Situated, Perceptual, Emotive and Cognitive Music Systems

A Psychologically Grounded Approach to Interactive Music Composition

Sylvain Le Groux

Tesi Doctoral UPF / 2011

Supervisada pel

Prof. Paul F.M.J. Verschure

Departament de Tecnologies de la Informació i les Comunicacions
To Nicole, Julia and Emilie
Abstract

This thesis introduces a novel situated interactive composition system called the SMuSe (for Situated Music Server) that is grounded on principles of modern cognitive science, provides perceptual control of sound synthesis and includes emotional feedback. It illustrates both a new music composition paradigm and a synthetic psychology approach to the study of music perception, emotion and cognition. The SMuSe is composed of cognitively plausible modules implemented as a hierarchy of musical agents and relies on distributed control, parallelism, emergence and embodiment. By interacting with its environment, which provides feedback via multiple sensors, the system generates complex adaptive affective musical structures. Focusing on the micro-level of sound generation, we present two complementary techniques that give high-level perceptual control over low-level sound synthesis parameters. In a first implicit approach, a support vector machine algorithm learns to automatically map perceptual features such as loudness, pitch and brightness onto additive synthesis parameters. In a second approach, a physically-inspired synthesis model provides explicit access to perceptual and physical parameters such as pitch, loudness, brightness, attack time, inharmonicity and damping. Moving from the study of music generation and control towards the study of the musical experience itself, we then evaluate how the music generated by the SMuSe influences the listeners’ emotional responses. A first psychoacoustics experiment shows the significant influence of structural (scale, register, harmony), expressive (velocity, tempo, articulation) and timbre (brightness, attack time, spectral flux) parameters on the emotional scales of valence, arousal and dominance. An additional large scale experiment involving dementia patients (an illness known to induce cognitive and affective deficits) shows that specific sound features (e.g. low loudness, low brightness) provoke specific emotional responses within the patients (e.g. low stress). Moreover, the patients’ emotional responses differ from the age-matched control group, and the analysis shows an increased emotional sensitivity to sounds as the severity of the disease increases. These results suggest that sound-based therapy and diagnosis for dementia are possible. Finally, the maturity and flexibility of the SMuSe music system are demonstrated by a series of real-world applications including the sonification of a mixed-reality space, a study on physiologically-based musical interaction, a neurofeedback musical interface, a closed loop system based on reinforcement learning of emotional feedback, and a large scale multimedia performance using brain-computer interfaces. A situated, perceptive, emotive and cognitive approach to the design of musical systems paves the way for new applications for therapy but also for interactive gaming and novel physiologically-based instruments. Our approach provides a well-grounded paradigm to develop advanced synthetic aesthetics system that can inform our understanding of the psychological processes on which they rely.
Résumé

Cet livre présente un système situé pour la composition interactive appelé SMuSe (Situated Music Engine) basé sur les sciences cognitives modernes, le contrôle perceptuel de la synthèse sonore et le feedback émotionnel. L’architecture de SMuSe repose sur une hiérarchie d’agents musicaux et s’appuie sur les principes de parallélisme, d’émergence et d’incorporation cognitive (embodiment). SMuSe est à la fois un nouveau modèle de composition musicale et une approche synthétique de l’étude de la perception, de la cognition et des émotions musicales. Ce système permet de générer et de moduler de manière intuitive et en temps réel des structures musicales complexes et adaptatives. SMuSe a été conçu pour des applications dans les mondes réel et virtuel et a été testé dans divers environnements sensoriels tels qu’un espace de réalité mixte, des systèmes de biofeedback, des jeux 3D ou des installations multimédia. Au niveau de la génération du son, deux techniques complémentaires sont proposées pour obtenir le contrôle perceptuel des algorithmes de synthèse sonore. Dans une première approche implicite un algorithme d’apprentissage par machine a support vecteur apprend automatiquement la relation qui relie les caractéristiques perceptive, comme le volume, la hauteur et la brillance à des paramètres de synthèse additive. Dans une deuxième approche explicite, un modèle de synthèse modale permet d’accéder directement à des paramètres comme la hauteur, le volume, la brillance, le temps d’attaque, l’inharmonicité et de le taux d’amortissement. En ce qui concerne l’expérience musicale a proprement parler, une première série d’études psychoacoustiques a montré l’influence significative des paramètres de SMuSe en termes de structure (registre et mode) d’expressivité (vitesse, rythme, articulation) et de timbre (brillance, temps d’attaque, flux spectral) sur les réactions émotionnelles des participants. Une seconde expérience impliquant 77 patients à différents stades de démence de type Alzheimer a montré que les spécificités sonores (faible volume, faible brilliance) induisaient des réactions émotionnelles significativement différentes chez les patients (par exemple faible niveau de stress), et que leurs réponses étaient différentes de celles du groupe de contrôle. L’analyse a révélé que plus la maladie était grave, plus la sensibilité émotionnelle des patients augmentait. Ces résultats suggèrent que la thérapie et le diagnostic de la démence de type Alzheimer basés sur les sons sont possibles. Enfin, la maturité et la flexibilité de SMuSe sont mises en évidence par une série d’applications telles que la sonification d’un espace de réalité mixte, une étude sur les réponses physiologiques à la musique, une interface de neurofeedback, un système d’apprentissage par renforcement du feedback émotionnel, et une performance multimédia qui utilise des interfaces cerveau-ordinateur. L’utilisation de systèmes musicaux affectifs, adaptatifs et interactifs tels que SMuSe ouvre de nombreuses perspectives a des fins thérapeuthiques mais aussi pour les jeux interactifs et le design de nouveaux
instruments de musique. Notre approche fournit un modèle solide pour le développement de système esthétiques synthétiques qui peut éclairer notre compréhension des processus psychologiques sur lesquels elle s’appuie.
List of Related Publications

Book Chapters


Journal Articles


Conference Proceedings


Le Groux, S., Manzolli, J., Verschure, P. F. M. J., October 2007a. Interactive sonification of the spatial behavior of human and synthetic characters in a mixed-reality


Conference Presentations


Invited Talks and Demos


Le Groux, S., Inderbitzin, M., Frohlich, M., October 2010. This is you brain on brain date. In: Brain and Mind Series at Swissnex San Francisco. Swissnex, San Francisco, CA.

Acknowledgments

I would like to thank my supervisor Prof. Paul F.M.J. Verschure for welcoming me at the Synthetic, Perceptive, Emotive and Cognitive Systems (SPECS) group of Universitat Pompeu Fabra. He gave me the opportunity to be part of a unique and stimulating research environment and always encouraged my curiosity, turning these four years at SPECS into a rich and productive experience.

I want to express my gratitude to the staffs and patients of the day hospital and care center of the Fundacio ACE, Institut Catala de Neurociences Aplicades in Barcelona for their collaboration and participation.

I am thankful to my colleagues at SPECS for their scientific and moral support and for making those days (and nights) of hard work at the lab such an enjoyable and fruitful experience.

I am grateful to all my friends in Spain, France and California (a special thought to Berkeley’s I-House diaspora) for sharing wonderful moments of life.

Finally, I am thankful to my family for their continuous support and for always being there for me.

Part of this dissertation was carried out at the Center for Computer Research in Music and Acoustics (CCRMA) at Stanford University under a BE-DGR grant from Generalitat de Catalunya.
# Contents

Abstract v

List of Related Publications ix

Contents xv

List of Figures xxi

List of Tables xxxiii

1. Preface 1
   1.1. Motivations ........................................ 1
   1.2. Organization ....................................... 2
   1.3. Contributions .................................... 3


2. Introduction 7
   2.1. Computer-based Music Composition .................. 7
      2.1.1. Sequencing ..................................... 8
      2.1.2. Algorithmic Composition ....................... 8
      2.1.3. Interactive Music ............................... 10
   2.2. Interactive Music Systems .......................... 11
      2.2.1. Taxonomy ...................................... 11
      2.2.2. Limitations .................................... 13
   2.3. Designing Modern Interactive Music Systems: .......... 14
      2.3.1. A Cognitivist Perspective ..................... 14
      2.3.2. Human in the Loop ............................. 17
      2.3.3. “Verum ipsum factum”: the Synthetic Method .... 22
   2.4. Robot-Based Music Generation ........................ 23
   2.5. Summary ........................................... 25

3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System 27
   3.1. Perceptual and Cognitive Models of Musical Representation ........ 27
      3.1.1. Event Fusion: the Micro Level of Sound .......... 29
## Contents

3.1.2. Musical Expressivity ................................................. 34  
3.1.3. Grouping: the Melodic and Rhythmic Level .................. 35  
3.1.4. The Form Level .................................................. 37  
3.1.5. Auditory Memory ................................................. 37  
3.1.6. Music Processing Modules ....................................... 39  
3.1.7. Implications for the Design of SMuSe’s Architecture ........ 40  
3.2. Computational Model: A Society of Musical Agents .......... 42  
3.2.1. Agency .......................................................... 42  
3.2.2. Data-flow Programming ......................................... 44  
3.2.3. Distributed Control ............................................. 45  
3.2.4. Concurrent and On-the-fly Control of Musical Processes .... 46  
3.3. Human in the Loop .................................................. 49  
3.3.1. Musical Interaction .............................................. 50  
3.3.2. Interaction Based on Spatial Behavior ....................... 51  
3.3.3. Interaction Based on Biosignals ............................... 51  
3.4. Musical Mappings .................................................... 55  
3.4.1. Taxonomy ........................................................ 56  
3.4.2. Explicit Mappings .............................................. 56  
3.4.3. Adaptive Mappings ............................................. 58  
3.5. Artistic Realizations ................................................ 62  
3.6. Conclusions ........................................................ 63  

## II. Composing Sounds: Perceptually Grounded Sound Synthesis  

4. From Sound Description to Sound Synthesis ....................... 67  
4.1. Introduction ....................................................... 67  
4.2. Sound Description ................................................ 68  
4.2.1. Music Information Retrieval: An Engineering Approach .... 68  
4.2.2. Psychoacoustics ............................................... 69  
4.2.3. Schaeffer’s Typology: a Composer’s Approach .............. 70  
4.3. Sound Synthesis ................................................... 71  
4.3.1. Abstract Models ............................................... 72  
4.3.2. Processed Recording ......................................... 72  
4.3.3. Spectral Models .............................................. 73  
4.3.4. Physical Models ............................................. 73  
4.4. Perceptual Mappings: From Sound Description to Sound Generation .................................................. 74  
4.5. Conclusions ....................................................... 75  

5. A Machine Learning Approach to Perceptual Synthesis ............ 77  
5.1. Sound Analysis and Synthesis Framework ....................... 77  
5.1.1. Harmonic Additive Model ..................................... 77  
5.1.2. Principal Component Synthesis ................................ 79
## Contents

5.2. Introduction to Machine Learning ......................................... 82
5.3. Feature extraction .................................................................. 84
  5.3.1. Controls: Perceptual Features ....................................... 84
  5.3.2. Targets: Time-Varying Principal Component Trajectories .... 86
5.4. Training and results ............................................................... 89
  5.4.1. Support Vector Mapping Machine .................................. 89
  5.4.2. The Dataset .................................................................. 90
  5.4.3. Model selection ............................................................. 91
  5.4.4. Evaluation ................................................................. 92
  5.4.5. Results ...................................................................... 93
5.5. Applications ............................................................................. 94
5.6. Conclusions ............................................................................ 95

6. Explicit Control of Perceptual Parameters with Physically Inspired Synthesis ................................................................. 99
  6.1. Physically-Inspired Sound Synthesis .................................... 99
    6.1.1. Modal Synthesis ....................................................... 99
    6.1.2. Computer Implementation ........................................ 100
  6.2. Perceptual Control of the Synthesis Parameters .................... 100
    6.2.1. Tristimulus ............................................................. 101
    6.2.2. Damping ............................................................... 103
    6.2.3. Time Envelope ...................................................... 104
  6.3. Conclusions ...................................................................... 104

III. Composing Emotions: Audio Kansei Engineering and Healthcare Applications ................................................................. 107

7. Introduction .............................................................................. 109
  7.1. Theories of Emotion ........................................................... 110
  7.2. Music-Induced Emotions ..................................................... 111
  7.3. Determinants of Musical Emotion ........................................ 111
    7.3.1. Musical Structure .................................................... 112
    7.3.2. Musical Performance ............................................... 113
    7.3.3. Timbre ................................................................. 113
  7.4. Time-course of Emotional Responses to Music ....................... 113
  7.5. Underlying Mechanisms of Musical Emotions ....................... 115
  7.6. Measuring Emotional Responses ......................................... 116
    7.6.1. Self-reports ........................................................... 116
    7.6.2. Physiology ........................................................... 119
  7.7. Summary .......................................................................... 120
Contents

8. A Synthetic Approach to the Study of Musically-induced Emotions 121

8.1. Methods ................................................................. 122
  8.1.1. Stimuli ............................................................. 122
  8.1.2. Procedure ......................................................... 124
  8.1.3. Participants ........................................................ 125

8.2. Results ................................................................. 125
  8.2.1. Musical Structure ................................................. 126
  8.2.2. Expressive Performance Parameters ......................... 127
  8.2.3. Timbre ............................................................. 132

8.3. Conclusions ............................................................ 136

9. Towards Sound-Based Therapy and Diagnosis for Dementia 139

9.1. Introduction ............................................................ 140

9.2. Methods ................................................................. 141
  9.2.1. Participants ......................................................... 141
  9.2.2. Stimuli ............................................................. 141
  9.2.3. Procedure .......................................................... 144

9.3. Results ................................................................. 145
  9.3.1. Global between-groups analysis (control vs. patients) ....... 145
  9.3.2. Detailed between-group analysis (control vs. patients) ....... 147
  9.3.3. Within-subject analysis for the control group ................. 148
  9.3.4. Within-subject analysis for the dementia group ............... 149
  9.3.5. Between group (disease stage) analysis for dementia patients ... 151
  9.3.6. Detailed between group analysis (disease stage) ............. 152

9.4. Discussion .............................................................. 155
  9.4.1. Methodological considerations ................................. 155
  9.4.2. Towards affective sound design for music therapy .......... 157
  9.4.3. Subcortical processing of sonic emotions in dementia ........ 158
  9.4.4. Towards a sonic-based diagnosis of dementia ................. 158

9.5. Conclusions ............................................................ 159

IV. The SMuSe in the flesh: Real-World Applications 161

10. XiM-SMuSE: Sonification of Spatial Behaviors in the Mixed-Reality Environment XiM 163

10.1. Introduction ........................................................... 163
  10.2. XIM: a mixed-reality space ....................................... 164
  10.3. System Architecture ............................................. 166
  10.4. Overview ............................................................. 167
    10.4.1. Sensing ......................................................... 167
    10.4.2. Neuromorphic analysis of user behavior .................... 168
  10.5. Music Generation ................................................ 169
  10.6. Mapping Spatial Behavior to Emotion and Music Generation .... 171
10.7. Conclusions ......................................................... 173

11. Neuro-SMuSe: Neurofeedback Through Musical Interaction 177

11.1. Introduction ..................................................... 177
11.2. Background .................................................... 178
  11.2.1. Electroencephalography ................................. 178
  11.2.2. Brain Waves ............................................... 178
  11.2.3. Neurofeedback ........................................... 179
  11.2.4. EEG-based Music Composition ......................... 180
11.3. The EEG System ............................................... 180
  11.3.1. Hardware ................................................. 180
  11.3.2. Software ................................................. 181
11.4. Interactive Music Generation ............................... 183
  11.4.1. Macro Structure: I-Ching ............................... 183
  11.4.2. Micro Structure: Perceptually Grounded Sound Generation Based on the Tristimulus Timbre Space ......................... 184
11.5. Neurofeedback Training with Music ....................... 185
11.6. Conclusions ................................................... 186


12.1. Introduction .................................................... 187
12.2. Background .................................................... 188
12.3. Adaptive Music Generator .................................. 188
12.4. Emotional Mappings ......................................... 190
12.5. Method ......................................................... 190
  12.5.1. Participants ............................................... 190
  12.5.2. Apparatus, Materials and Data Collection .......... 191
  12.5.3. Stimuli .................................................. 191
  12.5.4. Experimental Design ................................... 191
12.6. Results ........................................................ 192
12.7. Conclusions ................................................... 192

13. RL-SMuSe: A Closed-loop Interactive Music System Based on Reinforcement Learning 195

13.1. Introduction .................................................... 195
13.2. A Hierarchy of Musical Agents For Music Generation .... 197
13.3. Musical Tension as a Reward Function .................... 198
13.4. Pilot Experiment ............................................... 198
  13.4.1. One-Dimensional Independent Adaptive Modulation of Dynamics, Articulation and Tempo ............................ 199
  13.4.2. Two-Dimensional Modulation of Tempo and Dynamics .... 199
  13.4.3. Three Dimensions: Adaptive Modulation of Volume, Tempo and Articulation ................................................. 202
Contents

13.5. Conclusions ................................................................. 203

14. The Brain Orchestra ......................................................... 205
  14.1. Introduction ............................................................... 205
  14.2. System Architecture ................................................. 206
    14.2.2. Brain Computer Interfaces ...................................... 207
    14.2.3. The Situated Music Server ..................................... 210
    14.2.4. Wii-mote Conductor Baton ..................................... 211
  14.3. XMotion: a Brain-based Musical Performance ................... 212
    14.3.1. Emotion, Cognition and Musical Composition .................. 212
    14.3.2. Musical Material ................................................ 212
  14.4. Conclusions .............................................................. 214

15. Conclusions and Outlook ................................................. 217

Bibliography ........................................................................... 221

Index ....................................................................................... 243
List of Figures

2.1. **The Logic sequencer** by Apple allows to manipulate audio, MIDI events and effects via a graphical score-like representation. ............... 8

2.2. **Lejaren Hiller**, one of the pioneers of computer-based algorithmic music, used the Illiac, University of Illinois’ super computer to write one of the first programs able to generate simple cantus firmi. ....................... 9

2.3. **M**, a popular algorithmic composition software by Cycling 74, uses probabilities entered by the user via a graphical user interface to generate MIDI events. ......................................................... 10

2.4. **Mari Kimura** interacting with a computer-based system composed of a real-time audio analyzer listening to her performance on the violin, and a controller driving LEMUR’s GuitarBot (Singer et al., 2003) in real-time. 10

2.5. **Dan Overholt’s Overtone Violin** preserves the tradition of violin technique while augmenting its potential by adding extra sensors and controllers that drive a computer-based music system (Overholt, 2005) .... 12

2.6. **Rowe’s classification** of interactive music systems relies on the three axes of performance, interaction and playing modes (Rowe, 1993). ....... 13

2.7. **Human machine interaction** must also include a cognitive component at the machine side to be really considered interactive (adapted from Bongers 2000) .............................................................. 14

2.8. **Sense-think-act** interpretation of standard robotics tasks ................ 15

2.9. **An excerpt of Music of Changes (1951)** by composer John Cage. In this piece, most of note durations, dynamics, tempi and density are chosen out of the symbols of the I Ching divination system (also called a book of changes) by threefold coin flipping. ................................. 16

2.10. **The score of “In C”** by Terry Riley. In this piece, each musician can choose when to start playing and repeating the different patterns. .... 17

2.11. **Flow** is influenced by perceived challenge and skill levels (adapted from Csikszentmihalyi 1991) .................................................. 18

2.12. A) **“IanniX”** is a graphical editor of complex multi-dimensional and poly-temporal musical scores (Codys and Ferry, 2004a). B) **“The Hands”** converts movements of the hands, fingers and arms into sound (Krefeld and Waisvisz, 1990) C) **“The Very Nervous System”** transforms dance movements into sound (Rokeby, 1998). D) **“The Reactable”** transforms the manipulation of tangible objects into sound (Jorda et al., 2005). .... 19
List of Figures
2.13. MPTrain is a mobile phone based system that monitors the user’s physiology (namely heart rate and pace with ECG and accelerometer sensors
(Cf. Figure A) and select the music according to a desired exercise goal
(Cf. Figure B). (Adapted from Oliver and Flores-Mangas 2006). . . . . .

