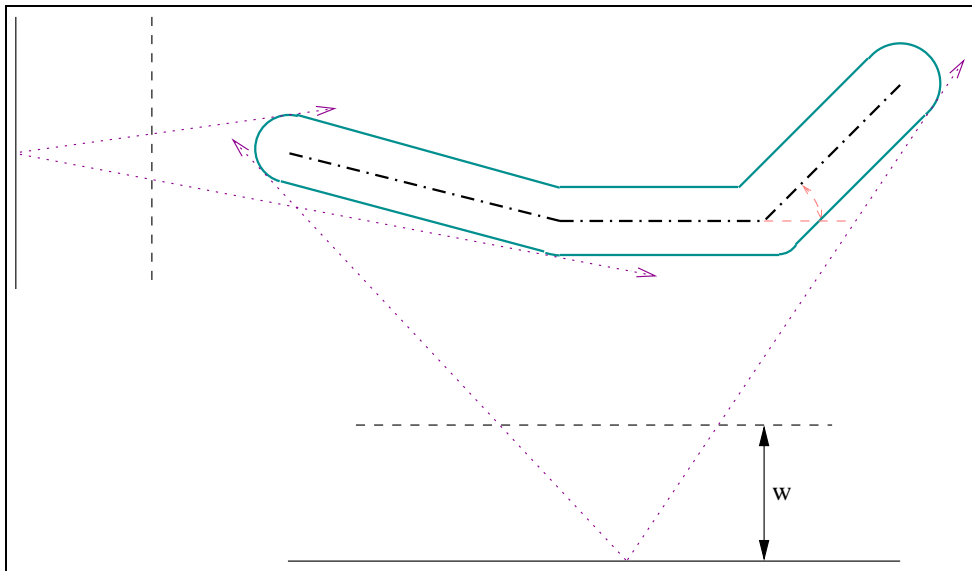


Figure 5.5: Articulated Segment Fitting.

# Chapter 6

## Further Work and Conclusion

### 6.1 Further Work

#### 6.1.1 Completion

The constants of the physical model have to be measured, and the articulate thick segment fitting algorithm has to be fully developed. Further generalisations are refinement will follow.

#### 6.1.2 Animation

A simple visual animated hand output (as well as hand posture, position and orientation data) should be provided for neat closed-loop real-time feedback and verification of results.

#### 6.1.3 Audio/Musical Parameters

As mentioned in the introduction, the intended use of this system is for audio and musical applications. Thus, the next stage of research is to investigate the mapping of manual gestures to the control of real-time audio parameters and to the control of higher level musical parameters.

### 6.2 Conclusion

This project investigates a bottom up approach to building a hand posture tracker for real-time control of computer music.

The on-going “which platform” issues were explored, and prototype and test programs have been developed for the various stages involved.

A new edge ranking technique was developed for image pre-processing in computer vision (which, as a corollary, provides accurate connected perimeter extraction). It is superior to the more common use of silhouettes due to its greater noise tolerance, greater accuracy and greater independence from ambient lighting fluctuations.

A working demonstrable system is well on its way, and may soon be downloaded by anonymous FTP from [11], along with a later edition of this report.

There has already been considerable interest expressed amongst the gesture capture community in both the approach and availability of this system.

# Bibliography

- [1] Claudia Lomelí Buyoli and Ramón Loureiro, editors. *Workshop on Current Research Directions in Computer Music*, Barcelona, Spain, Nov 2001. Audiovisual Institute, Pompeu Fabra University. ISBN: 84-88042-37X.
- [2] Antonio Camurri, Paolo Coletta, Massimiliano Peri, Matteo Ricchetti, Andrea Ricci, Riccardo Trocca, and Gualtiero Volpe. A real-time platform for interactive dance/music systems. In *Proc. Int. Computer Music Conf.*, Berlin, Germany, Aug 2000. ICMA.
- [3] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6), Nov 1986.
- [4] H. S. M. Coxeter. *Projective Geometry*. University of Toronto, Canada, 2nd edition, 1974.
- [5] O. Faugeras. *Three-Dimensional Computer Vision - A Geometric Viewpoint*. Artificial intelligence. M.I.T. Press, Cambridge, MA, USA, 1993.
- [6] Intel. Open computer vision library. WWW.  
<http://www.intel.com/research/mrl/research/opencv/>.
- [7] A. Kendon. Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key, editor, *The relation between verbal and nonverbal communication*, pages 207–227. The Hague, Mouton, 1980.
- [8] David McNeill. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, 1992. ISBN: 0-226-56132-1.
- [9] R. J. Millar and G. F. Crawford. A mathematical model for hand-shape analysis. pages 235–245, York, UK, Mar 1996. Springer. ISBN: 3-540-76094-6.
- [10] Declan Murphy. Extracting arm gestures for VR using eyesweb. In Buyoli and Loureiro [1]. ISBN: 84-88042-37X.

- [11] Declan Murphy. A hand posture recognition system. Anonymous FTP, 2002. <ftp://ftp.diku.dk/diku/users/declan/hand/>.
- [12] “Nemosoft”. Linux drivers for Philips USB webcams (including some Askey, Samsung and Logitech cameras). WWW site. <http://www.smcc.demon.nl/webcam/index.html>.
- [13] Claudia Nölker and Helge Ritter. Detection of fingertips in human hand movement sequences. In Wachsmuth and Fröhlich [21], pages 209–218. ISBN: 3-540-64424-5.
- [14] Claudia Nölker and Helge Ritter. GREFIT: visual recognition of hand postures. volume 1739 of *LNAI*, pages 61–72, Gif-sur-Yvette, France, Mar 1999. Springer-Verlag. For video clips and examples of hand gesture applied to audio synthesis, see also <http://www.techfak.uni-bielefeld.de/~claudia/vishand.html>.
- [15] J. G. Semple and G. T. Kneebone. *Algebraic Projective Geometry*. Oxford Science Publication, 1952.
- [16] Nobutaka Shimada, Yoshiaki Shirai, Yoshinori Kuno, and Jun Miura. Hand gesture estimation and model refinement using monocular camera – ambiguity limitation by inequality constraints. In *Proc. The 3rd Int. Conf. on Automatic Face and Gesture Recognition*, pages 268–273, 1998.
- [17] B. Stenger, P. R. S. Mendonça, and R. Cipolla. Model-based 3D tracking of an articulated hand. In *Proc. Conf. Computer Vision and Pattern Recognition*, Lihue, USA, Dec 2001.
- [18] Leonello Tarabella and Graziano Bertini. Wireless technology in gesture controlled computer generated music. In Buyoli and Loureiro [1], pages 102–109. ISBN: 84-88042-37X.
- [19] Leonello Tarabella, M. Magrini, and G. Scapellato. A system for recognizing shape, position and rotation of the hands. In *Proc. Int. Computer Music Conf.*, pages 288–291, San Francisco, USA, 1997. ICMA.
- [20] Etsuko Ueda, Yoshio Matsumoto, Masakaza Imai, and Tsukasa Ogasawara. Hand pose estimation using multi-viewpoint silhouette images. In *Proc. Int. Conf. on Intelligent Robots and Systems (IROS2001)*, pages 1989–1996, Maui, USA, Oct–Nov 2001. IEEE/RSJ.
- [21] Ipke Wachsmuth and Martin Fröhlich, editors. *Gesture and Sign Language in Human-Computer Interaction: Proc. Int. Gesture Workshop*,



volume 1371 of *LNAI*, Bielefeld, Germany, Sep 1997. Springer-Verlag, Berlin, Heidelberg. ISBN: 3-540-64424-5.

- [22] Alan Wexelblat. Research challenges in gesture: Open issues and unsolved problems. In Wachsmuth and Fröhlich [21], pages 1–12. ISBN: 3-540-64424-5.

## The Votion Project

**Bendik Stang (DTU), Erling Tind (DIKU), Declan Murphy (DIKU)  
Jens Arnsfang (DIKU), Kristoffer Jensen (DIKU),  
Anna-Mette Bach Jensen (DDS), Catherine Beyer (DDS), Michel Gugliemi (DDS)**

**A multidisciplinary artistic virtual reality project**



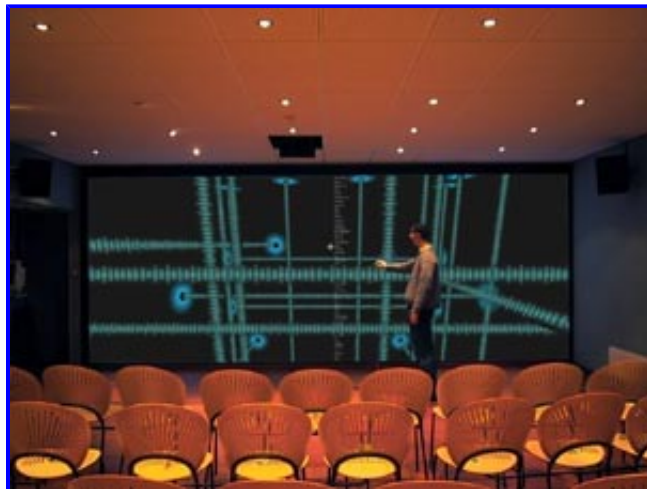
### Description of the project

Votion is an artistic interpretation of the workflow over the internet. The words are from a search engine, and can be accessed through a websp. The application will allow the user to immerse himself into a virtual world where there the words are flowing through space in long lists.

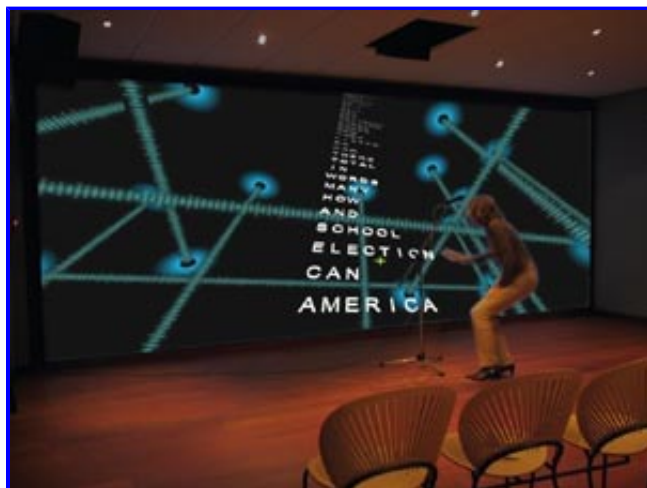
This representation allows the user to navigate between the lists of words, and interact with the world through motion and audio.

The user can grab a word, and hold it. While pronouncing the word the word will change shape and color, and then drift away in space in the appointed direction.

The user is able to use a 3D tracker (mouse) to maneuver in space, while pointing the yellow X (cursor) to select the desired words.



The objective is to grab a word out of the wordlist, and then use the microphone to articulate the selected word.



This is a report on a project made in cooperation between students from:

- **Denmark's Technical University**  
*Institute of Mathematical Modelling (DTU-IMM)*  
[Bendik Stang](#) 3D modelling and programming

---

- **University of Copenhagen**  
*Department of Computer Science (DIKU)*  
Erling Tind Audio systems and network programming.  
Declan Murphy Optical Motion Detection  
Jens Arnsparng Tutor  
Kristoffer Jensen Tutor

---

- **Danish School of Design.**  
[Anna-Mette Bach Jensen](#) Concept and Graphical Design  
Catherine Beyer Concept and Graphical Design  
Michel Gugliemi Tutor

---

- **Others**  
August Engkilde Sound design

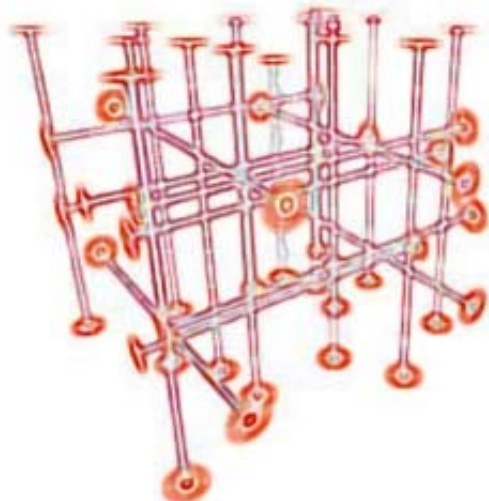
---

## Description overview

---

### Design features

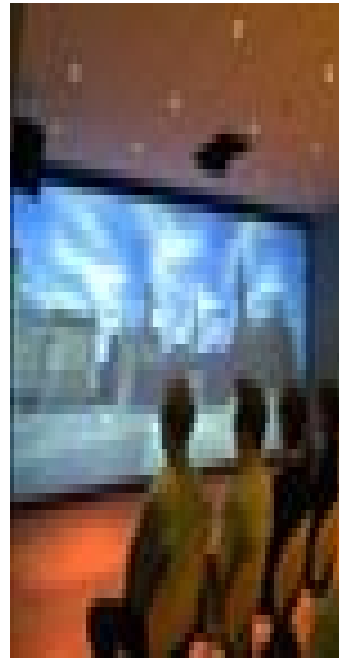
- [Interface](#)
- [Development](#)  
Tasks
- **Communication**
- **Tests**
- **Conclusion**



---

### Technical Features

- **3D Graphics**  
Used hardware - SGI  
[Used software - Preformer](#)
- **Audio I/O**  
Used hardware - Lake Huron  
Used software - Eyes Web
- **Motion detection**  
Used hardware - Webcam  
Used software - Eyes Web



## Interface

### Interaction - Interface

The interaction between the user and the computer can be divided into the following categories.

#### Human abilities

- Read - words with eyes
- Interpret - Pronounce the words
- Interact - Move around
- Choose - Select words
- Hear - sounds
- Consciousness - feel immersed

#### Computer features

- Collect - Get words from internet
- Record - input
  - Sound input
  - Mouse input
  - Motion input
- Display - Visualize VR
- Play Sounds - 3D sound

This is one way of presenting the link between the real world and the virtual universes of the workflow over the internet.

## Development

### Tasks

There were several tasks that had to be defined. After the general idea was conceived, a rather painful process of finding the realistic goals began. Each of the participants had very different backgrounds. This caused quite a communication gap between the technical group and the design group.

In order to find the final design, the designers would have to ping-pong the conceptual design with the programmers. In general the designers had no idea about what could be done or how long it would take, while the programmers would have no idea about the terms used by the designers.

Below are some of the general differences that caused this communication gap.



- Imported .flt objects
- pfText objects
- pfBillboards
- pfTexture (animated uv's)
- [pfLOD \(Level of Detail\)](#)
- pfSky (background colors)
- pfSegSet (colision detection)
- Stereo setup



This program was made using [openGL - Performer](#) by SGI.

---

# An Improved Edge Detection and Ranking Technique

Declan Murphy\*

Computer Science Dept, Copenhagen University

declan@diku.dk

## Abstract

A method is presented for refining the Canny edge detection algorithm, and subsequently ranking the resulting edges relative to their distance (in a topological sense) from the perimeter. This facilitates the selection of the appropriate level of detail for computer vision, particularly of articulated objects.

The first issue is the way in which the existing Canny algorithm chooses only one path where edges cross each other, resulting in erroneous gaps and spurious doubling of edges. An efficient pragmatic post-processing of the Canny technique is presented, which systematically bridges these gaps appropriately. The next issue is the automatic ranking of the edges.

## 1 Motivation and Introduction

The rationale and motivation behind this work was the requirement for a connected perimeter extractor with a variable amount of edge detail, tracing in from the perimeter. This need arose as an image pre-processing stage of a hand posture tracking system, in order to facilitate a deterministic, geometrical technique for reconciling image views with an articulated model. (For further details of this, see [6].) It should serve useful in any situation in which perimeter extraction of the foreground image, along with a specifiable level of edge detail in from the perimeter, would be advantageous.

Section 2 briefly describes the Canny edge detection algorithm. Section 3 covers the edge ranking with a description of the perimeter extraction, §3.1, and gap bridging, §3.2. Section 4 points towards how this technique serves the further stage of processing of reconciling image views to a model of an articulated 3D body. Section 5 winds up with a summary and conclusion.

---

\*Most of this work was carried out while the author was visiting the CART lab of CNR, Pisa, with subsequent refinement and write-up at the InfoMus lab of DIST, University of Genoa and at DAIMI, University of Århus.

## 2 Canny Edge Detection

One standard, or simple, approach to finding the perimeter of an object is to take its silhouette, which is very easy to compute if (as in the case motivating this work) there is a single foreground object of interest with a known or controlled background. However, the exact position of the silhouette border varies with the thresholding level used to compute it, and with both the source direction and the level of the ambient lighting.

A more accurate approach – almost immune to these weaknesses – is to use edge detection. As we may be dealing with low resolution images (to enhance real-time performance), such accuracy is desirable to offset coarse discretisation inaccuracy. Another significant advantage is the ability to make out some of the inner detail instead of perceiving just a single blob. This becomes particularly important if an articulated object is being tracked, as explained in §4.

*Edge detection*, in computer vision, is defined as the process of assigning a value to each pixel of an image in proportion to the likelihood that the pixel is on the boundary between two regions of different intensity values.

The Canny edge detection algorithm[1] is used – as implemented in the Intel Open Computer Vision Performance Library (OpenCV)[2]. This was chosen for its sharp results (in comparison to rival edge detection techniques), and for its efficient implementation (which was readily available).

### 2.1 How it Works

First the image is smoothed by Gaussian convolution; then a simple 2D first derivative operator is applied to the smoothed image to highlight regions of the image having high first spatial derivatives<sup>1</sup>. Edges give rise to ridges in the gradient magnitude image. The algorithm then tracks along the top of these ridges and sets all pixels that are not on top of the ridge to zero, so as to give a thin line in the output.

Two thresholds limit the tracking:  $t_1$  and  $t_2$ , with  $t_1 < t_2$ . Tracking can only begin at a point on a ridge higher than  $t_2$ . Tracking then continues in both directions out from that point until the height of the ridge falls below  $t_1$ . This hysteresis helps to ensure that noisy edges are not broken up into multiple edge fragments.

### 2.2 The Y-Junction Effect

One problem with the basic Canny operator is that of so-called ‘Y-junctions’: places where three (or more) ridges meet in the gradient magnitude image. Such junctions occur where an edge is partially occluded by another object.

---

<sup>1</sup>*Spatial derivative* simply refers to how much the image intensity values change per change in image position. See e.g. [3, ch. 5] or [4] for further background on edge detection.



The tracker will treat two of the ridges as a single line segment, and the third one as a line that approaches, but doesn't quite connect to, the other two. See figure 1.

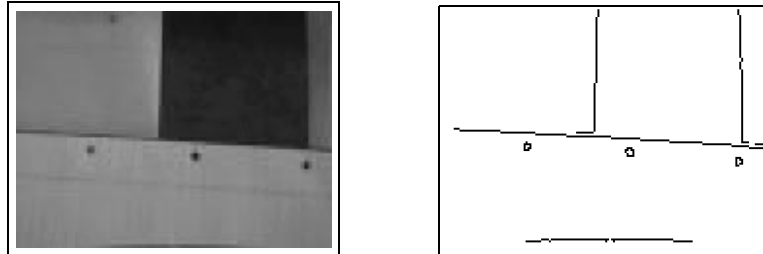


Figure 1: *Canny Edge Detection and its 'Y-Junction' effect. Left: camera image; right: Canny edge image. Note how the two vertical lines approach but fail to contact the central horizontal line, resulting in a spurious doubling up of a length of the horizontal edge.*

### 3 Edge Ranking

A rank (low non-negative integer) is associated with each edge according to the following rule:

- 0 corresponds to edges on the perimeter,
- 1 corresponds to edges terminating on the perimeter which are not of rank 0,
- $n$  corresponds to edges terminating on an edge of rank  $n - 1$  which are not of rank  $< n$ .

See figure 2.

Later stages of processing have the option of having an image frame composed of all edges up to a specified rank, thereby selecting the appropriate level of detail.

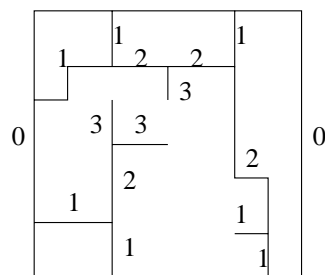


Figure 2: *An example of some edge ranks.*

### 3.1 Perimeter Extraction

Taking the edge detection from above, §2, the perimeter is in turn extracted. While this may sound trivial – as indeed it is to our human eyes – it is not so for the computer: there may be (and in general will be) some spurious noise outside of the object of interest, and the perimeter will not be continuous in general.

First the image is scanned, from an outer edge of the frame inwards, for any “blob”<sup>2</sup> in such a manner that any enclosed shape will be approached from the outside (even if this shape is not connected<sup>3</sup>). (Actually, there is a very remote possibility that the scan may first encounter the object of interest from the inside, but in fact this does not interfere with the success of the algorithm. At worst, it will only marginally slow the process in the case that only low rank edges are required.)

Next, the perimeter of the found shape is traced, paying careful attention to always keep to the *outside* of the shape while considering the next pixel along the perimeter. This is achieved by:

1. maintaining the concept of “the last outer pixel”, for each pixel as we trace along the perimeter,
2. rotating, from the last outer pixel to the next pixel of the shape, in a direction (always clockwise or always anticlockwise) consistent with the angular polarity of the trace.

In other words, every time we locate another perimeter pixel, we record the last non-perimeter pixel (on the outside). In order to find the next perimeter pixel, we rotate about the immediate neighbouring pixels – starting with the next pixel after the last outer pixel, in a consistent direction of rotation. See figure 3.

We are assuming from here on that edges outside of the object of interest (if any) have been removed or are disregarded in the processing. This may be achieved by such techniques as gauging the length, whether it lies close to other edges, etc., as described in, for example, [5]. This has not been a problem in the prototype.

### 3.2 Bridging Gaps

Now if we try to trace along the outside of the shape we have found, we will find that, very often, the trace only runs around a section of the desired object perimeter because of gaps left by the Canny edge detection (§2.2).

---

<sup>2</sup>Here, a “blob” simply means a number of pixels all above a certain intensity value, each one of which is adjacent to at least one other such pixel.

<sup>3</sup>The term *connected* is borrowed from topology: here it simply means that the shape in question can be entirely traced through adjacent pixels.

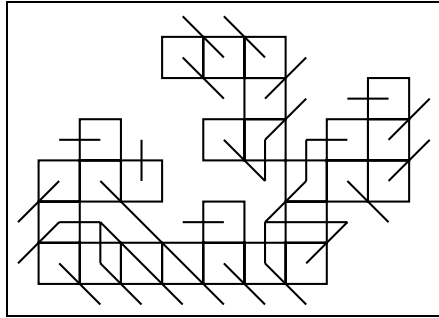


Figure 3: *Maintaining the “last outer pixel”. Each pixel of the perimeter is represented by a square, with a straight line segment from its centre to the centre of its last outer pixel, rotating anticlockwise.*

Such gaps come in various forms, so that a general technique to bridge them all is required. Some examples taken from observation of real data appear in figure 4. The technique found to cover all cases, tested with real (hand) data, is as follows:

1. Starting with the first perimeter pixel found, continue tracing anti-clockwise.
  - (a) If trace describes a full circuit, then we are finished.
  - (b) Otherwise record the end pixel and the last outer pixel of the second<sup>4</sup> from end pixel.
2. From the initial pixel, trace clockwise.
  - (a) Record the end pixel and the second<sup>4</sup> from end last outer pixel.
3. For all ends, try to join them.
  - (a) Rotate from outside to inside, scanning along a radius of the appropriate size<sup>6</sup>.
  - (b) If another segment is located, join to the nearest pixel on it.

---

<sup>4</sup>We use the last outer pixel of the second to end pixel because that of the end pixel will have swung around into the inside. Recovery of the outside may be made by rotating back<sup>5</sup>(if the segments tend to be long) or by placing the last outer pixels in a two cell Last In First Out (LIFO) buffer (if the segments tend to be very short).

<sup>5</sup>In rotating backwards, the temptation to take the short cut of rotating on through the edge must be resisted! See figure 5.

<sup>6</sup>There is a certain radius (or aperture) size observed to be both large enough to bridge all Y-junctions and small enough to not interfere with other edges in the vicinity, for a given image resolution and a given distance from camera to subject.

- (c) If this new segment is already on the list of ends, merge it (on the list) with its new partner; otherwise, trace the extent of the new segment, updating its end data.
4. If any segments remain (implying that all segments end in spurious diversions), try to join them. (This case occurs only very rarely in practice.)
- (a) Back-trace along the segment away from the end, testing within the aperture from outside to in, and otherwise proceed as in 3.
  - (b) If no join is found, delete end from list (consider as noise).

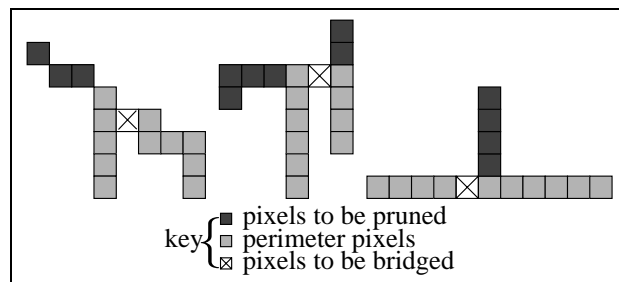


Figure 4: *Bridging the perimeter gaps. Internal bridging is similar.*

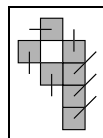


Figure 5: *A counter example of why rotation must be backwards, not onwards through the edge (which would get stuck in the hole!).*

### 3.3 Ranking

If the desired rank is greater than zero, then all perimeter ranks are first assigned their zero rank. Then, using the above gap bridging technique (but without further heed to maintaining a sense of outside, and joining to all other edges within the seek radius), all adjoining edges are given rank one. This process is iterated until the desired rank is attained: all edges adjoining those of the current rank,  $n$  say, which have not yet been assigned a positive rank, are now given rank  $n + 1$ .

## 4 3D Reconstruction of Articulated Objects

In trying to reconcile a model of an articulated body with camera images, it is generally useful to take note of the edges arising from creases along the articulation joints, thus yielding cues to the location of the joints, and facilitating recognition of the angles of flexion of the joints. Finer level detail, however, is generally undesirable at this stage, as it only introduces noise or needless complication into further stages of processing.

Moreover, in viewing the projection of articulated (or, indeed, non-concave<sup>7</sup>) three dimensional bodies to two dimensional images, some (or many) of the external edges of the body in three dimensions become internal edges after projection. See figure 6.



Figure 6: *Note how the edges of the fingers (external to the three-dimensional hand) become internal edges in the two-dimensional image. In this image, all the visible edges of the fingers are of rank 0–3, while the white sheen of the four distal interphalangeal joints and the dark crease of the distal interphalangeal joint of the thumb are of rank 2 or 3. Thus, an image of all edges of rank  $\leq 3$  is most appropriate.*

The doubling up of edges can be dealt with by judicious setting of the the two thresholds ( $t_1, t_2$  from §2.1).

Thus armed, with this technique for accurately extracting perimeters and edges of a given rank, and knowing the rough position/orientation of the body (e.g. from previous frames, velocity and model constraints), it becomes relatively straight-forward to employ a technique for selecting those edges which correspond to the perimeter of the body in three dimensions, and to have articulation creases and other such desirable features at will.

## 5 Summary and Conclusion

The technique is demonstrably (see figure 7) neat and efficient enough to be simply appended to existing edge detection algorithms with negligible real-time performance penalty. By supplying edges with (or up to certain) ranks, the appropriate level of detail is immediately brought into focus (including, by way of corollary, perimeter extraction), thereby directly facilitating the reconciliation of camera images with models of articulated objects and significantly reducing redundancy in further computer vision processing.

---

<sup>7</sup> $X \subset \mathbb{R}^3$  is concave means that  $\mathbf{x}_1, \mathbf{x}_2 \in X \Rightarrow \{(1 - \lambda)\mathbf{x}_1 + \lambda\mathbf{x}_2 : 0 \leq \lambda \leq 1\} \subset X$

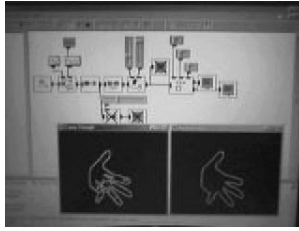


Figure 7: *Screen-shot of an EyesWeb patch having this technique embedded as an image pre-processing block. Note how the sub-image on the left (Canny edges) is cleaned showing only the perimeter in the right sub-image.*

The source code in C++ may be freely (GPL) downloaded from [7] both as an EyesWeb block and as standalone code. A report of the emerging hand posture tracker may also be found there.

## References

- [1] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6), November 1986.
- [2] Intel Corporation. Open source computer vision library. WWW. <http://www.intel.com/research/mrl/research/opencv/>.
- [3] E.R. Davies. *Machine Vision: Theory, Algorithms and Practicalities*. Academic Press, 1997. 2nd edition, pp. xxxi + 750.
- [4] Robert Fisher, Simon Perkins, Ashley Walker, and Erik Wolfart. The hypermedia image processing reference. [http://www.dai.ed.ac.uk/HIPR2/hipr\\_top.htm](http://www.dai.ed.ac.uk/HIPR2/hipr_top.htm).
- [5] Declan Murphy. Extracting arm gestures for VR using EyesWeb. Barcelona, Spain, Nov 2001. Audiovisual Institute, Pompeu Fabra University. ISBN: 84-88042-37X.
- [6] Declan Murphy. Building a hand posture recognition system from multiple video images: A bottom-up approach. Technical report, interim, cART lab, CNR, Pisa, Italy, March 2002. <http://www.diku.dk/users/declan/pub/cart.pdf>.
- [7] Declan Murphy. A hand posture recognition system. Anonymous FTP GPL downloadable software, updates, reports, papers, and links to applications as they become available, 2002. <ftp://ftp.diku.dk/diku/users/declan/hand/>.

# Extracting Arm Gestures for VR using EyesWeb

Declan Murphy

Music Informatics Group, Computer Science Department, Copenhagen University  
ComputerART lab of the Italian National Council of Research (C.N.R.), Pisa  
declan@diku.dk      <http://www.maths.tcd.ie/~dec/dec.html>

## Abstract

This paper describes research into a technique for extracting three dimensional arm, hand and head coordinates from a user in a virtual reality (VR) environment. Hand gestures are in turn extracted as a major part of the interactive user interface. Then some expression is extracted from these gestures, which is used to modify the sound-scape and graphical texture appropriately.

Development was carried out on the EyesWeb prototyping environment. Two cameras are employed: one giving a frontal view and the other in profile. The extended arm and hand are extracted from both digitised images, and pointing coordinates are passed to the graphics interface. A grabbing gesture (used for selecting the item pointed to) is also recognised, and 3D hand coordinates (of the grabbed item) are also passed to the graphics interface.

Efficient and robust algorithms are developed for the arm segmentation and grabbing gestures. Although developed for a dedicated VR environment, the techniques are no less applicable to real-time gestural control of computer music via video camera(s).

## 1 Introduction

The next section, §2, overviews the background to the VR installation, in the light of which, §3 outlines the choice of gestures, cameras, lights, etc. The following sections, §4–§6, describe the particular techniques developed for recognition of the particular gestures. The final section, §7, concludes and points towards future work.

## 2 Project Background

This virtual reality project is the result of a collaboration between the Music Informatics research Group at DIKU, Copenhagen University[1], the Danish Design School[2], and the Virtual Reality Center (VR•C) of the Technical University of Denmark[3].

The installation center, VR•C, is equipped with a large screen displaying 3D images and sophisticated Hi-Fi surround sound, all powered by a network of different computers.

### 2.1 Concept

Briefly, the concept is to immerse the user (and, we hope, also the audience) in a world composed of graphical streams of words which come from live samples of search text from WWW search engines. The user is free to reach out and grab words of his/her choice, and to speak the grabbed word. The manner of grabbing

and pronouncing determine the graphical texture of the word and also the three-dimensional sound-scape.

The task at hand was to develop the user gesture interface, one that would have further open applications for computer music.

### 2.2 Choice of Setup

It was decided that the use of video camera(s) in the auditorium was preferable to using the headsets with position sensors or mouse. This carries the advantages of allowing a large audience to spectate (there are only four headsets), avoiding the user having to be physically encumbered by being wired up with sensors, and – perhaps most importantly – facilitating a more natural interface whilst increasing the multi-modality, thereby lending towards greater immersion in the virtual world.

### 2.3 EyesWeb

EyesWeb was chosen as the development platform both for the video processing, and for the audio streaming.

EyesWeb[4] is a prototyping environment, primarily for processing of digital video and audio information, developed by the DIST[5] lab and MOSART node in Genoa. Care is taken to support real-time processing, as it was developed with a view to extracting expression from live video images. The EyesWeb software is actively supported, and freely downloadable (subject only to certain licensing conditions) from the project web-site at [6].

It is built around the Intel Performance Libraries[7] which provide optimised image analysis tools (for Intel based computers). It is also built around the DCOM[8] protocol by Microsoft: developers' own routines may be embedded in the program by registering them as COM components.

## 3 Methodology

### 3.1 Which Gestures

One of the first hard decisions which had to be made was to fix the “vocabulary” of gestures for which recognition would be developed. These became as follows:

- pointing,
- grabbing,
- hand tracking and discarding, with expression,
- gaze following,

where discarding could be throwing away or blowing away.

### 3.2 Lighting

Lighting then became a concern, as the user and audience require quite low ambient lighting in order to see the screen adequately, whereas the camera requires reasonably strong lighting in order to make out much detail. Consideration (including some experimentation) was given to using infra-red camera(s), to arranging spot lights, and to getting the user to wear distinctly brightly coloured gloves and hat, or light sources at appropriate parts of the body. However, it was found that the ceiling lights in the auditorium could be adjusted in banks (so as it was possible to illuminate just the “stage space” – the floor area between the audience and the screen – to a variable amount), and that they bevel and pivot individually in their mountings (so that they can be physically turned to highlight the user, and little else). With this lighting appropriately configured, it was found that the user's bare arm was clear enough for segmentation (particularly if they are otherwise darkly dressed), even using a simple webcam.

### 3.3 Cameras

The next concern was where to position the camera(s), and how many would be necessary. Coordinates of pointing and hand tracking were required with some accuracy in three dimensions, so it was decided to have at least two cameras. One camera viewing the user's profile, and another giving a frontal view of the user, were considered sufficient. However, the frontal camera could not be placed directly in front of the user since this would interfere with the general view of the screen for everybody present. A frontal camera placed

on the floor saw the user unclearly in severe silhouette of the lights from above; a frontal camera placed at ceiling height was considered to be satisfactory (the resulting aberration of the image could be compensated for, as necessary).

## 3.4 Procedure

Next, a working library of samples of these gestures being performed was recorded on location at the VR center. Conditions were varied, including lighting level, camera position, image resolution, the expression in the gestures and the people performing them.

For each of the following gesture recognition algorithms, two versions have been developed: “monitor” versions and “performance” versions. The monitor versions display their performance and output, and facilitate the calibration and/or optimisation available. The performance versions accept the same calibration and optimisation settings, but are otherwise stripped bare of unessential features, and are intended for optimum real-time performance.

## 4 Arm Segmentation

The first stage of image processing for the pointing, grabbing and hand tracking gestures, is to segment the pointing arm from the rest of the image. For each video frame, the following algorithm is applied.

### 4.1 Squiral

After conversion from colour to gray-scale, with optional background-subtraction<sup>1</sup>, the image is scanned for pixels above a certain luminosity threshold. In order to keep up with our performance goal of real-time processing, time is not wasted by starting in one corner of the image and iterating through adjacent pixels in subsequent lines in the usual manner of iterating through each pixel of an image.

We know in advance that the object we are looking for (the users outstretched arm) is

1. more likely to be somewhere near the center of the frame, and
2. is solid and of a certain width and length.

Accordingly, the scan

1. starts in the center of the frame and works outwards, and
2. tests pixels at regular intervals, greater than one but less than the arm width.

To be precise, the pixel test sequence describes an outwardly spiraling square shape, or “squirrel”<sup>2</sup>.

<sup>1</sup>Somewhat surprisingly, the arm segmentation method works fine (and requires rather less calibration/adjustment) without any background subtraction.

<sup>2</sup>The term “squirrel” is coined by Jamie Zawinski in his attractive



## 4.2 Blob Expansion

Having found a pixel of interest (*i.e.* above the luminosity threshold), it is “expanded” into the region of all recursively adjacent<sup>3</sup> pixels also above the threshold. Again, an efficient approach is devised.

### 4.2.1 Naïve Approach

This time, iterating over the pixels in order makes much less sense: it is not difficult to show that this would entail a search of order  $O(|N|d(M))$ , where  $|N|$  is the number of pixels in the frame  $N$ , and  $d(M)$  is the diameter of the final region  $M$  defined by

$$d(M) = \max_{\mathbf{x}_1, \mathbf{x}_2 \in M} |\mathbf{x}_1 - \mathbf{x}_2|.$$

The disastrousness of this approach is apparent if you consider that  $|N|$  is greater<sup>4</sup> than  $d(M)$  by three orders of magnitude for the arm, and by four for smaller blobs of noise we wish to mask out.

A more sensible approach is to start with the given pixel, and to successively work outwards – considering only pixels adjacent to those already expanded – until the entire region is covered.

### 4.2.2 Depth/Breadth First Search

We can identify the frame of pixels,  $N$ , as a graph by

1. considering each pixel to be a node, and
2. considering an edge to exist between adjacent<sup>3</sup> pixels which are
  - (a) above the threshold, and
  - (b) have not already been visited.

This allows us to apply the well-known Depth First Search (DFS, see any good text book or lecture notes on graph theory or algorithm analysis, *e.g.* [9, ch. 6]) or Breadth First Search (BFS) graph node traversal algorithms. Given any node (pixel), they step through all other (recursively) connected nodes (*i.e.*, all nodes of the same connected component to which the given node belongs), which is exactly what we want. They have the desirable property of being optimally efficient for graph node traversal in the sense that, if a connected component consists of  $|M|$  nodes and  $|V|$  edges, then they are of order  $O(|M| + |V|)$ . In our implementation,  $|V| = 4|M|$ , so that we have simply  $O(|M|)$ .

The difference between DFS and BFS lies in the order in which they traverse the connected component

XScreenSaver available from

<http://www.jwz.org/xscreensaver/>

<sup>3</sup>For recursive adjacency here, it is sufficient to consider only adjacency at the four edges (*i.e.* ignoring the corners). This greatly reduces the amount of redundant pixel testing, yet still covers the whole connected region. The only other consequence is the loss of connectivity between certain (highly unlikely) pathological regions touching only at certain corner points.

<sup>4</sup>Here  $|N| = 320 \times 240 = 76800$ .

(which is of no consequence to us), and in their implementation. Having to compute the adjacency conditions 2(a) and 2(b) on the fly, is not a problem for either implementation. (In fact, 2(b) is normally an integral part of both algorithms.) DFS has a very neat implementation using either recursion or a stack, whereas BFS involves the slightly greater overhead and complexity of the memory management of a First-In-First-Out (FIFO) buffer. For this reason, DFS is preferred, even though BFS ordering is more intuitively obvious for this application.

## 4.3 Size and Shape

At this stage we have a connected component of interest (CCoI), and we want to establish whether or not it is the arm we are looking for. We do this by gauging its size and shape. (This relies upon the hypothesis that there may be at most one CCoI in the frame having the appearance of an arm. The user is instructed to have one bare arm (which has been arranged to catch the available light, *cf* §3.2), whilst otherwise being dressed darkly. (This relies further upon the hypothesis that the arm in question is both caucasian[10] and not too tanned, which is the case amongst the people involved in the project.))

### 4.3.1 Area

The size condition is dealt with first, since the data structure used to store the CCoI stores the size (which corresponds directly to the area), and so it does not have to be computed. The constraint of having a reasonably<sup>5</sup> fixed distance between the camera and the user allows us to assume that the area of the arm will not vary, within reasonable limits. These limits are set beforehand, and the CCoI is rejected if it does not lie within them.

### 4.3.2 Shape

The first stage is to calculate the mean of the CCoI.

Next, the major axis of the CCoI is calculated as follows.

1. Let  $r$  = the largest radius we would expect (*i.e.* half the length of an arm in pixels) plus an adjustable margin).
2. Describe<sup>6</sup> a circle of radius  $r$  in the frame around the mean point. Let  $p$ , say, be this point tracing out the circle.

(a) Let  $p'$  be the point diagonally opposite  $p$ .

<sup>5</sup>The user may move about by a couple of steps, but that is approximately an order of magnitude less than the mean distance between the camera (which is fixed) and the user.

<sup>6</sup>Important details for an implementation, such as using the most appropriate data structures (and conversions between them), and dealing with the event of iterating out of the bounds of the frame, have been omitted for the sake of clarity.

- (b) If  $p, p' \in \text{CCoI}$  then
  - i. Let major-axis =  $|p - p'|$ .
  - ii. Return.
- 3. Decrement  $r$ .
- 4. Goto 2.

(The definition of major axis implicit in this method of calculation, is not guaranteed to be unique. This is not a problem, however, when the major axis is considered in combination with the other criteria. Cf §4.4. Its length, in particular, is unique within the margin of error.)

The minor axis (which is, in fact, only used for calibration) is in turn calculated as the intersection of the line (in the frame) perpendicular to the major axis, and the CCoI.

### 4.3.3 Gauging

First of all, if the major axis is too short then the CCoI is rejected.

Otherwise, a (hopefully) bounding rectangle is described with the arm's mean as center, and having its long edges parallel with the arm's major axis. The "hopefully bounding" dimensions are preset. If, at any stage of the perimeter of this rectangle, there is intersection with the CCoI, then the CCoI is rejected.

The actual dimensions and margins used for gauging differ between the profile and frontal cameras.

### 4.3.4 Time

Adjustable margins are used at several stages of this procedure so far. Utilising the knowledge of the location of the arm in the previous frame, tighter margins are used where the arm is not expected, and looser margins where the probability of finding it is high.

## 4.4 Rationale

The advantages of squirling are twofold. First, the search starts where it is most likely to find what it is looking for, proceeding (roughly) in order of greater likelihood. Secondly, it searches at sample points which are coarsely distributed in order to increase the speed the search several times over, and yet finely distributed enough to be guaranteed of success. (There is also the consequential benefit that some very local noise will be filtered out.)

The condition of the arm appearing as a connected region is ensured by the lighting and the camera pre-adjustments, and may be enhanced by some median filtering. (Median filtering also has the desired effect of tending to remove local noise.)

It is reasonable to assume approximate concavity<sup>7</sup> for a CCoI if it turns out to be the arm. (If it happens not

<sup>7</sup>A subset  $M$  of a linear vector space  $X$  over  $\mathbb{R}$  is said to be *concave* iff

$$\mathbf{x}, \mathbf{y} \in M \Rightarrow \{(1 - \lambda)\mathbf{x} + \lambda\mathbf{y} : \lambda \in \mathbb{R}, 0 \leq \lambda \leq 1\} \subset M.$$

to be the arm, then this assumption does not hamper its rejection: a non-concave CCoI will almost surely have a different ratio of major axis to area, and will almost surely fail the bounding rectangle test.) This justifies the use of the mean, and the method of calculation of the major and minor axes.

It is also expected that the forearm we are looking for has an aspect ratio of approximately 1:6.

Thus, given the CCoI's connectedness, if it has a satisfactory area and major axis, then it must be something which (at least, at a first glance) looks somewhat like an arm. If it then fails the bounding rectangle test, it must have been either an arm that was not outstretched in pointing, or something else masquerading as an arm.

This technique, as presented so far, tends to err on the safe side, in the sense that it will fail to find an arm rather than erroneously finding something else. (Of course, this situation may be reversed by sufficient loosening of the adjustable margins.) The net effect of this, is that occasional frames may lose the trace of the arm's motion. Although missing  $\frac{1}{25}$  second every once in a while is not a big problem in performance, it can nevertheless be brought towards perfection by incorporating temporal information. We know that the arm moves continuously through space, never exceeding certain velocities, which puts a tight limit on where we can expect the arm to be, as we move from one frame to the next, and so the margins are adjusted accordingly.

Therefore, if the CCoI passes the series of gauge tests, then it is reasonable to take it for the arm we are looking for.

## 4.5 Coordinates

In order to interface the user's gestures with the graphics on screen, coordinates of where the user is pointing to are passed to the network client. The mean position of the arm has already been calculated; in combination with the finger tip<sup>8</sup>, the direction of pointing (for each camera view) is determined. Resolution of the pointing directions from both camera views with the screen graphics is relatively straight-forward (being mostly a matter of camera calibration and user-space to graphics-space protocol). Thus, the task reduces to determining the location of the finger tip: which is simply the furthest distance along the major axis in the direction of the screen

## 4.6 Hand Tracking and Discarding, with Expression

As a corollary of the last subsection, §4.5, the position of the hand (for each frame) has been determined. Thus, after the user selects a word by reaching out and

A CCoI may be made "more nearly" concave by median filtering.

<sup>8</sup>The "finger tip" is defined here to be the furthest point of the hand away from the body, and as such may belong to a closed fist.

grabbing it (*cf* §5), he/she is able to move it about in space as desired.

Armed<sup>9</sup> with the location of the hand in 3D, its trajectory in space can be followed. The location of the head is also known (*cf* §6). If the hand goes behind the head, then a discard is signalled when the hand comes momentarily to rest – corresponding to the user throwing the word away behind their back or over their shoulder. If the hand comes to rest just in front of the head, then – upon the appropriate audio queue of white noise – a blow is signalled.

Some attempt at extracting higher level expression parameters is also included, in order to appropriately give colour, texture and shape to the graphical representation of the grabbed word, and to cue appropriate changes in the sound-scape. The velocity of the hand is extracted, and from this its speed and its rate of change of direction are supplied as expression parameters.

## 5 Grabbing

A grabbing gesture can be recognised from a binary outline of the hand.

The position of the hand's extremity, furthest from the body, is known. So is the direction of the rest of the hand. Within a certain distance from the extremity, in the direction of the arm mean, there is the characteristic constriction of the wrist – coinciding<sup>10</sup> in both views.

This distance, from the wrist to the finger tip, is approximately twice the length for a pointing hand, than for a closed fist. Transition from a pointing or open hand to a closed fist is signalled as a grab.

## 6 Gaze Tracking

There are documented examples of techniques for gaze tracking in the literature; in particular, the technique in [11] was implemented. Only the frontal camera is used.

First, the face has to be located in the frame. One cue is the (left) arm location we already have: the face is located, within certain margins, above and to the left of the arm. Other cues include skin tone, brightness, size (area) and shape. Shape is simply a bounding rectangle, given the area, with a rationale similar to that above for the pointing arm shape, §4.4.

There is an initial calibration shot of the face looking directly forward. An expanded<sup>11</sup> copy of the center of the face region is stored as a template. For each subsequent frame, the face region is located and similarly expanded, and then convolved with the template. The location of best match yields the direction of gazing.

<sup>9</sup>Please forgive the dreadful pun!

<sup>10</sup>The distance from finger tip to wrist is equal in pixels after compensating for the different camera-to-user distance and perspective of the frontal camera

<sup>11</sup>The face regions are expanded before convolution due to their small area in the frame. The severe discretisation of facial detail due to the low resolution is softened by expanding the region, giving a more robust convolution result.

(This is based on the assumption that, for the range of head movement we are interested in (a rotation of  $\pm \frac{\pi}{6}$ ), the camera image of the face changes relatively little, even though it is turning. At any rate, such change is small enough that it has negligible effect on the maximum match according to the convolution.)

## 7 Conclusion and Future Work

A brief sketch of how gesture capture can be incorporated into a VR environment was drawn. In particular, efficient and robust real-time segmentation techniques for parts of the body were developed. Development was towards some initial applications to expression in music and graphics, and is ripe for further such development.

Presently, related work continues at the Computer-ART lab in Pisa, and is to develop there over the next few months in the following directions:

- development of the Handel posture recognition system based on video silhouette spectra,
- extension of Handel system into 3 dimensions,
- further extension to two handed, and whole body, posture,
- including the time dimension in the model,
- study of semantic gestural expression,
- study of musical gestures,
- matching of physical and musical gestures,
- further applications to computer music composition and real-time performance.

## References

- [1] “Datalogisk Institut Københavns Universitet”, <<http://www.diku.dk/>>, 25 Sep 2001.
- [2] “Danmarks Designskole”, <<http://www.danmarksdesignskole.dk/index2.htm>>, 10 Sep 2001.
- [3] Danmarks Tekniske Universitet, “DTU’s hjemmeside”, <<http://www.dtu.dk/>>, 05 Jun 2001.
- [4] Antonio Camurri, Paolo Coletta, Peri Massimiliano, Ricchetti Matteo, Andrea Ricci and Trocca Riccardo, “A Real-Time Platform for Interactive Dance and Music Systems”, Proc. ICMC-2000, Berlin.
- [5] “Laboratorio di Informatica Musicale”, <[http://musart.dist.unige.it/laboratorio\\_en.html](http://musart.dist.unige.it/laboratorio_en.html)>.

- [6] InfoMus, “The EyesWeb Project”,  
<[http://musart.dist.unige.it/sito\\_inglese/research/r\\_current/eyesweb.html](http://musart.dist.unige.it/sito_inglese/research/r_current/eyesweb.html)>, 28 Sep 2001.
- [7] Intel, “Intel Performance Libraries”,  
<<http://developer.intel.com/software/products/perflib/>>,  
1 Aug 2001.
- [8] “Distributed Object Component Model”,  
<<http://www.microsoft.com/com/tech/DCOM.asp>>, 30 Mar 1998.
- [9] Michael T. Goodrich and Roberto Tamassia, “Algorithm Design: Foundations, Analysis and Internet Examples”, John Wiley & Sons, 2002.
- [10] Moritz Störring and Erik Granum, 2001, “Constraining a statistical skin color model to adapt to illumination changes”, Proc. Dansk Selskab for Automatisk Genkendelse af Mønstre, 30–31 Aug 2001, Copenhagen.
- [11] Rick Kjeldsen, “Facial Pointing”, Proc. Gesture Workshop 2001, London.



**IHP Network HPRN-CT-2000-00115 MOSART  
Music Orchestration System in Algorithmic Research and  
Technology**

**MOSART Task 5:**

**Symbolic Recognition of Musical Patterns and  
Recomposition.**

**Edited by Gerhard Widmer**

**Deliverable d25:**

Definition of Musical Patterns to be recognized automatically

**Table of Content**

**Patterns in music and music performance**

Gerhard Widmer

Page 248

## PATTERNS IN MUSIC AND MUSIC PERFORMANCE

*Report compiled on behalf of the consortium by*

Gerhard Widmer

Austrian Research Institute for Artificial Intelligence, Vienna

### 1. INTRODUCTION

In the original Work Plan, the goals of task T5 (“*Symbolic Recognition of Musical Patterns and Recomposition*”) were rather generically formulated as having to do with “musical pattern detection algorithms”. In the course of the first half of the MOSART project, the notion of musical patterns was clarified in several joint discussions, and two major lines of inquiry were defined that should guide the research activities in task T5.

The first line of research relates quite directly to the original understanding of “musical pattern recognition and recomposition”, as it guided the writing of the relevant passages of the project proposal: here we are interested in defining classes of patterns in the *structure of the music* that are both structurally important and perceptually salient. Computer programs that can detect such structures in the music and can manipulate them (or permit the user to manipulate them) in flexible and musically meaningful ways would be extremely helpful in many domains, including composition, live performance, and education. In the following, we will refer to this class of patterns as *structural patterns*.

The second class of patterns that was identified as deserving focused investigations are patterns in *performed music*. Music performance is a central aspect of the act of music making and has become an important focus of contemporary musicology. Computer programs capable of analysing, visualising, controlling, or even generating well-structured performances of given pieces of music would have enormous potential in many application scenarios. In fact, the MOSART consortium unites some of the top European research laboratories involved in computer-based studies of music performance, and this accumulation of expertise will be exploited and further enhanced in our second line of inquiry, which deals with what will henceforth be called *performance patterns*. Indeed, it was decided that a large part of our efforts will be invested in this second line of inquiry.

These two classes of patterns will be specified and discussed in more detail in sections 2 and 3. In the course of that presentation, it will turn out that the two approaches are not mutually exclusive. Quite to the contrary, the recognition of patterns in the music and the ability to correspondingly shape a performance of the music according to these patterns, are two complementary aspects of any computer system that is to deal with music in an intelligent way. So we expect the two lines of research to converge eventually.

The purpose of the present report is to more clearly define the classes of musical patterns to be investigated, and to sketch the research strategies that allow us to identify, characterise, and analyse these patterns. In addition, we briefly present current work and preliminary results already achieved along these lines. That is the contents of sections 2 and 3. Section 4 then sketches the main research directions to be followed in the next phase of task T5.

### 2. STRUCTURAL PATTERNS

#### 2.1. Motivation

Music is not simply a sequence of events in time. Hearing and understanding a piece of music as a structured object made up of a variety of structural elements is a crucial part of the experience of music. Tonal music, in particular, is a complex system of explicitly or implicitly codified rules and conventions, and tonal music compositions are richly structured objects that exhibit patterns along a variety of dimensions (rhythm, melody, harmony, motivic and phrase structure, etc.). Accordingly, computer systems that are to be useful in a practical musical context — be it in analysis, composition, performance, or education — must be able to recognise and manipulate such musical patterns or classes of musical structures in an intelligent way.

To focus the investigations, the MOSART group has defined a specific target application, namely, a prototype system for the manipulation of music in terms of its pattern structure via a gestural interface. The purpose of the system is:

1. To develop a more abstract (than existing) method of automatic musicological analysis based on macro level (pattern) structure, and to provide a composition tool/platform by using this with facilities for graphical representation, gestural manipulation and structural recombination.
2. To be a performable instrument. Recent discussions on the performance of computer music point out the inadequacy of the mapping paradigm, and suggest that an ability to compose while performing — an instrument that plays with structure instead of audio — would be an advantageous way forward. This system proposes to be just such an instrument. (Whether real-time capabilities will be achievable within the time frame of the MOSART project is currently unclear; however, that does not compromise the research goals set in the MOSART project.)
3. To be a platform upon which research on the perceptibility of musical patterns can be carried out. Indeed, perceptibility is to be one of the criteria associated with a pattern and in the limiting of the analysis algorithm. It is established that some note level patterns are more perceptible than others, and seminal pioneering work has been done on auditory scene analysis. However, much remains to be done regarding quantifying, or establishing rules saying which patterns are more perceptible than others amongst all those extractable.

The planned prototype system, tentatively called *Pattern Play*, will incorporate a number of models and software already developed by various MOSART partners, such as the rhythm quanti-

sation routines from NICI and the expressive mapping rules from KTH. The gestural interface will be developed with DIST's Eye-Web system.

## 2.2. Definition of Patterns

Rather than trying to give an exhaustive definition of the classes of patterns to be dealt with in the envisioned system (which is impossible at this stage), the present section sketches the major *processing steps* that will be realised in the planned system and that will detect and manipulate musical patterns.

Generally, the pattern identification part of the above mentioned *Pattern Play* system is going to be a two-step process. First features are extracted; then a generic pattern analysis takes place.

### Feature Extraction

At the time of writing, the list of *features* is the following:

1. note rhythm intervals (considered as a string of independent objects)
2. note pitch levels (considered as above)
3. rhythm patterns (considered across the entire section, at all intervals)
4. pitch contour (considering relative up/down progression, no. of scale steps)
5. chord progressions
6. note degree in chord (tonic, supertonic, etc. of underlying chord)

Thus, features 1 and 3 pertain to rhythm, 2 and 4 to pitch, and 5 and 6 to harmony. Features 1 and 2 examine the data purely as a string of information, without heed to any of the relationships or degrees of similarity between them (i.e., equality is the only criterion). Rhythm data has the strict ordering (only monophonic melodies are being considered here) of a time sequence and interval values which are nominally simple whole number multiples of each other, and so we subsequently consider the resulting higher level inter-relationships with feature 3. Pitch data has the strict ordering of higher/lower frequency with perceivable steps, and so these melodic contour patterns are extracted from feature 4. Even without explicit accompaniment, monophonic melodies clearly describe chord progressions (notwithstanding possible ambiguity)—hence feature 5 — with the melody tracing its way through them: how exactly it does this gives rise to feature 6.

Based on these basic features, a number of *higher-level patterns* will be identified. Rather than trying to define the kinds of patterns beforehand, automatic pattern recognition methods will be applied in order to discover those structural patterns that are significant in an information-theoretic sense.

### Pattern Analysis

The above-mentioned features are then analysed for *all* patterns above a variable entropy threshold (as compared with 'traditional' pattern recognition techniques which try to match data to a particular pattern which is known in advance).

First, each feature is extracted and subjected to its own pattern analysis. Much of the magic of music, however, arises out of the way in which multiple levels of patterns occur simultaneously. (For example, at a basic level, every note has both duration

and pitch, confirming and/or modifying rhythmic and harmonic expectations every time. There are also higher and lower level multiplicities of greater subtlety and complexity.) The approach of this system considers this parallel multiplicity essential to musical patterns.

It is in this multiplicity that the analysis algorithms used differ from compression algorithms: both seek to find neater ways in which to represent the data, but compression is only interested in a single optimum strategy whereas here all compressions above a certain entropy threshold are considered interesting (perceivable) and are recorded. On top of this, the final stage of analysis is a cross-feature pattern extraction.

The entire analysis is being formalised mathematically, algorithmically and in hard computer code. It will also have its own notation (being published).

## 2.3. Research Strategy

The research on *structural patterns* is a joint research effort by the MOSART partners DIKU, cART CNR, DIST, DAIMI, and AUE, Esbjerg. Much of the practical work (development of software etc.) will be performed by Declan Murphy, a MOSART young researcher, in the context of a PhD thesis supervised by Prof. Jens Arnsparng at DIKU (now: AUE, Esbjerg). The project incorporates video gesture capture techniques developed at DIKU Copenhagen, cART CNR Pisa, and DIST Genova, whose development will continue at DIKU, DAIMI, and Esbjerg.

The general two-phase research strategy has already been outlined in the previous section. The methods developed will be incorporated and tested in prototype system. The following is a brief list of the main components of this prototype:

- analysis algorithms for extraction and manipulation of patterns of musical features
- formal mathematical representation of such patterns
- score notation representation of such patterns
- GUI for pattern display and manipulation
- video hand tracking interface for manipulation
- inference rule system for consistent rearrangement of patterns

A preliminary description of the planned system and the current state of this work can be found in [45].

## 3. PERFORMANCE PATTERNS

### 3.1. Motivation

The central goal of the second line of investigation within task T5 is the identification, characterisation, and quantification of patterns in musical *performances*. That involves the development of computerised methods for extracting expressive patterns from performances, analysing these patterns with respect to various criteria, classifying them, studying their musical, perceptual, and pragmatic significance, and applying meaningful patterns to new pieces of music to achieve the desired effects, musical or otherwise.

After having been neglected for a very long time (possibly due to the problems related to actually measuring the details of a performance), music performance is now one of *the* current topics in music research. Understanding the fundamental principles behind expressive music performance is not only of theoretical interest,



but also promises lots of practical potential, from new, interactive instruments and novel paradigms of performance control [6] all the way to virtual musical settings like adaptive karaoke [25].

Music performance research is one of the definite strong points of many partners in the MOSART consortium. The planned investigations will continue and broaden the partners' work in this area. The research on performance patterns will make up the major part of the research activities related to task T5.

### 3.2. Definition of Patterns

Musical performance is an extremely complex phenomenon, with many dimensions, functions, and with many levels at which one might want to study and describe it. It is thus impossible to simply give a definitive list of patterns that together somehow 'characterise' or 'constitute' the phenomenon of expressive music performance. Instead, what will be given here is a list of different *dimensions* along which musical performance patterns can fruitfully be classified. To be more precise, some of the classification dimensions proposed in the following relate to performance patterns proper, while others pertain more to the research approaches applied to discover and study these patterns.

The first and most obvious choice is to classify performance patterns by the **parametric dimensions** that are controlled or shaped by a performer. Depending on the instrument (and the style of music being played), the performer can vary certain parameters within certain limits, the most important ones being *tempo and timing* (that concerns both the choice of global tempo and local tempo variations), *dynamics*, and *articulation*. Other important parametric dimensions available with some instruments include *vibrato*, *intonation*, and various other effects.

Obviously, all of these are crucial in explaining the effect of a performance, and in generating musically meaningful performances of new pieces. In fact, it will not be sufficient to study each of these dimensions in isolation; it is quite obvious that there are complex interrelations between these dimensions — the musical effect of a good performance is a result of the *interaction* of all these parametric dimensions.

Performance patterns can also be characterised with respect to their **continuous or discrete nature**. Some performance effects are inherently discrete and relatively easy to measure (e.g., the articulation or the timing (discrete onset point) of individual notes), while others are the results of continuous processes (e.g., vibrato) and require different measurement methods and different types of models.

Somewhat related to this is the issue of **musical range or scope**: some expressive patterns will tend to be of a highly local nature (e.g., affect only single notes, as in a *sforzando*), while others affect larger musical structures (such as in a gradual *crescendo* applied to a melodic line). Again, different types of models will be appropriate to adequately describe these patterns and the corresponding performance strategies. Indeed, music performance is known to be a *multi-level* phenomenon; any model, computational or otherwise, that claims to do justice to the full complexity of expressive performance will have to take account of performance patterns at different structural levels, and of their interactions.

Furthermore, expressive music performance is known to serve different **functions or purposes** [44]. Some of the observed performance patterns may reflect conscious strategies used by the performer to highlight the *musical structure* of a given piece, or to disambiguate a passages with several possible structural interpre-

tations. (In some cases, the performer may even choose to consciously increase the sense of ambiguity.) Other patterns can be more directly linked to the *dramatic narrative* developed by a performer, or to the *emotional content* of the music and the performer's intention to express and evoke certain emotions. Yet another class of patterns may be simply due to *physical constraints* or problems (e.g., fingering problems, discontinuities at melodic jumps, etc.). In any case, this dimension has a strong influence on the research strategies that are most appropriate to study performance (see also section 3.3 below).

Another classification dimension is the **description or resolution level** chosen — that relates not so much to the nature of the performance patterns themselves, but to the *granularity* with which one wishes to study and characterise them. For instance, in some contexts it may be sufficient to make categorical predictions regarding, e.g., staccato vs. legato, while in others the precise level of staccato will be relevant. That determines what kinds of models and what kinds of model building (and pattern recognition) algorithms are appropriate.

And finally, the **source of measurements** is relevant. Again, this dimension relates not to the patterns themselves, but to the underlying data source; it determines what features can be extracted from the data, and at which level of resolution certain patterns can be studied. Clearly, MIDI instruments offer the advantage of being able to precisely measure the discrete aspects (e.g., onset and offset times) of each individual played note. On the other hand, inherently continuous effects like intonation and vibrato will require audio recordings and audio processing methods.

The above discussion should give an indication of the complexity of the subject. It should be clear from this that it will be impossible to explore the entire space of possible performance patterns and explanatory models within the limited lifetime of a project like MOSART. We will have to be selective, while still striving for an integration of the results of our investigations. Examples of some of the specific classes of patterns and questions currently studied within MOSART can be found in section 3.4 below.

### 3.3. Research Strategies

In a public panel discussion<sup>1</sup> organised to discuss and clarify basic research issues regarding empirical, computer-based performance research, it became very clear that there is no single optimal research strategy. What strategy (including what methods of evaluation of the results) is most appropriate depends very much on the *purpose* and *goal* of the research. In preparing the panel, several different possible types of targets were identified [63]:

1. Computational models of performance that accurately describe the patterns and regularities observed in expert performances and can make predictions regarding aspects of expert performances;
2. 'Cognitively adequate' computational models that, through their very structure and conceptual design, reflect an observed or hypothesised cognitive reality;

<sup>1</sup>held in the context of the *MOSART Workshop on Current Research Directions in Computer Music*, Barcelona, Nov. 15-17, 2001; participants were Xavier Serra (UPF; moderator), Giovanni De Poli (DEI), Henkjan Honing (NICI), Johan Sundberg (KTH), and Gerhard Widmer (OFAI). A summary of the panel discussion (including the written position statements of the participants) can be found in [43].

3. Computational models of performance that provide optimal user control over the expressive renderings of performances;
4. Computational models of performance that produce well-sounding musical results and thus are useful to the music (software) industry.

All these goals are legitimate in the sense that they are both scientifically interesting and may lead to practically useful results. They do, however, place different emphasis on different qualities of the desired models, and they require both different research approaches and different strategies for evaluating the relevance and utility of the resulting models.

In the process of the discussion, four rather different and complementary research strategies were identified that have been (in part) and will be pursued by different partners in task T5:<sup>2</sup>

**Analysis by Synthesis:** In the context of performance research, the analysis-by-synthesis strategy has been pursued at KTH for more than two decades [7, 50, 36, 38, 40, 41]. The essence is to codify one's hypotheses regarding meaningful performance patterns and strategies in a computational model (e.g., a set of performance rules), to apply the model to musical material, and to test the adequacy of the results by analysing the musical quality of the resulting music (e.g., via listening tests). The principal goal is to arrive at a model that produces musically reasonable results (and thus also provides at least indirect evidence for the adequacy of the assumptions coded in the rules). The analysis-by-synthesis strategy could thus be related particularly to goals 3 and 4 listed above.

**Cognitive Modeling:** The goal of this type of approach is to better understand music cognition (cf. goal 2 from above). The method is to start with hypotheses concerning cognitive processes, representations, and constraints related to a musical task, to formalize them in the form of algorithms, to validate the predictions with experiments, and, often, to adapt the model (and theory) accordingly. Advocators of this approach stress the importance of controlled experiments and a (near-)exhaustive investigation of the space of possibilities (the 'performance space'), rather than experiments based on large corpora of real music performances [28]. Within the MOSART consortium, this approach is best exemplified by the work performed at NICI (e.g., [54, 56, 57, 64]).

**Analysis by Machine Induction:** An alternative way of building computational models of a phenomenon is to start from empirical data, rather than from hypotheses, and to use methods of computational data analysis to derive models from the data [63]. In the context of performance research, this approach has been developed particularly by MOSART partner OFAI (e.g., [58, 60, 62]), where large corpora of real music performances are collected and precisely measured, and methods from the field of Artificial Intelligence (in particular, Inductive Machine Learning and Data Mining) are used to directly extract statistically significant patterns and regularities from the data; these patterns can be described

as general performance rules and constitute (partial) predictive models of performance. This research strategy clearly relates to goal 1 from above.

**Acoustic and Perceptual Analysis:** This approach, which has been developed and pursued mainly by DEI, Padova, in the past few years (e.g., [14, 26, 19, 16]), consists in measuring performances under variety of conditions (often in a controlled setting) — e.g., performances of a particular piece in a variety of different 'moods' —, extracting informative features from these performances (usually directly from the acoustic sound signal), and performing statistical analyses in order to identify those parameters or features that seem to explain most of the variance related to a particular dimension of interest (e.g., the 'mood' intended by a performer, or inferred by listeners). This approach relates most directly to goal 3 from above; indeed, DEI has recently developed a conceptual framework (based on an abstract 'control space') for controlling and 'morphing' the expressive content of musical performances [21, 14, 26, 19, 20, 15, 17, 18].