21

2.14. Roboser sonifies the behavior of a real-world robot equipped with sensors
(A) in an arena composed of wall (obstacles to avoid) and light sources
(targets) and colored patches (B, C) (Manzolli and Verschure, 2005). . .

24

3.1. An excerpt from Mycenes Alpha’s (1978) score composed by Iannis
Xenakis. It provides a graphical representation of the time-varying audio
spectrum that needs to be generated by a computer-based synthesizer. .

28

3.2. The diﬀerent levels of sequential grouping for musical material:
event fusion, melodic and rhythmic grouping and formal sectioning (from
Snyder 2000) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

29

3.3. Frequency of the pitches for a diatonic scale starting from note C1
(1 semitone=100 cents). The horizontal grid lines show multiples of frequency C1 (harmonics). . . . . . . . . . . . . . . . . . . . . . . . . . . .

31

3.4. Loudness: to a first approximation, the perception of loudness is logarithmic. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

32

3.5. Time clumping map: the three sides of the triangle represent continuous time IOI intervals (physical time) while the colors represent the
rhythmic categories (perceived time) perceived by the participants. Gray
lines are boundaries, and darker shades indicate a higher proportion of
participant identifying the rhythmic category (“Rhythms” legend at the
right) (from Desain and Honing 2003). . . . . . . . . . . . . . . . . . . .

33

3.6. A multidimensional model of timbre space of musical instruments that exhibits three perceptually relevant axes corresponding to
attack time, spectral centroid and spectral flux. Each point in space corresponds to a specific musical instrument (from McAdams et al. 1995a). .

34

3.7. Transition matrix: Representation of a first-order Markov illustrating
the probability of playing one note at time t + 1 depending the current
note at time t (here with a set of 3 notes {C, E, G} only). . . . . . . . .

39

3.8. Memory structure: Short-term memory sequences (STM) are indexed
and stored in long-term memory (LTM) for later recall. These sequences
can both represent musical parameter values directly (e.g. a sequence
of pitch classes {0 5 8 10 2} or probabilities of occurrence of parameter
values (e.g. an histogram {(0, 0.5) (5, 0.2) (8, 0.1) (10, 0.1) (2, 0.10)}) (Cf.
Section 3.2.4 for a description of the syntax used in SMuSe for memory
management) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

40

xxii


3.9. **Modularity of music processing**: there is strong evidence from neuroscience literature for a modular view of musical perception. First, acoustic information is transformed into neural activity in the cochlea and brainstem. Acoustic features are then extracted in the auditory cortex (ERP around 100ms). MMN correspond to operations of the sensory memory reflected in mismatch negativity. ERAN stands for early right anterior negativity and RATN right anterior-temporal negativity, which occur when violating music syntax. N400 correspond to a semantically related response. (adapted from Koelsch and Siebel 2005). .......................... 41

3.10. **The SMuSe’s architecture** is based on a hierarchy of musical agents that deal with tempo, rhythm (*when* musical events occur) and pitch, dynamics and articulation material (*what* musical event occur) (the voices). Specific rhythms and melodies are retrieved and stored in the memory by the musical agents. The symbolic musical events are then sent to the synthesizers for audio rendering. The simulation of the room size and sound source location are approximated by a reverb and spatialization module (based on VBAP). Expressive performance is achieved by modulating parameters such as tempo, articulation, dynamics and timbre. .......................... 43

3.11. **FM synthesis** can easily be implemented in the Max/MSP programming language as a graph representing the flow of data passing through basic computational units (at control or audio rate) (Zicarelli, 2002). .......................... 45

3.12. **The SMuSe’s client-server architecture** allows for distributed control over the music agents. Different clients on the network can simultaneously modulate the music generated by the server. .......................... 46

3.13. **SMuSe’s environment**: the SMuSe can interact with its environment through different sensors such as biosignals, camera, gazers, lasers, pressure sensitive floor, MIDI, audio, but also via OSC commands sent from client applications (such as console terminal, IQR, Iannix graphical score, Torque game engine, etc.) to the music server over the network. .......................... 50

3.14. **The audio analysis** module extracts estimations of pitch, brightness, noisiness, loudness and onsets in real-time from the audio signal recorded via a microphone (Puckette et al., 1998; Jehan and Schoner, 2001). .......................... 50

3.15. **XIM (the eXperience Induction Machine)** is an intelligent space which has its own goals and expresses internal states. XIM is provided with a set of sensors and effectors. Sensors allow for the tracking of people in the space. Effectors are used to influence the emotional behavior of the visitors (Bernardet and Verschure, 2010). .......................... 52

3.16. **The nervous system** is divided into the central nervous system (brain and spinal cord) and the peripheral nervous system (sensory-somatic nervous system and autonomic nervous system). .......................... 53

3.17. **The Autonomic Nervous System** is a part of the peripheral nervous system that controls visceral functions (from Brown and Benchmark 1995) .......................... 54
List of Figures

3.18. **Biosignals** are used as an input to the SMuSe for direct control of musical parameters (BCI paradigm) or as emotional feedback (Cf. Chapters 14 and 11). The picture on the left shows the EEG cap of the ENOBIO system. The picture on the right shows the G-tec system setup for heart rate measurement. .......................................................... 55

3.19. **The reinforcement-based music system** is composed of three main components: the music engine (the SMuSe), the reinforcement learning agent and the listener who provides the reward signal. .......................................................... 58

3.20. **The agent-environment interaction** in the reinforcement learning framework (from Sutton and Barto 1998) .......................................................... 59


4.1. A) **The Buchla 200e** analog modular synthesizer, designed by electronic music pioneer Don Buchla allows to generate new timbres by patching different modules together (waveform generator, oscillator, filters, ADSR envelopes, ...). B) **The Alchemy Additive Synthesizer** software, by Camel Audio, allows to synthesize new sounds by manually editing the parameters of an additive synthesis algorithm (amplitude and frequency of partials). Both synthesizers require a high level of technical expertise to generate new interesting sounds. .......................................................... 68

4.2. **The timbre space** of Grey and Gordon (Grey, 1975). Dimension I is the spectral energy distribution. Dimension II is timing of the attack and decay. Dimension III is the amount of inharmonic sound in the attack. (BN - Bassoon C1 - E flat Clarinet C2 - B flat Bass Clarinet EH - English Horn FH - French Horn FL - Flute O1 - Oboe O2 - Oboe (different instrument and player) S1 - Cello, muted sul ponticello S2 - Cello S3 - Cello, muted sul tasto TM - Muted Trombone TP - B flat Trumpet X1 - Saxophone, played mf X2 - Saxophone, played p X3 - Soprano Saxophone) .......................................................... 70

5.1. **The SMS analysis** provides a robust methodology for the extraction of synthesis parameters such as amplitude, frequency and phase from an audio file (Serra and Smith, 1990). .......................................................... 80

5.2. **The time-varying partial** amplitudes from matrix A represent the first 50 partials of a 5 notes solo saxophone. Each color represents a different partial (one row of the matrix). .......................................................... 81

5.3. **The time-varying envelopes** represent trajectories in the Principal Component bases. .......................................................... 82

5.4. **PC bases** represent “principal spectra”. .......................................................... 83

5.5. **3D representation of the trajectories** in the PC bases space. .......................................................... 84

5.7. **Yin fundamental frequency estimation**, with noisy data points removal and smoothing. .......................................................... 85

xxiv
5.8. **The three-dimensional control vector** includes estimations of pitch (fundamental frequency), loudness (loudness summation) and brightness (spectral centroid). ................................................................. 86
5.9. **Pareto chart** of the number of Principal Components (PC) against the percentage of variance expressed. For this database, most of the variance in the data can be expressed by only 3 PCs. ................................. 87
5.10. **Test set trajectory** in a three-dimensional PC space. .................. 87
5.11. **Reconstructed partial trajectories** from the three-dimensional PC space. ................................................................. 88
5.12. **Kernel methods** map the training data non linearly into a higher-dimensional feature space an construct a separating hyperplane with maximum margin. ................................................................. 90
5.13. **The continuous perceptual input controls** are pitch, loudness and brightness. ................................................................. 91
5.14. **Partial estimation results. A)** Target and estimated first PC trajectory of the training set B) Target and estimated first PC trajectories of the test set. ................................................................. 93
5.15. **The correlation between target and estimate** on the test data set shows that the SVR gives a good estimation of the trajectory when presented with unknown inputs. ................................................................. 94
5.6. **Estimation of the time-varying amplitudes** for the first six partials of the spectrum. ................................................................. 97

6.1. **Modal analysis:** In the modal analysis paradigm, a complex vibrating structure is modeled as a sum of simpler vibrating modes. ................. 100
6.2. **Implementation of the modal Synthesizer:** The sounds are produced by an impulse that injects energy in a bank of exponentially-decaying resonant filters with parameters Frequency (F), Amplitude (A) and decay rate (d), the output of the filter bank is then modulated by an ADSR time envelope. ................................................................. 101
6.3. **Tristimulus representation of timbre:** The tristimulus timbre space can be represented in 2D as a triangle. The arrow represents a specific time-varying trajectory of a sound in the space (Pollard and Jansson, 1982). 102
6.4. **The modal synthesizer’s graphical interface** allows for graphical control over A) the tristimulus values (left) and automatically updates and displays the amplitudes of the corresponding partials (right).B) the damping parameters $\alpha_g$ and $\alpha_r$ C) the piano-like inharmonicity law D) the log-attack time via the Attack Decay Sustain Release (ADSR) envelope E) The amount of noise in the input signal as well as the ration even/odd partials. Additionally, the synthesizers provides a real-time visualization of estimation of pitch, loudness, brightness, noisiness and spectral flux (F). The synthesizer can be controlled by any midi device (G). The configuration of parameters can be saved and retrieved for later use (H). ................................................................. 103
6.5. **The ADSR time envelope** allows for direct control of the Log-Attack Time (LAT) each time a note is triggered. ........................................... 104

7.1. **Three key theories of emotion:** 1) the James-Lange theory emphasizes the role of bodily changes on emotions 2) The Cannon-Bard theory proposes that the stimuli is evaluated independently of the bodily reactions 3) The cognitive theory models emotions as a two factor process composed of both physiological arousal and cognition ................................. 112

7.2. **The Self Assessment Manikin (SAM) scale** evaluates emotional responses on the scale of Valence (top), Arousal (middle) and Dominance (bottom) (Lang, 1980) ................................................................. 118

7.3. **The circumplex model of affects** represents emotional responses as a point on the dimension of arousal and valence (Russell, 1980) .................. 118

8.1. **SMuSe client’s graphical user interface** is used to generate a well-parameterized musical stimulus by modulating macro-level parameters such as articulation, tempo, velocity, density of notes, mode, register and instrumentation. Those parameters are sent to the SMuSe server from the GUI via OSC messages (Cf. Section 3.2.4). .................................................. 122

8.2. **Musical structure samples:** 2 structural parameters (Register and Mode) are modulated over 9 sequences (3*3 combinations) ......................... 123

8.3. **Musical performance samples:** 3 performance parameters were modulated over 27 musical sequences (3*3*3 combinations of Tempo (BPM), Dynamics (MIDI velocity) and Articulation (duration factor) levels). . . . . . . 123

8.4. **Timbre samples:** 3 timbre parameters are modulated over 27 samples (3*3*3 combinations of Attack (ms), Brightness (tristimulus band), Damping (relative damping \(\alpha\))). The other parameters of PhySynth were fixed: decay=300ms, sustain=900ms, release=500ms and global damping \(\alpha_g = 0.23\). .................................................. 124

8.5. **The presentation software PsyMuse** uses the SAM scales (axes of Dominance, Arousal and Valence) (Lang, 1980) to measure the participant’s emotional responses to a database of sounds. ......................... 125

8.6. **Effect of mode and register** on the emotional scales of arousal, valence and dominance. A) shows mean and standard error values B) shows the statistically significant effects. ................................. 126

8.7. **Influence of structural parameters (register and mode) on arousal and valence.** A) A musical sequence played using random notes (level 3) and using a minor scale (level 2) is perceived as significantly more negative than a sequence played using a major scale (level 1). B) A musical sequence played in the soprano range (high register level 1) (respectively bass range, low register level 2) is significantly more (respectively less) arousing than the same sequence played in the tenor range (medium register level 3). Estimated Marginal Means are obtained by taking the average of the means for a given condition. ................................. 127
8.8. **Structure: interaction between mode and register for arousal, valence and dominance.** A) When using a random scale, a sequence in the tenor range (level 3) becomes less arousing B) When using a minor scale (level 2), a sequence played within the soprano range (level 1) becomes the most positive. C) When using a random scale, bass and soprano sequences are the most dominant whereas tenor becomes the less dominant. 128

8.9. **Effect of articulation, tempo and dynamics** on self-reported emotional responses on the scale of valence, arousal and dominance. A) shows mean and standard error values B) shows the statistically significant effects. 129

8.10. **Effect of performance parameters (Articulation, Tempo and Dynamics) on Arousal.** A) A sequence played with articulation staccato (level 2) is more arousing than legato (level 1) B) A sequence played with the tempo indication presto (level 1) is more arousing than both moderato (level 3) and lento (level 2). C) A sequence played forte (respectively piano) was more arousing (respectively less arousing) than the same sequence played mezzo forte (level 3). 130

8.11. **Effect of performance parameters (Articulation, Tempo and Dynamics) on Valence.** A) A musical sequence played staccato (level 2) induce a more negative reaction than when played legato (level 1) B) A musical sequence played presto (level 1) is also inducing a more negative response than played moderato (level 3). C) A musical sequence played forte (respectively piano) is rated as more negative (respectively positive) than a sequence played mezzo forte. 131

8.12. **Effect of performance parameters (Tempo and Dynamics) on Dominance.** A) A musical sequence played with a tempo presto (respectively lento) is considered more dominant (respectively less dominant) than played moderato (level 3) B) A musical sequence played forte (respectively piano) is considered more dominant (respectively less dominant) than played mezzo-forte (level 3) 132

8.13. **Effect of brightness, damping and attack** on self-reported emotion on the scales of valence, arousal and dominance. A) shows mean and standard error values B) shows the statistically significant effects. 133

8.14. **Effect of timbre parameters (Brightness, Damping and Attack time) on Arousal.** A) Brighter sounds induced more arousing responses. B) Sounds with more damping were less arousing. C) Sounds with short attack time (level 2) were more arousing than medium attack time (level 3). D) Interaction effects show that less damping and more brightness lead to more arousal. 134
8.15. **Effect of timbre parameters (Brightness, Damping and Attack time) on Valence.** A) Longer attack time are perceived as more negative B) Bright sounds tend to be perceived more negatively than dull sounds C) Interaction effects between damping and brightness show that a sound with high damping attenuates the negative valence due to high brightness. ................................................................. 135

8.16. **Effect of timbre parameters (Brightness and Damping) on Dominance.** A) Bright sounds are perceived as more dominant than dull sounds B) A sound with medium damping is perceived as less dominant than low damping. ................................................................. 136

9.1. **Parametric tree** describing the sound feature space consisting of sound source, speed, consonance, sound level and brightness. ................................. 142

9.2. **The three voice patterns** were simple melodies sang by a female voice. 142

9.3. **Parametric definition of consonance:** Pattern 9.3a was considered the highest consonance, since it has unison, a minor third, unison and octave. Pattern 9.3b was considered the medium consonance, since it has unison, a major second, a minor second and unison. Pattern 9.3c was considered the lowest consonance, since it has unison, a minor second, a minor second and octave. ................................................................. 143

9.4. **Zblast**-The stimulus delivery and data acquisition system’s Graphical User Interface: The user rates the sound samples on a scale from happy to sad based on her emotional state. The user can click the box “no lòse” (I do not know) or “pasar al siquiente” (pass to the next) to skip to the next dialog. The blue bar at the top indicates the duration that the sample has played and will play. ................................................................. 145

9.5. **Global distributions of the emotional responses from the patient and control groups.** The histograms on the Valence (A), Pleasantness (B), Dominance (C) and Arousal (D) scales show that dementia patients (in blue) have a tendency to be more extreme in their ratings than the control group (in green). ................................................................. 146

9.6. **Global between-group analysis.** Overall, patients rate the sound samples as sadder (A) and less arousing (B) than controls. Ratings on the the dominance axis exhibit many outliers (C), especially for the control group. This suggests the dominance axis might not be relevant for musical stimuli. ................................................................. 147

9.7. **Detailed between-group analysis of sound properties.** Synthetic (A) and medium loudness (B) samples feel more relaxing for patients, while they rated medium consonance (C) as less dominant and slow tempo (D) as less pleasant. ................................................................. 148

9.8. **Within-subject analysis (control group):** Responses to loudness on the arousal scale (low=1, medium=2, high=3) show that low sound levels are less arousing. ................................................................. 149
9.9. **Within-subject analysis (patient group):** water sound samples are perceived as happier (A) and more relaxing (B). Voice samples are evaluated as being sadder (A), more arousing (B) and more dominant (C).  ... 150

9.10. **Global between-group analysis (dementia disease stages):** The more the disease progresses, the more the patients tend to rate the sound samples as being happier (A) and more pleasant (B)  ... 151

9.11. **Detailed between-group analysis (dementia disease stages) for sound types:** the more the dementia progresses (from stage 3 to 5), the more water (A) and voice (B) sound samples are rated as happier and synthetic (C) and water (D) sounds are rated as more pleasant  ... 152

9.12. **Detailed between-group analysis (dementia disease stages) for loudness:** When the dementia progresses (from stage 3 to 5), samples with high loudness (A) and medium loudness (B) are rated as happier and more pleasant respectively (C, D)  ... 154

9.13. **Detailed between-group analysis (dementia disease stages) for brightness:** When the dementia progresses (from stage 3 to 5), samples with low brightness (A) are perceived as more dominant and more pleasant (B)  ... 155

9.14. **Detailed between-group analysis (dementia disease stages) for consonance:** medium consonance was perceived as happier at higher stage of the dementia  ... 156

9.15. **Detailed between-group analysis (dementia disease stages) for tempo:** When the dementia progresses (from stage 3 to 5), samples with high tempo are perceived as happier (A) and more pleasant (B)  ... 157

10.1. **The Persistent Virtual Community** consists of the real-world physical installation (called the Experience Induction Machine), the ClubHouse, the P-Garden and Avatar Heaven. Users can visit the physical installation XIM, connect over the Internet (remote users) or visit a CAVE installation  ... 165

10.2. **The architecture the mixed-reality music system** couples sensing with neuromorphic processing and musical responses  ... 166

10.3. **XIM** is provided with a set of sensor and effectors. Sensors allows for the tracking of people in the space. Effectors are used to influence the emotions and behavior of the visitors  ... 167

10.4. **IQR** provides a user-friendly interface to design multilevel neural models. Models in IQR are organized at different levels. The top level is the system containing an arbitrary number of processes, and connections. This example illustrates an IQR module where two processes are interconnected. Each process contain a neuronal circuit made up of three neuronal groups  ... 168

10.5. **Behavior analysis:** IQR allows to graphically design a module that extracts behavioral information from the tracking system. The cell activity of the different interconnected neuronal groups is related to the activity of avatars and human interacting in the mixed-reality environment  ... 169
10.6. **Musical Cells:** A group of related musical cells defines a certain style file which specific structural and musical parameters are associated with a mode of interaction of the XIM. When a style file is enabled, we cycle through the different cells whose properties are modulated in real-time via OSC control messages. ............................... 170

10.7. **Three layers mapping** from the spatial behavior of virtual and human characters in the mixed reality space to the semantics of emotional states we want to express, and finally to the musical parameters illustrating such emotional states. ............................... 171

10.8. **One user interaction:** This spectrogram (x is time in seconds and y frequency in kHz) illustrates the sonic result of a simple interaction scheme. One user is in the space and moves from one corner to the other. When approaching the corner she triggers a percussion sample. She then moves away from the corner. As the motion of the user is accelerating the rhythmic pattern of the current cell is played at a faster tempo. The duration of each note is rather long, and most of the energy is in the low frequencies. ............................... 174

10.9. **Six users interaction:** This spectrogram (x is time in seconds and y frequency in kHz) illustrates an interactive session where six users are moving around in the space. We observe that the density of event is high. There is also a large amount of energy in the high frequency spectrum. ............................... 175

11.1. **Brainwave patterns** ............................... 178

11.2. **The ENOBIO** is a wearable wireless dry electrode electrophysiology recording system. ............................... 181

11.3. **The 10-20 system:** Standard location and nomenclature of the electrodes for EEG recording according to the 10-20 system (American Electroencephalographic Society) ............................... 182

11.4. **Real-time EEG frequency analysis and visualization:** The upper left display represents the spectrogram in the 8 to 12 Hz frequency range (alpha wave). The upper right side is the discrete byte information sent from the server. The bottom left side is the interpolated version of the data. The bottom right display shows the alpha wave power as a bar plot. There is also the possibility to amplify and calibrate the raw signal for better viewing. ............................... 183

11.5. **The tristimulus synthesizer** allows control over tristimulus parameters, ADSR envelope, noisiness, loudness, inharmonicity and vibrato. ............................... 184

11.6. **Real-time spectral analysis and visualization of the tristimulus synthesizer:** the upper display represents the weighting of the harmonics as defined by the tristimulus controller. The middle display represents the time varying harmonics after synthesis. The bottom display represents the audio signal as well as the spectrogram and spectral density. ............................... 185
12.1. **The musical material** consists of the real-time modulation of precomposed musical cells. The SMuSe’s GUI represents one musical cell (rhythmic values and chords) for two voices that communicate with OSC (Cf. Chapter 3 for more details). .................................................. 189

12.2. **Tristimulus synthesis flowchart**. .......................................................... 190

13.1. **The Reinforcement Learning framework.** The system is composed of three main components, namely the music engine (SMuSe) that generates music, the reinforcement learning agent that modulates musical parameters based on feedback from the participant, and the listener who provides the reward signal based on the musical output generated by the SMuSe. ................................................................. 196

13.2. **Learning dynamics**: the RL agent automatically learns to map an increase of perceived tension, provided by the listener as a reward signal, to an increase of the dynamics gain. Dynamics gain level is in green, cumulated mean level is in red/thin, reward is in blue/crossed and cumulated mean reward is in red/thick. .................................................. 200

13.3. **Learning articulation**: the RL agent learns to map an increase of perceived tension (reward) to longer articulations. ................................. 201

13.4. **Learning Tempo**: the RL agent learns to map an increase of musical tension (reward) to faster tempi. ................................................. 201

13.5. The RL agent learns to map an increase of musical tension (reward in blue/thick) to faster tempi (parameter 1 in green/dashed) and higher dynamics (parameter 2 in red/dashed). ............................................. 202

13.6. A piano roll representation of an interactive learning session at various stage of learning. At the beginning of the session (left), the musical output shows no specific characteristics. After 10 min of learning, excerpts where low tension (center) and high tension reward is provided by the listener (right) exhibit different characteristics (cf text). The length of the notes correspond to articulation. Colors from blue to red correspond to low and high volume respectively. ............................................. 203

14.1. **The performance setup** for the world premiere of MBO at FET European conference in Prague, July 2009. .................................................. 206

14.2. **The Multimodal Brain Orchestra is a modular interactive system** based on a client-server architecture using the OSC communication protocol. See text for further information. ............................................. 207

14.3. **The P300 Speller.** **A)** Rare events triggers an ERP (Target X) 300 ms after the onset of the event indicated by the green arrow. **B)** A 6 by 6 symbol matrix is presented to the P300 orchestra member who can potentially trigger 36 specific discrete sound events. ................................. 209

14.4. Two members of the orchestra connected to their SSVEP-based BCI interfaces (Prueckl and Guger, 2009) ............................................. 209
List of Figures

14.5. **The real-time visualizer** displays a visualization of the two P300 systems output (the two upper rows show combinations of shapes and colors) and the two SSVEP system output (the two lower rows show the transitions from the 4 different states) ................................................................. 211

14.6. **The wii-baton module** analyzed 3D acceleration data from the wii-mote so the conductor could use it to modulate the tempo and to trigger specific sections of the piece ................................................................. 211

14.7. **Russel’s circumplex model of affect** represents emotions on a 2D map of Valence and Arousal (Russell, 1980) ................................................................. 213
### List of Tables

3.1. **Equivalence table** between pitch class and notes .......................... 31

3.2. **Equivalence table** between dynamics musical notation, MIDI note velocity for standard software instruments and psychoacoustic scales. .......................... 32

3.3. **Example of symbolic representation of note durations.** The same logic applies to all types of notes and silences (e.g. 32th dotted note are notated 32nd). ................................................................. 33

3.4. **The action selection agents** choose to play specific musical elements stored in the current working STM following deterministic or stochastic selection principles. ................................................................. 49

3.5. **The relationship between musical parameters and perceived emotion.** A summary of the main results. (adapted from Gabrielsson and Lindström 2001) ................................................................. 57

4.1. **Sound Synthesis Taxonomy** (adapted from Smith 1991) ................. 72

5.1. **Correlation and RMS** results on a 2 min saxophone solo. ............... 93

7.1. **Mappings between musical parameters and perceived emotion**
   (adapted from Gabrielsson and Lindström 2001) ..................................... 114

10.1. **Relationship between a set of standard musical parameters and semantics of musical expression** (adapted from Gabrielsson and Lindström 2001) ................................................................. 172

10.2. **Relationship between a reduced set of sound synthesis parameters and musical expression.** (adapted from Gabrielsson and Lindström 2001) ................................................................. 173

10.3. **Relation between spatial behaviors of human and avatars in the XIM and sonification parameters.** ................................................................. 173

11.1. **Brainwaves** and their associated mental states ............................... 179

12.1. **Influence of sound feature levels** on self-reported Valence and Arousal. 
   High Brightness and Loudness levels significantly increase perceived arousal. 192
12.2. The correlation table between self reports (valence, arousal) and physiology (heart rate, heart rate variability and electrodermal activity) shows significant correlations between heart rate and valence (p<0.05), heart rate variation and arousal (p<0.05) and arousal and EDA (p<0.001) ................................................................. 193

14.1. Relationship between sound quality and emotion expression on the scales of valence and arousal. ................................. 214
1. Preface

1.1. Motivations

Music listening, mediated by technologies such as radio broadcasting, podcasts, television, mobile phones or MP3 players as well as music making mediated by computers and new interfaces have now become common practice. Portable devices allow to listen to music anywhere, complex algorithms that run on standard laptop computers are able to generate virtually any kind of sound, and sophisticated music production software has become widely available. Yet, these advances in music technology, by focusing on practical engineering applications, have mostly neglected the psychological aspects of the musical experience. One could argue that we have reached a stage where the creation of new computer music is no longer limited by technology, but by our understanding of the reaction of the human brain to music (Park, 2009). In this thesis, we postulate that music technology and psychology are mutually beneficial, and that by developing psychologically grounded real-time interactive systems, we can address fundamental questions at the intersection of perception, cognition, performance and therapy.

One fundamental question is that of music representation and processing. Thanks to rapid advances in computing and interface technologies, computer-based music systems have evolved from simple computer-aided composition tools, which transposed traditional music composition techniques into the digital realm, to complex feedback music composition systems that allow for rich multimodal interactions. In parallel, the progresses made by cognitive science, neuroscience and biomimetic robotics have improved our understanding of perceptive, emotive and cognitive systems. Yet, most interactive music systems still rely on an obsolete classical cognitive science paradigm emphasizing internal representations, centralized control and disembodiment. What are the basic musical structures needed to generate music? What could modern cognitive science bring to the design of interactive music systems? Those issues are the object of the first Part of this dissertation (Cf. Part I).

On a more technical note, modern audio signal processing techniques and processors now allow for efficient real-time sound synthesis and direct transformation of audio material at the microlevel of sound. However, building new interesting sounds with current synthesizers require a high level of technical expertise. Intuitive and perceptually-grounded control over the many dimensions of musical timbre still remains an open problem. How can we reconcile complex synthesis algorithms with intuitive, perceptually grounded control of sound generation? We tackle these problems in Part II of this thesis.
1. Preface

From, the listener and performer’s perspectives, emotions are a crucial aspect of musical experience as attested by a variety of self-report, physiological, and observational means. However, the relationship between musical parameters and specific emotional responses is still not clear. The role of human emotion is in fact rarely taken into account in the design of new interactive music systems. What are the perceptual determinants of musically-induced emotion? How can we build a music system with emotional feedback? These problems are investigated in Part III and IV of the dissertation.

We propose to address all these questions by introducing a novel situated interactive composition system called the SMuSe (for Situated Music Server) that is based on principles of modern cognitive science, perceptual control of sound synthesis and emotional feedback.

1.2. Organization

This thesis is divided into four main parts. Part I deals with the design and making of a perceptual, cognitive and emotive music system. In Chapter 2, we propose a conceptual framework for interactive music composition based on embodied cognitive science and synthetic psychology. In Chapter 3 we describe a software implementation called the SMuSe that is built on a cognitively plausible architecture modeled as a hierarchy of musical agents. The SMuSe is integrated into a variety of real and virtual world sensate environments which can provide emotional feedback and modulate the generation of complex adaptive musical structures in real-time.

Part II of this thesis introduces two models that allow to synthesize a sound’s timbre from high-level perceptual parameters. Chapter 5 proposes an implicit machine learning approach based on Support Vector regression and additive synthesis, while Chapter 6 presents an explicit approach based on modal synthesis.

In Part III, we focus on experiments that investigate emotional responses to musical stimuli. The psychoacoustics experiments of Chapter 8 confirm that the SMuSe is able to evoke specific emotional responses by modulating its structural, expressive and timbre parameters. Chapter 9 describes a large scale experiment involving patients at different stages of dementia. It shows that dementia patients’ emotional responses to specific sound features differ from the aged-match control group. Moreover, the patient’s sensitivity to sounds increases as the severity of the disease increases. These results lay the foundations for sound-based therapy and diagnosis of dementia.

Part IV puts forward several detailed examples of real-world applications that rely on the SMuSe to generate music. They range from the sonification of human and avatar’s spatial behavior in a mixed reality space (Chapter 10) to experiments involving physiology and emotion (Chapter 11, 12 and 13) and multimedia artistic realizations (Chapter 14). Our paradigm of situated interactive music composition implemented in the SMuSe has been central to all these examples. It provides a well grounded approach towards the
1.3. Contributions

Contributions of this thesis to the literature can be summarized as follows:

• A conceptual framework for interactive music composition based on situated cognition and emotional feedback (Chapter 4).

• An agent-based software implementation of the framework called the SMuSe (Chapter 3).

• The integration of the SMuSe within real-world sensate environments (Chapter 3 and Part IV).

• Two methods for high-level perceptual control of a sound’s timbre (Part II).

• An experiment showing the ability of the SMuSe to evoke specific emotions (Chapter 8).

• A large scale experiment involving dementia patients that emphasize the importance of subcortical processes in musical emotions and suggests applications to sound-based diagnosis and therapy (Chapter 9).

• Extended case studies of real-world applications using the SMuSe that include sonification of a mixed-reality space, biofeedback experiments, reinforcement learning, and artistic installations and performances (Part IV).

development of advanced synthetic aesthetic systems and a further understanding of the fundamental psychological processes on which it relies.

Sound examples, code snippets and binaries for each chapter can be found online at http://www.dtic.upf.edu/~slegroux/phd/index.html.
Part I.

Composing Music Systems: A Situated Cognition Approach
2. Introduction

Externalism considers the situatedness of the subject as a key ingredient in the construction of experience. In this respect, with the development of novel real-time real-world expressive and creative technologies, the potential for externalist aesthetic experiences are enhanced. Most research in music perception and cognition has focused on tonal concert music of Western Europe and given birth to formal information-processing models inspired by linguistics (Lerdahl and Jackendoff, 1996; Narmour, 1990; Meyer, 1956). These models do not take into account the situated aspect of music although recent developments in cognitive sciences and situated robotics have emphasized its fundamental role in the construction of representations in complex systems (Varela et al., 1991). Furthermore, although music is widely perceived as the “language of emotions”, and appears to deeply affect emotional, cerebral and physiological states (Sacks, 2008), emotional reactions to music are in fact rarely included as a component to music modeling. With the advent of new interactive and sensing technologies, computer-based music systems evolved from sequencers to algorithmic composers, to complex interactive systems which are aware of their environment and can automatically generate music. Consequently, the frontiers between composers, computers and autonomous creative systems have become more and more blurry, and the concepts of musical composition and creativity are being put into a new perspective. The use of sensate synthetic interactive music systems allows for the direct exploration of a situated approach to music composition. Inspired by evidence from situated robotics and neuroscience, we believe that, in order to improve our understanding of compositional processes and to foster the expressivity and creativity of musical machines, it is important to take into consideration the principles of parallelism, emergence, embodiment and emotional feedback. In this chapter, we provide an in depth description of the evolution of interactive music systems and their limitations (Section 2.1 and 2.2), and propose a novel synthetic and situated approach to music composition based on modern situated cognitive science principles (Section 2.3). Finally, we emphasize the importance of emotion-aware interfaces (Section 2.3.2.3) and of a perceptually-grounded approach to music generation, both at the micro and macro levels of sound (Section 2.3.2.4).

2.1. Computer-based Music Composition

One of the first goals of computer-aided composition, based on a rich history of classical music theory and teaching, was to help the composer during the creative process. With the advent of new interactive technologies, computer-based music systems evolved from
2. Introduction

sequencers to algorithmic composers, to complex interactive systems. Consequently, the frontiers between composers, computers and autonomous creative systems have become more and more blurry, and the concepts of musical composition and creativity are being put into a new perspective.

2.1.1. Sequencing

One of the most widespread computer-aided composition paradigm is probably still that of the music sequencer. This model is somehow a continuation of the classic composition tradition based on the writing of musical scores. Within the sequencer paradigm, the user/composer creates an entire piece by entering notes, durations or audio samples on an electronic score (Figure 2.1). Due to its digital nature, this score can later be subjected to various digital manipulations. Within the sequencer paradigm, the computer is “passive”, and the composer produces all the musical material by herself. The human is in control of the entire compositional processes and uses the computer as a tool to lay down ideas and speedup specific tasks (copying, pasting, transposing parts).

Figure 2.1.: The Logic sequencer by Apple allows to manipulate audio, MIDI events and effects via a graphical score-like representation.

2.1.2. Algorithmic Composition

In contrast with the standard sequencer paradigm, computer-based algorithmic composition relies on mathematical formalisms that allows the computer to automatically generate musical material, usually without external output. The composer does not specify directly all the parameters of the musical material, but a set of simple rules or
input parameters, which will be taken into account by the algorithm to generate musical material. In this paradigm, the computer does most of the detailed work and the composer controls a limited set of initial global parameters.

Some mathematical formula provide simple sources of quasi randomness that were already extensively used by composers before the advent of computers. In fact, Fibonacci sequences and the golden ratio have been inspiring many artists (including Debussy, Bartok or Stravinsky) for a long time, while more recent models such as chaotic generators/attractors, fractal, Brownian noise, and random walks are exploited by computer technologies (Ames, 1987) (Figure 2.2).

![Image](image.jpg)

Figure 2.2.: **Lejaren Hiller**, one of the pioneer of computer-based algorithmic used the Illiac, University of Illinois’ super computer to write one of the first program able to generate simple cantus firmi.