We feel that this diversity of approaches is appropriate and fruitful in catering for the diverse questions that need to be studied. Having such diverse and complementary research programmes within MOSART (and specifically within T5) is an asset that we plan to exploit further, not only by continuing to pursue each of them independently, but also by combining and integrating various aspects of these approaches in further joint research.

### 3.4. Performance Pattern Research: Preliminary Results and Currently Ongoing Work

Research on the detection and modelling of patterns in expressive performance is already actively being pursued at the T5 partners' sites. Quite substantial work has been done both on somewhat auxiliary, but fundamental issues (like the experimental determination of the level of reliability of various ways of measuring performances) and on some of the central aspects of performance pattern detection and modelling. A brief summary of this research (which is currently ongoing) is given here:

#### Tests concerning precision and reliability of measurement methods:

Basic experiments were carried out to quantify the reliability and precision of some of the currently available methods for measuring the details of expressive performances. In a cooperation between KTH (Stockholm) and OFAI (Vienna), precise measurements of both the measuring and the reproduction precision of computer-monitored grand pianos were performed [42]. Alternative grand pianos are currently under investigation. At OFAI, experiments were also carried out to quantify the reliability of onset detection from audio signals [33].

**Performance databases:** The MOSART T5 partners have accumulated substantial databases of expressive performances, some recorded under 'natural' circumstances, some produced under controlled conditions. Among these are collections of MIDI recordings of several Mozart piano sonatas by a two concert pianists (OFAI), high-level measurements of tempo and dynamics from recordings by famous artists (OFAI), a database of vibrato examples for a variety of instruments (NICI), and a database of grace notes and ornaments in piano performance (NICI), which contains hun-

<sup>2</sup>For reference, these different approaches have been given short and rather arbitrary names that may not reflect precisely how the individual research partners might characterise them.

dreds of recorded and modeled performances of sixteen pianists at seven different tempi. Maintaining and extending these data collections is a major effort.

**Models of rhythm quantisation and tempo tracking:** In order to be able to make sense of the rhythmic structure of a live performance, the computer needs to be able to quantise the time durations in the performance to their score-bound note duration categories. This process (also referred to as categorisation [29, 30]) is tightly intertwined with the ability to follow tempo changes and the detection of beat and/or meter [4]. Both at NICI and OFAI further progress has been made in developing tempo trackers [34, 22]. At the NICI a database of Beatles performances was collected (containing 216 performances) to evaluate existing tempo trackers. This database is shared among the network partners [5]. It was used for evaluations described in [34] and [22]. Quantisation and beat tracking are also a prerequisite for being able to extract tempo and timing-related features from recorded performances (see below).

**Feature extraction for music expression analysis:** Research on ‘patterns’ in music performance presupposes the existence of parametric dimensions over which these patterns can be defined. That in turn requires that meaningful low-level features be extracted from the performed music. Algorithms have been and are being developed to process low level information and to extract performance-relevant features from audio recordings. These features (or ‘cues’) can then be used to understand several characteristics of the performances by means of, for example, statistical analysis or machine learning techniques. This work is done mostly by T5 partners DEI (Padova), DIST (Genova), and KTH (Stockholm) (e.g., [31, 32]).

**Identification of performers based on global performance features:**

Regarding the utility and empirical validity of such features, OFAI has recently managed to show in preliminary experiments that given the definition of appropriate musical features and musical pattern recognition capabilities, computers are capable of automatically recognizing and classifying different performers [51, 52, 53]. A set of basic features were identified that permit the automatic discrimination and recognition of different performers based on timing, dynamics, and articulation. In initial experiments with real performance data, the methods developed achieved an astoundingly high level of classification/recognition accuracy — in fact, a level of accuracy unlikely to be matched by human listeners under comparable experimental conditions.

**Estimation of parameters in performance rule systems:** A new mathematical method was developed for the estimation of optimal parameter settings in the currently best-known and most influential rule-based performance model, KTH’s ‘Director Musices’ performance grammar [37, 41]. In earlier experiments [38], it had been shown how one particular parameter could be optimised. Newly developed methods at DEI [65, 66] make use of the theory of Hilbert space, by representing a performance or a rule as a vector in a ‘performance space’ in which distances are defined according to the perceptive characteristics of the human ear and the fitting is obtained with an orthogonal projection. Preliminary results confirm this methodology

and give a numerical idea of how near the selected rule system can approach a real human performance. The method also permits a comparison between the synthesis produced by different rule systems.

**Discovery of patterns in piano articulation:** In joint research by KTH and OFAI, detailed empirical (statistical) studies on articulation patterns were performed on OFAI’s Mozart sonata corpus [10]. The result is a set of new articulation (staccato) rules that were integrated in KTH’s ‘Director Musices’ performance grammar [8]. These results also constitute a chapter in Roberto Bresin’s Ph.D. thesis at KTH (2000) [7].

**Discovery of patterns in note-level timing and dynamics:**

Based on a new inductive rule learning algorithm developed at OFAI [59], a set of robust, general performance patterns in timing, dynamics, and articulation, in the form of note-level performance rules, were discovered. In experiments with performances by different concert pianists and music of different styles, it was shown quantitatively that these rules generalise surprisingly well to other performers and other styles and thus seem to represent truly fundamental principles of note-level performance [61].

**Models of vibrato:** Vibrato is a key feature in life-like synthesis of a variety of instruments and their expressive control [27]. At NICI, much progress was made in developing fundamental frequency extraction methods essential for the study of the use of vibrato in music performance. This resulted in an overview and implementation of state-of-the-art f0-extraction techniques [46]. It was shown that with the use of explicit knowledge more reliable f0-trajectories could be obtained [47, 48]. This also resulted in a large database (ca. 400 performances) of audio recordings with fundamental frequency and amplitude information for a wide variety of instruments. At KTH, a computational model of violin vibrato was developed, based on measurements of Schubert violin performances, and integrated into the Director Musices system at KTH [49].

**Models of the performance of melodic ornaments:** Ornaments like grace notes, trills, etc. play an important role in much of classical music. Also at NICI, a large study on music performance was completed concerning the modeling of grace notes in piano performance. Next to a number of empirical studies [28, 54, 55, 56, 57, 64], this research resulted in a model and web demonstration that can be used as an exploratory and explanatory device for music students, musicians, and researchers in music performance [2, 3].

**Extraction of expressive content from multimodal information:**

Joint work between DEI and DIST is currently under way that deals with the development of models and algorithms for the extraction of high level, qualitative information to recognize expressive content from multi-modal input derived, among others, from a gesture recognition interface [13, 12]. The analysis of expressive gestures from the performers has to be performed from a multimodal perspective (e.g. how to use information coming from the analysis of expressive content in human movement to perform a better and deeper analysis of expressive content in music performances and vice versa).

**Parameter patterns related to emotional effects:** Quite some work is being done on the identification of performance cues that seem directly related to emotional effects of the music, both at DEI, Padova, and KTH, Stockholm. Performers were asked to play pieces of music with different ‘intentions’ (i.e., to convey different ‘moods’), and statistical analysis techniques identified some rather basic features in timing, dynamics, and articulation that seem to account for a large proportion of the variance related to these different artistic intentions. Based on this, an abstract ‘control space’ was defined at DEI for controlling and ‘morphing’ the expressive/emotive content of musical performances [21, 14, 26, 19, 20, 15, 17, 18]. At KTH, this research has led to the development of a set of new, specialised ‘rule palettes’ within the Director Musices system [9].

**Relation to patterns in human motion:** KTH is exploring different relations between biological motion and music. The direct relation between gesture and tone were studied by comparing arm and hand movements of drummers with produced timing variations of simple drum patterns [23, 24]. Current work includes also investigations of the relation between musicians’ movements and the musical expression as well as motion patterns of the turntable produced by scratchers performing typical patterns [35]. The indirect relation between tempo curves and common human motion curves were investigated in a couple of studies. The shape of the slowing down in the end of a piece (final ritardando) was found to be similar to how people stop from running [39]. When a hand gesture shape was used for changing the tempo, listeners rated the music as more gestural, human, musical and expressive than when a simple tempo shape was used [44]. The final ritardando curve was then transferred back to human motion by applying it to computer models of human walk [11].

All of this is currently ongoing work, and it is planned not only to continue the research along these lines, but also to try to integrate some of these specialised results into more comprehensive models of expressive performance.

#### 4. PLANS FOR PHASE II OF TASK T5

During the second half of the project, the efforts at discovering, formalizing, and empirically testing the validity of relevant patterns in musical performance (both at the sound level and at symbolic levels) will be intensified. The problem can broadly be divided into a number of major steps; all of these will be tackled by the partners in T5, some in isolation by those partners best equipped for the task, some in cooperation.

1. Definition of relevant expressive dimensions and classes of patterns. That is the topic of the present report and can be regarded as solved.
2. Acquisition and preparation of substantial amounts of empirical data (detailed measurements of actual performances by various performers on various instruments) on which these studies could be carried out. We are well on our way towards this goal (see above).
3. Development of audio and music analysis algorithms for extracting the relevant features related to patterns of interest. This requires expertise in audio / signal processing and

is being studied by the consortium partners specialized in that area (e.g., DEI).

4. Search for quantifiable regularities and patterns. This is the central task and will be pursued according to the different research strategies outlined in section 3.3 above. As the preliminary results listed in section 3.4 above indicate, a wealth of new discoveries and models can be expected.
5. Formulation of quantitative models of these patterns (in the form of mathematical models, predictive rules, etc.).
6. Systematic evaluation of these patterns in large and representative corpora of musical test data. Again, the evaluation may take different forms, depending on the research approach taken (cf. section 3.3 above).
7. Based on this, development of novel computer-based techniques for the analysis, control, and synthesis of ‘expressive’ performances. The MOSART partners DEI and KTH are already actively involved in such research.
8. Identification of opportunities and goals for future research in this area, not least with a view to establishing this research field (in which Europe has a definitive competitive edge) as a topic of importance in the framework of European Community research. This will be a joint effort.

The MOSART consortium expects that automatic pattern discovery and classification in music will become increasingly important, especially with the growing importance of content-based manipulation techniques in the field of multimedia. It is our goal to turn this into a generally recognised research area. The young researchers financed by the MOSART network will play an important role in these efforts.

#### 5. REFERENCES

- [1] *The Mystery of Vibrato* (1999). Documentary of Dutch National Television (VPRO Noorderlicht science series) showing vibrato research at NICI and IRCAM. Video.
- [2] Grace Note WWW-demo, see <http://www.nici.kun.nl/mmm> under “demos”.
- [3] Grace Note CD-ROM (in press).
- [4] “*Give me the beat*” (1999). Documentary of Dutch National Television (TELEAC ‘Wetensnappen’ educational science series) on rhythm perception research. Video.
- [5] Data base of Beatles performances (2002). See [22] for more info and <http://www.nici.kun.nl/mmm/>.
- [6] Brazil, E. (2002). *Proceedings of the International Conference on New Interfaces for Musical Expression*, Media Lab Europe, Dublin, Ireland, May 24-26, 2002.
- [7] Bresin, R. (2000). *Virtual Virtuosity: Studies in Automatic Music Performance*. Doctoral Dissertation, Royal Institute of Technology (KTH), Stockholm, Sweden (KTH Report TRITA-TMH-2000:9).
- [8] Bresin, R. (2001). *Articulation Rules for Automatic Music Performance*. In *Proceedings of the International Computer Music Conference (ICMC'2001)*, La Habana, Cuba.
- [9] Bresin, R. and Friberg, A. (2000). Emotional Coloring of Computer-Controlled Music Performances. *Computer Music Journal* 24(4), 44–63.

- [10] Bresin, R. and Widmer, G. (2000). Production of Staccato Articulation in Mozart Sonatas Played on a Grand Piano. Preliminary Results. Speech Music and Hearing Quarterly Progress and Status Report, 4/2000, KTH, Stockholm.
- [11] Bresin, R., Friberg, A., and Dahl, S. (2001). Toward a New Model for Sound Control. In *Proceedings of DAFx01*, 45–49.
- [12] Camurri, A., De Poli, G., and Leman, M. (2001). MEGASE: a Multisensory Expressive Gesture Applications System Environment for Artistic Performances. *CAST01 Conference*, GMD, Bonn, 21-22 Sept 2001.
- [13] Camurri, A., De Poli, G., Leman, M., and Volpe, G. (2001). A Multi-layered Conceptual Framework for Expressive Gesture Applications. *Workshop on Current Research Directions in Computer Music*, Barcelona, Nov 15-16-17, 2001.
- [14] Canazza, S. (2002). Analisi e morphing del contenuto espressivo di un' esecuzione musicale. *Suoni in corso: percezione ed espressione dell'uomo tecnologico*. Mitterfest Editore. Cividale del Friuli, Febbraio 2002, pp. 31-46.
- [15] Canazza, S., Cestonaro, F., De Poli, G., Drioli, C., and Rodà, A. (2000). Symbolic and audio processing to change the expressive intention of a recorded music performance. *Proc. of CIM 2000*, L'Aquila, 2-5 September, pp. 71-74.
- [16] Canazza, S., D'Arduini, G., and Rodà, A. (2000). Analysis of the influence of expressive intention in piano performance of classical music. *Proc. of CIM 2000*, L'Aquila, 2-5 September, pp. 53-58.
- [17] Canazza, S., De Poli, G., Drioli, C., Rodà, A., and Zamperini, F. (2000). Real-time morphing among different expressive intentions in audio playback. *Proc. of ICMC*, Berlin, 27 august-1 september pp. 356-359.
- [18] Canazza, S., De Poli, G., Drioli, C., Rodà, A., and Vidolin, A. (2000). Audio morphing different expressive intentions for Multimedia Systems. *IEEE Multimedia*, July-September, Vol. 7, N 3, pp. 79-83.
- [19] Canazza, S., De Poli, G., Rodà, A., Vidolin, A., and Zanon, P. (2001). Kinematics-Energy space for expressive interaction in music performance. *Proc. of MOSART Workshop on current research directions in Computer Music*, Barcellona, November 15-17, pp. 35-40.
- [20] Canazza, S., De Poli, G., Drioli, C., Rodà, A., and Vidolin, A. (2001). Expressive morphing for interactive performance of musical scores. *Proc. of First International Conference on WEB Delivering of Music*, Florence, Italy, 23-24 november, pp. 116-122.
- [21] Canazza, S., De Poli, G., Rodà, A., and Vidolin A. (2002). *Too marvelous for words: An abstract control space for non-verbal communication*. In preparation.
- [22] Cemgil, T., Kappen, B., Desain, P., and Honing, H. (2001). On Tempo Tracking: Tempogram Representation and Kalman Filtering. *Journal of New Music Research* 29 (4), 259–273.
- [23] Dahl, S. (2000). The Playing of an Accent – Preliminary Observations from Temporal and Kinematic Analysis of Percussionists. *Journal of New Music Research*, 29(3), 225–233.
- [24] Dahl, S. (2002). *Playing the Accent – Comparing Striking Velocity and Timing in an Ostinato Rhythm Performed by Four Drummers*. Submitted.
- [25] De Boer, M., Bonada, J., Cano, P., Loscos, A., and Serra, X. (2000). Singing Voice Impersonator Application for PC. *Proceedings of the International Computer Music Conference (ICMC'2000)*, Berlin. San Francisco: International Computer Music Association.
- [26] De Poli, G., Canazza, S., Drioli, C., Rodà, A., Vidolin, A., and Zanon, P. (2001). Analysis and modeling of expressive intentions in music performance. *Proc. of International Workshop on Human Supervision and Control in Engineering and Music*. Kassel, Germany, September 21-24,
- [27] Desain, P. and Honing, H. (1996). Modeling Continuous Aspects of Music Performance: Vibrato and Portamento. In *Proceedings of the International Music Perception and Cognition Conference*.
- [28] Desain, P., Honing, H. and Timmers, R. (2001). Music Performance Panel: NICI / MMM Position Statement. *MOSART Workshop on Current Research Directions in computer Music*, Barcelona, Nov. 2001.
- [29] Desain, P. and Honing, H. (2002) Modeling the Effect of Meter in Rhythmic Categorization: Preliminary Results. *Journal of Music Perception and Cognition*. Sapporo: JSMPC.
- [30] Desain, P. and Honing, H. (2002) *The Perception of Time: The Formation of Rhythmic Categories and Metric Priming*. Submitted.  
See <http://www.nici.kun.nl/mmm/time.html>.
- [31] Dillon, R. (2001). Extracting Audio Cues in real time to understand musical expressiveness. *Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*, Barcelona, November 2001.
- [32] Dillon, R. (2002). Expressive performance classification by audio cues analysis. *Stockholm Music Performance Symposium* (in press).
- [33] Dixon, S. (2001). Learning to Detect Onsets of Acoustic Piano Tones. In *Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*, Barcelona
- [34] Dixon, S. (2001) An Empirical Comparison of Tempo Trackers. *Proceedings of the 8th Brazilian Symposium on Computer Music*, Fortaleza, Brazil.
- [35] Falkenberg Hansen, K. and Bresin, R. (2002). Scratching: From Analysis to Modeling. In *Models and Algorithms for Control of Sounding Objects*, Deliverable 8, EU-IST Project no. IST-2000-25287, 15-48 (<http://www.soundobject.org/papers/deliv8.pdf>).
- [36] Friberg, A. (1991). Generative Rules for Music Performance: A Formal Description of a Rule System. *Computer Music Journal* 15(2), pp.56–71.
- [37] Friberg, A. (1995). *A Quantitative Rule System for Musical Performance*. Ph.D. dissertation, Department of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH), Stockholm.
- [38] Friberg, A. (1995). Matching the Rule Parameters of Phrase Arch to Performances of Träumerei: A Preliminary Study. In A. Friberg and J. Sundberg (eds.), *Proceedings of the KTH Symposium on Grammars for Music Performance*, May 27, 1995, pp.37–44.

- [39] Friberg, A. and Sundberg, J. (1999). Does Music Performance Allude to Locomotion? A Model of Final ritardandi Derived from Measurements of Stopping Runners. *Journal of the Acoustical Society of America* 105(3), pp.1469–1484.
- [40] Friberg, A., Bresin, R., Frydén, L., and Sundberg, J. (1998). Musical Punctuation on the Microlevel: Automatic Identification and Performance of Small Melodic Units. *Journal of New Music Research* 27(3), pp.271–292.
- [41] Friberg, A., Colombo, V., Frydén, L. and Sundberg, J. (2000). Generating Musical Performances with Director Musices. *Computer Music Journal* 24(3), 23–29.
- [42] Goebel, W. and Bresin, R. (2001). Are Computer-controlled Pianos a Reliable Tool in Music Performance Research? Recording and Reproduction Precision of a Yamaha Disklavier Grand Piano. In *Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*, Barcelona.
- [43] Grachten, M. (2001). Summary of the Music Performance Panel. *MOSART Workshop on Current Research Direction in Computer Music*, Barcelona, Nov. 2001.
- [44] Juslin, P.N., Friberg, A., and Bresin, R. (2002). Toward a Computational Model of Expression in Performance: The GERM Model. *Musicae Scientiae*, special issue 2001-2002, pp 63-122.
- [45] Murphy, D. (2002). Pattern Play. *2nd International Conference on Music and Artificial Intelligence (ICMAI'2002)*, Edinburgh, Scotland (poster). To appear in On-line Technical Report Series, University of Edinburgh.
- [46] Rossignol, S., Desain, P., and Honing, H. (2001a). State-of-the-art in fundamental frequency tracking. In *Proceedings of the Workshop on Current Research Directions in Computer Music*.
- [47] Rossignol, S., Desain, P. and Honing, H. (2001b). Refined knowledge-based f0 tracking: comparing three frequency extraction methods. In *Proceedings of the International Computer Music Conference*, San Francisco: ICMA.
- [48] Rossignol, S., Desain, P., and Honing, H. (2002). Refined knowledge-based f0-tracking. Manuscript, in preparation.
- [49] Schoonderwaldt, E., and Friberg, A. (2001). Towards a Rule-based Model for Violin Vibrato. In *Proceedings of the Workshop on Current Research Directions in Computer Music*, Barcelona, 2001.
- [50] Sundberg, J., Friberg, A., and Bresin, R. (2001). Music Performance Panel: Position Statement. *MOSART Workshop on Current Research Directions in Computer Music*, Barcelona, Nov. 2001.
- [51] Stamatatos, E. (2001). A Computational Model for Discriminating Music Performers. In *Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*, Barcelona.
- [52] Stamatatos, E. (2002). *Quantifying the Differences between Music Performers: Score vs. Norm*. Submitted to International Computer Music Conference (ICMC'2002).
- [53] Stamatatos, E. and Widmer, G. (2002). Music Performer Recognition Using an Ensemble of Simple Classifiers. In *Proceedings of the 15th European Conference on Artificial Intelligence ECAI'2002*, Lyon, France.
- [54] Timmers, R. (2001). Context-sensitive evaluation of expression. In *Proceedings of the workshop on Current Research Directions in Computer Music*, 75-78.
- [55] Timmers, R. (2002). On the contextual appropriateness of performance rules. In *Proceedings of the SRPMME conference on Music Performance*. London.
- [56] Timmers, R., Desain, P., Honing, H., and Trilsbeek, P. (2002). Introducing a model of grace note timing. In *Proceedings of the Workshop on Music, Motor Behavior and the Mind*. Ascona.
- [57] Timmers, R., Ashley, R., Desain, P., Honing, H., and Windsor, L. W. (2002). Timing of ornaments in the theme of Beethoven's Paisiello Variations: Empirical Data and a Model. *Music Perception* 20 (1).
- [58] Widmer, G. (2001). Using AI and Machine Learning to Study Expressive Music Performance: Project Survey and First Report. *AI Communications* 14(3), 149-162.
- [59] Widmer, G. (2001). Discovering Strong Principles of Expressive Music Performance with the PLCG Rule Learning Strategy. In *Proceedings of the 12th European Conference on Machine Learning (ECML'01)*, Freiburg. Berlin: Springer Verlag.
- [60] Widmer, G. (2001). The Musical Expression Project: A Challenge for Machine Learning and Knowledge Discovery. Invited talk / paper. In *Proceedings of the 12th European Conference on Machine Learning (ECML'01)*, and in *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01)*, Freiburg, Germany. Berlin: Springer Verlag.
- [61] Widmer, G. (2002). Machine Discoveries: A Few Simple, Robust Local Expression Principles. *Journal of New Music Research* 31(1).
- [62] Widmer, G. (2002). Trying to Explain a Creative Act: Studying Expressive Music Performance with Learning Machines. In *Proceedings of the ESCOM Conference on Musical Creativity*, Liège, Belgium. (Extended version to appear in the journal *Musicae Scientiae*, 2003).
- [63] Widmer, G., Dixon, S., Goebel, W., Stamatatos, E., and Toubudic, A. (2001). Empirical Music Performance: FAI's Position. Panel Discussion Paper, *MOSART Workshop on Current Research Directions in Computer Music*, Barcelona, Nov. 2001.
- [64] Windsor, W. L., Desain, P., Aarts, R., Heijink, H., and Timmers, R. (2001). The timing of grace notes in skilled musical performance at different tempi: a case study. *Psychology of Music* 29, 149-169.
- [65] Zanon, P., Canazza, S., and Rodà, A. (2000). A method for an objective comparison of rule systems for musical performance. *Proc. of CIM 2000*, L'Aquila, 2-5 September, pp. 193-196.
- [66] Zanon, P., De Poli, G., and Rodà, A. (2001). Estimation of parameters in rule systems for expressive rendering in musical performance. *Computer Music Journal* (submitted).



**IHP Network HPRN-CT-2000-00115 MOSART  
Music Orchestration System in Algorithmic Research and  
Technology**

**MOSART Task 6:**

**Computer Music Composition Tools**

**Edited by Barry Eaglestone**

**Deliverable d26:**

Definition of composition tools, to be supported by data bases

**Table of Content**

**Requirements specification for a composition tools system**

Page 258

Barry Eaglestone, Guy Brown, Nigel Ford, Adrian Moore, Ralf Nuhn



# MOSART REPORT

## REQUIREMENTS SPECIFICATION FOR A COMPOSITION TOOLS SYSTEM

Barry Eaglestone (editor), Guy Brown, Nigel Ford, Adrian Moore, Ralf Nuhn,  
University of Sheffield, UK

*email: B.Eaglestone@sheffield.ac.uk*

### Abstract

This report constitutes the first deliverable of the “Composition Tools” Task (T6) of the MOSART Research Network Project and an Interim Report on a Qualitative Analysis of Composers as Work conducted at the University of Sheffield in collaboration with Mosart. In it we document work towards defining a set of requirements for enhanced support for electroacoustic music composers. Electroacoustic music composition tools and systems selectively attempt to provide composers with services they require for music generation, e.g., for accessing, generating, organising and manipulating audio (and other) objects, which constitute the composition. However, a primary aim of composition software is also to create conditions in which composers can be creative in the use of these services. The evolution of composition software with respect to the former requirements has been dynamic and innovative, stimulated by continuing development of new programming and audio synthesis paradigms and techniques. However, we believe that the aim of providing a fertile environment for creativity has been largely ignored. Specifically, we believe there to be a need to establish a research base for enhancement of support within composition software for creativity. It is this aspect that is addressed by this report. In the first part of the report (chapters 2 and 3), we address the two questions: do research results exist already that can provide an effective research base, and if not, how should such research be conducted? Accordingly, chapter 2 provides a critical review of previous studies of composition and software to support composers, and the research methods used. In the light of this review, it then considers which research methodology will be most appropriate for further research into music generation systems. It concludes that there is need for studies towards a research base for enhancement of electroacoustic music composition software, such that creativity is better supported. Further, given the preliminary nature of such research, the methods used should be qualitative, with the aim of identifying the sensitising concepts. This preliminary study should involve observation of composers at work in naturalistic situations. A second phase is then envisaged which will involve both qualitative and quantitative methods used in tandem. A future aspiration is to devise software solutions to improve support for creativity, in which engineering research methodology will become appropriate. The second part of the report documents our experimental study of composers, in which we have applied the research methodology proposed. Chapters 3 and 4 respectively document and analyse a series of three experiments, conducted with progressive refinements to our research methodology. In each, composers were observed working in naturalistic setting over protracted periods of time. The main notions that emerged from the analysis of the rich data collected in this way are reiterated in chapter 5, the third part of the report, which concerns the software requirements motivated by this study. These concern: the value of the unexpected; the value of diverse, unfamiliar and idiosyncratic tools; the need for support for composers both as users and

programmers; the key role of new tools; the need for personal and shared know-how; the need for physical and visuo-spatial control of audio; support to allow users to multitask; and the support for holistic approaches to composition. The requirements which address these are then set out from the perspectives of: system architectures; object and process management; data and knowledge management; and composer-computer interfaces. The report concludes with a summary and identification future work toward elaborating the theories in this document and validating them through prototyping and evaluating composition systems and tools at Sheffield and DIKU.

## **ACKNOWLEDGEMENT**

This ongoing research has been partly funded by the Mosart Research Network programme (<http://www.diku.dk/research-groups/musinf/MOSART.html>), funded under the EU Framework 5

## Contents

Chapter 1 - Introduction	4
1.1 Introduction	4
1.2 The Nature of the Problem	5
1.3 An Overview of the Requirements Analysis Exercise	6
Chapter 2 Research Approach and Methodology	7
2.1 Introduction	7
2.2 Related Work	7
2.3 Methodological Issues	11
2.4 Methodology and Experimental Design	12
2.5 Conclusions	14
Chapter 3 Experiment Design	15
3.1 Introduction	15
3.2 First Experiment	15
3.2.1 Verbal Protocol	15
3.2.2 Video camera recording	16
3.2.3 Computer data	16
3.2.4 Interviews	16
3.2.5 Experimental Setting and Constraints	16
3.2.6 Reflections on Experiment One	17
3.3 Second Experiment	17
3.4 Third Experiment	18
3.5 Conclusions	19
Chapter 4 A Qualitative Analysis of Composers at Work	20
4.1 Introduction	20
4.2 Preliminary Results and Discussion	20
4.2.1 A taxonomy of Electroacoustic music composers	20
4.2.2 Creativity through diversity	21
4.2.3 The impact of visual senses	21
4.2.4 The impact of the semantic gap	22
4.2.5 Limitations to the role of the computer in supporting creativity	22
4.2.6 The impact of inappropriate interfaces	23
4.2.7 The need for accessible individual and community knowledge and dialogue within and beyond the community	24
4.3 Validation of our model of creativity	24
4.4 Summary	25
Chapter 5 Composition System Requirements	27
5.1 Introduction	27
5.2 Summary of the main notions from our analysis	27
5.3 Implied software requirements	28
5.3.1 System Architectures	28
5.3.2 Object and Process Management	30
5.3.3 Data and knowledge management	30
5.3.4 Composer-computer interfaces	31
Chapter 6 Conclusions And Future Work	32
6.1 Summary	32
6.2 Future Work	33
References	34
APPENDIX A – Catalogue of audio-visual media	37
APPENDIX B – The Pattern Play System	38

# CHAPTER 1 – INTRODUCTION

## 1.1 Introduction

This report constitutes the first deliverable of the “Composition Tools” Task (T6) of the MOSART Research Network Project (<http://www.diku.dk/research-groups/musinf/mosart/>).

T6 is a study of composition tools, through analysis of the state of the art, a requirements analysis, and the formulation and validation of new theories through the construction and evaluation of prototype composition tools. In this report we set out requirements, derived from the requirements analysis phase, which will form the basis for the subsequent theory testing through prototyping and evaluation of composition tools.

The requirements presented are designed to support electroacoustic music composition, and have emerged through a qualitative study of electroacoustic composers at work, carried out mainly by the Sheffield Partner, through experiments and analysis conducted by Ralf Nuhn, a Research Associate funded by MOSART. The experiments have involved observing composers at work. Data collected has been analysed using grounded theory (Ellis 1993), and the implications of the theory have been worked through with respect to software requirements.

Aspects of this work have been facilitated by the Stockholm and Barcelona Partners. Specifically, Stockholm have provided video recording of composers at work, and Barcelona has facilitated data collection, by Mr Nuhn, through interviews with composers working there. More general contribution to this task have been made through the MOSART Network, through presentation and discussion of the research and its results at the three MOSART Workshops, held, respectively in Copenhagen, Padova and Barcelona. In particular, a Panel Session was held at the Barcelona Workshop to present and discuss views and issues relating to T6. Other partners, specifically, at DIKU Copenhagen, cART CNR Pisa, InfoMus DIST Genoa, DAIMI Århus, and Esbjerg, to various extents have contributed to the development of this report, by responding to the content of the initial draft from their various perspectives.

Future contributions to this study are anticipated from complementary projects ongoing at various MOSART partner laboratories. Specifically, at DIKU there are three projects which have relevance to this study of composition tools. These are: the Pattern Play System (Murphy 2002) (see Appendix B); the Timbre Engine project (Jensen 1999; Marentakis & Jensen 2001); and a “digital turntable”, which is being researched by Tue Haste Andersen. The Pattern Play system is a composition tool that addresses directly many of the requirements that have emerged from the study documented in this report (see chapter 5). Compositions have been commissioned for the latter two systems, thus providing potential for further case study data.

Other research within MOSART of relevance to T6 includes development of synthesized “conventional” instruments with enhanced physical and perceptual modifications for use by performers and composers. This includes work on synthesized piano (Bensa, Gibaudan, Jensen, & Kronland-Martinet 2001; Bensa, Jensen, Kronland-Martinet, & Ystad 2000) at DIKU and CNRS-LMA.

Note that the problem of specifying software to support music composition is not a conventional one, in that there does not yet exist a significant research base for this area, and the application is a creative and open-ended one. Thus, a key aspect of the study is to understand better the needs of creative workers in this area, and how creativity can be better supported by software environments. Consequently, there is no clearly defined problem to solve. Accordingly, we present speculative requirements, inferred from the sensitising concepts (Olaisen, 1991: 254) that have emerged through analysis of qualitative data. Thus, these embody grounded theories (Ellis 1993) for further research.

The report is organized as follows. Chapter 2 reviews related research and discusses research methodology issues. Our research has been conducted through a series of experiments in which we have observed composers at work in naturalistic settings. The design of these experiments is discussed in chapter 3 and a catalogue of the data collected is provided as Appendix A. A preliminary analysis of the experimental data is given in chapter 4. Finally, the software requirements for composition

systems, derived from this analysis in chapter 4, are presented, followed by conclusions and recommendations for future work. Finally, a brief description of a complementary research project, the Pattern Play system (Murphy 2002), is given in Appendix B

In this introductory chapter we establish the nature of the problem, (section 1.2) and overview the requirements analysis documented in this report.

## 1.2 The Nature Of The Problem

The nature of the problem of providing software support for electroacoustic music composers has two facets, since electroacoustic music composition software serves two purposes. Firstly, it must make available to the composer services by which she can create a composition, i.e., services to retrieve, manipulate and combine musical artefacts. Secondly, the software must provide an environment within which those services can be used creatively.

Research into digital signal processing and the artist's use of sounds is ongoing. Consequently, services relating to musical artefacts are volatile and evolving as new techniques and paradigms are integrated into composition software. This is evident, for example, in the proceedings of the annual International Computer Music Conference series (published by the International Computer Music Association), since these provide a showcase for new research in this area, including new audio-related techniques and composition software.

The software environment within which those services are used creatively has largely been under-researched. Instead, software developers have applied current wisdom on what constitutes good software engineering. Accordingly, the evolution of composition software has largely paralleled the evolution of paradigms in software technology. Thus, early systems, e.g., csound (Moore, 1990) and cmusic (Vercose, 1985), supported asynchronous use and resembled assembler programming languages, whereas subsequent languages and systems first introduce higher-level abstractions, and later object orientation. Similarly, there has been a move from asynchronous to synchronous systems, and from text to graphical user interfaces.