Different approaches to algorithmic composition inspired by technical advances have been proposed and tested. The main ones are statistical methods, rule-based methods, neural networks and genetic algorithms (Papadopoulos and Wiggins, 1999; Nierhaus, 2009).

Among the wealth of mathematical tools applied to algorithmic composition, Markov chains play a unique role as they are still a very popular model probably thanks to their capacity to model and reproduce the statistics of some aspects of musical style (Ames, 1989). Markov-based programs are basically melody-composing programs that choose new notes (states) depending on the previous note (or small set of notes). The Markov state transition probabilities can be entered by hand (equivalent to entering a priori rules), or the rules can be extracted from the analysis of statistical properties of existing music (Assayag and Dubnov, 2002) (Figure 2.3).

One of the most refined and successful example of a style modeling system is EMI (Experiments in Musical Intelligence) by David Cope. It analyzes a database of previous pieces for harmonic relationships, hierarchical information, stylistic traits, and other details and manages to generate new music from it (Cope, 1996). Nevertheless, EMI does not explicitly take into account feedback and interaction in its compositional process.
2. Introduction

![Image of software interface]

Figure 2.3.: M, a popular algorithmic composition software by Cycling 74, uses probabilities entered by the user via a graphical user interface to generate MIDI events.

2.1.3. Interactive Music

With the advent of new programming languages, communication standards and sensing technologies, it has now become possible to design complex real-time music systems that can foster rich interactions between humans and machines (Rowe, 1993; Winkler, 2001; Zicarelli, 2002; Wright, 2005; Puckette, 1996). (Here we understand interaction as “reciprocal action or influence” as defined by Jewell et al. 2001 in the Oxford New Dictionary of American English). The introduction of a perception-action feedback loop in the system allows for real-time “evaluation” and modulation of the musical output that was missing in more traditional non-interactive paradigms. Nowadays, one can “easily” build sensate music composition systems able to analyze external sensor inputs in real-time and use this information as an ingredient of the composition (Figure 2.4).

![Image of Mari Kimura interaction]

Figure 2.4.: Mari Kimura interacting with a computer-based system composed of a real-time audio analyzer listening to her performance on the violin, and a controller driving LEMUR’s GuitarBot (Singer et al., 2003) in real-time.
These two aspects (real-time and sensate) are fundamental properties of a new kind of computer-aided composition systems where the computer-based algorithmic processes, can be modulated by external real-world controls such as gestures, sensor data or even musical input directly. Within this paradigm, the composer/musician is in permanent interaction with the computer via sensors or a musical instrument. The control over the musical output of the system is distributed between the human and the machine.

Emblematic recent examples of complex interactive musician/machine music systems are Pachet’s Continuator (Pachet, 2006b), which explores the concept of reflexive interaction between a musician and the system, or the OMax system based on factor oracles (Assayag et al., 2006) that allows a musician to improvise with the system in real-time.

2.2. Interactive Music Systems

Interactivity has now become a standard feature of multimedia systems that are being used by contemporary artists. As a matter of fact, real-time human/machine interactive music systems have now become omnipresent as both composition and live performance tools. Yet, the term “interactive music system” is often used for many related but different concepts.

2.2.1. Taxonomy

Early conceptualization of interactive music systems have been outlined by Rowe and Winkler in their respective books that still serve as key references (Rowe, 1993; Winkler, 2001).

For Rowe, “Interactive computer music systems are those whose behavior changes in response to musical input. Such responsiveness allows these systems to participate in live performances of both notated and improvised music (Rowe, 1993).”

In this definition, one can note that Rowe only takes into consideration systems that accept musical inputs defined as “huge number of shared assumptions and implied rules based on years of collective experience (Winkler, 2001)”. This is a view founded on standard traditional musical practice. Many examples of augmented or hyperinstruments (Machover and Chung, 1989; Overholt, 2005; Paradiso, 2002) are based on these premises (Figure 2.5).

In this context, Rowe provides a useful framework for the discussion and evaluation of interactive music systems (Rowe, 1993). He proposes a taxonomy along the three main axes of performance type that ranges from strictly following a score to pure improvisation, musical interaction mode which goes from sequenced events to computer-generated events, and playing mode that illustrates how close to an instrument or a human player the system is (Figure 2.6).
2. Introduction

Dan Overholt’s Overtone Violin preserves the tradition of violin technique while augmenting its potential by adding extra sensors and controllers that drive a computer-based music system (Overholt, 2005).

2.2.1.1. Models of Musical Performance

Score-driven systems rely on predetermined events that are triggered at fixed specific points in time depending on the evolution of the input, whereas performance-driven systems do not have a stored representation of the expected input.

Winkler extends Rowe’s definition and proposes four levels of interaction (Winkler, 2001). The conductor model, where the interaction mode similar to that of a symphony orchestra, corresponds to a situation where all the instruments are controlled from a single conductor. In the chamber music model, the overall control of the ensemble can be passed from one lead instrument to another as the musical piece evolves. The improvisational model corresponds to a jazz combo situation where all the instruments are in control of the performance and the musical material, while sharing a fixed common global musical structure, and the free improvisation model is like the improvisation model but without a fixed structure to rely on.

2.2.1.2. Modes of Interaction

Once the musical input to the interactive system is detected and analyzed, the musical response can follow three main strategies. Generative methods apply different sets of rules to produce a musical output from some stored original material, whereas sequenced methods use prerecorded fragments of music. Finally, transformative methods apply transformations to the existing or live musical material based on the change of input values.
2.2. Interactive Music Systems

Figure 2.6.: Rowe’s classification of interactive music systems relies on the three axes of performance, interaction and playing modes (Rowe, 1993).

2.2.1.3. Playing Modes

In the instrument mode, the performance gestures from a human player are analyzed and sent to the system. In that case, the system is an extension of the human performer. On the other hand, in the player mode, the system itself has a behavior of its own, a personality.

2.2.2. Limitations

2.2.2.1. Reaction vs. Interaction

The interaction between a human and a system or two systems is a process that includes both control and feedback, where the real-world actions are interpreted into the virtual domain of the system (Bongers, 2000). If some parts of the interaction loop are missing (for instance the cognitive level in Figure 2.7), the system becomes only a reactive (vs. interactive) system. In most of the human/computer musical systems, the human agent interacts whereas the machine reacts. As a matter of fact, although the term interactivity is widely used in the new media arts, most systems are simply reactive systems (Bongers, 2000).

2.2.2.2. Multimodal Inputs

Within Rowe and Winkler’s frameworks (Rowe, 1993; Winkler, 2001), the emphasis is put on the interaction between a musician and the interactive music system. The interaction is mediated either via a new musical interface or via a pre-existing musical
2. Introduction

Figure 2.7: **Human machine interaction** must also include a cognitive component at the machine side to be really considered interactive (adapted from Bongers 2000)

Instrument. This approach is anchored in the history of Western classical music performance. However, with new sensor technology, one can extend the possibilities of traditional instruments by creating new interactive music systems based on novel modes of musical interaction. These systems can generate musical output from inputs which are not necessarily musical (for instance they could be gestures, colors, spatial behaviors, etc).

2.2.2.3. Sense-Think-Act

The framework proposed by Rowe to analyze and design musical systems relies mainly on what he calls the sensing-processing-response paradigm. This corresponds to what is more commonly called the sense-think-act paradigm in robotics and cognitive science (Pfeifer and Scheier, 2001). It is a classical cognitive science approach to modeling artificial systems, where the different modules (e.g. perception, memory, action) are studied separately. Perceptual modules generate symbols representing the world, those symbols are stored in memory and some internal processes use these symbols to plan actions in the external world. This approach has since been challenged by modern embodied cognitive science, which emphasizes the crucial role of the perception-action or sensory-motor loop as well as the interaction of the system with its environment (Pfeifer and Scheier, 2001; Brooks, 1990; Verschure et al., 2003).

2.3. Designing Modern Interactive Music Systems:

2.3.1. A Cognitivist Perspective

A look at the evolution of our understanding of cognitive systems put in parallel with the evolution of composition practices (which do not necessarily rely on computer technology), gives a particularly interesting perspective on the limitations of most actual interactive music systems.
2.3. Designing Modern Interactive Music Systems:

![Diagram of Sense-think-act interpretation of standard robotics tasks]

2.3.1.1. The Classical View: Internal Representations, Centralized Control, Disembodiment

The classical approach to cognitive science assumes that external behavior is mediated by *internal representations* (Fodor, 1975) and that cognition is basically the manipulation of these mental representations by *sets of rules*. It mainly relies on the *sense-think-act* framework (Pfeifer and Scheier, 2001), where future actions are planned according to perceptual information.

Interestingly enough, a parallel can be drawn between classical cognitive science and the development of classical music which also heavily relies on the use of formal structures. It puts the emphasis on *internal processes* (composition theory) to the detriment of the environment or the body, with a *centralized control* of the performance (the conductor) Dawson et al. 2010.

*Disembodiment* in classical music composition can be seen at several levels. Firstly, by training, the composer is used to compose in his head and translate his mental representations into an abstract musical representation: the score. Secondly, the score is traditionally interpreted live by the orchestra’s conductor who “controls” the main aspects of the musical interpretation, whereas the orchestra musicians themselves are left with a relatively reduced interpretative freedom. Moreover, the role of audience as an active actor of a musical performance is mostly neglected.

2.3.1.2. The Modern View: Emergence, Parallelism, Embodiment

An alternative to classical cognitive science is the connectionist approach that builds biologically plausible systems using neural networks. Unlike more traditional digital computation models based on serial processing and explicit manipulation of symbols, connectionist networks allow for fast *parallel computation*. Moreover, it does not rely on explicit rules but on *emergent phenomena* stemming from the interaction between simple neural units. Another related approach, called *embodied cognitive science*, put the emphasis on the influence of the *environment* on internal processes. In some sense it replaced the view of cognition as a representation by the view that cognition is an active process involving an agent acting in the environment. Consequently, the complexity of a generated structure is not the result of the complexity of the underlying system only,
but partly due to the complexity of its environment (Simon, 1981; Brooks, 1990; Dawson et al., 2010).

Musical counterparts of some of these ideas can be found in American experimental music and most notably in John Cage’s work. For instance, the famous 4’33” silent piece transposes the focus of the composition from a strict interpretation of the composer’s score to the perception and interpretation of the audience itself. The piece is shaped by the noise in the audience, the acoustics of the performing hall, the reaction of the environment. Cage also made heavy use of probabilities, and chance operations to compose some of his pieces. For instance he delegated the “central control” approach of traditional composers to the aleatory rules of the traditional Chinese I Ching divination system in Music of Changes (Figure 2.9).

Figure 2.9.: An excerpt of Music of Changes (1951) by composer John Cage. In this piece, most of note durations, dynamics, tempi and density are chosen out of the symbols of the I Ching divination system (also called book of changes) by threefold coin flipping.

Another interesting aspect of American experimental music is how minimalist music composers managed to create complexity from small initial variations of basic musical material. This can be directly put into relation with the work of Braintenberg on robot/vehicles which appear to have seemingly intelligent behaviors while being governed by extremely simple laws (Braitenberg, 1984). A striking example is the use of phase delays in compositions by Steve Reich. In Piano Phase, Reich mimics with two pianists the effect of de-phasing two tapes playing the same material. Even if the initial pitch material and the phasing process are simple, the combination of both gives rise to the emergence of a complex and interesting musical piece mediated by the listener’s perception.

A good illustration of situatedness, distributed processing, and emergence principles
applied to music composition is the piece *In C*, by Terry Riley. In this piece, musicians are given a set of pitch sequences composed in advance, but each musician is left in charge of choosing when to start playing and repeating these sequences (Figure 2.10). The piece is formed by the combination of decisions from each independent musician who makes her own decision based on the collective musical output that emerges from all the possible variations.

![The score of “In C” by Terry Riley. In this piece, each musician can choose when to start playing and repeating the different patterns.](image)

Following recent evolution of our understanding of cognitive systems, we want to emphasize the crucial role of *emergence, distributed processes* and *situatedness* (as opposed to rule-based, serial, central, internal models) in the design of interactive music composition systems.

### 2.3.2. Human in the Loop

The feedback loop between the environment and the system is a fundamental aspect of embodied systems’ modeling. In the interactive music paradigm, the interaction takes place between a music generation system and a human participant (whether she is actively playing with the system, or simply listening to its musical output).
2. Introduction

2.3.2.1. Flow, Creativity, and Well-being

In the context of an interaction between a music system and the user, one relevant aspect is personal enjoyment, excitement and well-being as described in the theory of flow by (Csikszentmihalyi, 1991). As a matter of fact, flow and creativity have been found to be related as a result of musical interaction (MacDonald et al., 2006; Pachet, 2006b).

Csikszentmihalyi’s theory of Flow (Csikszentmihalyi, 1991) is an attempt at understanding and describing the state of Flow (or optimal experience) experienced by creative people. It takes a subjective viewpoint on the problem and describes creativity as a personal feeling of creating something new and interesting in a specific context of production. One interesting aspect of the theory of Flow is that it relates creativity to a certain emotional state obtained through an interactive process (Fritz and Avsec, 2007; Csikszentmihalyi and Hunter, 2003).

This raises the question of the nature of the human feedback that is injected in a given interactive music system. Indeed, Csikszentmihalyi’s theory suggests that the feedback should convey information about the emotional state of the human interactor in order to create an interesting flow-like interaction. This indicates that the design of appropriate feedback that takes into account the mental states of the participant plays a major role in the success of an interactive creative system (Figure 2.11).

![Flow Diagram](image)

Figure 2.11.: Flow is influenced by perceived challenge and skill levels (adapted from Csikszentmihalyi 1991)

2.3.2.2. Interfaces for Musical Expression

**Gesture-based Interfaces** The advent of new sensing technologies has fostered the development of new kinds of interfaces for musical expression. Graphical user interfaces, tangible interfaces, gestural interfaces have now become omnipresent in the design of live music performance or compositions (Paradiso, 2002). For example, graphical-based software such as IanniX (Coduys and Ferry, 2004b) or IXI (Magnusson, 2005) propose...
new types of complex multidimensional multimedia scores that are controlled with a computer mouse (Figure 2.12). Manipulating musical material with a mouse can be somewhat counter-intuitive and a wealth of novel gesture-based interfaces for music expression have been devised over the last decades. One famous example is the project “The Hands”, created by Waisvisz. “The Hands” is a gestural interface that converts movements of the hands, fingers and arms into sound (Krefeld and Waisvisz, 1990). Similarly, “The Very Nervous System” created by Rokeby transforms dance movements into sonic events (Rokeby, 1998). Machover in his large scale Brain Opera project devised a variety of novel interfaces used for the Mind Forest performance (Paradiso, 1999). More recently, tangible interfaces, such as the Reactable (Jorda et al., 2005), which allows a user to interact with digital information through physical manipulation, have become increasingly popular. Most of these interfaces involve some kind of gestural control and require explicit and conscious control of body movements from the user. They are usually not designed to exploit implicit emotional states of the user (even if this could be done for instance by a specialized and detailed analysis of bodily gestures (De Gelder, 2006)).

Figure 2.12: A) “IanniX” is a graphical editor of complex multi-dimensional and poly-temporal musical scores (Coduys and Ferry, 2004a). B) “The Hands” converts movements of the hands, fingers and arms into sound (Krefeld and Waisvisz, 1990) C) “The Very Nervous System” transforms dance movements into sound (Rokeby, 1998). D) “The Reactable” transforms the manipulation of tangible objects into sound (Jorda et al., 2005).
2. Introduction

**Biosignals** Although the idea is not new (Knapp and Lusted, 1990; Rosenboom, 1989), the past few years have witnessed a growing interest from the computer music community in using physiological data such as heart rate, electrodermal activity, electroencephalogram and respiration to generate or transform sound and music. Thanks to the development of more robust and accurate biosignal technologies, it is now possible to derive implicit emotion-related information from physiological data and use it as an input to interactive music systems.

Heart activity measurement has a long tradition in emotion and media research, where it has been shown to be a valid real-time measure for attention and arousal (Lang, 1990b). Attention evokes short-term (phasic component) deceleration of heart rate, while arousing stimuli accelerates heart rate in longer term (tonic component). Heart rate change has been also shown to reflect stimuli valence. While the heart rate drops initially after presentation of the stimuli due to attention shift, the negative stimuli result in a larger decrease of a longer duration (Bradley and Lang, 2000).

Similarly, the study of brainwaves has a rich history, and different brainwave activities have been shown to correlate with different states. For instance, an increase of energy in the alpha wave frequency typically correlates with states of relaxation (Nunez, 2005).

In the literature, we distinguish three main trend in using biosignals. First is the use of physiology to modulate pre-recorded samples, to directly map physiological data to synthesis parameters, or to control higher level musical structures with parameters extracted from the physiology. Examples of the first category are Microsoft’s MPTrain (Figure 2.13) and Fraunhofer StepMan sensing and music playback device that adapt the tempo of the music being played to the speed and rhythm of joggers’ step, calculated from biosensoric data (Bieber and Diener, 2005; Oliver and Flores-Mangas, 2006). While this approach appears efficient and successful, it allows control over only one simple musical parameter. The creative possibilities are somewhat limited. In other work by (Arslan et al., 2006), the emphasis is put on the signal processing chain for analyzing the physiological data, which in turn is sonified, using ad-hoc experimental mappings. Although raw data sonification can lead to engaging artistic results, these approaches do not use higher-level interpretation of the data to control musical parameters. Finally, musicians and researchers have used physiological data to modulate the activity of groups of predefined musical cells (Hamilton, 2006) containing pitch, meter, rhythm and instrumentation material. This approach allows for interesting and original musical results, but the relation between the emotional information contained in the physiological data and the composer’s intention is usually not clearly investigated. Yet, providing emotion-based physiological interface is highly relevant for a number of applications including music therapy, diagnosis, interactive gaming, and emotion-aware musical instruments.
2.3. Designing Modern Interactive Music Systems:

Figure 2.13.: **MPTrain** is a mobile phone based system that monitors the user’s physiology (namely heart rate and pace with ECG and accelerometer sensors (Cf. Figure A) and select the music according to a desired exercise goal (Cf. Figure B). (Adapted from Oliver and Flores-Mangas 2006).