There are exceptions to the above, where composition software diverges radically from conventional software engineering wisdom. For example, Think (<http://www.hitsquad.com/>

smm/programs/Think/) generates sounds from audio files using granular synthesis techniques, and boasts that the user has "no control whatsoever" over the process. Thus, this tool can be used to generate unanticipated material to solve "writers blocks without them having to think at all". However, such tools are the exception, and for the main part, composition software has followed a more conservative approach.

We believe that this conservative approach to composition software design is flawed, since there is an inherent tension between principles of conventional software engineering and the requirements of creative composers (Clowes, 2000; Eaglestone et al., 2001). This tension can be explained in terms of models of creativity in the literature. Creativity is often characterized by the notion of "divergent" as opposed to "convergent" thinking (Guilford, 1967), the latter being associated with relatively predictable logical activity and outcomes, the former with less logical and predictable activity and outcomes. Early pioneers of flight provide a simple example. These presumably initially worked by analogy with bird flight, and tried to devise effective flapping wings - a relatively convergent approach. The breakthrough was to conceive of a solution in terms of a fundamental reconfiguration of elements of the problem, i.e. to drive air over a fixed wing - a relatively divergent approach. The term "relatively" is important here, since a further breakthrough may diverge from the relatively convergent idea (now in common currency) of driving air over a fixed wing. Similarly, for example, a musical idea that is initially innovative (perhaps even shocking) can rapidly become a tired cliché once it becomes an established technique. The extent to which an idea may be thought of as "creative" is therefore time- and context-dependent.

Instances of creativity are often thought to occur as a sudden perception or realisation, occurring when the person is not intensely focused on the particular problem. As Gregory (1987:171) has noted:

"...our brains are at their most efficient when allowed to switch from phases of intense concentration to ones in which we exert no conscious control at all."

De Bono (1987) has described the first stage of thinking as the perception stage - how we look at the world, and the concepts and perceptions that we form; and the second stage as the processing stage - what we do with the perceptions. He

considers that logic can only be used in the second stage since it requires concepts and perceptions to work upon.

It may at first sight appear that computers are irrelevant to creativity - and particularly to De Bono's perception stage (De Bono, 1987) - in that they are better at convergent than divergent information processing tasks, and therefore have little if any role to play in supporting creative thinking. Arguably, to provide relatively direct computer-based support for creativity (as opposed to the more indirect forms of support currently available), we need further to develop knowledge representations and processes at a high level of abstraction, possibly entailing perceptual pattern recognition and matching to complement logical processing<sup>1</sup>, and entailing some element of "non-control", e.g. randomness and serendipity.

### 1.3 An Overview of the Requirements Analysis Exercise

Having discussed the nature of the problem, this subsection now overviews the way in which we have attempted to solve it.

We have identified a clear requirement for a research foundation upon which environments that better support creativity can be based. However, determining how best to support creativity in general, and electroacoustic music composition in particular, is problematic. As Laske observes,

“the kind of musical knowledge that, if implemented, would improve computer music tools is often not public or even shared among experts, but personal, idiosyncratic knowledge...the elicitation of personal knowledge and of action knowledge still awaits a methodology...”  
(cited by Polfreman, 1999:31).

Thus, our initial research aim was to determine what research methodologies are most likely to be effective in establishing that research base.

This study of research methodology is documented in chapter 2. The chapter provides

an extensive review of research into the creative process of composition, from the perspective of the methodologies used. The conclusions drawn are that there does not currently exist a significant research base in this area on which to ground composition software. Consequently, research in this area should be viewed as preliminary, and should be based upon qualitative methodology in order to determine sensitising concepts (Olaisen, 1991: 254) of the problem. Further, we concluded that this study should be based upon empirical data collected through a study of professional composers working in naturalistic settings.

The experiments that were conducted to collect data are described in chapter 3. Chapter 4 presents the subsequent data analysis, and composition system requirements are derived from the analysis in chapter 5.

---

<sup>1</sup> The importance of perceptual pattern recognition and matching in composition is also the focus of ongoing research within MOSART, specifically through experimentation with the “Pattern Play System” (Murphy 2002). A brief system description is given in Appendix B.

# CHAPTER 2 – RESEARCH APPROACH AND METHODOLOGY

## 2.1 Introduction

This chapter explores issues relating to research methodology for the elicitation of software requirements for electroacoustic music composition tools. Specifically, section 2.2 reviews research that already exists in which the nature of composition has been studied. The conclusions drawn are that this research is largely inadequate for our purposes, and thus we must conduct preliminary research towards establishing a research base upon which composition software can be based. Further, conclusions are drawn on which methodological research approaches may provide a more mature and appropriate research base. The main conclusion is that this research base can be provided through qualitative research in which a constructivist approach is taken. Specifically, this research should be a study of composition involving experienced composers working in naturalistic settings. The chapter concludes by defining the research approach and methodology that we have followed in deriving the requirements, presented in the proceeding chapters.

This study of methodological issues is also presented in (Eaglestone, Ford, Nuhn, Moore, Brown, 2001)

## 2.2 Related Work

In seeking to identify requirements for composition software, we have sought related research upon which we can build. However, we have found that research into software support for creativity in electroacoustic (timbre-based) music composition is a largely neglected area. The few articles that do contribute to this area of study are discussed later in this section.

However, there is a small, but more substantial body of research into the composition process for conventional (pitch-based) composition. This work is largely from a musicological or

educational perspective. An excellent and comprehensive review of this work is given in Collins (2001). Collins identifies four theoretical perspectives taken in this work: stage theory; emerging-systems theory; information-processing theory; and Gestalt theory. Stage theories model the composition as a staged process, e.g., preparation, incubation, illumination, and verification (Wallas, 1926), and mainly rely on biographical or autobiographical material. Thus, this approach is usually dependent upon “after the event” introspections or second source interpretations. Emerging-systems theories view composition as an evolutionary processes over a protracted period, and typically rely on case studies to gain insights. More controversially, information-processing theories take a modelling approach, whereby composition is explained or understood by constructing computational systems, which produce characteristic outcomes. Thus this approach is used for analysis through re-synthesis. In each of these cases, the approach reduces composition to a procedure. Though this may be useful in understanding specific compositions in retrospect, we believe that it cannot form a basis for composition software. In general, it is clearly inappropriate for composition software to constrain composers by imposing “best practice” or characteristic working methods and procedures. However, we concur with Collins (2001), that Gestalt theories provide a more promising basis, since the focus is on organizational aspects of problem components, and creativity through re-configuration of those components. This approach is therefore compatible with the model of creativity as divergent associations through flashes of inspiration, discussed in the previous section. However, most work in the area has focused on perception of music, rather than composition.

Empirical studies are widely used in audio and music perception research, but rarely to research the composition process (Sloboda, 1995). Also, the instances in which this has been applied, e.g., using think-aloud protocols and computer-based

data collection, reveal the difficulty in obtaining valid results. Again, Collins (2001) provides an excellent survey of empirical studies in this area. This work falls mainly into two categories, the collection of case study data, e.g., through think-aloud protocols as a professional composer composes (Reitman, 1965), or “close to the event” documentation (Eaglestone et al. 1993, Collins 2001), and collection of data through controlled experiment e.g., by setting a group of people an exercise, such as composing a monophonic tune. The subjects of the latter type of study were typically students, sometimes children and often musically untrained (e.g., Bamberger, 1977; Davidson & Welsh, 1988).

Collins also identifies the development of data collection techniques, “from interviews with, and anecdotal evidence of, composers through to a reliance upon the use of verbal protocol techniques, the use of audio-tapes, and most recently computer-based data collection systems”, but stresses the following limitations to the scope and depth of the empirical studies. There has been very little “time-based” analysis, or studies in naturalistic settings. Collins’ explanation of the latter limitation is that “so-called scientific objectivity, claimed by researchers, has been flavoured by their background as experimental psychologists rather than musicians”. Very few studies go beyond observation of trivial composition exercises, e.g., short monophonic composition, the studies in Reitman (1965) and Collins (2001) being the exceptions. Finally, most studies have used only crude and simple sound sources.

The nature of creativity in electroacoustic music composition is largely unresearched, though there exist a number of reflective and introspective papers on the requirements of individuals, often with proposals for future research directions. An example is Emmerson (1989). Emmerson elaborates a model of composition, the core of which is a cycle of action, i.e., generation of a musical artifact, and test, i.e., evaluation of the artefact. His elaborations introduce an awareness of the importance of a community and individual “knowledge base” and sources of inspiration. He uses this model to motivate a manifesto for future research, centred on the idea that the electroacoustic music community should establish an experimental analysis as a partner to experimental composition.

Electroacoustic music composition provides an interesting instantiation of the creative process. Further, the music community has embraced new

technologies throughout the ages, and in particular, composition systems attempt to provide environments within which creative thought is likely to occur. However, there is a lack of empirical studies in this area. Two rare examples of early work are the Tema project (Eaglestone, 1994; Eaglestone et al., 1993) and the survey of composers’ attitudes to the software they use, in Clowes (2000). Both studies attempt to address general difficulties of investigations in this area. These include lack of standards in electroacoustic composition (e.g., notations, languages, modalities, instruments, etc.), the libertarian, individualistic and revolutionary culture, the often aggressively individualistic, secretive, idiosyncratic and subversive psychology of composers, and the volatile, multi-platform and experimental nature of the technology.

The Tema project (Eaglestone 1994, Eaglestone et al. 1993) was an analysis of composition methods used to create Tema, a piece of electroacoustic ballet music composed in 1986 by Tamas Ungvary and choreographed by Peter Rajjka. Music and choreography were created using the Nuntius music-dance system (Ungvary et al. 1992), a file-base system with subsystems for composer and choreographer, and architectural features and interfaces to allow them to work co-operatively. The composer's system comprises a library of software tools for sound creation, manipulation, analysis and auditioning. However, Nuntius lacked direct support for the creative process, as is still typical of current composition systems (see proceedings of the ICMC throughout the years).

Data for this study was Ungvary’s diary in which he recorded technical and “human” details, including all computer interactions with composition software during the composition of Tema, and the motivations and method behind each program evocation. Thus, Ungvary provided unique insights into the creative process. The scope of this study was clearly limited, since it is impossible to generalise from a single case study how composers compose, or even how Ungvary composes. However, given the rarity of such material, the exercise provides valuable insights.

A modeling approach was taken, using methods associated with software engineering (Chroust 1989), i.e., analysis of transformations to identify objects, events and transitions, followed by generalization of activity descriptions to derive an object-oriented conceptual model. This was in effect a Gestalt-like approach in so far as it



related to the re-configuring of problem components, as opposed to a process model. This model thus provided a framework for a particular composition style. However, with hindsight, the researchers now believe that the main value of this study was the qualitative research through study of electroacoustic composition in a naturalistic setting.

The diary was analysed using methods associated with software engineering (Chroust 1989). Transformations were analysed to identify objects, events and transitions. Activity descriptions were generalised, and finally an object-oriented conceptual model was derived to provide a general framework for a particular composition style.

The analysis was from the perspective that there are similarities between artistic design and engineering, and therefore features of engineering support technology may be usefully adapted for use by artists. However, it became clear from the study that intuition rather than methodology directed the composition process, often within the context of methodical techniques, such as problem decomposition. Further, the design method was phase-oriented (sounds were assembled, then textured and reverberated).

The analogy apparent was of a painter mixing materials to form new colours and applying them to the canvas to form the artwork. It was also apparent that transitions between object states were of interest as well as the objects themselves. These constituted the composition techniques and skills (the artist's technical know-how), which were reused when results were satisfactory. This reflected the experimental nature of the creative process. Ungvary was often "feeling for" that which worked through trial and error, rather than simply implementing his conception. As Ungvary stated:

"Errors will often produce the most artistically interesting results!" (Ungvary, Personal communication).

Finally, we observed that both materials and techniques were frequently reused. This reflects an aspect not identified in our model of creativity, i.e., that of accumulating individual and community know-how which provides both knowledge of resources and proficiency in using the tools and techniques.

It was also observed that Ungvary experienced problems of object and process management that are effectively addressed by engineers' support systems. He experienced difficulty in keeping

track of and retrieving objects, and in operating the various user interfaces that had to be navigated. These problems are a distraction from the creative process, and enhanced object and process management support is likely to be beneficial.

The results of this initial study (Eaglestone et al. 1993) were an object model of the composition process, a prototype support system with abstract workspaces configurable to fit different ways of working, and a repository of artifact versions, processes and techniques. In particular, these aimed at better support for ad hoc experimentation, by retaining versions of objects and recording the processes applied. Also, the data model used reflected our conception of creativity as a divergent process, since it allowed high-level design objects (e.g., sounds, tools, etc) to be associated by composition events, in an ad hoc manner.

The second study (Clowes 2000) is a recent follow-up to the Tema project, in which attitudes of composers to electroacoustic music composition tools were surveyed using qualitative and quantitative research methods, including questionnaires, interviews and the mining of Internet discussion group archives. Here we summarise a part of that survey, that specifically relates to the generalised model of creativity in section 2.

The study explored the research question, "is composition of electroacoustic music made a more intuitive process through parameters that correlate with human perception, rather than phenomena suggested by spectral analysis, and input devices that are closer to the musical background of the composer?" A motivation was also to review findings of the Tema project in light of recent technological advances, and to broaden the scope of the study. Whereas the Tema provided an analysis through an in-depth study of a single case study, Clowes attempted to test the validity of its conclusions by surveying a wider population of composers. Study methods included in-depth interviews, questionnaires (returned by 30 composers), and mining of Internet-based discussion groups for electroacoustic composers to identify issues raised.

Of the composers surveyed, it was noted that they are often skilled software engineers and programmers for whom control over the whole composition process is more important than user-friendly interfaces. Clowes anticipated that they would be predominantly intuitive or methodological, and the latter would experience

greater frustration with current composition software if it did not enable them to apply their predetermined methods. However, in practice the subjects rejected this crude classification. The consensus was that composition is a mix of intuition and methodology (as in Tema). In fact we felt that the different modes of composition were better characterised by an alternative classification, proposed on an internet mailing list for electroacoustic music composers (cec.discuss), as "pure realisation" and "voyage of discovery". These respectively correspond, for example, to those traditional composers who transcribe the conception straight to the score and those who "doodle", for example, at the keyboard, seeking ideas and inspiration.

A further preconception was that the semantic gap between conceptualisation (in the mind) and realisation (in the software) would be a major impediment to creativity. However, this was found to have positive connotations and reinforced a theme that emerged throughout the study, that software limitations can be turned to advantage. This argument was nicely summarised by three of the subjects:

"...computer systems are not neutral...The holy grail of a universal representation is a fantasy and has nothing to do with creativity"

"a good representational system challenges one to broaden one's experiential knowledge and thereby creates correlations of its own"

"something that doesn't relate directly is more likely to produce results that will surprise me (the composer) - something that I wouldn't have thought of on my own. This is very important to me".

One subject expressed this desire for inspiration through "accident" in a more general sense:

"being an intuitive composer I have often found that my best results have happened while playing with software that I did not fully understand, adding a random element of mystery to the outcome".

This contrasts interestingly with the traditional concept of composer as the "virtuoso player" of an orchestra!

Subjects also saw software environments as open and extensible in a modular manner. Thus limitations of one tool or system are potentially overcome by other complementary software:

"Operational limitations are important because with the right collection of operationally limited stuff you can envisage a job that needs doing and do it. It is a modular way of operating"

"...in electroacoustic composition one is not limited to a single instrument - what I can't accomplish using NoteWorthy I can often do using CoolEdit..."

There also emerged a tension between the underlying principle that compositions are essentially conceived in the mind and realised in audio signals, and the influence on the composer of specific software. The former purist view was succinctly expressed by the following two quotes:

"...relating to a computer during the compositional ... phases of music creation is anathema and inevitably destroys my muse";

"The ability to execute whatever idea one conceives ...is the ability to get away from tools...this requires an understanding of the underlying mathematics and signal processing and computer science so that one can devise the algorithms and software needed..."

Whereas, the latter pragmatic acknowledgement of the real situation was expressed as:

"The overall modus operandi of music software and its idiosyncratic ways of executing tasks cannot fail to have an influence on the composer. It is important to have an awareness of such influences ...to maintain any sense of compositional freedom".

Other issues surveyed related to interfaces and visual representations of musical artifacts. There was evidence that composers wished to move beyond the constraints of conventional interface hardware - a number of subjects invented novel hardware. Similarly, there was evidence of dissatisfaction with representations, though subjects did find spectrum and waveform visualisations useful abstract representations of macro-structures.

The above two studies should be seen as representing a modest start to investigations in this area. However, they do provide experience in use of methodology, in addition to their tentative research results.

In summary, creativity in composition is largely under researched, particularly from the perspective of composition software. Analysis of composition according to Gestalt principles provides an approach that we believe warrants further investigation. In particular, future research should establish knowledge through study of composition in naturalistic settings. Data collection methods have evolved, but observation from multiple perspectives and using multiple media may well introduce analysis problems, and a need for the development of new techniques, for example for time synchronization of multiple data sources.

A variety of methods have been used in the research reviewed, which poses the questions, which have the greatest efficacy, in which situations should they be used, and how can they be best used in combinations to achieve triangulation. We explore these questions in greater depth in the following section.

### **2.3 Choice of methodological approach for the study**

The preceding review has highlighted that, although there exists research into the composition process, this largely fails to address the problems of how it is best supported by software. Hence, there is a need for further studies of composition, but from a software development perspective. This in turn begs the question, how should such studies best be conducted?

In general, lack of knowledge, which necessitates the conduct of research, may arguably be thought of as a curtain preventing us from viewing the reality beyond – or (if we do not accept such an objective notion of “reality”) the relativistic perceptual and conceptual constructions – that we seek to understand. Our knowledge may range between two extremes, which to some extent map broadly onto so-called “quantitative” and “qualitative” research approaches (we use these terms as shorthand for complex clusters of often-associated research parameters, acknowledging that to do so risks over-simplification). Once knowledge, which establishes the nature of the problem, is in place, further engineering approaches then become appropriate for invention of technological solutions. These tend to fall into the modelling and empirical research categories. In the current situation our concern is primarily to elicit

knowledge of composition, the invention phase being a future aspiration.

There are two further aspects to the problem of acquiring new knowledge: how does one understand the instance, and how does one gain generic knowledge? Quantitative approaches may be thought of as scattered pinpricks in the curtain, allowing clear and deep, but narrow and unconnected views through to the reality (or relativistic constructions) beyond. This limitation stems from the fact that much quantitative research is based only upon that which is controllable and measurable, and thus omits, for example, the complex human dimension. Thus, in the music domain, quantitative approaches may establish simplistic generalisations on composition through identification of statistically verifiable correspondences between the activities of composers, but will not deliver a deep understanding of how any individual composes.

Qualitative approaches may be characterised as more extensive areas where the curtain is thinned, allowing complex, inter-connected but hazy shapes to show through, inviting us to trace them onto the curtain, elaborating their detail to depict what we imagine them to represent. Thus, such techniques will deliver in-depth understanding of instances of music composition, in much of its human complexity, but without necessarily leading to reliable generalities.

Therefore, the results of qualitative and quantitative approaches are complementary and should often be used in tandem, recognising the merits of combining the advantages of different research approaches. For example, there may be advantages in combining a quantitative study of the work of a sample of composers, with a qualitative case study (as in (Clowes, 2001)). This will allow generalizations to be validated against the characteristics of instances, and vice versa. This methodological pluralism may take forms ranging from a belief – following Feyerabend (1975) – that any one paradigm is as good as any other (methodological relativism), to one in which differences between approaches may be used positively – for example, to map different paradigms onto different types of problem; to encourage critical constructive dialogue on common phenomena from the different perspectives afforded by different paradigms; or to blend different paradigms within a single study. However, at the same time we acknowledge that there are counter arguments to the above idea of triangulation. Burrell and Morgan (1979), for example, suggest that:

"Contrary to the widely held belief that synthesis and mediation between paradigms is what is required, we argue that the real need is for paradigmatic closure."

Olaisen (1991) considers the difference between problems that may be characterised in terms of "what we know that we don't know" as opposed to "what we don't know that we don't know". The former category represents what may be described as "taking the next logical step" in a research area, as opposed to approaches in which the bounds of the problem are surrounded by more uncertainty, and the results are less susceptible to precise anticipation. Such a focus arguably entails relatively convergent thinking in comparison with the type of thinking that is appropriate to the second type of problem.

Olaisen (1991) considers that the "what we don't know that we don't know" type of problem is characterised by high-complexity, an emphasis on social-intuitive as opposed to more logico-mathematical analysis, and "sensitising" as opposed to "definitive" concepts. Sensitising concepts (Olaisen, 1991: 254) are somewhat tentative and speculative concepts that: "... offer a general sense of what is relevant and will allow us to approach flexibility in a shifting, empirical world to 'feel out' and 'pick' one's way in an unknown terrain." Approaches geared towards this type of problem, and to developing sensitising concepts arguably entail relatively divergent thought or what we have referred to above as creativity. In an extreme form, this approach represents what de Bono (1987) has termed "lateral" thinking. It must, however, be acknowledged that creative thinking may also entail levels of convergent, as well as divergent thought (Ford, 1999), and that assuming any exclusive creative/divergent correspondence would be over-simplistic.

We argue that the state of our knowledge in the field of musical creativity – at least at the level of complexity that forms the focus of our research interest – is not yet well endowed with definitive concepts offering the certainty, control and anticipation appropriate to relatively quantitative research approaches. Rather, the field is characterised by the need to establish exploratory and relatively tentative sensitising concepts. Our belief is that the complementary use of the strengths of different approaches used in combination will be both desirable and necessary. Our argument here is that the integration of relatively quantitative approaches will be more appropriate at a later stage. An

analogy may be seen with software engineering. The Multiview methodology is a combination of soft systems methodology and structured techniques, based on the theory that the former relatively qualitative technique must precede the latter modelling approach in order to establish the nature of the problem.

We are aware that either a qualitative or quantitative approach if used in isolation risks its own characteristic limitations. Quantitative research without qualitative mediation may often produce highly reliable answers to highly meaningless questions. But without some element of quantitative testing, the subjective analysis of introspections, typical of qualitative research, may often supply highly meaningful questions with highly unreliable answers. Some balance and integration must be achieved between the two extremes.

In summary, this section has argued for research based upon both qualitative and quantitative approaches to establish a research base for composition software support for creativity. However, it has also identified a need for a preliminary phase of study which is predominantly qualitative, so as first to identify the sensitising concepts for this domain, and thus better establish the parameter of the research area and research "tools". Further, taking into account our conclusions to the review of related work, we identify an immediate need for qualitative studies of professional and expert composers "at work".

In the following section we illustrate and evaluate the application of the above strategy by describing and justifying the research methodology for our on-going research.

## 2.4 Methodology and Experimental Design

Our choice of research methodology to establish a research base for composition software support for creativity was problematic (Eaglestone et al, 2001). As Laske observes,

"the kind of musical knowledge that, if implemented, would improve computer music tools is often not public or even shared among experts, but personal, idiosyncratic knowledge...the elicitation of personal knowledge and of action knowledge still awaits a methodology..." (cited by Polfreman, 1999:31).

From our study of related work and research methodology, we concluded that both qualitative and quantitative approaches are appropriate. However, there is a need for a preliminary phase of study that is predominantly qualitative, so as first to establish the basic parameters of the research area and research “tools”. Further, we identify an immediate need for qualitative studies of professional and expert composers at work. We then envisage a second phase, in which both qualitative and quantitative methods will be used in tandem. Research-based invention of software devices to improve support for creativity are therefore a future aspiration, at which stage engineering methodology will also become appropriate.

A naturalistic and holistic approach was taken to the research on both theoretical and pragmatic grounds. At a theoretical level, arguably (see Ford, 1999 for a review) a mapping can be made between (a) a relatively holistic approach to perception and information processing and (b) creativity or divergent thought. At a practical level, this approach is particularly appropriate to the investigation of problems and phenomena which are not clearly understood and do not benefit from a large body of existing theory. Arguably what is required most urgently in this field is what Olaisen (1991: 254) has termed sensitising (as opposed to more definitive) concepts. These are somewhat tentative and speculative concepts that:

“... offer a general sense of what is relevant and will allow us to approach flexibility in a shifting, empirical world to 'feel out' and 'pick one's way in an unknown terrain.”

Research aimed at discovering sensitising concepts is particularly appropriate for discovering “what we don't know that we don't know” as opposed to “what we know that we don't know”, the latter arguably benefiting more from the relatively analytic and atomistic research more characteristic of the physical sciences.

The study sought to investigate the phenomenon of creativity as a complex holistic interaction of factors - including the “natural ecology” of the phenomenon as it takes place within a broader relatively naturalistic context. Multiple perspectives of the phenomenon under investigation were sought.

The research design of the first phase of our investigation evolves around in-depth case studies of only a few composers. Our approach follows the naturalistic paradigm as described by

Lincoln and Guba (1985), which stresses the existence of multiple constructed realities and the need to remain true to context.

A central aspect of this naturalistic paradigm is the triangulation of different data gathering methods. According to Erlandson et al. (1993), triangulation leads to credibility of the naturalistic inquiry, and hence increases the truth value of the study. The term credibility replaces the notion of “internal validity” in a more conventional inquiry.

An equally important feature of the naturalistic mode of inquiry is the absence of a clearly defined hypothesis before the data collection begins. Consequently there is no predefined goal how to analyse the data, but data collection and data analysis are an interactive process, and in an ideal situation, theory will emerge from the data alone.

The data generated was extremely rich, and in particular the multi-source computer data posed difficulties in terms of manageability and synchronization. In particular it was difficult to examine data with the intent of reconstructing the composition process in a procedural way. Instead, we chose to analyse data in a non-linear way by coding different sections of all data types produced and placing them into different categories. We then established relationships between the different categories that formed the basis for our attempt to derive models of the compositional process. Thus, we applied the spirit of a qualitative grounded theory approach, as expounded by Ellis (1993):

“The model derived should organize the features or the data in a coherent form that relates both to the perceptions and concepts of those studied and to the viewpoint that the researcher is developing. In that sense, although the concepts are derived from the data, they are not simply a restatement of the data. In developing the model with its attendant categories, properties, and relations, the researcher embodies the perceptions and activities of those studied in the model but in a way that allows them to be understood in other terms.”

Accordingly, rather than attempting a procedural or comprehensive model of creativity in timbre-based computer composition, the following analysis identifies sensitising concepts which more clearly establish parameters of the problem. Further, we concentrate on those emerging aspects that have not yet been looked at in previous studies in favour of more obvious and

previously discussed aspects of creativity, such as the role of serendipity. However, where appropriate we discuss findings of this study with reference to previous studies in the electroacoustic area.

## 2.5 Conclusions

In this chapter we have identified the need for a research base for enhancement of electroacoustic music composition software, such that creativity is better supported, and discussed how this research arguably should be conducted. Weaknesses of most related studies of conventional (pitch-based) composition have been the reliance on “after the event” introspections and observations of composition, or experimental data collected from unnatural settings through over-simplistic composition exercises, involving non-expert composers. Further, these have been from educational and musicological perspectives, and have not addressed software requirements. The procedural explanations derived in this work fail to provide a realistic basis for software support, since this should be enabling, rather than prescriptive. We therefore identified a need for future research into electroacoustic music composition involving expert composers in naturalistic settings, possibly taking a Gestalt perspective.

Further, we argue that given the preliminary stage of this research, a qualitative research approach should initially be taken, in order to identify sensitising concepts which will more clearly establish the parameters of the problem and methods that should be used. We then envisage a second phase, in which both qualitative and quantitative methods will be used in tandem. Research-based invention of software devices to improve support for creativity are therefore a future aspiration, at which stage

engineering methodology will also become appropriate.

Finally, the paper has described methodological issues relating to our on-going research aimed towards establishing a basis for composition software enhancement, and specifically, the research by which the software requirements presented in this report were derived. This study involves observation of professional composers at work. Though we have been able to derive a number of conclusions concerning the nature of software support for composition, and hence the software requirements, we still view our research as currently at a piloting stage, in which much attention is being given to the refinement of the methodology and research instruments. A number of principles are emerging from this pilot study. These include: observation from multiple perspectives is necessary, since a single view, e.g., computer interactions only, can give misleading clues; a more complete picture is given by observation through multiple complementary views, but care must be taken to use data collection methods adapted to the needs of the composer, since data collected in an unnatural or uncomfortable setting may be meaningless; the relationship between the researcher and subject is of great importance, since it will determine the openness of composers, both conscious and subconscious, to the observer; and finally, our initial work has very strongly confirmed to us that getting inside the mind of a composer is a very difficult research problem.

The experiments conducted using the above research methodology are described in the following chapters. Complementary research is also underway, primarily at DIKU, into perceptibility of music structures, and associated audio and music representational and control issues.

## CHAPTER 3 – EXPERIMENT DESIGN

### 3.1 Introduction

This chapter describes in turn the design of the first three experiments of our ongoing research conducted so far. Further, we critically review the design and research methods of each experiment and justify the changes made to the design during the course of our research. The data collected in the course of these experiments is catalogued in Appendix A.

These experiments constitute the first phase of our on-going research into electroacoustic music composition. In this work we are applying the strategy argued in the preceding chapter, i.e., we are taking a qualitative approach to analysis of data collected through observation of composers at work. In subsequent phases we will test the theories that emerge from analysis of the data collected in these experiments, through the construction and evaluation of prototype composition software.

This first phase of our investigation consists of in-depth case studies of only a few composers. Our approach follows the naturalistic paradigm as described by Guba and Lincoln (1985), which stresses the existence of multiple constructed realities and on the need to remain true to context.

As mentioned in chapter two, all three experiments evolve around ‘on-site’ observations of professional electroacoustic composers. In line with the naturalistic paradigm, our experiment design did not constitute a static, inflexible framework but has evolved over the period of our research and several adjustments have been made to optimise the data gathering and analysis procedure. In short, data analysis and data gathering are interactive processes.

“The principle of interaction between data collection and analysis is one of the major features that distinguishes naturalistic research from traditional research [...]. The human instrument responds to the first available data and immediately forms very tentative working hypotheses that cause adjustments in interview questions, observational strategies, and other data

collection procedures. New data obtained through refined procedures, test and reshape the tentative hypotheses that have been formed and further modify the data collection procedures.” (Erlandson 1993, p. 114)

### 3.2 First Experiment

Based on results of previous research and personal experience of the researchers involved, we tried to benefit from the notion that electroacoustic composers often regard externally imposed limitations as welcomed challenges around which to design their compositional strategies (e.g., in Eaglestone et al. 2001). Our brief for the composers was therefore to compose a piece of electroacoustic music in a single day, using their own familiar configurations of computer hardware, software and audio systems. We saw that as a fairly natural and acceptable imposition which would allow us to be present with the composers during the entire composition process. However, we use the term “fairly nature” advisedly, since the requirement to compose a piece over a single day in itself may have introduced time pressures, leading to different working practices.

With the rationale to make the process of verbalisation as unobtrusive and natural as possible, in the first data collection exercise, two composers working together on a single composition were observed.

#### 3.2.1 Verbal Protocol

Sloboda (1995) strongly favours think-aloud protocols - “to have a living composer speaking all his or her thoughts out loud to an observer or a tape recorder while they are engaged in composition” - as the most effective data gathering method to capture the nature of the compositional process.

However, Sloboda himself, as well as other researchers, point out that, the think-aloud protocol is not without dangers. In particular when utilising a concurrent protocol, verbalisation can affect the compositional

process in as much as it can negatively interfere with the creative flow.

With the rationale to make the process of verbalisation as unobtrusive and natural as possible, in the first data collection exercise, two composers working together on a single composition were observed. The subjects were selected because they are used to composing in collaboration with others, and also they have complementary approaches, one being methodically oriented and the other intuitive, or to put it in different terms, one being an academic composer the other one a non-academic composer.

This idea does not only reflect the notions of dialectic synthesis and maximum variety sampling pertinent to the naturalistic paradigm, but it also echoes our ideological position, which is to bridge the gap between academic and non-academic sectors of timbre based music. For a comprehensive analysis of the cultural gap between these two factions the reader is referred to Cascone (2000).

### 3.2.2 Video camera recording

In addition to the use of verbal protocols, discussed above, we also videotaped the composition scenario. The recording of the physical interaction between the two composers as well as between composers and computers (i.e. physical computer input devices) addresses Ericsson's and Simon's findings, that when subjects are involved in physical manipulation they "appear to lack a mediating symbolic representation that can be readily encoded into the verbal code" (cited by Collins, 2001:102). Hence, the videotaping should be regarded as a complementary method to the verbal interaction protocol.

### 3.2.3 Computer data

The composition tools themselves produce the most obvious source of data in computer music composition. In prior process-based research of conventional music composition, computer data have mainly been used in the form of MIDI save-as files, audio files and screenshots.

We believe that this snapshot approach is too fragmentary and also puts an unnecessary burden on the composer's mind, because she has to constantly consider what is a noteworthy development within the composition process. We have therefore introduced a permanent capturing

of the composition process by simply connecting a video recorder to the video output port of the computer. This provides a remarkably rich record, not only of the evocation of software processes, but also of some of the nuances apparent in the intervening manipulations of input devices, e.g., hesitation and hovering over icons, and their relative timing.

### 3.2.4 Interviews

At the end of the one-day composition scenario we conducted an unstructured, reflective interview where topics and incidents that have arisen during the task were reflected on.

Further, we resorted to semi-structured interviews, which took place on a separate meeting before and after the one-day observation. These interviews provided background information about the composers and covered more general composition related areas, such as the relationship between composer and computer, and attitudes to computer music composition. In particular we tried to shed light on those lengthy periods of time that are difficult to monitor, where composers gain inspiration for compositions.