### 2.3.2.3. Emotional Feedback

Music and its effect on the listener has long been a subject of fascination and scientific exploration from the Greeks speculating on the acoustic properties of the voice (Kivy, 2002) to Musak researcher designing “soothing” elevator music. It has now become an omnipresent part of our day to day life, whether by choice when played on a personal portable music device, or imposed when diffused in malls during shopping hours for instance. Music is well known for affecting human emotional states, and most people enjoy music because of the emotions it evokes. Yet, the relationship between specific musical parameters and emotional responses is not clear. Curiously, although emotions seem to be a crucial aspect of music listening and performance, the scientific literature on music and emotion is scarce if compared to music cognition or perception (Meyer, 1956; Gabrielsson and Lindström, 2001; Le Groux et al., 2008b; Krumhansl, 1997a; Bradley and Lang, 2000; Le Groux and Verschure, 2010b). We believe that in order to be complete, the design of a situated music system should take into consideration the emotional aspects of music. Biosignal interfaces in this respect provide valuable information about the human interactor to the system.

### 2.3.2.4. Perceptually-grounded Control of Music Generation

An important decision in the design of a music system is the question of relevant representations. How do changes in technical parameters relate to an actual change at the perceptual level for the listener? Even if macro-level musical parameters such as pitch, melodies, intensity, rhythm and tempo are quite well understood, the microstructure of a sound, its timbre, is not as easy to handle in an intuitive way. A cognitive
psychology approach to system design allows to define structures, representations and time scales that are perceptually salient and cognitively meaningful. While macro-level musical parameters can be defined, represented and processed quite explicitly, the definition, representation and processing of timbre remains a difficult task that can be tackled thanks to advanced audio signal processing methods.

One of the most important shift in music technology over the last decades was the advent of digital signal processing techniques. Thanks to faster processors, the direct generation of sound waves from compact mathematical representations became reality. Recent years have seen the computer music community focus its efforts in the direction of synthesis of sonic events and the transformation of sound material. Personal computers are now able to synthesize high quality sounds, and sound synthesis software has become largely accessible.

Nevertheless, the use of these tools can be quite intimidating and even counter intuitive for non-technically oriented users. Building new interesting synthesized sounds, and controlling them often requires a high level of technical expertise. One of the current challenges of sound and music computing is to find ways to control synthesis in a natural, intuitive, perceptually meaningful manner. Most of the time the relation between a change of synthesis parameter and its effect on the perception of the synthesized sound is not predictable. Due to the high dimensionality of timbre, the automated control of sound synthesis in a generic interactive music system remains a difficult task. The study of the relationship between changes in synthesis parameters and their perceptual counterpart is a crucial question to address for designing “meaningful” interactive systems.

2.3.3. “Verum ipsum factum”: the Synthetic Method

“Verum et factum convertuntur” or the “the true and the made are convertible” is the motto of the synthetic approach proposed by Giambattista Vico (Vico, 1862) an early eighteenth-century philosopher. The synthetic approach states that meaning and knowledge is a human construction and that the manipulation of parameters and structure of a man-made synthetic artifact helps to understand the underlying model. For Vico the building process itself is a source of knowledge (“understanding by building” (Pfeifer and Scheier, 2001; Verschure, 1998)), as it forces us to think about the role of each element and its interaction with the other parts of the system. In the synthetic approach, the construction of a model precedes behavioral analysis.

Applying the synthetic approach to engineering (sometimes called “forward engineering”) is not as common as the “reverse engineering” methodology, but is a good method to avoid the so-called frame of reference problem (Pfeifer and Scheier, 2001; Verschure, 2002, 1997, 1998). As a matter of fact, when functional analysis (or reverse engineering) is performed, usually all the complexity is assumed to pertain to the cognitive processes, while the role of the environment is underestimated. This is the frame of reference
problem. As a result, it has been argued that theories that are produced via analysis are often more complicated than necessary (Braitenberg, 1984).

“Analysis is more difficult than invention in the sense in which, generally, induction takes more time to perform than deduction: in induction one has to search for the way, whereas in deduction one follows a straightforward path” (Braitenberg, 1984).

When a complex behavior emerges, the synthetic approach allows the researcher to generate simpler explanations because she knows the properties of the components of the system she built. This motivates the choice of a synthetic approach to the study of music perception, cognition and emotion.

2.4. Robot-Based Music Generation

An interesting approach that applied principles of embodied cognition and robotics to music generation was proposed in a project called Roboser (Manzoli and Verschure, 2005). Roboser tackled the problem of music generation using a real-world behaving device (e.g. a robot equipped with sensors) as an input to a MIDI sequencer called Curvasom. This way, the musical output somehow illustrated how the robot experienced the world (Figure 2.14). The robot behavior was controlled by the Distributed Adaptive Control model (DAC) (Verschure et al., 2003), a model of classical and operand conditioning, which is implemented using the real-time neuronal simulation environment IQR (Bernardet et al., 2002). DAC consists of three layers of control, namely the reactive layer, the adaptive layer and the contextual layer. While the reactive layer is a set of prewired reflex loops, the adaptive layer associates co-occurring stimuli. Finally, the contextual layer provides mechanisms for short and long-term memory that retain sequences of perceptions/actions that led to a goal state (for instance reaching a light source). Specific neural states such as exploration, collision or light encounter are used to trigger MIDI tracks or modulate expressive MIDI parameters (volume, tempo, velocity, pitch bend).

The aim of Roboser was to integrate sensory data from the environment in real-time and interface this interpreted sensor data combined with the internal states of the control system to Curvasom. The variation in musical performance was provided by the operational states of the system. The more the robot behaved in the environment, the more it learned about this environment, and started structuring its behavior. In this way, unique emergent behaviors were generated and mapped to musical parameters. Experiments showed that the dynamics of a the real-world robot exploring the environment induced novelty in the fluctuations of sound control parameters (Manzoli and Verschure, 2005).

While the Roboser project paved the way for a new type of interactive music systems based on emergence, parallelism and the interaction with the environment, some important ingredients were still missing in the light of our proposal for a new kind of situated music system. One weakness of Roboser is that the “structure generator”, i.e. the robot,
2. *Introduction*

![Image](image.jpg)

Figure 2.14.: **Roboser** sonifies the behavior of a real-world robot equipped with sensors (A) in an arena composed of wall (obstacles to avoid) and light sources (targets) and colored patches (B, C) (Manzolli and Verschure, 2005).

controlled by DAC (Verschure et al., 2003), behaving in the real world, did not take into account any musical feedback. In this paradigm, from the robot’s perspective, the learning of perception/action sequences depended on the structure of the robot arena only, not at all on the musical output. The music was driven by a fixed one-way mapping from spatial behavior to a reduced set of expressive MIDI parameters (volume, tempo, velocity, pitch bend). There was no interaction between the behavior of the robot and the musical output, as there was no music-related feedback sent to the robot. Hence, the musicality or expressive quality of the musical result was not taken into account by the system. The human listener was not part of this model either and did not contribute any emotional, musical or behavioral feedback. The real-world robot, Curvasom MIDI sequencer, and the listener were somewhat connected but did not really interact. Moreover, Curvasom internal data structures and architecture did not rely on cognitively or perceptually plausible models. It consisted of a simple sequencer that could read fixed, pre-composed MIDI events entered by hand in a text file following an ad hoc protocol. It did not allow for control over the micro level of sound (sound synthesis), nor did it allow to interactively change the basic musical content as the piece evolved. For each session, the musical sequences to be played were precomposed and fixed once for all. Curvasom did not support polyphonic voices, which means musical concepts such as chords or harmonic consonance could not be defined at the level of a single voice. These limitations put some restrictions on the expressive power of the system, as well as on the musical styles that could be generated. The system heavily relied on pre-composed
material. At a more technical level, Roboser was not multi-platform, and the music sequencer Curvasom could only be controlled from the neural simulator IQR (Bernardet and Verschure, 2010), which added unnecessary complexity in situations where specific neural modeling was not deemed necessary. Those limitations are addressed by our new system called the SMuSe described in Chapter 3.

2.5. Summary

The evolution of computer-based music systems has gone from computer-aided composition which transposed the conventional paradigms of music composition to the digital realm, to complex feedback systems that allow for rich multimodal interactions (Section 2.1). Yet, most current composition systems still rely on classic cognitive science paradigms such as internal representations, centralized control and disembodiment (Section 2.3.1). Besides, the crucial role of human emotional feedback is rarely taken into account (Section 2.3.2.3) in the interaction loop. Following a synthetic psychology approach (2.3.3), we believe that current knowledge about music perceptual and cognitive processes should drive the design of interactive composition systems. We propose to build a system that relies on cognitively grounded representations of musical knowledge. In this context, the control of perceptually relevant musical parameters is crucial. Notably, the control of high dimensional timbre remains a difficult open problem (Section 2.3.2.4). Situated robotics research (Section 2.4) is a primary source of inspiration that serves as the basis for the work that is described in the following chapters.
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

We propose to build a novel situated synthetic interactive composition system called the SMuSe (for Situated Music Server) based on basic principles of modern cognitive science and robotics such as parallelism, emergence, situatedness and emotional feedback. We first give a detailed description of the perceptually and cognitively relevant musical features and processes that have been put forward by psychoacoustics and cognitive neuroscience research (Section 3.1). This informs the choice of appropriate musical representations and process architecture. Based on these premises, we propose a software implementation that relies on a hierarchy of musical agents which offers intuitive and distributed control over the memorization, manipulation and generation of musical parameters (Section 3.2). The influence of the environment is taken into account via the integration of the system to diverse complex sensing environments (Section 3.3). Finally, in order to close the interaction loop, we propose different mapping schemes that link the musical output of the system to the emotional responses of the users (Section 3.4). This paradigm is applied to real-world problems and large-scale artistic performances (Section 3.5), and opens interesting perspectives for a synthetic approach to the study of music perception, cognition and emotion (Cf. Chapters 8 and 9).

3.1. Perceptual and Cognitive Models of Musical Representation

Over the last centuries, most composers from the Western classical music tradition have been relying on a representation of music known as staff notation or score (which was first proposed by Guido D’Arezzo in the 11th century). The score traditionally specifies standard musical dimensions such as tempo, meter, notes, rhythm, expressive indications (e.g. crescendo, legato, etc) and instrumentation. Yet, recent advances in music practice and technology have extended the vocabulary of musical expression beyond standard musical dimensions. Indeed, powerful computer music algorithms running on a consumer-level laptop can now directly manipulate the properties of a sound wave. Moreover, the usage of extended playing modes (e.g. the use of subharmonics on a violin (Kimura, 1999) or the use of the instrument’s body for percussion) has become common
practice in contemporary music. As the amount of information needed to describe subtle music modulations or complex production techniques increases, musical scores get more sophisticated; to the point where some scores explicitly specify the synthesis parameters needed for the production of the sound waveform itself (Figure 3.1).

Figure 3.1.: An excerpt from Mycenes Alpha’s (1978) score composed by Iannis Xenakis. It provides a graphical representation of the time-varying audio spectrum that needs to be generated by a computer-based synthesizer.

It is not uncommon for contemporary composers to reinvent a new notation system for each new composition. This raises the question of a generic musical representation. What are the most relevant dimensions of music? What minimal information should be given to a person (or system) so that she can effectively produce a convincing musical piece? Here, we take a cognitive psychology approach to find a set of parameters that is perceptually salient and cognitively meaningful.

Music is a real-world stimulus that is meant to be appreciated by a human listener. It involves a complex set of perceptive and cognitive processes that take place in the nervous system. Research in the neuroscience of music over the past twenty years have taught us that music processes are partly interdependent, are integrated in time and involve memory as well as emotional systems (Koelsch and Siebel, 2005; Peretz and Coltheart, 2003; Peretz and Zatorre, 2005). The study of these processes shed light on the structures and features involved that stand out as being perceptually and cognitively relevant.

One of music’s specificity is that it varies over time. In fact, music perception happens at three different time scales (Snyder, 2000), namely the event fusion level when basic musical events such as pitch, intensity and timbre emerge (∼ 50ms); the melodic and rhythmic grouping when pattern of those basic events are perceived (∼ 5s), and finally the form level (from 5s to 1 hour) that deals with large scale sections of music (see
3.1. Perceptual and Cognitive Models of Musical Representation

Figure 3.2.: **The different levels of sequential grouping** for musical material: event fusion, melodic and rhythmic grouping and formal sectioning (from Snyder 2000)

(Snyder, 2000) for a review of music and memory processes illustrated in Figure 3.2). This hierarchy of three time scales is the basis of the SMuSe’s architecture.

3.1.1. Event Fusion: the Micro Level of Sound

We can distinguish five separable components that contribute to the perception of musical sounds. Namely the **pitch**, the **loudness**, the perceived **duration**, the **timbre**, and the spatial **location** of the sound (Rasch and Plomp, 1999). These perceptual attributes of music have been extensively studied and are associated to fundamental physical properties. For instance, **pitch** relates to the perception of the fundamental frequency of the stimulus, **loudness** relates to the perception of the intensity of the sound wave, **duration** relates to the perception of the time over which the sound is heard, and **location** relates to the interaural delays (difference in arrival time of a sound between two ears) of the sound source. **Timbre** is a more complex multidimensional attribute defined by the American Standards Association as the perceptual attribute of sound that allows a listener to distinguish among sounds that are otherwise equivalent in pitch, loudness and subjective duration (American Standards Association, 1960).

The first level of music perception is the extraction of a sound’s low-level features that arise from the original acoustical signal by mechanism of **feature extraction** and **perceptual binding**. During feature extraction, continuous auditory data is converted into impulse trains at the level of the inner ear, and acoustical features are extracted from this activity by specific neural networks (Bharucha, 2002). Then, during perceptual binding, those
acoustic features are bound together to form a perceptually coherent auditory event (Bregman, 1994) (in terms of pitch, loudness and timbre). This happens at the early processing stage of the auditory pathway (Koelsch and Siebel, 2005) and the time scale is that of the auditory echoic memory (around 50 ms) (Cf. Figure 3.2).

Pitch  The pitch is at the same time related to the fundamental frequency of the audio signal (pitch height) and to its place in a musical scale (chroma). Pitch height vary directly with the frequency in the audible frequency range. Pitch chroma, on the contrary, models the perceptual phenomenon of octave equivalence that is to say the fact that two sounds at different octave are perceived as sharing a common perceptual property. A pitch class is a set of pitches that share the same chroma. The limits of human audition for audible frequencies are approximately from 20 Hz to 20 kHz, but in a musical context, pitch perception is not accurate below 60 Hz or above 5 kHz (Schubert, 1979). The perception of pitch is roughly proportional to the logarithm of frequency and can be modeled by psychoacoustic scales which describe perceptually meaningful pitch intervals. The most commonly used is the semitone scale that serves as the basis of conventional Western music notation (Figure 3.3). Other scales such as Mel, Bark and ERB rate are based on more detailed psychoacoustics models (Moore, 2003; Zwicker and Fastl, 1990).

For the twelve-tone equal temperament, the logarithmic perception of fundamental frequency can be approximated by the following mapping from fundamental frequency \( f \) to pitch \( p \):

\[
p = 69 + 12 \log_2(f/440)
\]  

with a fundamental frequency reference of 440 Hz (Note A4).

Here, pitch can be a real number (i.e. microtones are represented by this model). Integer numbers correspond to the pitches of the standard twelve-tone equal temperament.

In this context, the pitch classes are the pitches that have the same chroma i.e. all the numbers that are separated by multiples of one octave \((p + 12)\) (Table 3.1). A pitch is then defined by a pitch class and its register (e.g. C4 has a pitch class of 0 and a register of 4).

More detailed psychoacoustics models of pitch also take into account the interaction effects between frequency and intensity level of a sound. An increase in level induce a slight increase in pitch (Moore, 2003; Zwicker and Fastl, 1990).

Based on this information, the SMuSe uses a pitch class (chroma) / register (height) symbolic representation of pitch (Table 3.1).

Loudness  Loudness is a measure of the perceptual intensity of a sound. The intensity range of sounds is very large (from the order of \(10^{-10}\) atm to 1 atm). Perception of
3.1. Perceptual and Cognitive Models of Musical Representation

Figure 3.3.: **Frequency of the pitches** for a diatonic scale starting from note C1 (1 semitone = 100 cents). The horizontal grid lines show multiples of frequency C1 (harmonics).

<table>
<thead>
<tr>
<th>Note</th>
<th>C</th>
<th>C#</th>
<th>D</th>
<th>D#</th>
<th>E</th>
<th>F</th>
<th>F#</th>
<th>G</th>
<th>G#</th>
<th>A</th>
<th>A#</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Class</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 3.1.: **Equivalence table** between pitch class and notes.

Loudness is also roughly logarithmic (Figure 3.4). A first rough approach to loudness would be to follow the Fechner law that states that a sensation (loudness) is proportional to the logarithm of the excitation (pressure level of the audio signal).

\[
L_p = 20 \log_{10}(p/p_0)
\]  \hspace{1cm} (3.2)

where \(p_0\) is the reference and \(p\) is the root-mean-square sound pressure being measured.

More detailed psychoacoustics models take into account the interaction effects between features such as intensity, spectral content and time. The scale of phons compensates for the fact that hearing sensitivity varies with frequency (Fletcher and Munson, 1933), and the scale of sones was created to provide a linear scale of loudness (Stevens, 1936). For reference, the usual range of orchestral music varies from around 40 to 100 phons.

The SMuSe uses a numerical representation based on the sone scale to represent musical dynamics (Table 3.2).
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

Figure 3.4.: **Loudness:** to a first approximation, the perception of loudness is logarithmic.

<table>
<thead>
<tr>
<th>Dynamics</th>
<th>ppp</th>
<th>pp</th>
<th>p</th>
<th>mp</th>
<th>mf</th>
<th>f</th>
<th>ff</th>
<th>fff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velocity</td>
<td>16</td>
<td>33</td>
<td>49</td>
<td>64</td>
<td>80</td>
<td>96</td>
<td>112</td>
<td>126</td>
</tr>
<tr>
<td>Phons</td>
<td>40</td>
<td>50</td>
<td>60</td>
<td>65</td>
<td>75</td>
<td>80</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>Sons</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>12</td>
<td>16</td>
<td>32</td>
<td>64</td>
</tr>
</tbody>
</table>

Table 3.2.: **Equivalence table** between dynamics musical notation, MIDI note velocity for standard software instruments and psychoacoustic scales.

**Duration** Musical time is characterized by the use of discrete time intervals for note durations and the continuous modulation of timing for expressive purposes. Humans are able to extract discrete events from a musical performance without specific musical training, and there is evidence for categorical perception of time intervals on a discrete scale (Sternberg et al., 1982; Clarke, 1987). Research on perception of musical event duration has shown that the formation of perceived discrete rhythmic categories from musical signal is consistent within a listener if the duration ratios are simple (Desain and Honing, 2003) (Figure 3.5).

This supports the use of music notations representing duration categories by simple ratios. Consequently, the duration of sound events in the SMuSe is represented by alphanumerical symbols based on conventional Western music rhythm notation (Table 3.3).

**Timbre** Timbre, a perceptual feature notoriously hard to describe, is officially defined by the negative as the perceptual attribute of a sound that allows the listener to distinguish among sounds that are otherwise equivalent in pitch, loudness and subjective duration (American Standards Association, 1960). In spite of these definition issues,
3.1. Perceptual and Cognitive Models of Musical Representation

Figure 3.5.: **Time clumping map**: the three sides of the triangle represent continuous time IOI intervals (physical time) while the colors represent the rhythmic categories (perceived time) perceived by the participants. Gray lines are boundaries, and darker shades indicate a higher proportion of participant identifying the rhythmic category (“Rhythms” legend at the right) (from Desain and Honing 2003).

<table>
<thead>
<tr>
<th>Duration</th>
<th>quarter</th>
<th>eighth</th>
<th>sixteenth</th>
<th>silence</th>
<th>dotted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
<td>4n</td>
<td>8n</td>
<td>16n</td>
<td>-4n</td>
<td>4nd</td>
</tr>
<tr>
<td>Musical Notation</td>
<td><img src="image" alt="Musical Notation" /></td>
<td><img src="image" alt="Musical Notation" /></td>
<td><img src="image" alt="Musical Notation" /></td>
<td><img src="image" alt="Musical Notation" /></td>
<td><img src="image" alt="Musical Notation" /></td>
</tr>
</tbody>
</table>

Table 3.3.: **Example of symbolic representation of note durations**. The same logic applies to all types of notes and silences (e.g. 32th dotted note are notated 32nd).

Recent psychoacoustics research has gone a long way in improving our understanding of timbre and in clarifying its perceptual dimensions. One of the most common method used to define timbre is to describe perceptual dimensions of sounds from abstract descriptors. Those acoustic descriptors can be spectral, temporal or spectro-temporal, and generate a “timbre space” (McAdams et al., 1995b; Grey, 1975; Wessel, 1979). Psychoacoustics experiments determine the timbre space using multidimensional analysis on data taken from experiments where a listener has to judge of the dis/similarity of pairs of sounds. Those pairs of sounds differ only by their timbral characteristics. Multidimensional analysis techniques allow to extract a set of “principal axes” which try to explain well the answers of the listeners. Once the principal axes have been found, acoustical descriptors that correlate with those axes derived from multidimensional analysis have to be found.
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

![Diagram of multidimensional model of timbre space of musical instruments](image)

Figure 3.6: A multidimensional model of timbre space of musical instruments that exhibits three perceptually relevant axes corresponding to attack time, spectral centroid and spectral flux. Each point in space corresponds to a specific musical instrument (from McAdams et al. 1995a).

The most common descriptors unveiled by these psychoacoustics studies (for different musical instrument’s sounds) were the centroid, the spectral spread (Peeters et al., 2000), the log-attack time and the spectral flux (Figure 3.6).

The synthesis of a sound’s timbre from high-level perceptual control parameters is a difficult problem. In the SMuSe, sound synthesis is controlled from the three perceptual axes of brightness, attack time and spectral flux (as well as initial spectral content or modes). This is achieved via machine learning techniques and physically-inspired sound synthesis models described in detail in Part II of this thesis.

Auditory perception is complex, and many other factors are to be taken into account to fully understand the way human perceive sounds (For instance, masking phenomena in the frequency and time domain, and other properties such as those exploited in the MPEG-1 Audio Layer 3 codecs, more commonly know as mp3 format (Brandenburg, 1999)). Nevertheless those first definitions are a good starting point and take into account the most salient characteristics of the auditory system.

### 3.1.2. Musical Expressivity

Pitch, rhythm and form are perceptual constructions that have to do with categorization. They allow the listener to distinguish between different patterns. Still, the categorization process is flexible enough to allow for small deviation in the constituting features. As a
matter of fact, those small variations in pitch, loudness, durations are commonly used by musicians as a means of to add expressivity to a musical piece. This is why the same piece played by distinct musicians, although structurally identical, will sound differently. These nuances are fundamentally of a continuous type (unlike pitch or rhythm) (Snyder, 2000). They cannot be easily remembered by listeners and are typically processed at the level of echoic memory (Raffman, 1993). Musical expressivity is obtained in the SMuSe by real-time modulations of tempo, dynamics, and articulation (a duration factor $f$ that allows to go from legato ($f > 1$) to staccato ($f < 1$)).

3.1.3. Grouping: the Melodic and Rhythmic Level

Music is composed of a combination of multiple auditory events that form patterns, higher-order constructions such as melody, rhythm and harmony. Once auditory features have been extracted, auditory Gestalten are formed following principles such as similarity, continuity and proximity (Carlyon, 2004). In melodic grouping, sequences of pitch are grouped by similarity or range while in rhythmic grouping, sequences are grouped in terms of timing and intensity. The time scale of this mid-level of processing is in the order of 5 s. for a single grouping, i.e. the time limit of auditory short-term memory (STM). An additional constraint of short-term memory is that it can contain around seven different elements on average, which incidentally is the number of different notes in most musical scales (Snyder, 2000).

**Pulse, Tempo** Music is the art of composing with time and musical development relies on a time reference called tactus or pulse. Tactus is originally a Renaissance term that currently designates the perceived pulse of the music, or the rate at which listeners tap their feet to the music (also commonly called beat). It relates to the Western music notion of tempo which measures the speed of variation of the beat (usually defined in ‘beats per minute’).

Experimental studies have found that a majority of people are able to sing songs within 4 % of their nominal tempo (Levitin and Cook, 1996) (i.e. the just-noticeable difference (JND) for tempo variation is about 4%). Moreover, pulse appears to have a stable and accurate representation in the brain, probably involving the cerebellum that acts as a timekeeper (Hazeltine et al., 1997).

EEG studies have also shown that an induced brain activity corresponding to the pulse of the input stimuli could be observed in the gamma frequency range (20-60 Hz). This induced activity (in contrast to the evoked gamma activity) was found to persist even in the absence of the original stimuli, thus suggesting that induced gamma activity could play the role of some sort of independent metronome (Snyder and Large, 2005; Large, 1995).

In the SMuSe, all events are synchronized to a global pulse generated by the tempo / metronome / cerebellum module. Its role is to output pulses at regular intervals. The
value of these intervals is received as an input and defined in beat per minute (BPM) following Western music conventions. This tempo, a global property of the system, can be sent as a reference to all the children modules of SMuSe’s hierarchy (Cf. Section 3.1.7).

Rhythm and Meter  Rhythm and meter both refer to the organization of sound durations over time but are somewhat different notions.

“Meter is how you count time, and rhythm is what you count or what you play while you are counting” (London, 2004)

Perception of rhythm mainly depends on the inter-onset interval (IOI) which is defined as the time between two consecutive onsets (Honing, 2002; Desain and Honing, 2003). Meter is a related notion that can be inferred from information given by the combination of both rhythm and accentuation (accentuation can be emphasized for instance by sound intensity or note density). It relates to how tones are grouped together over time in cycles of strong and weak accents (measures) and induce a metrical hierarchy (Krumhansl, 2000).

Rhythms in SMuSe are defined as sets of durations and silences (e.g. \{8n 16n 16n -4n 4nd 8n 1n\})

![Rhythm Example](image)

Melody  A melody is a succession of tones that makes a coherent pattern. It is traditionally assumed that perception of melody involves some kind of Gestalt perceptual grouping based on relations of similarity, proximity and continuity of pitches. In fact, melodies often follow archetypical pitch contours called melodic schematas (axial contours fluctuate around a central pitch, arch contours first move towards a specific pitch and then move away from it, gap-fill contours start with a large pitch interval and continue with a succession of smaller intervales in the other direction) (Meyer, 1956). Several cross-cultural studies support the so-called implication-realization model which proposes a set of simple governing principles (such as registral direction, intervallic difference, registral return, proximity and closure) that regulate the formation of melodies (Narmour, 1990; Schellenberg, 1996; Krumhansl et al., 1999; Krumhansl, 2001).

Melodies in SMuSe are defined as combination of pitch classes and registers sets (e.g. \{0 5 7 10\} and \{4 4 5 5\})

![Melody Example](image)
3.1. Perceptual and Cognitive Models of Musical Representation

Harmony Harmony is the study of the relationships between simultaneous pitches. It is closely related to the notion of consonance that depends on the amplitude of different partials in the critical bandwidth (roughness), on the perceptual fusion stemming from partials that are close to an harmonic spectrum and on the familiarity to a specific chord (Plomp and Levelt, 1965; Terhardt, 1974). Moreover, since melodies are often heard in a specific harmonic context, listener that are musically trained also tend to infer the harmonic context (called implied harmony) of a melody when the harmonic accompaniment is not there (Holleran et al., 1995; Platt and Racine, 1994; Trainor and Trehub, 1994).

The SMuSe represents harmony with different notation conventions. Chords can be defined as a degree of a specific chosen mode (e.g. degree 2 5 1 of Ionian mode), or following jazz-like notation (e.g. Cb5#9). This allow to define various degrees of consonance or dissonance using well-known Western music notation (e.g. a chord progression for a polyphonic voice would be defined as {2 5 1} or {Dm7 G7 CM7}).

The melodic module in SMuSe can be biased by the harmonic context using an harmonic weight factor. The weight corresponds to the probability of having a note in the melody belonging to the chord being currently generated. When the weight is 1, all the notes in the melody belong to the chord. When the weight is 0, the melody is independent of the harmony implied by the chord.

3.1.4. The Form Level

This level of musical experience concerns large groupings of events over a long period of time (longer than the short-term memory). It deals with entire sequences of music and relates to the structure and limits of long-term memory (LTM). Those groupings are typically too long to be available at once in the working memory. Unlike sequences of melodies and rhythms at the mid-level, sequences at the form level are not necessarily remembered with a time order (e.g. one can retrieve the chorus or the verse of a song independently of their respective order in the song). Sequences in SMuSe can be indexed and retrieved at will.

3.1.5. Auditory Memory

We know that musical perception happens at the three different time scale of event fusion level (~ 50ms), melodic and rhythmic grouping (~ 5s), and form level (> 5s) that correspond to echoic memory, short-term memory and long-term memory respectively (Snyder, 2000).
Perceptual learning research has revealed that we are sensitive to the frequency of occurrence of stimuli encountered in our environments. As a matter of fact, it appears that in the auditory domain, listeners are learning simple statistical properties of pitch sequences. Saffran and colleagues (Saffran et al., 1999) demonstrated this by exposing new born infants to random sequences of six pitch patterns during 21 minutes. Each pattern was made of three notes. After this first exposure, infants had to listen to a new set of sequences, composed of both previously heard and completely new patterns. During this testing phase, the familiarity of the infant with the current pattern being played was checked using the head-turning paradigm (the newborn turns her head towards the stimuli when it sounds familiar). This experiment indicated that infants were more familiar with the pitch sequences they had been the most exposed to (Saffran et al., 1999), showing some evidence of statistical learning of pitch sequences. Similar results have been found concerning the reproduction of rhythmic patterns (Sadakata, 2006). There is evidence for a generalized perceptual learning mechanism for the processing of novel sound patterns based on probabilities (Loui et al., 2009).

It is often thought that listeners (without any specific perfect pitch ability) are not sensitive to absolute pitch, yet studies have found that some form of long-term memory at the pitch level exists. In a study by Levitin (Levitin, 1994), students were asked to sing their favorite songs, and their performance was then compared to the original performance. The study showed that the student’s performances were very close to the pitch level of the original recordings. Similar results were reproduced for tempo (Levitin and Cook, 1996), showing further evidence that auditory memory has an absolute component.

Hence, we used a generic computer data structure that is able to take into account both a probabilistic and absolute view on the memorization of musical features. Musical patterns can be input to the short term memory directly as sequences or as probabilities of occurrence, and it is possible to transfer specific short-term pattern to the long-term memory for later recall, update and deletion (Cf. Figure 3.8). Absolute patterns are represented by sequences of musical parameters values. Probabilistic patterns are represented by probabilities of occurrence of parameters values modeled as Markov chain of order 0, 1 or 2 (Figure 3.7). At a higher “meta” structural level, sequences of sequences are also memorized in the long term memory. These hierarchies of memory structures serve as the basic blocks from which an interactive composition is created using different sequence generation procedures from the stored musical memories (Cf. Section 3.2.4.3 for a description of the sequence generation processes).

Markov chains are stochastic processes which respects the Markov property, namely that future states depend only on present states and are independent of past states. Stochastic processes describe sequences of events that depend on time \( t \). The set of different events is called the state space, while the set of parameters is called the parameter space. If the number of states is finite, then the process is called a chain, and if the current state \( X_t \) of a random variable \( X \) only depends on its previous states, this is a Markov Chain. The transition probability of state \( X \) for a Markov Chain of order one (\( X_{t+1} \) depends
only on the previous state \(X_t\) can be written as such:

\[
P(X_{t+1} \mid X_t, X_3, X_2, X_1) = P(X_{t+1} \mid X_t)
\]  

(3.3)

Similarly, a Markov chain of order \(N\) (depending on \(N\) past observations) is written:

\[
P(X_{t+1} \mid X_t, \ldots, X_3, X_2, X_1) = P(X_{t+1} \mid X_t, X_{t-1}, \ldots, X_{n-N})
\]  

(3.4)

A Markov chain can also be represented as graph or as a transition matrix (Figure 3.7).

![Transition matrix](image)

**Figure 3.7:** Transition matrix: Representation of a first-order Markov illustrating the probability of playing one note at time \(t + 1\) depending the current note at time \(t\) (here with a set of 3 notes \{C, E, G\} only).

SMuSe can use Markov Chains of order zero (simple histogram), one and two to store and generate sequences of musical features. The transition probabilities of the Markov models can be defined a priori or learned from a corpus of musical patterns thanks to the MIDI analysis module (Cf. Section 3.3.1).

### 3.1.6. Music Processing Modules

Research on the brain substrates underlying music processing has switched in the last twenty years from a classical view emphasizing a left-right dichotomy between language (supposedly processed in left hemisphere) and music (respectively right hemisphere) to a modular view where different musical features are processed by different networks in both hemispheres (Altenmüller, 2001). There is some evidence that music processing modules are organized into two parallel but largely independent submodules that deal with pitch content (“What?”) and temporal content (“When?”) respectively (Peretz and Zatorre, 2005; Krumhansl, 2000) (Figure 3.9). This evidence suggests that they can be treated separately in a computational framework. Additionally, studies involving
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

![Diagram of memory structure]

Figure 3.8.: **Memory structure:** Short-term memory sequences (STM) are indexed and stored in long-term memory (LTM) for later recall. These sequences can both represent musical parameter values directly (e.g. a sequence of pitch classes \{0 5 8 10 2\}) or probabilities of occurrence of parameter values (e.g. an histogram \{(0, 0.5) (5, 0.2) (8, 0.1) (10, 0.1) (2, 0.10)\}) (Cf. Section 3.2.4 for a description of the syntax used in SMuSe for memory management)

music-related deficits in neurologically impaired individuals (e.g. subjects with amusias, who can’t recognize melodies anymore) have shown that music faculty is composed of a set of neurally isolable processing components for pitch, loudness and rhythm (Peretz and Coltheart, 2003). The common view is that pitch, rhythm and loudness are first processed separately by the brain to then later form (around 25-50 ms) an impression of unified musical object (Levitin and Tirovolas, 2009) (see Koelsch and Siebel 2005 for a review of the neural basis of music perception). This modularity as well as the three different levels and time scales of auditory memory (sound, groups, structure) form a set of basic principles for designing of a synthetic bio-mimetic music system.

3.1.7. Implications for the Design of SMuSe’s Architecture

These cognitive and perceptual constraints influenced the design of the SMuSe’s architecture. At the low *event fusion level*, SMuSe provides a set of synthesis techniques validated by psychoacoustics tests (Le Groux and Verschure, 2009b; Le Groux et al., 2008b) that give perceptual control over the generation of timbre (Cf. Part II) as well as the use of MIDI information to define basic musical material such as pitch, dynamics and duration. Inspired by previous works on musical performance modeling (Friberg et al., 2006), SMuSe also allows to modulate the expressiveness of music generation by varying parameters such as phrasing, articulation and tempo (Le Groux and Verschure, 2009b).
Figure 3.9.: **Modularity of music processing**: there is strong evidence from neuroscience literature for a modular view of musical perception. First, acoustic information is transformed into neural activity in the cochlea and brainstem. Acoustic features are then extracted in the auditory cortex (ERP around 100ms). MMN correspond to operations of the sensory memory reflected in mismatch negativity. ERAN stands for early right anterior negativity and RATN right anterior-temporal negativity, which occur when violating music syntax. N400 correspond to a semantically related response. (adapted from Koelsch and Siebel 2005).
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

At the medium melodic and rhythmic grouping level, the SMuSe implements various state of the art algorithmic composition tools (e.g. generation of tonal, Brownian and serial series of pitches and rhythms or Markov chains) (Section 3.2.4.2). The time scale of this mid-level of processing is in the order of 5s. for a single grouping, i.e. the time limit of auditory short-term memory.

The form level concerns large groupings of events over a long period of time (longer than the short-term memory). It deals with entire sequences of music and relates to the structure and limits of long-term memory. Influenced by experiments in synthetic epistemology and situated robotics, we rely on the interaction with the environment (Verschure et al., 2003; Verschure, 1998) and on reinforcement learning (Cf. Section 3.4.3 and Chapter 13) to achieve the longer term structure.

The modularity of the music processing chain is also reflected in different SMuSe modules that specifically deal with time (“when”) or material (“what”).

The SMuSe is built on a hierarchical, bio-mimetic and modular architecture. The musical material is represented at three different hierarchical levels, namely event fusion, event grouping and structure corresponding to different memory constraints (section 3.1.7). From the generative point of view, SMuSe modules are divided into time modules (“when”) that generate rhythmic pattern of events and content modules (“what”) that for each time event choose musical material such as pitch and dynamics (Figure 3.10).

3.2. Computational Model: A Society of Musical Agents

The architecture of SMuSe is inspired by neurological evidence. It follows a hierarchical and modular structure (Figure 3.10), and has been implemented as a set of agents using data-flow programming.