A distinctive feature of our interview technique is the encouragement of respondents to enrich their answers with metaphors. We believe that metaphors are a very powerful tool to mediate meaning between respondent and researcher or between a person's interior world and the exterior reality. We would like to join Eisner when he discusses metaphor:

"What is ironic is that in the professional socialization of educational researchers, the use of metaphor is regarded as a sign of imprecision; yet, for making public the ineffable, nothing is more precise than the artistic use of language. Metaphoric precision is the central vehicle for revealing the qualitative aspects of life." (cited by Janesick, 2000:380)

### 3.2.5 Experimental Setting and Constraints

The first experiment took place in a large studio space at Middlesex University, because both composers were familiar with and fond of that space and it also provided all the facilities needed for our experimental set-up, in particular two computers with a video output port.



As mentioned before, the compositional boundaries for the composers were set by time restriction of one day and we also asked the composers to use a collection of roughly 80 sound files provided by the researcher. This second imposition was introduced with the rationale of having a constant parameter that would make the triangulation of different similar experiments more effective and also to compensate for the short period of working time. In order to stay true to our naturalistic objective, the composers were involved in defining the situation within which they worked. In particular, we adopted their request to work on separate computers for the most part of the composition process.

However, we note that our naturalistic objective is an aspiration which is relative to the subjects, since any constraint imposed may conflict with the composers' chosen working situation. Thus, in this experiment, the naturalistic environment is compromised by both the time constrain and the requirements to work with pre-specified sounds.

### 3.2.6 Reflections on Experiment One

The working process between the two composers went extremely smoothly and both assured the researcher that they felt the situation was "pretty natural". They didn't feel they were taking part in a scientific experiment, but rather experiencing "an intense, but enjoyable situation".

They stated that even though there was not much verbal exchange, they had a fair amount of exchange by listening to each other's sonic output and also via the Ethernet file exchanges. However, the lack of verbal communication between the composers did pose an obvious problem to the researchers because an important data source was almost lost.

Moreover, the initial analysis of the data revealed that the lack of verbal data on the camera tape proved to be even more drastic than initially anticipated, because the picture quality of one of the computer output tapes was poor. This was due to low resolution quality of the computer's video output and made a reconstruction of the composition process of one of the composers extremely difficult. This experience had technical implications for the subsequent experiments, i.e. we had to ensure that computers used provide high quality video output.

It became apparent during the composition task as well as in the reflective interview, that the time restriction of one day, was initially an exciting challenge for the composers, but we all realized that this time period was simply not enough to bring the compositional process to a satisfying, and more importantly, naturalistic result.

The composers' request to work on separate computers during the first, and longest stage of the composition process, their obvious reluctance and difficulties to increase the verbal exchange during this first stage, and the shortage of time, strongly implied a rethinking of the methodological approach. In the reflective interview, one of the composers made clear that he would feel a lot more comfortable to talk concurrently about what he is doing while working on his own. The next observations therefore involved only one composer working within a protracted period of time.

## 3.3 Second Experiment

The second experiment stretched over five three-hour sessions during which a composer worked in his familiar environment at home. Because of difficulties in recruiting professional composers, i.e. freelance composers or professional academics, over such a long period of time we decided to recruit a research student studying for a Ph.D. in composition of electroacoustic music, who has also had over twelve years experience in composing electronic music, eight years experience of composing timbre focused pieces and for the past five years has used the computer as his primary composition tool.

Unfortunately it was not possible to observe the composer using his own set up, because his computer did not have a video output from which to capture the real time computer data. Instead he used a computer provided by the researcher and onto which we had imported all the audio software and sound files the composer said he would use during the sessions. As a further resource for sound material he also had his minidisc recorder from which he could import sounds of previous field recordings that had not yet been digitised.

In retrospect the imposition of an unfamiliar computer, so as to facilitate the recording of a video record, was problematic. This was because there were instances during the experiment in which the composer expressed the need for a physical mixing desk which is part of his usual

set-up but which could not be used with the computer provided by us.

In accordance with the composer's request, we dropped the idea of providing a collection of sounds for this experiment. Also, we felt that the initial idea of having a constant parameter would not yield any useful results, but instead imposed another unnecessary and unnatural limitation on the composer.

In order to avoid the poor verbalisation results of the first experiment, we emphasised to the composer the unconditional importance of this process. In order to maximize the verbalisation, the observing researcher constantly encouraged the composer to talk about and reflect on what he was doing. This form of intense and inquisitive observation was viewed by the composer in a surprisingly positive light and he assured us that instead of it being an intrusion he actually enjoyed this process of reflective and analytic working.

Because of the multitude of short interviews during the observations in the second experiment we felt there was no need to conduct separate interviews with the composer, i.e. all the relevant issues had been discussed during the observation process and were recorded on video camera.

It became apparent during the observation of the composer that it can be difficult to observe a single piece of composition from beginning to end because composers will often work on different pieces simultaneously and the composition of those pieces is interlinked. For example a sound created at a certain point in time could potentially be used for different pieces the composer is working on. It is therefore difficult to define what constitutes the composition of a single piece in the first place.

The intent to capture a particular composition in a procedural fashion was therefore sacrificed for the notion of observing the composer as he would normally work.

This new approach was viewed by the composer in a very positive light and the verbalisation during the composition process was far more prolific.

A preliminary analysis of the rich data gathered in this second experiment confirmed our assumption that there is no need to capture the composition process of a single piece from beginning to end in order to satisfy our objective of investigating how electroacoustic composers interact with their computer environment. Moreover we were overwhelmed by the richness of data that had been generated by our interrogative observation technique, which is

based on picking up on issues that arise during the composition process and actually interrupt the process for the benefit of an ad hoc discussion of the issues in question.

### 3.4 Third Experiment

Our experiences from the second experiment - in particular our decision that it is not necessary for our study to capture the composition process of a single piece from beginning to end, as well as the success of our interviewesque observation style - formed the basis of our third and arguably most radical design. In this third experiment, we took a snapshot approach to the observation of a well renowned and prize-winning electroacoustic composer at work. In fairness we would like to point out that to some extent this time condensed approach also reflected difficulties of recruiting professional composers over a protracted period of time.

The experiment consisted of a single three hour composition session during which the composer was asked to simply continue his composition of a piece in progress.

Unfortunately the computer in the composer's studio again, lacked a suitable video output. Due to the negative implications of the unnaturalistic imposition in the second experiment - where the composer had to work on a computer with video output provided by the researcher and therefore could not use his personal set-up, i.e. his mixing desk - we decided to do without the capturing of computer data in favour of the composer using his familiar set up, i.e. the computer in this set-up did not have a video output. This decision also reflected difficulties in analysing the twofold computer data (camera and computer data) of our previous experiments.

At the beginning of the session we asked the composer to briefly contextualise the piece he was working on. In the remainder of the session we simply followed the compositional progress and asked questions in an intuitive and spontaneous manner whenever an interesting issue arose out of the context of the composition process. As described in experiment two, where appropriate we would then try to go into more depth of some of the issues, even if that meant that the compositional process was interrupted.

We felt that the snapshot approach in this experiment was particularly feasible because we had previously identified the sensitising concepts for our research and had a good idea of what we were looking for. In this experiment we could therefore focus on issues that had become

relevant in our previous experiments. To put it in different terms, we did not try to expand our model of the compositional process but to look in more depth into areas that had previously been identified as relevant for our study.

The positive feedback of the composer, the ease of execution and the richness and relevance of data collected in this third experiment suggested to us that this approach would be ideal to increase the number of composers captured by our study and to verify our results from the case studies with a larger number of respondents. In short, this third experiment suggested a transition from a purely qualitative study into a more quantitative one, as has been suggested in the section on methodology of this report.

### **3.5 Conclusions**

In this third chapter we have described the original design and objectives of our study and its development over the period of our ongoing research.

In the last experiment design we have applied a research approach that offers the chance to extend our study and to observe a larger number of composer. This snapshot approach therefore enables us to look into more depth of the sensitising concepts identified in the large case study and to validate our qualitative data by

triangulation with results from a more quantitative approach

We would like to point out that we are confident about our new form of compositional observation based on the interruptive interrogation of a composer at work. We believe that this technique is a well-balanced compromise between our naturalistic obligations and our need to generate concurrent verbal data during the composition process. We have evidence that this form of observational interviewing delivers far richer and more relevant data than a 'dry', questionnaire-based interview. In particular, we have found that it is much easier for the composer to mediate his thoughts by talking about a concrete instance of relevant issues - which constantly arise while they are composing- than to think about his/her compositional process in an aloof interview situation.

However, our caveat is that the area of study is characterised by the idiosyncrasies and individualism of the subjects that we wish to study and also the secretive and internal nature of that which we wish to analyse, i.e., the creative process. Thus, as our research progresses we will continue to apply research method in a flexible way, continuously reflecting on the appropriateness of our methods for any given situation.

# CHAPTER 4 – A QUALITATIVE ANALYSIS OF COMPOSERS AT WORK

## 4.1 Introduction

This chapter presents a preliminary analysis of data collected through experiments described in the preceding chapter. The analysis is far from exhaustive, due to the large volume of rich data collected and constraints on our time and resources. However, it does provide some insights into an elusive aspect of computer music: how composers interact with computer-based composition systems when they are being creative. The value of this analysis stems from the rarity and authenticity of the data collected through our experiments in which we have observed composers at work in natural settings (Eaglestone et al. 2001).

The analysis presented here provide the basis for the penultimate chapter, in which these results are related to requirements for enhanced composition systems which might better support creativity. These constitute the grounded theories which we will validate in subsequent phases of our research, through the prototyping and evaluation of composition software systems.

## 4.2 Preliminary Results and Discussion

Data collected through the experiments described in the preceding chapter was analysed using naturalistic inquiry techniques entailing an inductive approach. Rather than seeking to impose pre-determined analytical categories on the data, the researcher categorised the data inductively. Conceptual categories were thus established, along with their properties and dimensions, then combined where appropriate to form higher order concepts entailing relationships.

Only after conceptual categories had emerged in this way from the data were they compared with concepts and models existing in the literature. In the sections below, emergent concepts are first

described, then compared where appropriate with existing literature.

Issues relating to a number of broad themes emerged from the analysis, which form the headings of the sections below.

### 4.2.1 A taxonomy of Electroacoustic music composers

In order to analyse creativity in the context of electroacoustic music it is essential to understand it not as simply another strand of classical music. Electroacoustic music is not merely concerned with its ‘musical’ outcome, but often equally important is the development of new tools or at least new ways of using the tools available. Hence creativity in electroacoustic music can not be determined by purely taking into consideration work with the sound material, but also must take into consideration other artefacts, such as home-made software, that have been created to produce those sound pieces. It seems that electroacoustic music is not so much judged, within its community, by what it sounds like, but by what made it sound. In some ways electroacoustic composition can therefore be better understood in terms of research rather than artistic design.

However, it would be too simplistic to generalise the above notion; since composers are very different in their approaches.

With regards to the dichotomy between creating music for the music’s sake and creating music as a showcase for new tools and techniques we have identified three classes into which composers can be loosely grouped.

- 1 Our first group comprises composers for whom creation of computer related tools (i.e. software, hardware interfaces) is a natural accompaniment to composition and is inseparable from the process of composing. Composers in this category also consider it normal to adapt their way of thinking to the way the computer is structured in order to mediate their ideas to the computer, even if

this means that “there is a constant battle going on.”

One composer succinctly characterised the problem which this group faces, as follows:

“...when the computer becomes an interest in itself, in terms of certain programming aspects, ... the computer takes over as an interest in its own right, over and above sounds...”

- 2 The second group are instinctively more concerned with engaging with the sounds themselves, i.e. composing with the tools made available to them, but seem to feel the need to deal with aspects of tool creation, i.e. computer programming etc., because of “peer pressure”. The importance of the structure of the social interactions surrounding computer use has already been highlighted by Ungvary when he discusses parameters of human computer interaction (Ungvary & Kieslinger, 1998).

There appears to be a strong notion within the (academic) composing community that the quality of a musical outcome is directly related to the complexity and idiosyncrasy of the processes involved in creating those sound pieces. The question remains whether originality of processing necessarily results in a high quality and originality of the sound material produced.

- 3 The third group of composers are similar to the second one, with regards to their natural preoccupation with sounds rather than with the tools. However, unlike the second group they do not worry about technical sophistication involved in the production of their audio pieces and hence are able to better focus on the sounds themselves.

“I don’t care about it [self-written software tools] ..., all that stuff is bullshit, ..., it’s just tools and you’ve just got to use whatever you feel comfortable with.”

This does not mean that they are not interested or do not critically engage with the tools they use, but will rather try to push the tools they know to new boundaries to create original sound events.

“I think probably more than anything I have tried to find interesting, say, audio events from mal-appropriation of existing programs and so on, and I tend to more take

an existing piece of software, just an average, ordinary piece of audio software, and try to enhance any idiosyncrasies.”

## 4.2.2 Creativity through diversity

We have found that the use of multiple audio applications during the compositional process is not only a phenomenon that composers have learned to live with, but also has an important positive impact on their compositional process and appears to support their creative behaviour.

On occasions, for example, when a particular sound was needed, the composer would quit the arrangement program she was working in and open up a more specialised application for the creation and transformation of sounds. This quite logical and problem-focused action would then often lead her astray from the original intentions, because the new sounds inspired her to wander off in a completely different direction and not return to the original arrangement for quite some time.

The switching of applications could be viewed as a hindrance for the composition process, as indeed it will often prevent the composer from focusing on the original problem (creation of new sound for a particular section in the arrangement). This view would certainly hold true if we assumed that electroacoustic composition was subject to any demands on (cost/time) efficiency. However, those criteria will enter a composer’s mind only in rare circumstances whereas usually their top premise will be to create interesting, new sounds – at whatever ‘cost’. Viewed from that perspective, the switching of applications is a stimulus for creativity because it frees composers from getting stuck on a particular problem. Instead the diversion catalyses the expansion of ideas and possibilities.

It has been suggested in a previous study (Clowes, 2001) that composers do not feel the need for a single audio application – the “holy grail” of audio software- that would facilitate them with all the processing and arranging power they could wish for and make the existence of multiple applications redundant. This study has found further support for this.

## 4.2.3 The impact of visual senses

Our observations have shown that paradoxically the visual senses do play a major role in what is

often referred to as acousmatic composition. On a very basic, perceptual level this even applies to the mere look of the hardware as well as software interfaces. One composer even tried to explain his adverse attitude towards command line based programs for aesthetic reasons!

“I think very visually when I think about sounds. Maybe that’s why I don’t like text based programs, because they look so awful.”

The same composer positively commented on the unusual look and feel of Metasynt (http://www.metasynt.com). In this software environment the normal desktop environment is completely hidden by a black canvas and the software application takes over completely – leaving the composer with only the waveform representation of the sound sample and various sound editing tools. The composer observed that this masking “helps me to concentrate on the task I am doing.”

At a procedural level we observed that even when it comes to editing sounds (cutting and pasting) or when placing sounds into the time-aligned arrangement environments composers are often led more by visual cues than by auditory ones.

“[...] sometimes when I edit, you know, obviously I can recognize where certain audio events are visually, and I cut according to that. I don’t even listen.”

Despite the fact that the visual representation of sound is generally regarded as an advantage by composers we got a strong impression by looking at the computer data that the visual score representation has the potential to negatively influence the compositional process. This is because it can give misleading information about sonic material and can also distract from the listening process. For instance, if there are many score events in a given section the composer might be misled by the event density to feel that the section is “busy” enough or “too busy”, even if the sonic material is actually quite “thin”, and vice versa.

To some extent the negative, or at least distracting, impact of visual representation of sound (events) is supported by the fact that composers would frequently request to listen to composition in a totally acousmatic situation, i.e. from minidisk over a hifi system.

The impact of vision on auditory perception is a known phenomenon in experimental psychology. For example, in the McGurk effect (McGurk and

MacDonald, 1976), the movement of a speakers face and lips has a large influence on the perception of speech. Similarly, visual stimuli influence the auditory localisation of sounds in space (Wallach, 1940). Clearly, what we hear is influenced by what we see - and composers may elect to work from a purely visual representation of sound, or to disregard visual cues in order to achieve a "pure" listening experience.

#### 4.2.4 The impact of the semantic gap

The semantic gap between conceptualisation (in the mind) and realisation (in the software) could, it was found, have positive connotations.

The availability of easy-to-use, heavily destructive (real-time) processing tools as well as the easiness to assemble a huge number of sounds over a short period of time seems to create a situation where composers are not aware anymore of the processes involved in their sound manipulation. This can lead to over-processing and over clustering of sounds and consequently results in sound pieces which do not refer to any common, shared experience but the experience within the composing community. Thus the accessibility of electroacoustic music to a non-expert audience is effectively denied. Maybe there should be some form of feedback from the computer regarding the amount of processing that is involved in certain operations so the composer has an objective basis on which to assess how much she is actually doing. The problem of over-processing is widely acknowledged by members of the composing community. One composer commented that he “likes the gap in command line based application, because in very user-friendly applications like ProTools he feels he often does too much.”

The research contradicted the notion that the semantic gap between conceptualisation and realisation may be a major impediment to creativity. The positive connotations found in the present study reinforced a theme prominent in previous studies (Clowes, 2001) that software limitations can be turned to advantage.

#### 4.2.5 Limitations to the role of the computer in supporting creativity

There was evidence in all observations that a lot of the creative process is happening away from the computer, e.g. between computer based composition sessions and during field

recordings. Also, a very short interruption from working on the computer can act as a huge inspiration for the compositional process, similar to the catalytic effect of switching between computer processes previously discussed. A good illustration of this, captured on video, occurred when one composer got out of his computer chair to pick up a metal tube nearby, recorded the sound of the hit tube into the computer and then continued to work on the computer. Even though the time spent away from the computer was less than 5 minutes, it became evident from the procedural protocol of the observation that in the following 15 minutes the composer went on to create the most “significant” (in his own judgment) sound structure of the whole 7 hour composition day.

In more substantial breaks composers reflected on their compositional process and made plans for the proceeding sessions. Some composers made lists on paper about tasks they intended to perform at the computer. Generally the composers followed those lists not very closely and a more immediate feed-in of compositional strategies and tasks from the physical paper note into the digital domain might be beneficial.

One composer printed out lists of all the sound files he had used and would possibly use in a particular piece and said that he “regularly spends a day just listening to his pool of sounds” in order to make notes about what is contained within the sound files and also to highlight relationships between different sounds. He would do this under the premise of “which sounds might go well together” and delete files that he thought were useless. The fact that the process of writing down qualities of and relationships between different sounds happened to a large degree on paper highlights the insufficiency of composition environments to allow for the possibility of expressing relationships between data and tools in a free associational manner.

#### **4.2.6 The impact of inappropriate interfaces**

The inappropriateness of existing GUI-based software tools to express free associations between data and tools was highlighted by the preference of one composer to group sounds logically in a custom made command line based application rather than having to engage with sounds on a visual interface where they would appear away from each other.

“...with an object orientated approach you can structure sounds hierarchically into groups. Or at least I could perceive a means of doing this, I have started rudimentary experiments with this, but even though that you have got these separate objects perhaps in Logic which belong to the same group of sounds in the way that you perceive the whole sound texture, they may actually be implemented on separate tracks – even a number of visual spaces away from each other – whereas they are actually acting [sonically] in combination, whereas in SuperCollider you could physically group them – or not physically – but you could group them logically together in a group so that they stayed as one unit.”

A further illustration of the mismatch of cognitive styles and available interfaces was provided by another composer. He made a metaphorical comparison between composing and building a house - “First you have to lay down the foundation, then build the walls and roof, then decorate the walls and put furniture in the rooms until you come down to the very fine details.” This begs the question, if a composer perceives his work in such a way, is it then appropriate to reflect the same metaphor in the software environment, rather than having a standard track based arrangement environment? In conversation with the composer, he indicated that he has already considered one possible way of partly achieving this. He suggested putting the first two tracks of the arrangement environment, which usually contain the sonic equivalent of the house’s foundation, at the bottom of the arrangement window. The compositional representation within the computer domain would then better fit with his mental image of the composition.

Further, the above accounts about expressing relationships between individual sound files and the desire to group sounds logically together as well as the holistic, image based analogy between composing and building a house, support our suggestion that parts of the compositional process can be explained by analogy with Gestalt concepts. Particularly relevant in this context is Gestalt theory’s concern with the importance of relationships between individual elements of a system as well as the principle of ‘grouping by similarity’.

Generally, approximate real-time manipulation with an intuitive graphical user interface was preferred over more accurate non real-time

manipulation. However, on occasions, accurate non real-time applications were also valued highly. The real-time use of hyperdraw whilst listening back to the developing composition, seemed particularly popular with the composers. In hyperdraw mode, which is available in most professional sequencing applications, one can place different envelopes, e.g. amplitude envelope, over an entire audio or auxiliary track. This development of the whole as a kind of envelope placed over the individual elements is again very much in line with Gestalt concepts (Reybrouck, 1997).

In the previous Tema and Clowes studies, there was no evidence of a perceived need for spatio/visual processing, though it was concluded that this may be a cause of the composers' frustration with conventional software interfaces. Certainly, a need for more direct, tactile means of seeking and manipulating sounds in composition and performance is apparent in much research (e.g., sentograph (Vertegaal & Ungvary, 1995), composer's cockpit (Vertegaal et al, 1996), etc). However, in contrast to our previous studies (Eaglestone 1994, Clowes 2001) strong evidence emerged that spatio/visual processing could serve as an aid to creativity<sup>2</sup>.

One composer expressed the need to enhance the translation of body movements he would perform while engaging with the computer into more representational computer data.

"An interesting thing I noticed as well, I developed new body movements as a consequence of using it (the computer), because, ... I am constantly using the mouse and I make cuts, and this kind of arm movement to the point where I damaged my shoulder and back, yeah (laughs), so I move my arm from left to right quite rapidly. It would be a case of maybe enhancing this."

A need for more direct, tactile means of seeking and manipulating sounds in composition and performance was expressed by the desire for malleable interfaces that would allow for a sculptural shaping of sounds. There seems to be

---

<sup>2</sup> Note also, that the Pattern Play System project (Murphy 2002), located primarily at DIKU, is also exploring efficacy of visuo-spatio representations, using a different approach (see Appendix B). We anticipate that this research should also contribute to our analysis of this aspect of composition systems.

a general desire to physically touch the sounds which implies the need for force feedback interfaces.

The need for physically engaging with the tools composers are working with feeds our overall impression that all composers had to cope with the distance between physical and virtual domains. This contrasts with areas of computer science in which this is considered an asset. For example, a principal of databases theory is data independence, whereby users are shielded from physical implementations.

#### **4.2.7 The need for accessible individual and community knowledge and dialogue within and beyond the community**

In the overall context of this report we feel it is necessary to point out that there was absolutely no indication that composers need more and new signal processing techniques. Though this contradicts what must be the underlying motivation for the bulk of research in the area of music informatics (see proceeding of the International Conferences on Computer Music). However, there are several pointers that indicate high demand for increased knowledge exchange and a "know-how" database. It was evident that the composers would have profited from collaboration with a wide community. On several occasions composers encountered a clearly defined problem which could have been helped with by a query of a knowledge base or to others who may have addressed the same problem.

Further there was a strong notion that composers value the judgement of a second person, expert or not, highly and it seems desirable to create a (internet based) platform where immediate feedback can be obtained for the work in progress.

### **4.3 Validation of our model of creativity**

To what extent does our data from the experiments described in chapter 3, and also the two previous studies the researchers have conducted (Tema and Clowes' survey) (see 2.2) support the model of creativity introduced in section 1.2, and elaborated in (Ford, 1999).

A key premise of our model is that creativity is primarily a divergent process. The studies



provide ample evidence of this. In each, the unexpected was deliberately sought and valued. Thus, there is an implicit requirement to allow composers to make free associations between musical objects that emerge from the composition process. This desire to create from a wide and varied range of musical ideas was apparent in the often evident desire to create musical artefacts through use of a diverse range of tools, often home-made.

Within this modular context, the constraints of specific systems were also useful, since they forced ingenuity. Whereas the literature on creativity describes the “flash of inspiration”, the first study illustrated that identification of the “correct artefact” can be through protracted consideration.

All studies provide evidence that divergent behaviour was complemented by convergent methodological behaviour, illustrating De Bono’s perception and processing phases. For example, Ungvary’s work involved methodical, phase-oriented refinement, and the composers surveyed gave a strong message that composition is a combination of the methodical and intuitive. Serendipity played a major role both in the Tema composition process, and for composers surveyed and observed. Errors and unfamiliar tools that produce the unexpected were found to be artistically useful.

The analysis of computational data in our observation of composers suggests that the reliance on serendipity decreases with the continuation of a particular piece and is replaced with more intentional and purposeful actions towards its completion. Finally, there seems to be a reciprocal relationship between divergence and serendipity as indeed random incidents will catalyse divergent progress.

Know-how, both individual and community, is a resource of the creative artist, but does not feature in the model. The desire to gain and develop sophisticated know-how was evident in all studies. For example, Ungvary demonstrated an interest in the processes (transformations) by which the musical artefacts were created and the ability to reuse effective processes. A lack of object and process management proved to be a hindrance to creative work, but could also be a catalyst for serendipity. Clearly, this management must be supportive, rather than prescriptive, since the notion of constraining composers to work within the rules of “good practice” is inappropriate! Also, this would be contrary to the frequently expressed desire of the composers who participated in the experiments,

to break out of any habitual strategies which would otherwise inhibit their exploration of new musical possibilities.

There is evidence in the analysis that creativity in music has a visual/spatial (as a complement to textual/logical) dimension. For example, composition ideas were expressed through body movement (4.2.6), and at times there was evidence of “sculpting and landscaping” of audio material by manipulating visual representations only (4.2.3.). This validates part of our model of creativity (1.2) which is elaborated by Ford as, “the initial process of similarity recognition, particularly at high levels of complexity, often entails the sort of holistic, parallel information processing characteristic of perceptual pattern matching, as opposed to relatively sequential logic-based processing.” (Ford 1999:532)

## 4.4 Summary

In this chapter, we have analysed data collected through the experiments described in chapter 3. The analysis has broadly followed the inductive research approach of grounded theory.

The main notions to emerge were that electroacoustic music composers are not a homogeneous group, but range from those who focus on the development of software tools as an integral aspect of composition, to those who see existing software tools as a means to an end, i.e. music. Further, the discipline is very much one of exploration, in which diversity, the unexpected and the idiosyncratic are valued. Thus, composers will choose to work with systems that challenge their preconceptions.

The visual representations of audio information are important, but have both positive and negative impacts on the composition process. For instance, current user interfaces can inhibit composition activities, because of the limitations of their representations. For example, interfaces may be limited by constrained inherent in the underlying paradigms and assumptions of composition systems, and thus may not be able to combine audio artefacts into visual groupings which make sense artistically. Also, the process and representation models of the interfaces may conflict with the composers’ approach to composition. However, diversity of representations of audio at all levels is often viewed positively, since it can effectively summarise both macro and micro detail of audio structures, sometimes to the extent where listening becomes superfluous.

Control over the whole composition process is more important to some composers than easy-to-use interfaces. Thus, they value access to audio object at all levels, ranging from raw DSP, to representations based upon perceptual and aesthetic properties of musical sound. In this respect, the requirements of these composers contradicts the notion, widely held in computer science, that interfaces should model users perceptions, and shield them from the complexities of computer implementations. However, there is also awareness that dealing with the minutiae of waveforms can become an obsessive and time-consuming activity that will detract from musical endeavour.

Much creative work currently occurs away from the computer. This may represent an inherent limitation to the role of computers in creative activities, and / or current limitation of current composition software.

Better explicit accumulation, representation and access to personal and community know-how, and active dialogue within and beyond the community, would be of value.

Finally, the chapter considers the extent to which the above notions that have emerged through inductive analysis support the model of creativity introduced in section 1.2. The conclusion drawn is that the model is well supported, since composition emerges as a process which relies considerably on divergent thought processes, and hence serendipity and randomness are valued. Convergent thought processes are also involved, but these are secondary to the divergent processes, since their role is to elaborate and refine those creative notions that are the result of inspirational (divergent) thought.

The above notions are further discussed in the following chapter with respect to the requirements of composition systems that might better address them.

# CHAPTER 5 – COMPOSITION SYSTEM REQUIREMENTS

## 5.1 Introduction

In this chapter we consider implications of the results of our data analysis presented in the preceding chapter, with respect to requirements of computer systems that could better support creativity in electroacoustic music composition.

These requirements are inevitably more controversial than the analysis. Whereas, the latter is grounded in the data collected and its trustworthiness stems from transparency of the research methodology applied, there is no deterministic process by which software requirements to address identified problems are derived. Alternative proposals will always exist. This has been born out whenever the substance of this report has been disseminated (e.g. in Eaglestone et al (2001a,b)). In each case, our analysis of tensions between composers and current composition systems has been well received by both computer science and music specialist audiences, but no apparent consensus has emerged on how those tensions can be resolved. Thus, given the preliminary nature of our research, and the problem characteristics (see chapters 1 and 2), these requirements should be viewed as speculative, and thus constitute theories to be validated in subsequent research.

The requirements are derived by interpreting results of the preceding analysis from a number of perspectives - systems architecture, object and process management, knowledge and data management, and composer-computer interfaces. In each case we present the motivation for the requirements that we specify, and discuss possible software design implications. However, our aim is to present requirements such that they address the main conclusions of the preceding analysis without pre-empting any future software design decisions

The chapter is structured as follows. Section 5.2 states the main notions that emerged from the analysis in the preceding chapter, as the basis for deriving software requirements. We then analyse the software requirements that address these

notions in section 5.3. This analysis is structured, so as to consider software requirements from different perspectives.

## 5.2 Summary of the main notions from our analysis

This section presents the key notions that have emerged from the analysis of electroacoustic music composition, presented in chapter 4.

- 1 **The unexpected is sought and valued by composers.** The importance of serendipity and random encounters as a source of inspiration has been a recurring theme (see 4.3), and clearly has implications in terms of the music generation facilities of composition systems.
- 2 **Composers value access to a diverse range of tools.** This second notion (see 4.2.2.) is not surprising, given the intrinsically experimental nature of electroacoustic music composition, but it does mediate against seeking any definitive composition systems standard. Our findings are that divergence is a key issue in composition, both in terms of creative thought (see notion 1, above) and the software tools that stimulate and facilitate it. Both this and the previous studies (Clowes 2001), have found that diversity of audio applications has a positive impact on the compositional process, and appears to support creative behaviour. Further, the limitations of specific tools can stimulate creative ingenuity, but also the need to go beyond those constraints is often achieved by using tools in combination in a modular fashion.
- 3 **Composers considered unfamiliar tools to be good, and their idiosyncrasies were particularly valued.** This conclusion (see 4.2.2.) is consistent with the previous two, since unfamiliarity and idiosyncrasies are valued for the new and unexpected

possibilities presented to the composer, and are achieved by ongoing experimentation with new and diverse tools and features. In particular, new (and hence unfamiliar) tools provide opportunities for experimentation; attempts to “break” such tools can often lead to creative advances.

- 4 **Composers interact with software at different levels; both as users and programmers.** Our classification of composers (see 4.2.1) is into those who seek innovation in the use of computer-generated sounds within music (group 3); and those who also seek innovation in sound generation tools (groups 1 and 2). This split is similar to that in computer science, between those who research advanced applications within existing theoretical models and paradigms, and those who seek new ones. Software environment requirements for using existing tools and for generating new tools differ. Accordingly, we should not be seeking a generic software environment for all persuasions of composers. This complements the above requirement for diversity of composition tools that we have already noted (see notion 2, above).
- 5 **Production and availability of new tools is key.** This follows from notions 1-4 and is facilitated by notion 6.
- 6 **The accumulation of personal know-how, sharing of community know-how and interaction within and beyond the community is advantageous to the composer** (see 4.2.7.).
- 7 **Direct physical control of audio materials is desired by composers.** In general interfaces emerged as an issue, and there were instances of clear frustration with existing interfaces. One clear requirement that emerged is for interfaces that put composers in touch with the physical nature of sounds (see 4.2.6.). This requires human computer interfaces beyond those conventionally used in computer systems.
- 8 **Visuo-spatial pattern recognition is valued by composers.** Examples of this are support for the expression of compositional ideas, through body movement recognition, visual pattern matching and the use of visual metaphors, (see 4.2.3 and 4,2,6.).

- 9 **Flexibility to deviate from the task in hand, and develop emerging ideas as they occur is valued by composers.** Unplanned working patterns became apparent in the analysis of the experiments (see 4.2.2. and 4.2.5.). Composers would frequently switch between tasks and compositions. Also, time spent away from the computer appeared to have considerable impact on the development of the compositions.
- 10 **Better support for holistic approaches to composition is valued by composers.** For example, the desire of composers to express freely relationships between sounds is not currently supported by composition systems. Also mental images of compositional structures are difficult to reflect in software tools. This was apparent in composers’ comments on inappropriate interface metaphors, and representations.

The above constitute what we believe are the more fundamental notions that have emerged from our analysis. In the following section we discuss their implications with respect to software requirements for composition systems.

### 5.3 Implied software requirements

The above notions imply limitations of some current composition software, and suggest generic requirements for overcoming these. However, we believe it inappropriate to attempt to list specific “hard” requirements, as is conventionally “good practice” in software engineering. This is because our study concerns a range of evolving applications, rather than a set problem. Also, the experimental nature of the area precludes any definitive solution.

In some cases, the requirements are a confirmation of what is already emerging in many composition systems. However, in other cases, the generic requirements are likely only to identify an area of future research aimed at developing composition systems that better support creativity.

#### 5.3.1 System Architectures

In general, **notions 1 to 5** suggest extensible environments for hosting collections of composition tools and for programming new

ones. Further, there should be environments appropriate to each of the categories of composers identified (see 4.2.1). This suggests two requirements.