3.2.1. Agency

The agent framework is based on the principle that complex tasks can be accomplished through a society of simple cross-connected self-contained agents (Minsky, 1988). Here, an agent is understood as “anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors (Russell and Norvig, 2009)” In the context of cognitive science, this paradigm somehow takes a stand against a unified theory of mind where a diversity of phenomena would be explained by a single set of rules. The claim here is that surprising, complex and emergent results can be obtained through the interaction of simple non-linear agents.

Agents can be of many kinds and associated to various tasks. For instance, reflex agents act on their environment based on their current percepts only, when the environment is
3.2. Computational Model: A Society of Musical Agents

The SMuSe’s architecture is based on a hierarchy of musical agents that deal with tempo, rhythm (when musical events occur) and pitch, dynamics and articulation material (what musical event occur) (the voices). Specific rhythms and melodies are retrieved and stored in the memory by the musical agents. The symbolic musical events are then sent to the synthesizers for audio rendering. The simulation of the room size and sound source location are approximated by a reverb and spatialization module (based on VBAP). Expressive performance is achieved by modulating parameters such as tempo, articulation, dynamics and timbre.
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

fully observable. A model based reflex agent is somewhat similar, but also keeps track of the current state of the world using an internal model that informs the action decision process. Utility-based agents store information concerning the utility of a certain state in order to reach a goal. Another interesting characteristic of an agent is its capacity to learn, improve its behavior over time, this is another class of agent called learning agent (Russell and Norvig, 2009).

The agent framework is particularly suited to building flexible real-time interactive musical systems based on the principles of modularity, real-time interaction and situatedness.

3.2.2. Data-flow Programming

We chose to implement SMuSe’s hierarchy of musical agents in a data-flow programming language called Max/MSP (Zicarelli, 2002). Dataflow programming conceptually models a program as a directed graph of data flowing between operations. This kind of model can easily represent parallel processing which is common in biological systems, and is also convenient to represent an agent-based modular architecture.

Max/MSP (named after one of the father of computer music Max Mathews) is a visual data flow programming environment that is specially designed for building interactive real-time music and multimedia applications. Its inventor, Miller Puckette, implemented the first version in the eighties while working at IRCAM, Paris. It was later released as commercial software by Opcode System (now Cycling 74\(^1\)). Later in the nineties, Miller Puckette released a very popular open source environment based on Max/MSP called Pure Data (Puckette, 1996).

One of the advantages of the Max/MSP environment is that it provides the programmer with a lot of well-documented common control and digital signal processing (DSP) operations. These operations can be combined visually via the use of object/boxes that are laid out in a patcher (Figure 3.11). Additionally, third party developers can also easily extend the language possibilities by programming so-called externals in C using Cycling 74’s API. Externals are compiled C libraries that are dynamically linked at runtime when the correspondent object is loaded in the Max environment. Consequently, SMuSe is built as a mixture of standard Max/MSP processes and ad hoc externals.

Interestingly enough, programming in Max/MSP encourages the programmer to think in a way that is close to how a brain might works. Firstly, since Max/MSP is based on a data-flow paradigm, processes can operate in parallel (e.g. pitch and rhythm processes). Secondly, thanks to the concept of patch abstraction (a Max “meta-patch” that abstracts or include another Max patch), one is able to easily build several layers of processing units (which is somehow similar to the different layers of the cortex). Finally, each process can be connected to every other in a variety of ways (like neurons). Of course, the comparison to neural processes is limited to higher level, organizational processes. For more detailed and realistic neural simulations, there exist dedicated simulators that

\(^1\)http://cycling74.com/
3.2. Computational Model: A Society of Musical Agents

Figure 3.11.: **FM synthesis** can easily be implemented in the Max/MSP programming language as a graph representing the flow of data passing through basic computational units (at control or audio rate) (Zicarelli, 2002).

allow modeling at the neural level (a good example of such a real-time neural simulator is IQR\(^2\) by Bernardet and Verschure 2010).

### 3.2.3. Distributed Control

All the music agents in the SMuSe communicate information via the Open Sound Control (OSC) protocol (Wright, 2005), which means they can be controlled and accessed from anywhere (including over a network) at any time. This gives great flexibility to the system, and allows for shared collaborative compositions where several clients can access and modulate the music server at the same time (Figure 3.12).

We can imagine a new collaborative composition paradigm where every performer builds on what the others have done. The result is a complex sound structure that keeps evolving as long as different performers contribute changes to its current shape. A parallel can be drawn with stigmergetic mechanisms of coordination between social insects like ants (Simon, 1981; Bonabeau et al., 1999; Hutchins and Lintern, 1995) (also see Section 2.3.1.2). In ants colonies, the pheromonal trace left by one ant at a given time is used as a means to communicate and stimulate the action of the others. Hence they manage

\(^{2}\text{http://iqr.sourceforge.net/}\)
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

Figure 3.12.: The SMuSe’s client-server architecture allows for distributed control over the music agents. Different clients on the network can simultaneously modulate the music generated by the server.

to collectively build complex networks of trails towards food sources. As the ant colony matures, the ants appear smarter, because their behaviors are more efficient. But this is because “the environment is not the same. Generations of ants have left their marks on the beach, and now a dumb ant has been made to appear smart through its simple interaction with the residua of the history of its ancestor’s actions” (Hutchins and Lintern 1995, p. 169).

Similarly, in a collective music paradigm powered by an OSC client/server architecture, one performer leaves a musical trace to the shared composition, which in turn stimulates the other co-performers to react and build on top of it.

3.2.4. Concurrent and On-the-fly Control of Musical Processes

We have proposed a biologically inspired memory and process architecture for the SMuSe as well as a computational model based on software agents. The OSC communication protocol allows to easily send text-based control commands to specific agents in SMuSe’s hierarchy. The SMuSe offers flexible and intuitive time-based concurrent and on-the-fly control of musical processes. Hereafter we give several examples of OSC-based control syntax for some of the most common musical parameters (a more detailed documentation, tutorials, examples and sound excerpts can be found on the SMuSe’s webpage\(^3\)).

3.2.4.1. Addressing

The different musical agents in the SMuSe all have a specific ID/address where to receive commands and data. The addresses are divided into /global (affecting the whole hierarchy), /voice/n/ (affecting specific voices), and /synth/n/ (affecting specific sound

\(^3\)http://www.dtic.upf.edu/~slegroux/thesmuse.html
3.2. Computational Model: A Society of Musical Agents

The OSC syntax supports regular expressions which allows to address several modules at the same time with a compact syntax.

For instance, a global property such as `tempo` can be sent as a reference to all the rhythm cells that are children in the SMuSe’s hierarchy of agents as such:

```
/global/tempo 120
```

If now we want to address a specific voice and have it slightly delayed (in ms) from the original global pulse, the syntax will be:

```
/voice1/delay 200
```

To send the same delay value to voices 2 to 5 we can use a regular expressions:

```
/voice[2-5]/delay 400
```

3.2.4.2. Memories of Musical Patterns

Patterns of perceptually-grounded musical features (carefully chosen and defined in Section 3.1) can be sent to the short-term (STM) and long-term memory (LTM) modules at any moment in time. By default, material is sent to the current STM or working memory (Cf. Section 3.1.5). When replacing an old STM by a new STM, events are automatically beat-synchronized on the next beat.

```
/* Fill up the STM */
/voice1/rhythm/pattern 4n 4n 8n 16n 16n 4n
/voice1/pitch/pattern 0 0 5 7 10 0
/voice1/pitch/register 4 5 4 4 4 5
/voice1/velocity 12 12 16 16 12 32
```

```
/* Store two pitch sequences in LTM with indices 1 & 2 */
/voice1/pitch/track 1 0 0 5 7 10 0
/voice1/pitch/track 2 9 2 9 7 5 0
/* Retrieve pitch class sequence 1 from LTM */
/voice1/pitch/track 2
```

```
\[\text{Music notation for STM and LTM sequences.}\]
```
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

3.2.4.3. Musical Sequence Generation

Once we have declared a global architecture of communication between agents and how we store the musical material in the system’s memory, the issue at stake is to generate musical sequences from this material. In the SMuSe, the generation of music is modulated by the interaction of the system with its environment (or human feedback) via explicit rules or learning (Cf. Section 3.4), but the system also makes use of autonomous internal agents that are able to retrieve chunks of memory information (whether absolute, or probabilistic events) to generate musical content from it. For this purpose, the action selection agents provide mechanisms that are available for every musical dimension.

Selection principles, a term inspired by the reflexions of Koenig on serial music and algorithmic composition (Laske, 1981), refer to the actions taken by the system to generate musical events using the available short and long term memory content. These actions can be deterministic (e.g. playback of a stored sequence) or based on probability of occurrence of specific events (series, Markov chains, random events). This allows for an hybrid approach to algorithmic composition where complex stochastic processes are mixed with more deterministic repeating patterns (Cf. Table 3.4). Additionally, the perceptual presence of specific voices can be modulated via a rest factor. When the rest factor is high, the voice is paused for longer times, and is perceived in the background. If the rest factor is low, the voice plays long phrases with only short pauses and is perceived as being in the foreground (Essl, 1995).

By default, the patterns in the current short-term memory are played sequentially in a loop and all the voices are automatically synchronized to the beat. When the command “start” is sent, the voice starts playing on the next beat.

/* Low rest factor: voice is in the foreground*/
/voice1/rest 0
/* Rhythmic events will be generated at random */
/voice1/rhythm/pattern 4n 8n 4n 16n
/voice1/rhythm/selection random
/voice1/start 1
/* Pitch events {2,5,7} are weighted by probabilities {0.2,0.6,0.2}*/
/voice1/pitch/selection markov0
/voice1/pitch/pattern 2 0.2 5 0.6 7 0.2

3.2.4.4. Modulators

Expressivity parameters such as articulation, tempo and dynamics can be continuously accessed and modulated:

/* voice1 articulation values are shortened by factor 0.2 (staccato)*/
3.3. Human in the Loop

<table>
<thead>
<tr>
<th>Selection Principle</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>The order of selection follows the initial sequence. (The sequence in memory is “played back”).</td>
</tr>
<tr>
<td>Inverse</td>
<td>The elements of the original are selected starting from the end. (The sequence in memory is played backward).</td>
</tr>
<tr>
<td>Markov</td>
<td>The elements of a sequence are chosen based on state transition probabilities</td>
</tr>
<tr>
<td>Series</td>
<td>Uniform random choice between elements of the pattern without repetition. If an element of the pattern has already been selected, it can’t be selected again until all the other elements have been selected.</td>
</tr>
<tr>
<td>Aleatory</td>
<td>Elements of the sequence are chosen randomly.</td>
</tr>
</tbody>
</table>

Table 3.4: The action selection agents choose to play specific musical elements stored in the current working STM following deterministic or stochastic selection principles.

/voice1/modulate/articulation 0.2
/* voice2 articulation values are lengthened (legato)*/
/voice2/modulate/articulation 1.7
/* voice1 dynamics are softened by a factor 0.3
/voice1/modulate/dynamics 0.3
/* tempo for all voices is divided by 2
/*/modulate/tempo 0.5

3.3. Human in the Loop

The SMuSe is based on perceptually and cognitively plausible principles and processes implemented in software. This allows for intuitive generation and storage of musical events as well as highly flexible and distributed control of musical material. One crucial aspect of situated system design is the interaction between a cognitive system such as the SMuSe and its environment (Cf. Section 2.3.1.2 and Section 2.3.2).

We tested the SMuSe within different sensing environments ranging from physiology sensors that can provide implicit emotional user interaction (heart rate, electrodermal activity, electroencephalogram), to virtual and mixed-reality sensors for behavioral interaction (camera, gazers, lasers, pressure sensitive floors) and finally MIDI and audio (microphone) for direct musical interaction (Figure 3.13). SMuSe integrates sensory data from the environment (that conveys information about the human participants interacting with the system) in real time and send this interpreted data to the music generation processes after appropriate fixed or learned musical mappings (Section 3.4).
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

Figure 3.13.: SMuSe’s environment: the SMuSe can interact with its environment through different sensors such as biosignals, camera, gazers, lasers, pressure sensitive floor, MIDI, audio, but also via OSC commands sent from client applications (such as console terminal, IQR, IanniX graphical score, Torque game engine, etc.) to the music server over the network.

Figure 3.14.: The audio analysis module extracts estimations of pitch, brightness, noisiness, loudness and onsets in real-time from the audio signal recorded via a microphone (Puckette et al., 1998; Jehan and Schoner, 2001).

The initial musical material generated by SMuSe is amplified, transformed and nuanced as the interaction between the system and the participant evolves.

3.3.1. Musical Interaction

SMuSe includes an audio analysis agent that can detect auditory events such as handclaps, background noise, as well as pitch, brightness and loudness of instruments. This agent uses state of the art audio analysis algorithms that estimate perceptually relevant attributes such as pitch, loudness, onset, brightness and noisiness from raw audio input (Puckette et al., 1998; de Cheveigné and Kawahara, 2002; Jehan and Schoner, 2001) (Cf. Figure 3.14).

Additionally, SMuSe is provided with a simple music listening agent that computes transition probabilities for the different dimensions of musical sequences (pitch, velocity,
duration) input to the system via the MIDI protocol. These probabilities can then be re-injected into the generative modules described in Section 3.2.4.3 in order to generate musical material that has the same probabilistic properties than the original sequences. Although basic direct musical interaction is possible, this is not the main focus of the system which focuses on interaction with the environment via emotional feedback.

### 3.3.2. Interaction Based on Spatial Behavior

The SMuSe has been integrated to a mixed-reality space called the XIM (for eXperience Induction Machine) (Bernardet et al., 2010). XIM is equipped with a number of sensors (overhead cameras, gazers, microphones, pressure sensitive floor) that are used to determine its internal states and to map the physical XIM onto its virtual representation. It uses its effectors to influence the behavior of humans and avatars in the mixed-reality space.

The sensors currently used are 3 overhead cameras and frame grabbers, 4 gazers, 3 microphones, and a pressure sensitive floor. They provide information for the accurate tracking of the visitors and can influence the emotions and behaviors of XIM. The effectors are 8 speakers, the SMuSe, 6 video projectors, 8 lightfingers and the floor (Figure 3.15). In that context, SMuSe is the “musical effector” of XIM. Higher level control of the room’s interaction with avatars and visitors is implemented by neuromorphic processes using the neural network simulator IQR (Bernardet and Verschure, 2010). XIM is capable of accurate and robust tracking of movement and people interacting with it (Mathews et al., 2007). The participants’ spatial behaviors in the environment are used by SMuSe as an indicator of the degree of activity in the room. The SMuSe then generates continuous music and soundscapes in real-time to stimulate the interaction between the participants and the mixed-reality space (Cf. Chapter 10 and (Le Groux et al., 2007c) for a more detailed description).

### 3.3.3. Interaction Based on Biosignals

One of the main objectives of SMuSe is to promote and investigate the relationship between perceptually meaningful acoustic cues, musical emotion and bodily and brain responses. It is a well known fact that emotions influence the physiological responses of a person as illustrated by changes in muscle activity, facial expression, skin conductance, heart rate and electroencephalogram. Hence, measuring these changes can provide information about the emotional state of a subject interacting with the SMuSe (Levensons, 1994; Levenson, 1992). Thanks to recent advances in interface technologies, these measurements can now be made non-invasively, without any lesion to the body and used as an input to an interactive system.

Here, we present a few biosignals that have been successfully integrated into the SMuSe in order to provide emotion-related feedback to the system (Figure 3.18). The SMuSe
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

Figure 3.15.: **XIM (the eXperience Induction Machine)** is an intelligent space which has its own goals and expresses internal states. XIM is provided with a set of sensors and effectors. Sensors allow for the tracking of people in the space. Effectors are used to influence the emotional behavior of the visitors (Bernardet and Verschure, 2010).

supports the following hardware: g-tec’s g.MOBIIab+, g.BCIsys and SSVEP, Starlab’s ENOBIO®, Emotiv’s EPOC® and Neurosky Mindset® via OSC communication modules.

Biosignals are measured from the various responses of the nervous system which is composed of the central nervous system, that includes the brain and the spinal cord, and the peripheral nervous system, itself composed of the sensory-somatic nervous system and the autonomic nervous system (ANS).

The somatic nervous system is involved in muscular activity, whereas the ANS, is responsible for glands and organs. The ANS is further composed of the parasympathetic nervous system (dominant when a person is at rest) and the sympathetic nervous system (dominant when a person is active) (Figure 3.17). The different responses of the nervous

4http://www.gtec.at/
5http://starlab.es/products/enobio
6http://www.emotiv.com/apps/epoc/299/
7http://www.neurosky.com/
3.3. Human in the Loop

The nervous system is divided into the central nervous system (brain and spinal cord) and the peripheral nervous system (sensory-somatic nervous system and autonomic nervous system).

The nervous system can be analyzed and used as indicators of specific emotional states.

3.3.3.1. Heart Rate (HR)

The main function of the heart is to pump blood to the rest of the body, but it can also be a source of interesting psychophysiological information. In fact, it has been shown that a relationship exists between heart activity and emotional reactions, motor activities, mental tasks and conditioning (Andreassi, 2006; Cacioppo et al., 2007). In particular, heart rate seems to be a good measure for attention and arousal as well as an indicator of valence of an emotional response (Lang, 1990b).

3.3.3.2. Electrodermal Activity (EDA)

Electrodermal activity corresponds to changes in the skin conductance and has been found to relate to emotional arousal, memory effect and novelty. This measure is often used in emotion and stress research. EDA is mediated by the sympathetic nervous system only and is a good indicator of small variations of arousal and can differentiate between positive and negative emotions (Dawson et al., 2000; Cacioppo et al., 2007).

Using the generative power of SMuSe, we have explored the potential of using physiological data to extract information about emotional states. We studied a set of well-defined sound parameters and showed that a variation of those parameters triggered significantly different physiological responses, corresponding to distinct affective states, which in turn could be used to control high-level musical parameters (Cf. Chapter 12 and Le Groux et al. 2008b for more details).
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

![Image](image.png)

Figure 3.17.: The Autonomic Nervous System is a part of the peripheral nervous system that controls visceral functions (from Brown and Benchmark 1995)

3.3.3.3. Electroencephalogram (EEG)

Electroencephalography devices measure the summed activity of post-synaptic currents in the brain. The electrical voltage of an individual neuron can’t be detected by an EEG electrode placed on the scalp, but a surface EEG reading is the summation of the synchronous activity of thousands of neurons. If a group of neurons fires in synchrony, the activity will result in the measurement of a large signal whereas asynchronous firing will trigger a smaller irregular signal.