**Requirement 1:** Composition software should be hosted within extensible software environments. Specifically, composition software systems should be extensible such that arbitrary collections of software tools and materials can be brought together for use within the composer's working environment.

**Requirement 2:** A single "standard" generic composition system is inappropriate for the domain. However, application programming interfaces (APIs) and coding standards, with open-source implementations, are likely to be of value, since they will increase the compatibility of composition tools.

Typically, extensibility in software is achieved through the provision of an API which allows third-party extensions to be developed for an existing application. Within the computer music field, a well-known example is the VST plug-in architecture. Object-orientation provides a natural framework within which to implement such extensible architectures, since it emphasizes the separation of the interface of a software object from its underlying mechanism. Additionally, object-orientation supports extension through inheritance, which allows software components to be created which are specialisations of existing components.

Within software engineering, there is also a move towards open-source development, in which source code is contributed from many individuals who are external to the core development team (often enthusiastic users of the application). Well-known examples of open source projects include the Linux operating system and the MySQL database.

This approach is slowly emerging within the music community and has led to an increase in the diversity and quality of software tools available. We believe that an enhancement and better organisation of the trend will be beneficial. Notions concerning the holistic approach of composers (**notion 10**), suggest a requirement for systems which provide control over the entire composition process, with freedom to configure artefacts in novel ways at all levels of abstraction (**notion 4**, also). Thus, a third requirement is:

**Requirement 3:** It is advantageous for the composition system to support manipulation of audio information at all levels of abstraction, ranging from the DSP level, to the aesthetic and perceptual levels, and to give composers control over the entire composition process.

Given the diversity of tools that may be hosted within composition systems, there is, at the highest level, a need to provide a tool-independent representation, such that associations between artefacts of different types may be made. Our fourth requirement is therefore:

**Requirement 4:** The representation and manipulation of artefacts utilised in the composition process should, at the highest level, be largely unconstrained by the types of specific artefacts. Thus, the system should allow composers to make free associations between these artefacts.

Requirement 4 is a consequence of **notions 1-5**, since the greater the diversity within a single system, the greater the requirement for heterogeneous combinations of artefacts.

Also, a high-level "untyped" representation of artefacts may increase the overall usefulness of the computer system within the composition process. The observed importance of "time away" from the computer system during composition (see 4.2.5.) has two possible interpretations. Either, it represent limitations of software being used, where the composer has to find some alternative form of support, or aspects which are essentially human and for which the computer has no role. Our analysis has revealed instances of both, each having significance for future composition systems. Specifically, the added compositional freedom provided by an untyped workspace may reduce the need for composers to resort to other means away from the system, such as pen and paper sketching and listing.

However, it is not clear if, and how, requirement 4 can be realised within an actual software system. This is an area for further research and will be addressed in the next phase of the study at Sheffield and also within the ongoing Patter Play System project at DIKU (Murphy 2002). One possible approach is to explore its realisation within the object-oriented paradigm and its notion of inheritance; associations can be

made between objects at the highest level of the hierarchy to create relationships between specialisations at lower levels of the hierarchy (e.g., an audio signal and a video clip). Thus, building on applications of object-orientation in many existing systems (e.g., OpenMusic (<http://www.ircam.fr/produits/logiciels/openmusic-e.html>)) and composition languages (e.g., CLOS (Taube 1989; Desain & Honing 1997)). Finally, we consider one aspect of system openness that appears to be desirable for composers, but which is contrary to software engineering “good” practice. That is, openness to encounters with the unexpected (**notion 1**). The importance of serendipity and randomness has been a constant theme in our study. The following requirement is a direct consequence of this.

**Requirement 5:** Composition systems should provide the conditions at all levels where serendipity and randomness can occur and lead to creative inspiration.

Future studies should include consideration of facilities for controlled “loss of control”, random local and global (i.e., Web-based) browsing, and stochastic generative features.

### 5.3.2 Object and Process Management

We now consider possible interpretations of the above notions with respect to support for object and process management. Clearly, this management must be supportive, rather than prescriptive, since the notion of constraining composers to work within the rules of “good practice” is inappropriate.

The following two requirements are implied by our observations concerning the multithreading working methods of the composers (**notions 1-3, 9 and 10**).

**Requirement 6:** Support for multitasking by a single user is likely to be advantageous to composers.

**Requirement 7:** Support systems require a long memory of composers’ previous interactions, and the ability to answer queries concerning those activities.

The above requirements address the observed “voyage of discovery” nature of composition,

whereby refinement of an artefact (convergent thought) may lead to new inspiration (divergent thought) and unanticipated compositional activity. This has software implications, since when composers multitask their activities in this manner it creates multiple incomplete, possibly long and complex concurrent transactions that must be managed. Support for concurrent transactions is standard for multi-user systems, but not for single user systems. Also, the conventional requirements for transaction correctness (i.e., the ACID criteria (Atomicity, Consistency, Independence and Durability)) do not hold, since these are concerned with semantic isolation and serialisability of transactions, such that they do not interfere with each other, whereas transaction of the above type can be tightly interrelated.

Requirement 7 is also a consequence of Requirement 6, since multiple transaction management requires rollback, resume and commit facilities, and also memory-aids for the composer, which in turn suggests some form of system memory, such as transaction logging, temporal database or versioning features, whereby process and entity histories are retained, and previous states can be queried and re-created.

### 5.3.3 Data and knowledge management

We now further consider requirements from the perspective of data and knowledge management. The following two requirements follow directly from **notion 6**.

**Requirement 8:** It is advantageous for composition software to make explicit the personal know-how accumulated through compositional experience, and also that of the community.

**Requirement 9:** It is advantageous for composition systems to support dialogue with others within and beyond the electroacoustic music community.

The above imply the creation and maintenance of repositories within which personal and community knowledge is stored and can be accessed. Again, it is not clear how this requirement can be satisfied, and it is therefore a topic for future research.

At an operational level, much knowledge of practical techniques and their use can be captured by logging and storing the composers' interactions, with composition systems, artefact versions, etc.

Further, the Web now provides low-cost global infrastructure which is already being used to facilitate dialogue and the communication and sharing of knowledge within and beyond the community. We believe that an enhancement and better organisation of use of the Web to this end will be beneficial. Other communities are currently developing ways of pooling data and computational resources through the creation of computational grids (e.g., Jeffries 2001). Also, there exist initiatives within such other communities to define metadata standards for describing information and artefacts, e.g., the XML-based RDF scheme ([www.ukoln.ac.uk/metadata](http://www.ukoln.ac.uk/metadata)). One advantage of such initiatives is to make recoded information more amenable to search engine-based retrieval. It is likely that the music community would benefit from similar initiatives.

### 5.3.4 Composer-computer interfaces

The following requirement is a direct response to **notion 7**.

**Requirement 10:** It is advantageous that composition systems should provide interfaces, which communicate the physical "feel" of sounds.

Our experiments reveal a clear requirement for interfaces that put composers in touch with the physical nature of sounds. In systems terms, this suggests the need for interfaces that provide low-

level, possibly tactile, feedback relating to waveforms and composers' perceptions of sound qualities, to complement the symbolic representations.

Further, our analysis identified a more general requirement for interfaces based upon visuo-spatial dimensions (**notion 8**). The next requirement is a direct response to this.

**Requirement 11:** It is advantageous to support the expression of compositional ideas through visuo-spatial dimensions and representations.

The preceding two requirements confirm and support current directions in interface research and design. We suggest that composition systems based upon interactive, immersive environments, e.g., the CAVE system ([www.evi.uic.edu/research/vrdev.html](http://www.evi.uic.edu/research/vrdev.html)) that trace users' body movements and provide immediate visual and acoustic feedback might be advantageous to composers.

**Notion 10** also has implications with regards to software interfaces, their representations and metaphors. There is no clear requirement which directly addresses this aspect of this notion. For example, two contradictory messages have emerged from the analysis. On the one hand, some composers expressed frustration with interfaces that did not mirror their mental images and perceptions, whereas there was also a clear requirement for interfaces which provided inspiration by challenging their pre-conceptions. There is clearly a need for ongoing research into these HCI issues. Specifically, following our earlier argument (see 4.2.4.) a possible negative impact of audio visualisation should be considered in HCI design.

# CHAPTER 6 - CONCLUSIONS AND FUTURE WORK

## 6.1 Summary

In this report we have identified a need for research geared to developing a base for composition systems which better support creativity. Also, we have argued that preliminary qualitative research is necessary to establish the sensitising parameters for this area. The main contribution is the presentation and discussion of results from such a study, in which composers have been observed in natural settings, working on commissions.

The analysis and discussion of results in chapter 4 imply limitations of current composition systems, and suggest requirements for overcoming those, elaborated in chapter 5.

We have loosely classified composers into those who consider engagement with computer software and hardware at the lowest level as part of the compositional process and those who seek innovation in the (mis-)use of existing tools. Software environment requirements for using existing tools and for generating new tools differ. Accordingly, we should not be seeking a generic software environment for all persuasions of composers. This complements a requirement for diversity of composition tools.

The observed “voyage of discovery” nature of composition, whereby refinement of an artifact (convergent thought) may lead to new inspiration (divergent thought) and unanticipated compositional activity, has software implications. Consequently, composers may multitask their activities, thus creating multiple incomplete concurrent transactions. Support for concurrent transactions is standard for multi-user systems, but not for single user systems. Also, the conventional requirements for transaction correctness, i.e., the ACID (atomicity, consistency, independence and durability) test, do not hold, since these are concerned with semantic isolation of transactions, such that they do not interfere with each other.

A key interface issue is the importance and impact, both positive and negative, of the

visualization of sounds. This relates both to individual composition tools, and also to the environments within which they are used, since the latter must also represent compositions and their components such that the use of tools can be focused and integrated. Our observations suggest the need for further research into this HCI aspect, specifically focusing on visualization of compositions at both the macro and micro levels, so as to better communicate properties and quality of the audio content. Specifically, interfaces based upon visuo-spatial parameters and representation are likely to be beneficial.

Our analysis revealed that important parts of the composition process happen away from the computer. This is partially due to limitations of the computer system being used and partially can be explained by aspects for which the computer has no role. One clear implication from our observations is that some integral untyped workspace within which composers may make notes and freely sketch associations between musical artifacts may be valuable.

We have also identified tension between those associations that are important to composers and those that are visible in GUIs. Again, this suggests that it would be useful to have a space to represent what is perceptually important, in addition to representations concerned with “engineering” the composition.

Note that, both of the previous two speculations are consistent with the notion of creativity as a process of divergent thought, whereby associations are made at a very high level of abstraction.

Finally, our experiments reveal a clear requirement for interfaces that put composers in touch with the physical nature of sounds, and with the wider community. In systems terms, this suggests: (i) the need for interfaces that provide low-level possibly tactile feedback relating to waveforms and sound qualities, to complement the symbolic representations; and (ii) the need for a repository component within which a composer can accumulate personal know-how



together with access to the community's know-how, through the Web.

## 6.2 Future Work

The notion and generic requirement presented in this report will form the basis of future research into enhanced composition systems. Aspects of these findings will be elaborated and the theories will be tested through the construction and evaluation of a prototype system.

Also, we cite the Pattern Play System (Murphy 2002) (see Appendix B) as an approach which addresses many of the requirements presented in

this report. This project, primarily being conducted at DIKU, will contribute to the validation of the theories presented as requirements in this report. Specifically, we anticipate that the projects at Sheffield and DIKU will prove to be complementary, each providing the other with feedback, research and examples relating to efficacy and specification of a general composition tool-kit.

We also view electroacoustic music composition as a rich instance of the creative process and thus aim to derive general techniques for improved software support for creative individuals.

## References

- Bamberger, J. (1977) In search of a tune. In: D.Perkins & B.Leondar (eds.) *The Arts and Cognition*. Baltimore: Johns Hopkins Press.
- Bensa, J., Gibaudan, F., Jensen, K., Kronland-Martinet, R.: (2001). Note and hammer velocity dependence of a piano string model based on coupled digital waveguides. In *Proceedings of the ICMC*, Havana, Cuba, 2001.
- Bensa, J., Jensen, K., Kronland-Martinet, R., Ystad, S (2000): Perceptual and analytical analysis of the effect of the hammer impact on the piano tones. In *Proceedings of the ICMC*, Berlin, Germany, 2000.
- Bregman, A.S. (1990) Auditory Scene Analysis. Boston, MA: MIT Press.
- Burrell, G., Morgan, G. (1979) Sociological paradigms and organisational analysis. London: Heinemann, 397-8.
- Cascone, K. (2000). The Aesthetics of failure: "Post-Digital" Tendencies in *Contemporary Computer Music*. *The Computer Music Journal*, 24(4), 12-18.
- Chroust, G. (1989) Duplicate Instances of Elements of a Software Process Model, *ACM SIGSOFT Software Engineering Notes* 14:4, June 1989, pp 61-64.
- Clowes, M (2000) An investigation of compositional practices in the field of electro-acoustic music, with an evaluation of the main software environments currently in use. Dissertation, Master of Science in Information Management, The University of Sheffield.
- Collins, D. (2001) Investigating computer-based compositional processes: a case-study approach. Ph.D. Thesis. The University of Sheffield.
- Collins, D. (2001) Investigating computer-based compositional processes: a case-study approach. Ph.D. Thesis. The University of Sheffield.
- Davidson, L., Welsh, P. (1988). From collections to structure: the developmental path of tonal thinking In: J.A.Sloboda (ed.) *Generative processes in music; the psychology of performance, improvisation & composition..* Oxford Science Publications.
- De Bono, E. (1987). Oxford Companion to the Mind, Oxford: Oxford University Press
- Desain, P, Honing, H (1997): CLOSe to the edge? – Advanced object-oriented techniques in the representation of musical knowledge In *Journal of New Music Research*, 2, 1-16. 1997. ISSN: 0929-8215.
- Eaglestone, B.M., Davies, G.L., Ridley, M., Hulley, N. (1993) Implementation of an Artists Versions Model using Extended Relational Database Technology, *Advances in Databases, BNCOD-11*, Keele, UK, July 1993, Lecture Notes in Computer Science, Springer Verlag, pp 258-276.
- Eaglestone, B.M., Ford, N., Clowes, M. (2001) Do Composition Systems support Creativity: An Evaluation. Proceedings of the *International Computer Music Conference (ICMC 2001)*, Habana, Cuba, International Computer Music Association, pp. 22-25.
- Eaglestone, B.M, Ford, N., Nuhn, R., Moore, A., Brown, G.J. (2001): Composition Systems Requirements for Creativity: What Research Methodology? In *Proceedings of MOSART Workshop on Current Research Directions in Computer Music*, Barcelona, November 15-17, 2001, Audiovisual Institute, Pompeu Fabra University, Spain, pp. 7-16.
- Eaglestone, B.M. (1994) An Artistic Design System, *SOFSEM '94 Invited Talks*, Milovy, Czech Republic, Czech Eaglestone, Society of Computer Science, pp 15-37.

- Eaglestone, B.M., Davies, G.L., Ridley, M., Hulley, N. (1993) Implementation of an Artists Versions Model using Extended Relational Database Technology, *Advances in Databases, BNCOD-11*, Keele, UK, July 1993, Lecture Notes in Computer Science, Springer Verlag, pp 258-276.
- Eaglestone, B.M., Holton, R., Rold, L. (1996) GENREG: A Historical Data Model Based on Event Graphs, *7th International Conference on Database and Expert Systems Applications (DEXA'96)*, Zurich, September, 1996, Lecture Notes in Computer Science 1134, Springer, pp 254-263.
- Ellis, D. (1993) Modelling the information seeking patterns of academic researchers: A grounded theory approach. *Library Quarterly*, vol 63, no. 4, pp 469-486.
- Emmerson, S. (1989). Composing strategies and pedagogy. *Contemporary Music Review* 1989, Vol 3, Harwood Academic Publishers, UK, 133-144..
- Erlandson, D. A. et al (1993). Doing naturalistic enquiry. Sage Publications, London.
- Feyerabend, P. (1975). Against method: outline of an anarchist theory of knowledge. London: New Left Books.
- Ford, N. (1999) Ford, N. (1999) Information retrieval and creativity: towards support for the original thinker, *Journal of Documentation*, 55(5), 528-542
- Godsmark, D. & Brown, G. (1999). A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27, pp351-366.
- Gregory, R.L. (ed.) (1987) The Oxford companion to the mind. Oxford: Oxford University Press.
- Guilford, J.P. (1967). The nature of human intelligence, New York; McGraw-Hill.
- Guba, E. G., Lincoln, Y. S. (1985) Effective evaluation. San Francisco: Jossey-Bass.
- Janesick, V. J. (2000). The Choreography of Qualitative Research Design. In Denzin, N. K. & Lincoln Y. S. (eds.), *Handbook of Qualitative Research (2nd ed.)*. Sage Publications, Thousand Oaks, CA, USA.
- Jensen, K (1999): Timbre Models of Musical Sounds. Ph.D. dissertation, Department of Computer Science, University of Copenhagen, 1999. Report no. 99/7.
- Lincoln, Y.S. & Guba, E.G. (1985) Naturalistic inquiry. Sage Publications; California.
- Marentakis, G., Jensen, K.. (2001): Hybrid synthesizer: Progress report. In *Workshop on current research directions in computer music*, Barcelona, Spain, 2001.
- McGurk, H. McDonald, J. (1976) Hearing lips and seeing voices. *Nature* 264, 746-748.
- Moore, R.F. (1990) Elements of Computer Music, by F. Richard Moore, Prentice-Hall.
- Murphy, D. (2002): Pattern Play, poster at ICMAI, Edinburgh, Sep 2002. (to be published in University of Edinburgh on-line technical report series).
- Olaisen, J. (1991) Pluralism or positivistic trivialism: important trends in contemporary philosophy of science. In H.E. Nissen, H.K. Klein & R. Hirschheim (Eds.). *Information systems research: contemporary approaches and emergent traditions*. Amsterdam: Elsevier. pp. 235-265.
- Polfreman, R., (1999), "A task analysis of musical composition and its application to the development of Modalyser", *Organised Sound* 4 (1)
- Reybrouck, M. (1997) Gestalt concepts and music: limitations and possibilities. IN: M. LEMAN (ed.) Music, gestalt and computing. Lecture notes in artificial intelligence, no.1317. Springer-Verlag, Berlin:
- Reitman, W.R. (1965). Cognition and thought. New York:Wiley.
- Sloboda, J. (1995). Do psychologists have anything useful to say about composition? Paper presented at the *Third European Conference of Music Analysis*, Montpellier, France, 16-19 February. Courtesy of the author.
- Taube, H. (1989): Common Music: A Compositional Language in Common Lisp and CLOS. In T. Wells and D. Butler (eds) *Proc. Int. Computer Music Conf.* 1989 pp. 316-319
- Ungvary, T., Waters, S. and Rajka, P. (1992) Nuntius: A computer system for the interactive composition and analysis of

- music and dance. *Leonardo* (Pergamon Press, Oxford). No.1.
- Ungvary,T., Kieslinger, M. (1998) Creative and Interpretative Processmilieu for Live-Computermusic with the Sentograph. " In *Controlling creative processes in music* (Herausgegeben von R. Kopiez und W. Auhagen). Publisher : Peter Lang, Frankfurt am Main. (Schriften zur Musikpsychologie und Musikästhetik. ISBN : 3-631-33116-9
- Vercoe, B. (1985). *The Csound Music Synthesis Language* . Cambridge, MA: Media Lab, MIT  
ftp://  
sound.media.mit.edu/pub/Csound.
- Vertegaal, R. & Ungvary, T. (1995) *The Sentograph: Input devices and the communication of bodily expression. Proceedings of the International Computer Music Conference, San Francisco, Computer Music Association, pp 253-256.*
- Vertegaal,R.; Ungvary,T, Kieslinger, M. (1996) *The Musician's Cockpit: Transducers, Feedback and Musical Function. in the Proceedings of ICMC96- Hong-Kong. Published by Default\_XREF\_style REFICMA.*
- Wallach, H (1940) *The role of head movements and vestibular and visual cues in sound localization. J. Exp. Psychol. 27, 339-368.*
- Wallas, G. (1926). *The art of thought. London; Watts.*

## APPENDIX A - Catalogue of audio-visual media:

Media type & Index No.	Content	Date
Minidisk No.1	General Interview with Richard Thomas	20/09/2001
	Reflective Interview with Richard Thomas	24/09/2001
Minidisk No.2	General Interview with Martin Robinson	07/10/2001
Minidisk No.3	Audio Recordings from Experiment 1	21/09/2001
CD-R No.1	Audio files from Experiment 1	21/09/2001
CD-R No.2	Audio files and <i>LogicAudio</i> Arrangements from Experiment 2	13/12/2001
VHS No.1	Computer output Experiment 1 (Richard Thomas' PC)	21/09/2001
VHS No.2	Computer output Experiment 1 (Martin Robinson's and joint PC)	21/09/2001
VHS No.3	Camera recording I Experiment 1	21/09/2001
VHS No.4	Camera recording II Experiment 1	21/09/2001
VHS No.5	Computer output Experiment 2/Session 1	05/12/2001
VHS No.6	Camera recording Experiment 2/Session 1	05/12/2001
VHS No.7	Computer output Experiment 2/Session 2+3	06/12/2001
VHS No.8	Camera recording Experiment 2/Session 2	06/12/2001
VHS No.9	Camera recording Experiment 2/Session 3	06/12/2001
VHS No.10	Computer output Experiment 2/Session 4	12/12/2001+ 13/12/2001
VHS No.11	Camera recording Experiment 2/Session 4	12/12/2001
VHS No.12	Camera recording Experiment 2/Session 5	13/12/2001
VHS No.13	Camera recording Experiment 3	13/12/2001

## Appendix B – The Pattern Play System

The Pattern Play (Murphy 2002) system is intended to serve as a composition tool/platform although it is also aimed at being a platform for research into perception of musical structure and ultimately also a performance instrument.

Structure is represented in terms of patterns extracted from features, which may then be rearranged via a graphical and gestural user interface. Aspects from a given passage of music can be imposed upon, and mixed with, other sections of music. There is also the facility to capture expressive gestural user input: both for

immediate generation of phrases and for imposing phrasing interpretation onto passages.

The system includes a feature extractor and analysis engine, native score and feature pattern formats, an infrastructure for composing and rearranging music based on its pattern representation, a graphical user interface for display of pattern representations, and a gestural user interface for manipulation of these patterns and for capture of expressive gesture. See figure 1.

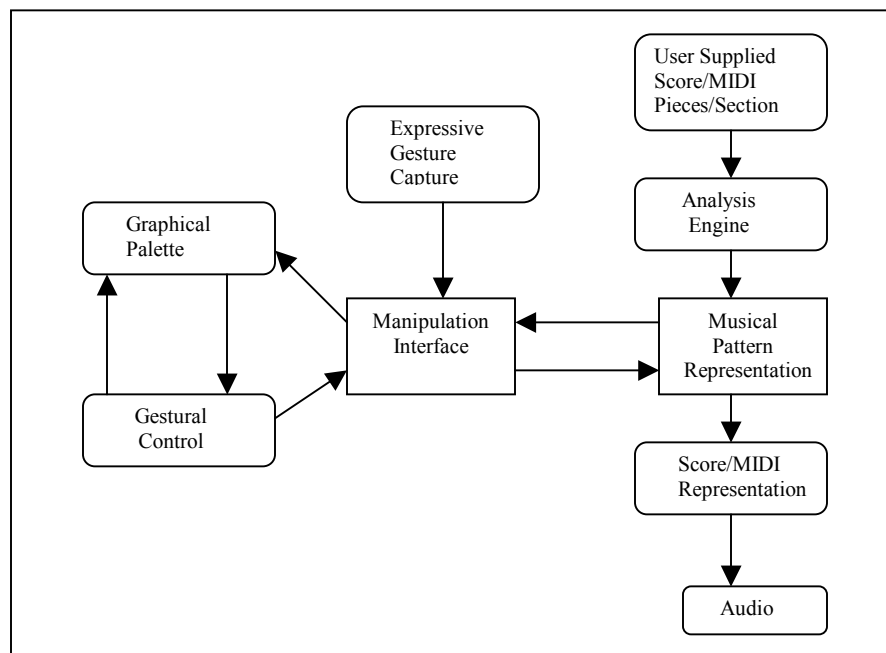


Figure 1: Patter Play Architecture

The internal pattern representation may be input to the system either by an analysis of a piece or passage supplied by the user, and/or by use of the Graphical and Gestural User Interface (GGUI).

- Once a pattern representation exists, it may then be manipulated via the GGUI. Such manipulations include:
- high-level re-arranging of the piece's structure,

- imposing abstract structural patterns or local phrasings from one piece onto a new piece, or
- beating out rhythms, describing melodic contours or phrasing interpretation by expressive hand gesture.

The gesture interface is used both for control of the pattern manipulation and for expressive gesture capture.



**IHP Network HPRN-CT-2000-00115 MOSART**  
**Music Orchestration System in Algorithmic Research and Technology**

**Summary of the Panel Discussions**

from

Workshop on Current Research Directions in Computer Music

Barcelona, Nov 15-16-17, 2001

Audiovisual Institute, Pompeu Fabra University

**Music Sound Modeling**

Keywords: Instrument modeling, Instrument recognition, Content processing.

**Summary of the Music Sound Modeling Panel**

Enric Guaus (editor)

Page 298

**Music Interfaces**

Keywords: Gesture Based Interaction, Mapping strategies, Multimodality.

**Summary of the Music Interfaces Panel**

Álvaro Barbosa (editor)

Page 302

**Music Performance**

Keywords: Quantitative Models of Performance, Cognitive Models of Perception and Production, Expressive Performance and Emotion.

**Summary of the Music Performance Panel**

Maarten Grachten (editor)

Page 306

**Empirical Music Performance Research: ÖFAI's Position**

Gerhard Widmer, Simon Dixon, Werner Goebel, Efstathios Stamatatos, Asmir Tobudic

Page 311

**Music Performance Panel: Position Statement: KTH Group**

Johan Sundberg, Anders Friberg, Roberto Bresin

Page 314

**Music Performance Panel: NICI / MMM Position Statement**

Peter Desain, Henkjan Honing and Renee Timmers

Page 318

**Music Generation**

Keywords: Algorithmic composition, Composition systems.

**Summary of the Music Generation Panel**

Rubén Hinojosa Chapel (editor)

Page 323



# Sound Modeling Panel

Enric Guaus  
enric.guaus@iua.upf.es

*Music Technology Group, Audiovisual Institute, Pompeu Fabra University  
Passeig de la Circumvalacio 8, 08003 Barcelona, Spain*

## Abstract

This paper presents an overview and conclusions of different topics discussed in the Sound Modeling Panel, in the MOSART Workshop - Workshop on Current Research Directions in Computer Music - that took place in Barcelona, November 16th of 2001. The chairman was Mr. Giovanni De Poli (DEI-University of Padova, Italy) and the invited members were: 1. Richard Kronland-Martinet (CNRS - LMA, France), 2. Stefania Serafin (CCRMA - Stanford University, USA) in substitution of Mark Sandler (Queen Mary University, Great Britain), 3. Jan Tro (NTNU Norwegian University of Science and Technology, Norway) and 4. Xavier Serra (Pompeu Fabra University, Spain). The panel had a duration of one hour and it was structured in two parts: The introduction from each one of the members, and the open discussion on the introduced ideas with the audience participation.

## 1 Introduction

This panel focuses on current research directions of sound modeling. Topics that will be discussed include:

### Signal vs. source vs. perceptual models:

Which are the best models for sound synthesis, understanding, coding? These models are often developed in different scientific communities. Are models developed for one specific application, useful in other domains? Can sound processing models (digital audio effects) also be useful as sound models? Transfer of sound and data compression procedures have been developed. Will there be other compression techniques based on heavily reduced parameter sets such as the ones developed in sound modeling applications?

**Timbre modeling and understanding:** Timbre can be interpreted as sound quality or as sound (source) identity. Are there general models for timbre representation? How is a timbre space organized for synthetic sounds? Which are the most relevant parameters for quality evaluation and for identity recognition? Do synthetic sounds possess an identity?

**Sound design:** How to find the right input parameters for a particular synthesis models? Can we model sound articulations? Can we separate sound design from performance issues? How important is the idea of sound identity when designing a synthetic sound? How can we map, in

real time environments, gesture to sound models? Can soft computing methods be useful in sound design?

**Model evaluation:** Are there objective criteria for assessing the effectiveness of a sound model? How perceptual issues can be exploited for evaluation? Can a subjective measure of distance be formalized? How to include sensitive, emotional aspects in the evaluation? Can usability tests be applied to sound design?

## 2 Opening Statements

### 2.1 Richard Kronland-Martinet

There is a lot of people that thinks that, in sound modeling, everything is done. It's not true. Sound modeling means to give a precise description of all the phenomena occurring in an instrument. Sound modeling in computer music is any technique which allows us generating sound. It means that we can use physical modeling, but also any kind of representation of signal. It's obvious that a lot of things have been done in sound generation, but a lot of things have to be done now. Probably, the problem of sound modeling should be addressed in a different way: it's very important to take into account what musicians want to do with sound models. That means that it should be a very close relation between sound modeling and control of sounds. If somebody wants making music with a violin, somebody that builds the violin is needed. Of course, when somebody is building an instrument, it's very important

taking into account that these instruments must be playable.

There are several very good models from a technical point of view (physical models, spectral models . . .) but the instruments built with these methods are unplayable. The problem is the relation between the control part of the instrument and the design of the technique. This relationship is not good enough, at least for musician.

## 2.2 Stefania Serafin

We can ask ourselves which model is the best one. There is no any best model in general, but there is a best model for each specific application. The goal of sound modeling is building a model as similar to the real instrument as possible without taking care with the computational cost and other. In the other hand, we can implement different algorithms of data compression in order to transmit data fastly without loss of information. With these two techniques (modeling and audio processing) we should be able to help musicians in their compositions.

Some years ago, we all were surprised with new systems that could implement hundreds of different sinusoid signals in real time, and it seemed that these techniques could give us the full control in sound modeling. Nowadays, the research community is looking for new techniques for reducing the number of parameters in sound implementations, for a better real-time Internet transmission, but at the same time, maintaining the sound quality. Of course, it depends on the specific application.

It is not very difficult building a good quality synthesizer (with spectral modeling is quite easy reproduce any kind of instruments), but we can discuss, for a long, time if this sound is good enough or if it's too synthetic. In the case of physical modeling, it's very important a good parameter choice for a realistic sensation.

## 2.3 Jan Tro

On the Music Generation panel (Mosart - November 15th of 2001), the discussion was that the music and technology, artists and engineers, must meet somewhere. In the history of technology it has never been easy to reproduce some realistic sounds. In the other hand, we are able to understand music from a technological point of view, i.e. any reverb model. But a lot of recordings uses exactly the same reverbs! We should be very careful when choosing models, specially the input data for these models. Note that, in psychoacoustics, there are no new studies in the last 40 o 50 years.

It's very difficult to separate the sound models and the perceptual models: The feedback from the

listener to the performer is a basic question. Not only the room, but the interpretation. There are a lot of parameters that are fully unknown in a music performance: the mental factor has not been described yet. It's very hard to choose the best model for an instrument, when we are evaluating different ones, as the best one is the instrument itself.

Data reduction mentioned above is very important, but we must go a bit further: Don't make a list of only parameters, let's go to find which is the brain process to feel the music as it is.

## 2.4 Xavier Serra

In the last twenty-five years, the driving force in the computer music community has been the sound modeling, and it has been a complete failure. If we measure the research activity by the impact that it has in different fields of the music (composers, industry . . .), specially in the past 15 years, it's zero: there has been practically no new devices built. If we look at the music, most of that *new* music uses technologies that were developed before these 15 years.

We could think that there is nothing to do now, everything it was useful to do is done. Perhaps the sound synthesis is dead. If we assume this, it would be a good starting point for reconducing our research activities. We have spent about 30 years trying to model and reproduce the classical instruments. We have to change our paradigms: we must meet new ways in the scientific community; the engineers have to look outside and realize what the society needs.

## 3 Open discussion

The open discussion started with a discussion about which one is the actual driving force in sound modeling. The poor need of knowledge is a very hard driving force. If the model is perfect then you know it all, so you don't need the model. But if we want survive in the business area, we must change the role and we must build and listen different models in order to find the best one. *Xavier Serra* answered that the model is just a scientific approximation to a problem. There is people of different disciplines, in this Mosart Workshop, and we realize that the scientific output is not enough: we really have to go forward and find new ways to work with that.

Another topic discussed was about the poor communication between engineers and composers. *Xavier Serra* answered, from an historical point of view, that in the first ICMC there was a good balance between industry and academics. Nowadays, at the ICMC, the industry is not present. In the other hand, in the last ICMC, there were no composers in the scientific sessions, because they were

in other sessions. In the last few years this division has been emphasized. There are some explanations for that. The first one is that our scientific research has evolved as much as musicians don't understand anything in that so specific field. When they use our techniques, they don't understand what's going on. The public asked why don't the engineers assist at the music conferences, and *Xavier Serra*: answered that sometimes they do because most of the engineers are musicians too. The idea that the engineers plagiarize the musicians has been quite common.

The discussion followed with the difficult communication between engineers and musicians. Nowadays, it's possible and quite easy to get a synthesis software and compose any kind of music. But some years ago, musicians had to learn computer science and the different synthesis techniques. Where is the difference? Nowadays, when we click at the plug-in button of any commercial software, there is a lot of people all over the world doing the same thing and using exactly the same reverb. It's a very good option from a social point of view, but from a musical point of view, perhaps it would be better a good knowledge of CSound for developing our own audio effects. Musicians are losing something when use these commercial techniques.

There was a reference to Iannis Xenakis who told that we can not be amused by the computers, we have to be amused by the ideas we have. As the composers get afraid with the mounts of information we manage (i.e. in this Mosart Workshop) the wall between composers and engineers is higher and higher. Perhaps the solution is a good communication between them.

At this point, there was an enthusiastic intervention with the future of computer music: the scientific part tries to understand how the sounds work and computer music is studying how a human can plays a instrument and applies the psychoacoustic information to understand this. In the other hand, there are very strong models, but the problem is how the humans communicate with these models. Perhaps we should think about how the interfaces must be done in order to make all these models more useful. *Richard Kronland-Martinet* answers that, although there are a lot of tools for engineers, there aren't tools for composers. Ten years ago, everybody was talking about how to create sounds, and these sounds must be create with CSound or Music V. Nobody was talking about the control of sound. We should think about sound modeling in a different way and it's obvious that modeling sound is modeling an instrument. There is an example, from Yamaha, of a very good synthesizer from a technical point of view, but nobody can use it. There are 2 problems: when you want to play a new instrument, all you need is time. If you want to play the piano you need

about 10 years, and so on, but now it's impossible: there are no scores for these new instruments. Everybody wants to play a new instruments, but nobody wants to put time in these new instruments. *Xavier Serra* added that there are very few examples of electronic instruments that have created a virtuoso school, and it took few years. Sound modeling misses the interfaces or composer aspects of the instrument, but there are many things that can be done for making them useful for musicians. The driving force for sound modeling is nowadays is not in the sound modeling community by itself but on the issues like performers. From a signal processing point of view, we are a little bit stuck. We have to go beyond that and include more ways for improving our work. Finally, *Jan Tro* added that it's important to emphasize that we actually use a lot of years training for playing a music instrument. We should take care with the knowledge on modeling instruments, because we need a lot of time and method to handle a instrument, even if it's an acoustic one or an electronic one. We should put that experience as a part of models, in the same way we have tried to implement the perceptual model.

At the end of this open discussion, the panel was invited to comment why the synthetic instruments like DX-7 or Moog succeed and became as reference instruments. *Xavier Serra* told that the DX-7 was a combination of a lot of things and it was so successful that everyone is still trying to imitate it. It was the combination of his relatively low price, it was general enough, and the synthesis technique is no important because only few people could program it. From a technological point of view there are a lot of things, like the microprocessors, the polyphonic features... that make it attractive for community. It was at the right time at the right place and pushed by the right company. The Moog is another history, but it seems that his high versatility and real-time performance on the stage characteristics made it successful. *Giovanni de Poli*: added that the important thing is that the instrument should be playable. These instruments have a good characteristic: his identity. The own identity of these instruments help themselves. About the correct time, there is a change of attitude in people towards new instruments or reproducing the traditional ones depending the time period of people. The portability of these instruments was also commented and, finally, the missing MIDI implementation on the Yamaha-DX7 was commented too.

## 4 Conclusions

The panel was about Sound Modeling, specially focussed on their future directions. One of the topics

discussed was about the sound modeling definition, and how relatively simple is generating quite realistic sounds. The problem of these techniques is the high number of parameters needed, that make them unplayable. Another topic discussed on the panel was the slow evolution of the computer music community in the last fifteen years, in sense that although there are a lot of papers and conferences in sound modeling, there isn't any repercussion on the society. But the main topic it was discussed was the big separation between the computer music community and the performers or musicians. Which one is the reason? Perhaps the models are so complex or perhaps the engineers need more and better communication with performers, and viceversa.

From my personal point of view, it seems that quite engineers involved in Sound Modeling know that their work has to focus, not only on the model by itself, but on the final user of this model: the musician. This is the unique way for creating new instruments and, in the future, listening some good enough recordings of these new instruments.

# Overview and conclusions of the Music Interfaces Panel Session at the MOSART Workshop (Barcelona, 2001)

Álvaro Barbosa

Email: alvaro.barbosa@tecn.upf.es

Music Technology Group, Audiovisual Institute, Pompeu Fabra University  
Passeig de la Circumvalació 8, 08003 Barcelona, Spain

## Abstract

*In this paper is presented an overview and conclusions of topics and ideas discussed in the e Music Interfaces panel that took place during the MOSART Workshop – Workshop on Current Research Directions in Computer Music - Barcelona, November 17<sup>th</sup> of 2001.*

*The invited members of Panel were: Antonio Camurri (DIST-University of Genova, Italy); Sergi Jorda (IUA-Pompeu Fabra University in Barcelona, Spain); Roger Dannenberg (Carnegie Mellon University, Pittsburgh, USA); Leonello Tarabella (CNUCE/C.N.R. in Pisa, Italy).*

*The Chairman for the Pannel was: Johan Sundberg (KTH-Royal Institute of Technology, Sweden).*

*The panel had the duration of approximately one hour and it was structured in 3 parts: An introduction to the theme of the panel by the Chairmen; A five minutes introductory open statement by each one of the members; An open discussion on the introduced topics and ideas opened to the audience.*

## 1. Introduction

This panel focuses on scientific as well as artistic research perspectives on interactive systems. Main issues that will be discussed include the following:

Interaction Metaphors and Mapping Strategies: from the "musical instrument" metaphor (interaction with an object, immediate cause-effect response), to "dialog" paradigms (interaction as a dialogue with human as well as virtual agents; virtual agents can be auditory, and possibly visual, robotic in a Mixed Reality scenario).

Can models of Expressiveness, Kansei, Emotion influence/improve interaction metaphors, integration of modalities, and mapping strategies? How can such models contribute to more effective interactive systems?

Which methodology and evaluation methods to verify and consolidate results from scientific research?

Are state-of-the-art sensor systems mature enough to capture the physicality, the

sensibility, the expressive content from music performers, dancers, and spectators...?

Reports on good examples and lessons learned from experiences with artists (composers, performers, dancers...) can be very useful as feedback for scientific research. Which models demonstrated successful from artistic productions?

## 2. Opening statements

The opening statements started with **Roger Dannenberg's** presentation, which focused mostly on his personal perspective of what are the main problems and new directions on music creation using interactive systems.

Some of the main difficulties are that enabling technologies have not yet reached a desirable point, real time systems are difficult to create mostly due to the constraints of existing tools like for instance the MAX/MSP software, programming languages and related devices are not easy to setup or the fact that as of now

it is not yet defined a standard I/O portable interface for low level music input and output. Some new directions that should be pursued are for instance the integration of the Look Ahead concept (always present in the way musicians plan and execute their compositions and performances) with sensor technology, the use of artificial intelligence techniques and again, how this could be intergraded with sensor technology, trying to improve existing technology with the use of visual content, the use of the web (on which the work of Sergi Jordá is a reference) <sup>[1]</sup> or the use of robotics.

The following opening statement was by **Antonio Camurri** that started of by planting the question:

*Why interactive performances are so bad?*

According to Antonio Camurri one possible answer is that so far composers and performers have managed to learn how to control sound in space, however it is still necessary to learn how to manage action in space.

This is an extension of musical language that is very difficult to manage, and the main issue is to find deeper correlation between music language and gesture.

There is consolidated research in the field of controlling action in space, like for instance the work presented in the MOSART workshop by Sergio Canazza e Giovanni de Poli <sup>[2]</sup>, however it is necessary to integrate this knowledge in musical performance.

Antonio Camurri also announced the Gesture Workshop for the spring of 2003 that will take place in Genova Italy and will be a good opportunity to consolidate the community of researchers interested in this field.

Following up **Leonello Tarabella** started his presentation by answering Antoio Camurri's opening question with another question:

*Why is computer music not so spread out as other forms of music? Maybe because real time performance is so bad.*

For Leonello Tarabella we are at the "Stone Age" of this culture, and that is quite clear because so far it is not easy to define how to use the computer as an instrument, the way it can be done with traditional instruments.

The effort that must be made is towards fulfilling musicians needs, realising that technology developments are the driven force to the creation of new movements in music, like it happened for instance with rock music that had its origins on the use of the electric guitar has a new instrument made possible by technological developments.

The final opening statement was by **Sergi Jordá** that started of bay claiming that he was moved to use interactive systems with computes to create music because he was a bad jazz performer.

Sergi Jordá agrees with the fact that there are many constrains in the existing languages to develop interactive systems, and this has the effect of increasing the complexity of the mapping.

One of the most common problems, caused by the fact that controlers are often separated from the synthesis engine, is that when designing such a system you cant build a sophisticated controller without previous knowledge of what you are about to control. To address this problem one has to think about parallel design of the input and the output of the system.

On the topic of current state of sensor technology, Sergi Jordá points out that the question is not whether sensor technology is mature enough to allow de creation of interesting music, but instead one should think that technology is never enough in any case, and so we should use what we have.

Computers are not instruments, but a paradigm to create instruments with infinite possibilities, and therefore we can have complexity in this area, however what we really need is simplicity.

Another important point is that the current state of technology allows the creation of instruments that can be performed collectively by several users, that often are untrained musicians.

In this case one should consider that the instruments should be designed to be simple and very constrained.

On the Sergi Jordá's opening statement, Roger Dannenberg commented that although he agrees that we don't need to wait for technology to start creating and that we should use what we have, one should be aware that we are close to the arrival of a revolution in sensor technology.

Sergi Jordá replied that this could lead to an overload, since we have not used what we have yet.

### 3. Open discussion

The Chairman of the Pannel, Johan Sundberg, introduced the open discussion session proposing as a theme the topic of mapping, stating as an example the difficulty of mapping voice which is the most common instrument. Is one to one or many to one the best strategy?

**Sergi Jordá** believes that at least the many to one mapping strategy is flexible enough, but the focus should also be on another critical point, which is the feedback that is always present in traditional music instruments at a physical and visual level and that hardly exists in sensor technology.

Another important point is that we don't have to think that with computer instruments we have to map physical gesture into sound, but we can map into higher-level musical forms.

**Leonello Tarabella** pointed out that the mapping is the critical point to actually create an instrument, given as example, that the "Twin Towers" device presented earlier in the conference<sup>[3]</sup> is a controller that can map to different synthesis engines, however a Theremin is a musical instrument since its coupled with a fixed synthesis engine.

**Roger Dannenberg** added up that also the concept of a musical instrument comes from that past and it's not a requirement that we must follow this model.

**The audience** addressed one final comment to the panel, pointing out to the fact that music has the ability to raise emotions, however for a new instrument to succeed it should not only be able to raise the emotions we have inside, but it should also create new emotions.

On this topic **Leonello Tarabella** states that at the current moment the direction being followed in new musical instruments design should be towards the traditional musicians performing paradigm, and only future generations will succeed in a different approach.

**Sergi Jordá** adds up that by performing mouse music, he found out that the interface he designed for the general community of Internet users is the one that he likes the most.

As a final statement **Roger Dannenberg** pointed out that when the composer designs the controller it has a tremendous influence over the final result of the piece and the performance.

## 4. Conclusions

From the discussion on this panel some strong ideas and questions that were raised could be considered as references and issues to keep in mind on future research work in this field.

One of the strongest ideas is that mapping is a crucial problem.

One could think of mapping as part of the composition or part of the instrument and therefore it has a tremendous influence in the performance.

Mapping strategies depend on the context and the purpose of the interface, and for general purpose musical instrument a simple mapping strategy is sufficient, however in a more complex situation, like controlling action in space during a performance, one needs, for instance, to integrate knowledge about musical performance, and therefore increasing the complexity of mapping. This knowledge should be incorporated on the mapping strategy as an extension to the musical language.

It is clear that the study of mapping strategies is, still at an early stage and that the definition of mapping strategies for different contexts, applying results from research work related with expressiveness analysis and control technology, has tremendous potential.

Another strong question that was raised, was if the environments to create music or music applications should or should not be constrained?

The overall tendency is that the more general is the users universe that will use the controller, the most constrained it must be. On the other hand controllers designed for smaller groups usually constituted by experimented musicians are much less constrained and are more flexible.

However it is clear that there is a need for more scalable environments, which can be constrained at an entry stage, when the user/performer/composer first starts to use the system, and that after a certain point can be configured to be unconstrained and flexible enough to allow an implementation close to the original concepts of the musician.

A final idea that was present in this panel session was that although current work in this field is aiming more towards the need to expand the existing paradigms for music performance and composition, the true potential of this media in order to create new forms of musical language could be unveiled by looking for new paradigms. These paradigms might be limited by the current state of technology development, but must not necessarily be alike the traditional musical instrument model that we know.

## References

- [1] Jordá, Sergi. Barbosa, Álvaro. *Computer Supported Cooperative Music: Overview of research work and projects at the Audiovisual Institute – UPF; Proceedings of MOSART Workshop on Current Research Directions in Computer Music*
- [2] Canazza, Sergio. De Poli, Giovanni. Rodà, Antonio. Vidolin, Alvise. Zanon, Patrick. *Kinematics-energy space for expressive interaction in music performance; Proceedings of MOSART Workshop on Current Research Directions in Computer Music*
- [3] Tarabella, Leonello. Bertini, Graziano. *Wireless technology in gesture controlled computer generated music; Proceedings of MOSART Workshop on Current Research Directions in Computer Music*
- [4] Camurri, Antonio. De Poli, Giovanni. Leman, Marc. Volpe, Gualtiero. *A Multi-layered Conceptual Framework for Expressive Gesture Applications; Proceedings of MOSART Workshop on Current Research Directions in Computer Music*



# Summary of the Music Performance Panel, MOSART Workshop 2001, Barcelona

Maarten Grachten

*Artificial Intelligence Research Institute, IIIA  
Spanish Council for Scientific Research, CSIC  
Campus UAB, 089193 Bellaterra, Catalonia, Spain  
maarten@iiaa.csic.es*

## Abstract

This paper presents a summary of the Music Performance Panel, held at the MOSART Workshop 2001, in Barcelona. The approaches of the represented research groups are described briefly, and an overview is given of the topics that were addressed.

## 1 Introduction

During the MOSART Workshop 2001, on current research directions in computer music, a discussion panel addressed some issues on the topic of music performance. The panel consisted of the following members: Gerhard Widmer (ÖFAI, Vienna), Henkjan Honing (NICI, Nijmegen), Johan Sundberg (KTH, Stockholm) and Giovanni de Poli (DEI, Padua). The main topics that were discussed are:

**Research Strategies** What are the relative strengths and weaknesses of different research strategies (theory-driven vs. data-driven, oriented towards cognitive plausibility vs. computational simplicity, perception-oriented vs. production-oriented, etc)? And could there be more synergy between these strategies?

**Functions of performance** Expressive music performance seems to fulfill several functions (e.g. expressing emotional content, but also clarifying structural aspects of a piece); how do these functions fit together? How do current models of performance account for these functions?

**Evaluation** Given that expression is a subjective notion and that there is no such thing as the “correct” interpretation of a piece of music, can we nevertheless develop quantitative and scientifically rigorous procedures and standards for evaluating the quality/significance/validity of proposed models of expression? Would it be worthwhile to try to col-

lect or construct ‘benchmark problems’ on which different models could be compared?

In the following sections, I will resume what was said with respect to the above questions during the panel. In section 2, an overview will also be given about the approaches that each of the research groups have been taking. This overview is partly based on the opening statements that each of the panel members made, and partly on the submitted position statement papers (included at the end of this paper).

## 2 Research strategies

As pointed out by Widmer et al. in [6], prior to the question of how research strategies relate to each other and how they can be compared, it should be clear what the *research aims* are. Three typical aims are respectively: 1) development of musical models that produce well-sounding musical results, 2) development musical models that produce results that maximally resemble (generalized patterns of) observed expert performances, and lastly, 3) development musical models that structurally resemble observed or hypothesized cognitive processes of musical performance. To capture the research of DEI in a category, I would like to propose a fourth category, in addition to these three categories. Namely: 4) development of performance models that provide optimal user control over the expressive renderings of performances. Depending on the aims of research, different methodological approaches may be appropriate.

**ÖFAI** ÖFAI's research is explicitly directed towards the second aim. More specifically, the aim is to discover regularities and patterns that can be found in the performance (particularly on piano) of musical pieces. This should result in models that as precisely and compactly as possible describe principles that emerge from performance data. Their strategy to achieve this aim can be described as analysis by machine induction. A large collection of musical data (expert performances of various complete Mozart piano sonatas) form the data for the inductive learning. These data are analyzed in terms of dynamics and tempo deviations. Structural representations (vz. transcriptions) of the performed music are also made. To this analysis of the data, the machine learning techniques are applied. The outcome of the learning process are the regularities that were found (co-occurrence of performance deviations and structural features of the music), typically in the form of prediction rules that predict expressive deviations based on the structural description of the music.

As the above suggests, rather than using the data to test preconceived hypotheses, the data are analyzed to generate hypotheses about music performances. In this respect, ÖFAI's research could be called bottom up, or data-driven. The knowledge that is gained in this way, is only about *what* is done during a performance, not *why* it is done; the intentions of the player are not accessible through this approach.

An advantage of ÖFAI's approach, being a data-driven approach, is that the results are independent of theories, hence independent of the need to verbalize and conceptualize in advance what will be analyzed. A data-driven approach may detect regularities in the data that were never noticed by human subjects.

**NICI** As opposed to the above approach, NICI tries to build *cognitive* models of music performance, rather than models that describe the performance itself. Thus, in terms of the three research aims mentioned above, the third aim is pursued by NICI. A major implication for methodology is that empirical data is obtained through controlled experiments (as is common in cognitive psychology), rather than by analyzing (large samples of) musical performances. In these experiments, the validity of constructs from musicological theories is tested. To do this, these theoretical constructs must first be formalized and implemented as algorithms. In this way, musicological theories are computationally modeled (cf. [3]). An important requirement for these models is that they are not primarily able to faultlessly reproduce human performances, but rather that parameters of the model are musically meaningful, i.e. correspond to musical concepts in the mind of the musician.

During the panel, Honing argued that controlled ex-

periments are preferred over the analysis of large corpora of music performances. In experiments it would be possible to discard unintended deviations, and to give the performer instructions and measure their effects on performances. This kind of interaction is not possible in the case of corpora-analysis.

Another point that was made, is that the focus is on the re-construction of a 'performance-space', rather than on studying only expert-performances. By also observing non-expert performances and performances with particular intentions, a more complete view on the range of meaningful deviations can be accomplished.

Furthermore, Honing mentioned the relevance of perceptual studies of expressive deviations. Through controlled listening experiments for example, it can be established how great deviations in the performance of a rhythm can become before it is perceived as a different rhythm. In this way it is possible to elucidate the constraints on expressive transformations.

Lastly, he argued that models of performance might very well benefit from a deeper understanding of the cognitive reality that accounts for the performances.

**KTH** The main research goal of KTH has been to gain a deeper understanding on music communication between musicians and listeners (possibly other musicians). Contrary to the idea that musical pieces can be performed in infinitely many equally well-sounding ways, the professional music performer Lars Frydén perceived that there are clear regularities in music performance. This gave rise to the idea of constructing a set of rules for musical performance (see [4] and [5]). To test the validity of the rules, an *analysis-by-synthesis* approach is adopted. This is to say that music performances are reconstructed from the score, by using the rules. The rules are applied one by one, each with an individual parameter that controls the magnitude of the effect of the rule, so the effect of each individual rule can be examined in detail.

In contrast to ÖFAI's approach, KTH's approach is primarily theory-driven. After one or more rules (which can be regarded as hypotheses) have been formulated, it's effect on performance is tested, by listening experiments where subjects rate the musical acceptability of performances that were generated by the rules. This way of evaluating is rather different from ÖFAI's evaluation method, where the performances generated by the system are not evaluated by listening, but by measuring their deviations from professional performances.

Sundberg mentioned some advantages of the analysis-by-synthesis method of testing the performance model. Firstly, the synthesis enables researchers to evaluate hypotheses under musically realistic conditions, namely by listening to performances that are generated under these hypotheses. Secondly, it is possible to test

hypothesized rules separately and tune them, one at a time. Thirdly, the situation of rule-tuning is similar to a teacher student setting, so that the person that tunes the rules (usually a musical expert) can rely on his/her pedagogical skills. Lastly, this approach is independent of training data and as such, it is apt to produce non-obvious interpretations of a piece, that nevertheless comply to the musical performance principles.

Some limitations of this method that Sundberg mentioned, were that the rules are a reflection of the expertise of just one individual. Also, the system will produce identical performances with identical rule palettes.

**DEI** Music performance research at DEI is primarily concerned with *control* of sound, in order to give music composers useful and usable tools for generating music from sound/instrument models. As De Poli noted, the problem of music performance is in between music generation and sound production. An important question here is how the sounds can be controlled in an expressive manner, on a slow varying time scale (e.g. on the level of musical phrases). In general, there are two strategies: one is control by gesture, the other is the use of models for controlling the sound production. The latter approach has been adopted by DEI.

Thus, performance research at DEI is aimed at building models that map expressive intentions (through the use of dichotomous labels like ‘hard’, ‘soft’, ‘bright’, ‘dark’), to low-level acoustic features of the performance. These models can be used to render nominal performances in expressive ways, with (real-time) high-level control over the expressive parameters.

To establish the relation between expressive labels and acoustic performance features, performing/listening experiments have been done. These experiments showed that listeners ordered music performances that were played with different expressive intentions, along two abstract dimensions: ‘kinematics’ and ‘energy’ [1]. A mapping was then established between coordinates in the kinematics-energy space and deviations of expressive parameters like tempo, legato and intensity. This mapping serves as a model for expressiveness, translating the points in the abstract control space to expressive deviations.

De Poli noted that an important aspect of models of performance is that they convey a multi-level abstraction from the score, that is, the highest level expressive concepts should not be directly mapped to the lowest level (acoustic) parameters, but via several hierarchically ordered abstraction levels, corresponding to different time scales.

Another important question is the generalizability of the expressive models. The models were constructed based on Western classical music, of which the practice

is relatively fixed. The practice of popular Western music, on the contrary, is less codified. This may imply that accurate expressive models are harder to build for this kind of music.

Future plans of DEI are to work in the reverse way; that is, instead of constructing performances based on expressive intentions, rather analyzing performances in order to derive the musicians expressive intentions from it.

### 3 Functions of performance

The question of how different approaches account for the functions of performance did not receive much direct attention during the panel. Nevertheless, some statements can be made about it, as to some extent, the stance toward the function of performance is inherent to a particular approach.

Firstly, in accord with Widmer’s remark that they study *what*, rather than *why*, it can be observed that ÖFAI’s research comprises only structural/syntactical analysis of performances. Performance elements are not related to anything external to the musical piece (like the performer’s intentions). For that reason, the functions of performance that can be investigated are bound to be about the performance itself; not e.g. communicative functions (like expressing emotions). Indeed Widmer noted that the function of performance they study, is performance as clarifying musical structure.

NICI’s research is not explicitly directed to accounting for the functions of performance. However, in studying expressive timing, the structural role of performance elements is identified as one of the factors that influence timing (see [7]).

KTH’s rule based approach also incorporates the function of performance as clarifying structure: this is evident by the categorizations within the rule base, where one category of rules is called ‘grouping rules’. These rules are intended to elucidate the boundaries between different structural units, like phrases.

On the other hand, the function of performance as communicating emotional content can also be modeled, as Sundberg noted, by the existence of magnitude parameters for the effect of each rule. Their hypothesis is that particular kinds of interpretations of a piece (‘sad’ or ‘happy’), correspond to particular settings of these parameters. This correspondence is nicely present in the metaphor of ‘rule palettes’, that Sundberg used, suggesting the possibility of ‘painting’ with the magnitude parameters.

Inherent to DEI’s approach to musical performance, is the function of communicating emotional content. As their aim is to render or transform music performances according to expressive labels like ‘dark’, ‘bright’, ‘light’

or ‘heavy’, it is clear that the focus is on performance as expressing intentions. As argued in [2], sensorial adjectives were preferred over emotional ones like ‘sad’/‘happy’, in order to limit the semantics under examination. The sensorial adjectives clearly have a more restricted meaning and related to performance more closely.

## 4 Evaluation

Finally, there was the question about the evaluation of performance models and the use and usefulness of benchmarks in the area of music performance research. There were quite diverse opinions about this among the panel members.

About the evaluation of their research, Widmer said that the focus was on two prime concepts: predictive accuracy and generality. Performance models should on the one hand predict the performance deviations of (expert) performers as accurately as possible, while on the other hand, the predictions should ideally be valid across different performers and musical styles. About the possibility of using benchmarks for evaluation of performance models, Widmer remarked that the use of benchmarks suggests that there is a set of pieces for which there is a ‘correct’ performance, which must be matched as close as possible by any good model. This is obviously not the case. Furthermore, using a standardized dataset as benchmark, introduces the risk of over-fitting the models to the benchmark data, as has apparently been the case in the area of machine-learning. This over-fitting should be avoided. Given these risks, Widmer supposes that, with much awareness, it could still be useful to propose a standardized test dataset. This dataset should at least be very diverse, different musical styles and performers should be represented.

At this point, I would like to note that the usefulness and justifiability of benchmarks for evaluation, is somewhat dependent on the goal of research. In ÖFAI’s case, where the goal is to match human expert performances as closely as possible, it makes more sense to use benchmarks, because the objective is purely quantitative: the deviation between test-data and predictions of the model should decrease to zero. If the goal is to produce musically acceptable results (as with the research of KTH), the use of benchmarks is less obvious, because the most important thing is that the performances resulting from the model should sound musically convincing in themselves, not that they are *similar* to musically convincing performances. When the aim is a truth-like cognitive model of music performance, a good use of benchmarks is neither very clear, because good cognitive models do not necessarily predict a particular set of actual

data very accurately. Rather they predict the constraints that hold for music performance in general.

A related remark, made by Honing, is that in the case of human performances of music, not all the information is contained in the data, but that there is a lot of information which is only *suggested* by the data, but actually is in the minds of the listeners (e.g. tempo and evoked emotion are not measurable in the data themselves). This perceptual information is an important aspect of performances, which is not covered by a straight-forward use of benchmarks to evaluate models. Hence, benchmarks are only partly relevant as an evaluation tool.

There was a reply to this from the audience (by Jan Tro), that although the information may not all be conveyed by the data, at least the ‘triggers’ for this external information, are embodied in the data. Although this is obviously true, I would like to add that it does not take away the need for perceptual research to music as well, in order to establish to what kind of perceptual phenomena these triggers map.

## 5 Other remarks made during the panel session

Sundberg raised the point that how often a performance principle applies, may not touch the essence of such a principle (a rarely used performance rule may nevertheless be musically important). A more relevant question would be what the meaning of the deviation is, that is, what is expressed by it?

Widmer answered that the two kinds of research on music performance (looking for musical meaning of expressive deviations on the one hand and looking for communal expressive patterns in performances on the other), might very well co-exist at the same time. They should be regarded as complementary, where the regularities found by the latter approach could form a useful point of departure for the former. He furthermore noted that it should be made explicit that it is a *hypothesis* that the central function of expression is to communicate meaning to the listener.

A remark from Honing was that there is no such thing as an ‘average performance’. Averaging over several performances of the same piece (let alone different pieces), will not result in a ‘typical’ performance, and will probably not convey much useful information. Sundberg agreed that using averages in a quantitative way, will tend to diminish the magnitude of the measured effects of performance principles.

A critical remark from the audience (by Werner Goebel) was that it should be realized that performances generated under a controlled experiment cannot be taken to be exchangeable with other performances, like live

performances or studio performances. Attempts to manipulate the performance by instructions may yield performances that are not representative, because it affects the performer in unnatural ways. Honing replied that through clever design of experiments it may be possible to manipulate the performers in unconscious ways. Sundberg noted that it could be interesting to study how performances in different settings differ.

A question from the audience was about the issue of instrument fingering. In what way does fingering affect the performance? Widmer suggested that fingering does not so much affect performance as fingering is chosen to achieve the expressive affect that is intended by the performer. Honing added that in addition to the effect of musical structure and emotion on expression, there is the effect of the instrument on expression. Typically musicians emphasize parts of a piece that are difficult to play on a particular instrument by a deliberate fingering. In general, Honing agrees with Widmer that fingering is chosen to maximize expressive control.

A final appeasing point made from the audience (by Roger Dannenberg) with respect to the problem of evaluation and collaboration of different performance models was that criticism and skepticism about the right way to proceed and combine research may be an obstacle for progression. It may be fruitful to share results and data, even with the limitations that hold.

## References

- [1] S. Canazza, G. de Poli, A. Rodà, A. Vidolin, and P. Zanon. Kinematics-energy space for expressive interaction in music performance. In *Proceedings of the Workshop on current research directions in computer music*, pages 35–40, Barcelona, november 2001.
- [2] S. Canazza, G. Poli, and A. Vidolin. Perceptual analysis of the musical expressive intention in a clarinet performance. In M.Leman, editor, *Music, Gestalt and Computing*, pages 441–450. Springer Verlag, Berlin, 1997.
- [3] P. Desain, H. Honing, and R. Timmers. Music performance panel: NICI/MMM position statement. MOSART Workshop on current research directions in computer music, november 2001.
- [4] A. F. Johan Sundberg and R. Bresin. Music performance panel: Position statement kth group. MOSART Workshop on current research directions in computer music, november 2001.
- [5] J. Sundberg, A. Friberg, and L. Frydén. Common secrets of musicians and listeners: An analysis-by-synthesis study of musical performance, 1991.
- [6] G. Widmer, S. Dixon, W. Goebel, E. Stamatatos, and A. Tobudic. Empirical music performance research: ÖFAI's position. MOSART Workshop on current research directions in computer music, november 2001.
- [7] W. Windsor, P. Desain, H. Honing, R. Aarts, H. Heijink, and R. Timmers. On time: the influence of tempo, structure and style on the timing of grace notes in skilled musical performance. In *Rhythm perception and production*, pages 217–223. Swets & Zeitlinger, Lisse, NL, 2000.

# Empirical Music Performance Research: ÖFAI’s Position

Gerhard Widmer, Simon Dixon, Werner Goebel, Efstathios Stamatatos, Asmir Tobudic

Austrian Research Institute for Artificial Intelligence (ÖFAI)

Schottengasse 3, A-1010 Vienna, Austria

{gerhard|simon|werner|stathis|asmir}@ai.univie.ac.at

## Abstract

This short paper presents our view on some general questions regarding empirical research on expressive music performance. The main direction of performance research going on at the Austrian Research Institute for Artificial Intelligence (ÖFAI) is briefly reviewed and positioned relative to three general issues, namely, different research strategies, different dimensions of performance, and the question of empirical evaluation of performance models.

## 1 Introduction

The *Music Performance Panel* held at the MOSART 2001 workshop is dedicated to three principal questions that try to put current research on expressive music performance into perspective: what are different *research strategies*, and what are their respective roles? what are different *functions* or *dimensions* of performance, and how are these accounted for by different research approaches? and how should formal, computational models of performance be *evaluated*?

We believe that when, and indeed before, trying to answer these questions it is crucial to define for oneself what the *goal* and *purpose* of one’s research is: (a) do we aim at computational models of performance that produce well-sounding musical results and thus are useful to the music software industry? or (b) do we aim at models that as much as possible fit the patterns and regularities observed in expert performance and can make predictions regarding aspects of expert performances? or (c) do we want models that, through their very structure and conceptual design, reflect an observed or hypothesized cognitive reality?

These are quite distinct goals. For instance, in the first case (a), we will probably not care about whether the model itself is cognitively adequate, or we will care about that only to the extent that a model expressed in more “intuitive” terms is also easier to use and control (cf. Desain et al.’s point on FM synthesis vs. physical modeling [4]). Also, the different goals will necessitate different strategies for evaluating the usefulness (a) or precision and generality (b) or plausibility (c) of proposed models.

Different research groups (some of which are represented in the MOSART consortium) capitalize on different goals, and thus both their approaches, theoretical and technical, and the way they present and evaluate their results, are different. In the rest of this paper, we will focus on our own research as it relates to musical performance, and will try to position it relative to the above issues.

## 2 Inducing Models from Large Collections of Expert Performances

### 2.1 Research Goals

ÖFAI’s immediate research goals focus on the second of the above three alternatives: we want to find *descriptive* and *predictive (partial) models* of certain aspects of expressive performance. These models should “explain” (i.e., fit) as much as possible of the observed phenomena, and they should be predictive in the sense that they generalize to other performers and possibly other types of music. The starting point for these investigations are large collections of “real-world” performances (in particular, performances by concert pianists made not specifically for research purposes).

To this end, we develop and use *Artificial Intelligence* and, more specifically, *Inductive Machine Learning* techniques to find computational models of typical performance strategies [10]. We take a strictly data-driven approach: expert performances are collected, quantitative details concerning expressive performance (timing, dynamics, articulation) are measured, and the resulting data are analyzed with the help of machine learning algorithms that try to find common patterns and regularities in these data. In this way, the computer is used as a tool or assistant in the process of inductive model building.

Cognitive adequacy of the resulting models is not an immediate goal; that would probably require a different kind of approach (and it would require expertise in cognitive psychology that we do not have). The main point of our research is to discover potentially new, general patterns that have hitherto been neglected in performance research. These may then be studied in more focused and controlled experiments.

## 2.2 Research Strategy

We see our approach as complementary to the research strategies followed by other performance researchers, be they based on systematic controlled experimentation (e.g., [13]), on ‘analysis-by-synthesis’ [8] or on purely statistical methods (e.g., [6]).

What distinguishes our work from most of the other work in empirical performance research is the use of computational learning and knowledge discovery methods and, connected with that, the strictly data-oriented approach. We use algorithms that can search for and discover complex dependencies and regularities in extremely large data sets, and can describe their discoveries to the user in intelligible terms [12].

A distinct advantage of such an approach is that the computer is free of any musical preconceptions and expectations and thus may more easily come up with novel and possibly surprising hypotheses [11]. These hypotheses may not necessarily always relate to a conceptual framework that musicians or musicologists find musically intuitive or cognitively plausible. In other words, they may not be directly interpretable as a model that reflects the musical reality of a performer. But the discovered patterns may point to interesting phenomena that have not been looked at so far and that can then be studied in more focussed and controlled experiments. In our view, that is the main role of this machine induction approach.

## 2.3 Aspects of Performance Studied

Starting from given collections of expert performances also has consequences on the types of things we can and cannot study. To put it simply, what can be hypothesized from given performances is *what* the performer did and what s/he is likely to do in other pieces, but not (or not directly, at least) *why* s/he did it (the performer’s musical or communicational intentions) or what effect the observed performance strategies have on the *listener* (the perception of performed music). The latter questions would require controlled experiments with performers and/or listeners, where performers are asked to play pieces under different conditions (as, e.g., in [9]) or with different kinds of ‘target emotion’ [3]. If we only take given performances, we cannot, for instance, make any quantifiable statements about emotional aspects, either in terms of production or perception. What we can hope to discover from large collections of precisely measured expert performances is general expressive patterns that seem to be common across a wide range of pieces and different performers and thus seem to indicate general performance strategies [11]. The same kind of material can also be used to study systematic *differences* between performers, again with inductive methods [7].

To widen the range of questions we can answer, and to clarify some very basic, but elusive notions (e.g., what really is “tempo?”), we have recently also started to perform controlled experiments with human subjects, both listeners and performers (e.g., [2, 5]). Here we can study certain specialized questions (e.g., the phenomenon of *melody*

*lead*) in more detail, but with a narrower data basis (because producing controlled experimental data with human subjects is expensive).

## 2.4 Evaluation Issues

In empirical research, testing inductively obtained hypotheses on independent data is essential. In order to make it possible to compare competing models and algorithms, they have to be tested on a common set of data of an appropriate level of complexity.

In the area of machine learning, for instance, this has led to the establishment of a database of common benchmark data sets on which new algorithms must be tested so that their results can be compared to the results of other methods. The database is maintained by a group at the University of California at Irvine [1] and is continually updated with new data sets contributed by members of the scientific community.

In the area of music performance research, establishing such a database of common test data would be an interesting (and laborious) task. Whether it would be worthwhile would depend on a consensus, within the research community, on a set of basic evaluation criteria. To prevent a possible misunderstanding, let us make clear that we do *not* mean that such a set of test performances would in any sense represent *the* “correct” interpretations, in the sense of an absolute benchmark. On the contrary, it should contain performances of the same pieces by different performers and possibly under different conditions.

What we consider crucial is that the *generality* of the models should be established experimentally, and that requires testing them on large sets of diverse musical situations. Using only a few hand-selected pieces for model building or testing always comes with the danger of *overfitting* (either by fitting the model too tightly to the data, or by (consciously or unconsciously) selecting the test data in such a way that they confirm the model).

Of course, working with large sets of training and test pieces makes it difficult to attend to all the details and possible artifacts that may be hidden in the data, and to have a fine control on all experimental conditions. On the other hand, the kinds of patterns we find with our data-driven approach have a certain empirical weight and generality simply by virtue of the fact that they are based on (and their predictive potential has been tested on) a large set of diverse musical pieces. We do believe that the size, complexity, and musical diversity of experimental test data can give a new kind of quality and validity to experimental results. To put it (overly) simply, in our current machine learning experiments, we sacrifice observation precision for stronger or broader empirical support.

## 3 Conclusions

It seems clear that no research approach alone will lead to complete models of expressive performance that do justice to the complexity of the phenomenon and that are

adequate from every possible point of view. More cooperation between different approaches will be needed (for instance, discovering novel types of patterns with our approach and then investigating these further in more detailed and controlled experiments). That requires first and foremost the definition of a common set of problems and evaluation criteria. This panel has at least made explicit some of the differences between current approaches, but has also revealed a lot of common ground that we can build on in future work.

## Acknowledgements

ÖFAI's research on expressive music performance is supported by a very generous START Research Prize (project no. Y99-INF) by the Austrian Federal Government, administered by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung (FWF)*, and by the EU project HPRN-CT-2000-00115 MOSART. The Austrian Research Institute for Artificial Intelligence acknowledges basic financial support by the Austrian Federal Ministry for Education, Science, and Culture.

## References

- [1] Blake, C. and Merz, C. (1998). UCI Repository of Machine Learning Databases. <http://www.ics.uci.edu/~mllearn/MLRepository.html>. Department of Information and Computer Science, University of California at Irvine, Irvine, CA.
- [2] Cambouropoulos, E., Dixon, S., Goebel, W., and Widmer, G. (2001). Human Preferences for Tempo Smoothness. In *Proceedings of the VII International Symposium on Systematic and Comparative Musicology, III International Conference on Cognitive Musicology*, Jyväskylä, Finland.
- [3] Canazza, S., De Poli, G., and Vidolin, A. (1997). Perceptual Analysis of the Musical Expressive Intention in a Clarinet Performance. In M. Leman (ed.), *Music, Gestalt, and Computing*. Berlin: Springer Verlag.
- [4] Desain, P., Honing, H., and Timmers, R. (2001). Music Performance Panel: Position Statement. *MOSART Workshop on Current Research Directions in Computer Music*, Nov. 2001, Barcelona.
- [5] Goebel, W. (2001). Melody Lead in Piano Performance: Expressive Device or Artifact? *Journal of the Acoustical Society of America* 110(1), 563-572.
- [6] Repp, B. (1992). Diversity and Commonality in Music Performance: An Analysis of Timing Microstructure in Schumann's 'Träumerei'. *Journal of the Acoustical Society of America* 92(5), 2546-2568.
- [7] Stamatatos, E. (2001). A Computational Model for Discriminating Music Performers. In *Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*, Nov. 2001, Barcelona.
- [8] Sundberg, J., Friberg, A., and Frydén, L. (1991). Common Secrets of Musicians and Listeners: An Analysis-by-Synthesis Study of Musical Performance. In P. Howell, R. West & I. Cross (eds.), *Representing Musical Structure*. London: Academic Press.
- [9] Timmers, R., Ashley, R., Desain, P., and Heijink, H. (2000). The Influence of Musical Context on Tempo Rubato. *Journal of New Music Research* 131-158.
- [10] Widmer, G. (2001). Using AI and Machine Learning to Study Expressive Music Performance: Project Survey and First Report. *AI Communications* 14 (in press).
- [11] Widmer, G. (2001). Inductive Learning of General and Robust Local Expression Principles. In *Proceedings of the International Computer Music Conference (ICMC'2001)*, La Habana, Cuba.
- [12] Widmer, G. (2001). *A Meta-learning Method for Discovering Extremely Simple Partial Rule Models*. Submitted. Available as Technical Report OEF AI-TR-2001-30, Austrian Research Institute for Artificial Intelligence, Vienna.
- [13] Windsor, L., Desain, P., Honing, H., Aarts, R., Heijink, H., and Timmers, R. (2000). On Time: The Influence of Tempo, Structure and Style on the Timing of Grace Notes in Skilled Musical Performance. In Desain, P. and Windsor, W. L. (Eds.), *Rhythm Perception and Production*. Lisse: Swets & Zeitlinger.



# Music Performance Panel: Position Statement

*KTH* Group

Johan Sundberg, Anders Friberg, Roberto Bresin

## Research Strategy

**The goal of our research on music performance is to gain a deeper understanding of music communication. Our research was initiated in the 1970s. Those days the general belief was that any piece of music could be performed in a number of widely differing and yet musically acceptable ways. Therefore, it was argued, there is no chance that a decent performance can be generated by rules. On the contrary, music performances are unique, and this is what makes them musically interesting and attractive. In this situation, a statistical analysis of, e.g., tone durations in a set of performance of a given piece did not seem promising.**

At the same time the idea that performances are completely independent of rules did not agree with the vast musical experience of the late musician Lars Frydén. During his violin playing in string quartet and in orchestras, he had found a number of regularities that he wanted to test.

In this situation it seemed advantageous to choose an analysis-by-synthesis strategy, i.e., to use the score transformed into a music file as the input for a rule system that converts it into a sounding performance. This allowed the testing of the performance rules that Frydén wanted to test. Each rule could be tested on a number of music examples, thus allowing him to listen to what extent the performance was improved by the rule and to find out if the effect was of an appropriate magnitude.

This analysis-by-synthesis strategy has certain unique advantages. Its main strength is the synthesis that allows the researcher to test hypotheses by listening to performances under musically reasonably realistic conditions. The method further allows a good control of the performance in the sense that one rule can be tuned and tested at a time. Also, it provides examples that are judged in a setting somewhat similar to the student-teacher setting, so that the musician can rely on his pedagogical expertise. The method also has the advantage of allowing a systematic build up of rules, in each state giving priority to salient effects. Hence, rules tend to be developed in the order of importance. The method also has some advantage over methods where statistical data on performance drive the research process. The analysis-by-synthesis strategy is independent of such data. It may generate quite unusual interpretations of a piece that still are musically acceptable and/or interesting.

The strategy also has limitations. The basic idea is that performance is determined by regularities. This implies that the machine will generate exactly the same performance effect each time its context conditions are met. In reality, however, musicians

may play a given piece quite differently. In particular, the same sequence of tones may be played differently the second time it appears in a piece. As the magnitude of the effect of each rule is controlled by a quantity parameter, the rule system can indeed generate differing performances of a piece, e.g., eliminating or exaggerating various rules.

Within our group we have an ongoing discussion regarding the perceptual relevance of random variation, but as yet, we have failed to reach a common view. Measurable random variation obviously occur, the crucial question being to what extent it is subliminal and whether or not it contributes to the esthetical quality of a performance.

The assumption that performances are controlled by regularities may prove to be unrealistic in the future. On the other hand, it seems wise to test the simplest models first and to abandon them only when their limitations have been clearly exposed.

Another limitation of our rule system is that basically it is a formalised description of the musical competence of one professional musician only. We have found this a minor concern, as our musician is generally acknowledged as an outstanding expert. Therefore, his competence must be, by and large, representative.

Research using the analysis-by-synthesis strategy is driven by data rather than by theory. Indeed, our results have sometimes driven theory. An example is the concept of melodic charge. Here, the playing of music examples demonstrated the need for variations reflecting the relation between the tone and the underlying harmony. We tried a number of different existing alternatives as control parameters for the dynamic variations, all with inappropriate result. Eventually we arrived at the relationship along an asymmetric version of the circle of fifths, that we called melodic charge. Thus, the playing of melodic lines void of characteristics that reflected this melodic charge seemed to lack an important aspect of an ideal rendering. The fact that the introduction of the melodic charge into the performance grammar improved the musical acceptability of the performance seems to imply that this novel concept is relevant to music perception.

Perturbation of tone duration is an important channel for musical expression. A remaining question is when such perturbation should be controlled by proportion or in terms of absolute duration. In some rules, such as the *inegalle*, we use proportions, while other rules work with absolute duration. Both alternatives seem relevant to music listening. For example, a tone appears to lose its autonomy and sound like a grace note as soon as its duration is shorter than about 100 ms.

An important task in our research is to sort out the roles of rules that operate at same level. For example, the phrasing rule should not operate on tone sequences treated by the final *ritard* rule. There are also other as yet not quite resolved interference between certain rules, such as the punctuation and leap articulation rules.

**Functions of performance** Our formulation of performance rules have yielded a generative grammar of music performance that has invited us to speculation regarding to function of performance, or, more precisely, regarding the function of the expressive deviations. Thus, we have seen that the rules can be divided into three major categories depending on their apparent function in music communication. One category seems to serve the purpose of differentiating tones belonging to different tone categories, i.e., to enhance the differences between pitch and interval classes and between note values. Another category seems to mark which tones belong together and where the structural boundaries are. In this way, the performer facilitates the listener's processing of the signal flow. Interestingly, the same two principles, differentiation and grouping, can be observed also in spoken communication. As yet, we have not tested these cognitive aspects of music performance, though. The third group concerns technical aspects of ensemble playing related to synchronisation of voices and tuning.

Music performances are also coloured emotionally. We have found that emotional colouring can be achieved by varying the rules' quantity parameters. Thus, by enhancing some rules and suppressing others, emotionally differing performances of the same piece can be generated. We have already constructed a set of palettes that add different emotional colours to performances (angry, sad, happy, scared, tender, and solemn) and we plan to build special rule palettes that will generate agitated and peaceful performances.

**Evaluation** Synthesised performances appear to represent a powerful tool for evaluating the perceptual relevance of research findings. It seems advantageous, however, to use expert listeners. We have had good experiences of listening tests where musicians were asked to adjust the quantity parameter to an optimum for different music examples. In these experiments, rules have been tested one by one. As zero is thereby an available choice the results show if the rule tested provides a desirable effect.

In case performance research relies on statistical data from real performances, the evaluation may be more problematic. It appears that synthesis will greatly facilitate the verification of such results.

**Future work/Remaining problems** The score we now use as input for the performance grammar is rudimentary in the sense that it contains information on nothing but pitch and duration. Thus phrase markers and chord symbols are introduced manually. Also, the realisation of conventional items like trill, point, and grace notes requires hand editing of the score. Our plan is to complement the input score with signs for such events. We also plan to implement Craig Sapp's algorithm for automated chord analysis.

Another planned improvement is to test the usefulness of a realtime control of rule quantity. This will be realised within the MEGA project; hopefully, this may solve the

problem that the grammar performs the same music material exactly the same way if it reappears in a piece.

We have lately been cooperating with Max Mathews and Gerald Bennett implementing the performance grammar in the Radio Baton system. This has been an informative experience, elucidating the boundaries between the musician's and the conductor's responsibilities in shaping a performance.

Basically the research method seems unproblematic. We do not regard the analysis-by-synthesis strategy as the only possible method. An exchange of data assembled by various methods will improve quality of research and promote progress. The Vienna material represents an extremely valuable resource for the further development of the performance grammar. A crucial condition, however, would be the use of synthesis, apparently representing an indispensable opportunity to test the perceptual relevance of findings. The MOSART project comprises exchange of research results and computer synthesis of music performance as two of its core aims and thus offers a perfect opportunity to proceed along these lines.

## Music Performance Panel: NICI / MMM Position Statement

Peter Desain, Henkjan Honing and Renee Timmers

*Music, Mind, Machine* Group

NICI, University of Nijmegen

mmm@nici.kun.nl, [www.nici.kun.nl/mmm](http://www.nici.kun.nl/mmm)

In this paper we will put forward our view on the computational modeling of music cognition with respect to the issues addressed in the *Music Performance Panel* held during the MOSART 2001 workshop. We will focus on issues that can be considered crucial in the development of our understanding of human performance and perception in its application to computer music systems. Furthermore, they were chosen such as to complement the issues brought forward by the other contributing institutes (i.e. OFAI/Vienna, KTH/Stockholm, and DEI/Padua). In summary these are:

- A computational model in agreement with music performance data is *starting point* of research, rather than an *end product* (cognitive modeling is preferred over a descriptive model)
- Importance of empirical data obtained in controlled experiments (rather than using individual examples of music performances)
- Preference for the concept of *performance space* (over the use of large corpora of music performances)
- Study performance through perception, focusing on the constraints of expression rather than studying the ideal or “correct” performance (as such avoiding the issue of performance style, and enabling the study of important aspects that are not directly measurable in the performance data itself, e.g., those of a perceptual and/or cognitive nature)

## **Research aims**

The panel addresses a number of dichotomies in the study of music performance, such as theory-driven vs. data-driven, oriented towards cognitive plausibility vs. computational simplicity, perception-oriented vs. production-oriented. The discussion aims to reveal research aims and methods, which are quite varied among research groups.

In our group, we study music perception and performance using an interdisciplinary approach that builds on musicology, psychology and computer science (hence the name *Music, Mind, Machine*). The aim is to better understand music cognition as a whole. The method is to start with hypotheses from music theory, to formalize them in the form of an algorithm, to validate the predictions with experiments, and, often, to adapt the model (and theory) accordingly. In other words, in the method of *computational modeling*, theories are first formalized in such a way that they can be implemented as computer programs. As a result of this process, more insight is gained into the nature of the theory, and theoretical predictions are, in principle, much easier to develop and assess. With regard to computational modeling of musical knowledge, the theoretical constructs and operations used by musicologists are subjected to such formalization. Conversely, with computational modeling of music cognition, the aim is to describe the mental processes that take place when perceiving or producing music, which does not necessarily lead to the same kind of models. As such, for us, a computational model that mimics human behavior is not enough. It in fact is more a starting point of analysis and research, than an end product (see [1] for an elaborate description).

## **Evaluation and validation of music performance models**

One of the key issues in developing algorithms and computational models is their validation on empirical data. In the case of the MOSART project, music that is artificially generated should respect human perception and performance such as to assure seamless interaction and intelligible control by its users. For evaluating and validating models of expression, it is problematic to search for a “correct”, general or

benchmark interpretation of music [2], to which the models can be compared. Though this approach is quite common in AI modeling, it is very unattractive for music cognition research. Not only is the notion of an ideal performance questionable, comparing the input-output relation between the model and the musical performance is also too limited an evaluation. A data-driven perspective might eventually result in an accurate description [2,3], it will, however, not be a model, in the cognitive sense. It needs to describe more than just an input-output transformation. In fact, a good model is a model for which changes in parameter settings that relate to manipulated aspects of the performance (e.g. by instruction to the performer) remains to show agreement between model and performance. As such step by step further validating the model.

As an illustration of the difference between a model and a good description from another domain, one can take difference between FM-synthesis and physical modeling. It is possible to generate very convincing sounds with FM synthesis (after careful selection of the parameters). However, the whole space of sounds is unintuitive and difficult to control. In contrast, physical models have more similarity with the human world and succeed in replicating the behavior of existing objects (e.g., made of tubes and strings) that are known to the user and are therefore easier to control, despite their more restricted expressive power.

In general, a computational model that captures important aspects of human perception and action will be more successful in computer music systems. Models that simply aim at an input-output agreement do not necessarily give us a better understanding of the underlying perceptual or cognitive processes, which is essential for the development of convincing and intuitive models for human interaction with machines (see [4] for a discussion on the psychological validation of models of music cognition). A solely data-driven approach ignores the fact that important aspects of music performance are not directly measurable or present in the data itself. For instance, tempo (or expressive rubato for that matter) is a percept, and cannot be directly measured. The same applies for syncopation and other temporal aspects of music that exist due to (violations of) listener's expectations.

With regard to the methodology of evaluating models of expression, we assign great importance to the systematic collection of empirical data, experimentally

manipulating the relevant parameters. For instance, in our research on expressive vibrato [5, 6], we explicitly control for global tempo to reveal how it is adapted to the duration of notes. And we record repeated performance to get a better grip on consistency (e.g. to be able to separate between intended and non-intended expressive information).

Similarly, in our studies on piano performances (e.g., [7]), only careful experimental manipulation of a few parameters (like global tempo, or the addition or removal of one note) will give a precise insight in the underlying mechanisms that we need to reveal in order to make better computer music editing software or music generation systems.

Blindly examining very large samples of music performance is clearly not an alternative to this.

And, finally, in our work in rhythm perception, we put quite some effort in developing methods that allow us to investigate the concept of *performance space*, abstracting from individual examples. The idea here is to consider all possible interpretations, including musical and unmusical ones, in a variety of styles. While currently we only applied this approach to relatively short fragments of music [8], we find this method a more systematic and insightful alternative for randomly grown corpora of music performances. In addition, studying the perception of rhythm is also a way to identify the constraints on expressive timing in music performance (instead of focusing on an ideal or unique performance) as such avoiding the notion of a “correct” performance, which is an important advantage that allows for models to be elaborated independent of performance style.

## References

- [1] Documents on <http://www.nici.kun.nl/mmm> under heading “Research”.
- [2] Sundberg, J., Friberg, A., and Frydén, L. (1991) Common Secrets of Musicians and Listeners: An Analysis-by-Synthesis Study of Musical Performance. In P. Howell, R. West & I. Cross (eds.). *Representing Musical Structure*. London: Academic Press.
- [3] Widmer, G. (2001) Using AI and Machine Learning to Study Expressive Music Performance: Project Survey and First Report. *AI Communications*, 14.



- [4] Desain, P., Honing, H., Van Thienen, H. & Windsor, L.W. (1998). Computational Modeling of Music Cognition: Problem or Solution? *Music Perception*, 16 (1), 151-16.
- [5] Desain, P. & Honing, H. (1996) Modeling Continuous Aspects of Music Performance: Vibrato and Portamento [ICMPC Keynote address], *Proceedings of the International Music Perception and Cognition Conference*. CD-ROM, Montreal: McGill University.
- [6] Rossignol, S., Desain, P. & Honing, H. (2001). State-of-the-art in fundamental frequency tracking. *Proceedings of the Workshop on Current Research Directions in Computer Music*. Barcelona: UPF.
- [7] Timmers, R., Ashley, R., Desain, P., Honing, H., and Windsor, L. (in press) Timing of ornaments in the theme of Beethoven's Paisiello Variations: Empirical Data and a Model. *Music Perception*.
- [8] Desain, P. & Honing, H. (submitted). *The Perception of Time: The Formation of Rhythmic Categories and Metric Priming*. See <http://www.nici.kun.nl/mmm/time.html>



## WORKSHOP ON CURRENT RESEARCH DIRECTIONS IN COMPUTER MUSIC

Barcelona, Nov 15-16-17, 2001  
Audiovisual Institute, Pompeu Fabra University

### MUSIC GENERATION PANEL (A critical review)

Rubén Hinojosa Chapel <[ruben.hinojosa@tecn.upf.es](mailto:ruben.hinojosa@tecn.upf.es)>  
Doctoral student in Computer Science and Digital Communication  
Pompeu Fabra University, Barcelona, Spain



## MUSIC GENERATION PANEL (A critical review)

In the frame of the [Workshop on Current Research Directions in Computer Music](#), a **Music Generation Panel** took place. The chair was **Henkjan de Honing** (NICI-University of Nijmegen, The Netherlands), who introduced and conducted the panel, whose members were (in order of appearance): **Barry Eaglestone** (University of Sheffield, United Kingdom), **Roger B. Dannenberg** (Carnegie Mellon University, Pittsburgh, USA), **Eduard Resina** (IUA-Pompeu Fabra University in Barcelona, Spain) and **Jens Arnsprang** (DIKU-University of Copenhagen, Denmark). Each of the panel members made a short intervention and, after Arnsprang's words, some questions came from the audience. The panellist's answers were focused on what they have talked about.

To make available the main topics discussed in the Music Generation Panel, to the community of computer science researchers, composers, musicians, students and everybody who is interested in Computer Aided Composition, we have made a summarize and have added a critical review at the end. As an introduction to the purposes of this panel, here is the

### Call to the Music Generation Panel

The abundance of music generation tools and systems is well documented. These range from AI-based systems for autonomous generation of musical ideas to conventional design tools, for example, for designing and rendering of sounds. However, emerging de facto standards have been short lived, generating frustration rather than satisfaction. This panel will focus on why this is so, i.e., the extent to which accumulated results of this effort fail to satisfy the aspirations of composers. Three specific aspects of music generation will be considered. These are:

- 1 - Representation and contents of the product, i.e., the composition;
- 2 - The nature of and support for the process, i.e., creativity and composition; roles of artificial intelligence;
- 3 - Representation and application of individual and community know-how, including the use of repositories and archives to accumulate a history of compositional techniques used, and the use of the Web to provide open access to community knowledge.

Each aspect will be considered from philosophical, conceptual and technological perspectives. The aim is to identify open questions and unsatisfied requirements that technology has the potential to address. Many of these are partly evident in ongoing research in this area. The outcome will form the basis of a proposed scientific agenda for future composition systems research.

## MUSIC GENERATION PANEL (A critical review)

### THE PANEL

A brief overview of the words by each of the panel members follows:

**-Dr. Barry Eaglestone** (University of Sheffield, United Kingdom)

*(Note: Dr. Eaglestone made his speech based on his paper "[Composition Systems Requirements for Creativity: What Research methodology?](#)", so we have kept some paragraphs from that paper, and some words from his speech.)*

Electroacoustic music composition tools and systems selectively attempt to provide composers with services they require for music generation, e.g., for accessing, generating, organising and manipulating audio (and other) objects, which constitute the composition. However, a primary aim of composition software also is to create conditions in which composers can be creative in the use of these services. We believe there to be a need to establish a research base for enhancement of support within composition software for creativity.

Research into digital signal processing and the artist's use of sounds is on-going, and consequently, services relating to musical artifacts are volatile and evolving as new techniques and paradigms are integrated into composition software.

The software environment within which those services are used creatively has largely been under researched. Instead, developments have followed those of software technology. Consequently, there has been a move from asynchronous to synchronous systems, and from text to graphical user interfaces.

We believe there to be an inherent tension between principles of conventional software engineering and the requirements of creative composers. This tension can be explained in terms of models of creativity, which is often characterized by the notion of "divergent" as opposed to "convergent" thinking; the latter being associated with relatively predictable logical activity and outcomes, the former with less logical and predictable activity and outcomes.

One of the most talked about and most researched area is the one which involves services for creating and manipulating musical artifacts, and there is a lot to say about that. However, we will be looking at the largely neglected area, which is the environment within which those services can be used creatively?

In our research towards this end, we are analysing data collected by observing composers at work in naturalistic settings, using methodologies ranging from software engineering through to the social sciences. What comes out is the tension between the composer's requirements and the conventional wisdom of software engineering.

Specifically, there appears to be a need for an un-typed workspace within which composition artefacts can be freely associated; support which enables composers to control the whole process, employing programming skills at the lowest DSP levels; support for interfaces which challenge the composers' conceptions, rather than reflect

## MUSIC GENERATION PANEL (A critical review)

them; facilities within which randomness and accidental encounters may occur; and the facility to accumulate both a personal and community repository of know-how.

The future: seeking a definitive composition system is a waste of time. It only generates dissatisfaction. Software developers need to understand composers better.

I suggest two worldwide projects to the community:

- 1- To develop a research base for better environments that support creativity.
- 2- To work with the community to establish some musical grid, network and know-how base.

### Software for Electroacoustic Music Composition

COMPOSITION TOOL    COMPOSITION TOOL    COMPOSITION TOOL

- Physical (Computational / Data / DSP)
- Logical (Tools / Materials / Composition objects)
- Perceptual (Freely associated untyped objects)
- Community know-how (A Grid)
- Personal Know-how (Archives / Knowledge base)
- Composition Support (What? / How? / Where? / Again? / Suggest? / Random?)

(Diagram by Eaglestone)

**Jens Arnsprang:** I think this is related with education. Composers become craftsmen, following the transition into the other side of the user environments that better support creativity. The long answer when I have the chance will be: new education.

**-Dr. Roger B. Dannenberg** (Carnegie Mellon University, Pittsburgh, USA)

I would like to propose that the composition has two components:

- 1- Methodological applications of standard practice.
- 2- Creative practice, which is anything but methodological.

So, what I mean by methodological application of standard practice are techniques, maybe know-how of good works, including using digital audio, synthesis hardware and software, music notation, publishing software also organizational materials. We record and use performances gestures and knowledge, and know about simple manipulations, such as: stretching, transposing and copying.

On the other hand, creative practice is where the music really comes from. The first rule is: break all the rules. The second is: whatever you start with, you want to think about going outside the boundaries. So, almost by definition, if there is a standard practice, you have to go outside to do something creative.

## MUSIC GENERATION PANEL (A critical review)

Maybe one of the most common things when people talk about creativity, is to combine things in new ways, often in unanticipated ways. So all of these things work against any kind of methodology.

In the work that we do, one thing is the very strong tendency, especially in computer science, to try to clarify standard practice and implement it, and by doing that, we ignore the creative practice side because that is in the future work. I think approaches in that way are almost useless because it just ignores the most important part. I think computer scientists have made the same mistake over and over again.

I think there are a lot of things we can do technologically to aid the creative practice. One is building interfaces at the right level. We need to build more open-ended systems; systems that can be invoked by other programs, have options such as text input and output, so when we get a creative idea to combine systems in unusual ways, we need some way to communicate. We can make systems more scriptable, so they can do things they were not originally designed to do.

There is also a need to provide functions at many levels. What I mean here is that maybe there is an application for doing some interesting kinds of synthesis. Maybe it is great, but it is an application, and maybe what I need is a plug-in, or maybe what I need is a function library, or maybe what I need is the source code.

The final point is cost of accessibility. I think this is very critical for helping people be creative. Composers and artists are people that cannot go out and buy every piece of technology. We have this revolution of personal computers and the Internet, which give people access to so much stuff, but it is limited by expensive software applications and proprietary software, you cannot get into the source code and do creative things. It would be good for this community and the whole computer industry, to think about ways to enable artists to get access to those things.

**-Eduard Resina** (IUA-Pompeu Fabra University, Barcelona, Spain)

What concerns me is basically what makes sound become music. Sound is a natural phenomenon, music is not. Music is an idea we impose on top of sound. Basically it has to do with the perception or the ability to perceive meaningful relationships between sonic events.

From the point of view of algorithmic composition, there are different trends. For instance, starting from some sort of mathematical logic that is not intrinsically musical, and then we want to make musical, in some way. This could be the case of fractal composition. Another trend would be just expressing some standard musical knowledge, traditional knowledge like traditional counterpoint, and implementing this into some system that make, more or less, automatic composition.

I think there is some sort of composition where you want to start from musical intuition, but you want to set your own rules, you want to define your own musical context. And then after that, you want to be able to implement this algorithmically. In this case you



## MUSIC GENERATION PANEL (A critical review)

### OUR FINAL REMARKS

Finally, we would like to make some critical comments. Firstly we disagree, to some extent, with the ambiguous use of the word “creative” made by Dr. Eaglestone. Implicitly, he divides composers in two groups: “creative composers and not creative ones”. We think composers are creators by definition. Every time a composer writes a composition, he makes something new, for him and maybe for the rest of the music history. He **always** creates an “object” that never existed before: his music composition. This music shares common elements with others, but **always** has “something new” that makes it different from the rest of music created before. When this “something new” is so little, is what Dr. Dannenberg calls “methodological applications of standard practice”. When this “something new” is not so little, is what Dannenberg calls, in a careful way, “creative practice”.

So, we would prefer to think about composers who explore new ways of music composition and lead towards new aesthetic concepts, and composers who keep themselves, more or less, in the tradition of music composition. This is, to our mind, what Eaglestone means when he talks about creative composers and, implicitly, about not creative ones.

On the other hand, and thinking in the same direction, we would like to comment the phrase: “*create conditions in which composers can be creative*”, and others very much alike.

We depart from a question such as: ¿is there any environment, composition software, tool, etc., where a composer can't be creative? Most music composed throughout history has been written with a pen. With this single pen a lot of composers have made contributions to music development. Great music compositions have been written, and new aesthetic concepts have been explored and developed using only a single pen.

So, should we accept that there is an environment, composition software, tool, etc., where a composer can't be creative? If we should, maybe is time to tell composers: “forget computers, you can't be creatives with them, go back to pen and paper times”.

This happen because creativity does not reside in any tool, creativity is owned only by the musician who uses that tool. We would prefer to talk about “create conditions which **stimulate** composers's creativity” and support different and open ways of handling musical objects. The limits of creativity are in the mind of the person seated in front of the computer; nevertheless, as Resina said: “*it is true that certain software or certain tools allow more creative things than others*”.

“*The future: seeking a definitive composition system is a waste of time. It only generates dissatisfaction*” (B. Eaglestone). If we think about a marvellous composition system with the amazing ability to create all kind of music composition, from the past, the present, and even from the future, we totally agree: that is impossible.

Throughout music history composers have developed techniques, most of them algorithmic procedures, to handle all the elements of music, say: melody, harmony, rhythm, timbre, articulation, form... What a composition system makes is to apply these techniques, and even new or not generalized ones, to the material provided by the user,



## MUSIC GENERATION PANEL (A critical review)

i.e., the composer. The ways to handle the elements of music are so many, almost infinite, so that is, from our point of view, impossible to find a definitive computer composition system.

On the other hand, as we said before, creativity does not reside in the system, but on the composer who uses that system. We could try to model every way of create music, and implement those models in a computer system, but who assures there will not be a composer who invent a new one?

*“Software developers need to understand composers better”* (B. Eaglestone). This is a plain truth. From our point of view, the unique way of achieving that is to learn the basics of music composition, and to work as close as possible with composers. If we want to develop medical applications, we should learn the basics (and maybe not only the basics) of medicine related to the software, and work with doctors. If we want to develop applications for astronomy research, we should learn the basics (and maybe not only the basics) of astronomy, and work with astronomers. And of course, if we want to develop music composition applications, we should learn the basics (and maybe not only the basics) of music composition, and work with composers. That is all. To understand composers better, we should think the same way they do.

## CONCLUSIONS

The aim of the Music Generation Panel was to identify open questions and unsatisfied requirements that technology has the potential to address. We think this aim was achieved to some extent. Panellists clarified some theoretical aspects and proposed some directions for future research. We would like to extract what, to our mind, were the main requirements that music technology has the potential to address, and what should form the basis of a proposed scientific agenda for future composition systems research:

**Barry Eaglestone:** *We believe there to be a need to establish a research base for enhancement of support within composition software for creativity. (As we have remarked before, “to create conditions which **stimulate** composers’s creativity and support different and open ways of handling musical objects”.)*

**Roger B. Dannenberg:** *I think there are a lot of things we can do technologically to aid the creative practice. One is building interfaces at the right level. We need to build more open-ended systems; systems that can be invoked by other programs, have options such as text input and output, so when we get to creative idea combining systems in unusually ways, we need some way to communicate. We can make systems more scriptables, so they can do things they were not originally designed to do.*

**Eduard Resina:** *It would be essential to develop software where you can really work with musical concepts. Software has to be more intuitive for musicians, and certain solutions have to be found in this direction.*

We hope our work would be useful. Please, feel free to send any feedback, they are welcome.