Scalp EEG activity can oscillate at different frequencies representing specific rhythmic, synchronized activity: the brain waves (Nunez, 2005), and they have been shown to correlate with different mental states such as sleep, meditation or relaxation.

Humans are able to easily perceive relaxed or agitated states in an auditory stream. In turn, our own state of agitation can now be detected via EEG technologies. We have explored both ideas in the form of a framework for conscious learning of relaxation through sonic feedback generated by SMuSe (Cf. Chapter 11 and Le Groux and Verschure 2009a).

**Brain Computer Interface (BCI)** EEG devices can also be used as interfaces for conscious control. These interfaces are based on the principle that mental activity can lead to observable changes of electrophysiological signals in the brain. These signals can be measured, processed, and later transformed into useful high level messages or...
commands (Wolpaw, 2007; Felton et al., 2007; Guger et al., 2009). Two such measures are the P300 and the SSVEP. The P300 is an Event Related Potential (ERP) that can be detected 300ms after an infrequent stimuli occurs and is used as a control signal. Another type of interface is provided by steady-state visually evoked potentials (SSVEP) triggered by flickering light. This method relies on the fact that when the retina is excited by a flickering light with a frequency > 3.5 Hz, the brain generates activity at the same frequency (Allison et al., 2008; Allison and Pineda, 2003).

The advent of brain-computer interfaces (BCI) now allows us to directly access subjective mental states and express these in the physical world without bodily actions. In the context of an interactive and collaborative live performance called the Multimodal Brain Orchestra (MBO), we have exploited these novel brain-computer technologies to achieve unmediated brain control over music generation and expression. We have developed a disembodied interactive system designed for the generation and modulation of musical material from brain signal only, and performed an interactive “brain quartet” piece based on novel brain computer interface technologies. The MBO shows how novel BCI technologies can be used in a multimodal collaborative context where the performers have volitional control over their mental state and the music generation process (Cf. Chapter 14 and Le Groux et al. 2010b).

3.4. Musical Mappings

We have built a situated cognitive music system that is sensitive to its environment via musical, behavioral and physiological sensors. Thanks to a flexible architecture, the system is able to memorize, combine and generate complex musical structures in real-time. We have described various sensate environments that provide feedback information from the human interactor and allow to close the interaction loop (Cf. Figure 2.7of
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

Section 2.2.2). As proposed in Section 2.3.2, we take an approach that focuses on emotion-related feedback.

3.4.1. Taxonomy

Different types of mappings from the environment to the music generation system parameters can be envisioned. A first distinction is the difference between explicit and implicit mapping (Hunt et al., 2000). Explicit mapping implies that the relationship between the environment and SMuSe’s parameters is directly expressed for instance by means of a mathematical expression. On the other hand, implicit mapping suggests that the mapping is implicitly learned, as the interaction between the user and the system evolves. Another distinction is the difference between dynamic and static mapping. Dynamic mapping is defined as a mapping that evolves in time, and adapt to the changes in data. Examples of these are given in the following paragraphs.

The relationships between the environment and the SMuSe can also be interpreted in terms of complexity. One can distinguish between complex mappings and simple mappings. Complex mappings are defined as mappings from many sensors to many music generation parameters (Hunt et al., 2000). On the contrary, simple mappings are straightforward one-to-one relationships between one control parameter and one sound generation parameter.

In the context of SMuSe’s emotional mappings, the relationship is mostly a complex one that associates a bi or tridimensional emotional space (on the scales of arousal, valence, dominance) to the many musical parameters of SMuSe (pitch, rhythm, articulation, density, tempo, etc.).

3.4.2. Explicit Mappings

From advanced behavioral and physiological interfaces, we can infer emotion-related information about the interaction with SMuSe. One possible way to take this feedback into account is to design fixed mappings based on previous results from experimental psychology studies that have investigated the emotional responses to specific musical parameters. This a priori knowledge can be used to drive the choice of musical parameters depending on the difference between the goal emotion to be expressed or induced, and the emotional state detected by the system via its sensors. A number of reviews have proposed generic relationships between sound parameters and emotional responses (Gabrielson and Juslin, 1996; Juslin and Sloboda, 2001; Juslin and Laukka, 2003). These results serve as the basis for explicit emotional mappings (Cf. Table 3.5) and have been confirmed in the specific context of SMuSe’s parameter space (Cf. Section 8). This explicit design approach is taken by (Le Groux et al., 2007c) and described in Section 10.
3.4. Musical Mappings

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
<th>Emotional expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amplitude envelope</td>
<td>Rounded</td>
<td>Disgust, sadness, fear, boredom, potency (Scherer &amp; Oshinsky, 1977), tenderness, fear, sadness (Justlin, 1997)</td>
</tr>
<tr>
<td></td>
<td>Sharp</td>
<td>Pleasantness, happiness, surprise, activity (Scherer &amp; Oshinsky, 1977), anger (Justlin, 1997)</td>
</tr>
<tr>
<td>Articulation</td>
<td>Staccato</td>
<td>Fear, anger (Justlin, 1997)</td>
</tr>
<tr>
<td></td>
<td>Legato</td>
<td>Tenderness, sadness (Justlin, 1997)</td>
</tr>
<tr>
<td>Harmony</td>
<td>Simple / consonant</td>
<td>Relaxation, tenderness (Lindström, 1997)</td>
</tr>
<tr>
<td></td>
<td>Complex / dissonant</td>
<td>Anger, Tension (Lindström, 1997), tension, fear (Krumhansl, 1996; Krumhansl, 1997)</td>
</tr>
<tr>
<td>Loudness</td>
<td>Loud</td>
<td>Anger (Justlin, 1997)</td>
</tr>
<tr>
<td></td>
<td>Soft</td>
<td>Fear, tenderness, sadness (Justlin, 1997)</td>
</tr>
<tr>
<td>Variation in loudness</td>
<td>Large</td>
<td>Fear (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>Happiness, pleasantness, activity (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td></td>
<td>Rapid changes</td>
<td>Fear (Krumhansl, 1997)</td>
</tr>
<tr>
<td>Melodic range</td>
<td>Wide</td>
<td>Fear (Krumhansl, 1997), joy (Bakewell &amp; Thompson, 1999)</td>
</tr>
<tr>
<td></td>
<td>Narrow</td>
<td>Sadness (Bakewell &amp; Thompson, 1999)</td>
</tr>
<tr>
<td>Mode</td>
<td>Major</td>
<td>Happiness (Scherer &amp; Oshinsky, 1977; Krumhansl, 1997)</td>
</tr>
<tr>
<td></td>
<td>Minor</td>
<td>Sadness (Krumhansl, 1997), disgust, anger (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td>Pitch level</td>
<td>High</td>
<td>Surprise, potency, anger, fear, activity (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>Boredom, pleasantness, sadness (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td>Tempo</td>
<td>Fast</td>
<td>Activity, surprise, happiness, pleasantness, potency, fear, anger (Scherer &amp; Oshinsky, 1977), happiness (Krumhansl, 1997)</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>Sadness, boredom, disgust (Scherer &amp; Oshinsky, 1977), sadness, tenderness (Justlin, 1997), sadness (Krumhansl, 1997)</td>
</tr>
<tr>
<td>Timbres</td>
<td>Few harmonics</td>
<td>Pleasantness, boredom, happiness, sadness (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td></td>
<td>Many harmonics</td>
<td>Potency, anger, disgust, fear, activity, surprise (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td></td>
<td>Soft</td>
<td>Tenderness, sadness (Justlin, 1997)</td>
</tr>
<tr>
<td></td>
<td>Sharp</td>
<td>Anger (Justlin, 1997)</td>
</tr>
</tbody>
</table>

Table 3.5.: The relationship between musical parameters and perceived emotion. A summary of the main results. (adapted from Gabrielsson and Lindström 2001)
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

The reinforcement-based music system is composed of three main components: the music engine (the SMuSe), the reinforcement learning agent and the listener who provides the reward signal.

3.4.3. Adaptive Mappings

In cases where the relationship between the change of musical parameters and the emotional response is not clear, explicit mappings is not possible. One solution in SMuSe is to use a Reinforcement Learning (RL) agent that learns to adapt the choice of musical parameters based on the interaction of the system with the environment (Sutton and Barto, 1998). Reinforcement learning is a biologically plausible learning algorithm particularly suited to an explorative and adaptive approach to mapping as it tries to find a sequence of parameter changes that optimizes a reward function (which for instance relates to a state of emotional stress) (Le Groux and Verschure, 2010a).

Reinforcement learning (as well as agent-based technology) has already been used in various musical systems and most notably for improving real time automatic improvisation (Assayag et al., 2006; Franklin and Manfredi, 2002; Thom, 2000; Collins, 2008). Musical systems that have used reinforcement learning can roughly be divided into three main categories based on the choice of the reward characterizing the quality of musical actions. In one scenario the reward is defined to match internal goals (a set of rules for instance), in another scenario it can be given by the audience (a like/dislike criterion), or else it is based on some notion of musical style imitation (Collins, 2008). Unlike most previous examples where the reward relates to some predefined musical rules or quality of improvisation, we are interested in the emotional feedback from the listener (Figure 3.19).

This approach contrasts with expert systems such as the KTH rule system (Friberg et al., 2006; Friberg, 2006) that can modulate the expressivity of music by applying a set of predefined rules inferred from previous extensive music and performance analysis. Here, we propose a paradigm where the system learns to autonomously tune its own parameters in function of the desired reward function (some emotional feedback) without using any a-priori musical rule.

Interestingly enough, the biological validity of RL is supported by numerous studies in
3.4. Musical Mappings

Figure 3.20.: The agent-environment interaction in the reinforcement learning framework (from Sutton and Barto 1998)

psychology and neuroscience that found various examples of reinforcement learning in animal behavior (e.g. foraging behavior of bees (Montague et al., 1995) or the dopamine system in primate brains (Schultz et al., 1997)).

3.4.3.1. Reinforcement Learning Formalism

Our goal is to teach our musical agent to choose a sequence of musical gestures (choice of musical parameters) that will optimize an emotional state perceived by the listener. This can be modeled as an active reinforcement learning problem where the learning agent must decide what musical action to take depending on the emotional feedback given by the listener in real-time (Figure 3.19). The agent is implemented as a Max/MSP external in C++, based on RLKit and the Flext framework.\(^8\)

The interaction between the agent and its environment can be formalized as a Markov Decision Process (MDP) where (Sutton and Barto, 1998):

- at each discrete time \(t\), the agent observes the environment’s state \(s_t \in S\), where \(S\) is the set of possible states (in our case the musical parameters driving the generation of music).
- it selects an action \(a_t \in A(s_t)\), where \(A(s_t)\) is the set of actions available in state \(s_t\) (here, the actions correspond to an increase or decrease of the musical parameter value).
- the action is performed and a time step later the agent receives a reward \(r_{t+1} \in \mathbb{R}\) and reaches a new state \(s_{t+1}\) (the reward is given by the listener’s emotional state).
- at time \(t\) the policy is a mapping \(\pi_t(s,a)\) defined as the probability that \(a_t = a\) if \(s_t = s\) and the agent updates its policy as a result of experience.

3.4.3.2. Returns

The agent acts upon the environment following some policy \(\pi\). The change in the environment introduced by the agent’s actions is communicated via the reinforcement\(^8\)

\[http://puredata.info/Members/thomas/flext/\]
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

signal $r$. The goal of the agent is to maximize not only the immediate reward it receives, but also the reward in the long run. The discounted return $R_t$ is defined as:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$  \hspace{1cm} (3.5)

where $0 \leq \gamma \leq 1$ is the discount rate that determines the present value of future rewards. If $\gamma = 0$, the agent only maximizes immediate rewards. In other words, $\gamma$ defines the importance of future rewards for an action (increasing or decreasing a specific musical parameter).

3.4.3.3. Value Functions

Value functions of states or state-action pairs are functions that estimate how good (in terms of future rewards) it is for an agent to be in a given state (or to perform a given action in a given state).

$V^\pi(s)$ is the state-value function for policy $\pi$. It gives the value of a state $s$ under a policy $\pi$, or the expected return when starting in $s$ and following $\pi$. For MDPs we have:

$$V^\pi(s) = E_\pi\{R_t|s_t = s\} = E_\pi\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s\}$$  \hspace{1cm} (3.6)

$Q^\pi(s, a)$, or action-value function for policy $\pi$, gives the value of taking action $a$ in a state $s$ under a policy $\pi$.

$$Q^\pi(s, a) = E_\pi\{R_t|s_t = s, a_t = a\} = E_\pi\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s, a_t = a\}$$  \hspace{1cm} (3.7)

We define as optimal policies the ones that give higher expected return than all the others. Thus, $V^*(s) = \max_\pi V^\pi(s)$, and $Q^*(s, a) = \max_\pi Q^\pi(s, a)$ which gives

$$Q^*(s, a) = E\{r_{t+1} + \gamma V^*(s_{t+1})|s_t = s, a_t = a\}$$  \hspace{1cm} (3.8)

3.4.3.4. Value Function Estimation

Temporal Difference (TD) Prediction  Several methods can be used to evaluate the value functions. We chose TD learning methods over Monte Carlo methods as they allow for online incremental learning. With Monte Carlo methods, one must wait until
3.4. Musical Mappings

the end of an episode whereas with TD, one need to wait only one time step. The TD learning update rule for \( V^* \) the estimate of \( V \) is given by:

\[
V(s_t) \leftarrow V(s_t) + \alpha [r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]
\]  
(3.9)

where \( \alpha \) is the step-size parameter or learning rate. It controls how fast the algorithm will adapt.

**Sarsa TD Control**  For the transitions from state-action pairs we use a method similar to TD learning called Sarsa on-policy control. On-policy methods try to improve the policy that is used to make decision. The update rule is given by:

\[
Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]
\]  
(3.10)

**Memory: Eligibility Traces (Sarsa(\( \lambda \)))**  An eligibility trace is a temporary memory of the occurrence of an event.

We define \( e_t(s, a) \) the trace of the state-action pair \( s, a \) at time \( t \). At each step, the traces for all states decay by \( \gamma \lambda \) and the eligibility trace for the state visited is incremented. \( \lambda \) represent the trace decay. It acts as a memory and sets the exponential decay of a reward based on previous context.

\[
e_t(s, a) = \begin{cases} 
\gamma \lambda e_{t-1}(s, a) + 1 & \text{for } s = s_t, a = a_t \\
\gamma \lambda e_{t-1}(s, a) & \text{if } s \neq s_t
\end{cases}
\]  
(3.11)

we have the update rule

\[
Q_{t+1}(s, a) = Q_t(s, a) + \alpha \delta_t e_t(s, a)
\]  
(3.12)

where

\[
\delta_t = r_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)
\]  
(3.13)

**Action-Value Methods**  For the action-value method, we chose a \( \varepsilon \)-greedy policy. Most of the time it chooses an action that has maximal estimated action value but with probability \( \varepsilon \) it instead select an action at random (Sutton and Barto, 1998).

This reinforcement learning formalism allows the system to dynamically adapts its parameter depending on the user feedback without any fixed a priori mapping (Chapter 13).
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

3.4.3.5. Applications

Within the SMuSe framework, we created an original algorithmic music piece that was modulated by parameters such as articulation, dynamics, tempo and note density, assumed to influence musical tension. The modulation of those parameters was autonomously learned in real-time by a reinforcement learning agent optimizing the reward signal based on the musical tension fed back to the system in real-time by the listener via a slider interface. This real-time learning of musical parameters provided an interesting alternative to more traditional research on music and emotion. We could observe correlations between specific musical parameters and an increase of perceived musical tension (Cf. Chapter 13 and Le Groux and Verschure 2010a for additional details).

3.5. Artistic Realizations

Inspired by previous works on Roboser and Ada, a human accessible space that was visited by over 500,000 people (Eng et al., 2003), we have further explored the purposive construction of interactive installations and performances using the SMuSe system. To name but a few, during the VRoboiser installation (Le Groux et al., 2007d), the sensory inputs (motion, color, distance) of a 3D virtual Khepera robot living in a game-like environment modulated musical parameters in real-time, thus creating a never ending musical soundscape in the spirit of Brain Eno’s “Music for Airports”. The robot was controlled via a joystick by a participant whose decisions were influenced by both the robot’s spatial behavior and the resulting musical output. This provided an essential human and musical feedback loop originally missing in the Roboser paradigm. In another context the SMuSe generated automatic soundscapes and music which reacted to and influenced the spatial behavior of human and avatars in the mixed-reality space called XIM (for eXperience Induction Machine) (Bernardet et al., 2010; Le Groux et al., 2007c) thus emphasizing the role of the environment and interaction on the musical composition. Based on similar premises, Re(PER)curso, an interactive mixed reality performance involving dance, percussion, interactive music and video was presented at the ArtFutura Festival 07 and Museum of Modern Art in Barcelona in the same year. Re(PER)curso was performed in an augmented mixed reality environment, where the physical and the virtual were not overlapping but distinct and continuous. The border between the two environments were the projection screen that acted like a dynamic all seeing bi-directional eye. The performance was composed by several interlaced layers of artistic and technological activities: e.g. the music controlled had three components: a predefined soundscape, the percussionist who performed from a score and the interactive composition system synchronized by SMuSe; the physical actors, the percussionist and the dancer were tracked by a video based active tracking system that in turn controlled an array of moving lights that illuminated the scene. The spatial information from the

[^9]: [http://www.k-team.com/]
3.6. Conclusions

Figure 3.21.: **Artistic realizations:** A) Re(PER)curso (Art Futura and MACBA, Barcelona 2007) B) The Multimodal Brain Orchestra (FET, Prague 2009) C) XIM sonification (SPECS, Barcelona 2007)

stage obtained by the tracking system was also projected onto the virtual world where it modulated the avatar’s behavior allowing it to adjust body position, posture and gaze to the physical world. Re(PER)curso was operated as an autonomous interactive installation that was augmented by two human performers. In 2009, the Brain Orchestra (Le Groux et al., 2010b), a multimodal performance using brain computer interfaces, explored the creative potential of a collection of brains directly interfaced to the world. During the performance, four “brain musicians” were controlling a string quartet generated by the SMuSe using their brain activity alone. The orchestra was conducted by an “emotional conductor”, whose emotional reactions were recorded using biosignal interfaces and fed back to the system. The Brain Orchestra was premiered in Prague for the FET 09 meeting organized by the European Commission.

3.6. Conclusions

The SMuSe illustrates a novel situated approach to music composition systems. It is built on a cognitively plausible architecture that takes into account the different time frames of music processing, and uses an agent framework to model a society of simple distributed musical processes. It takes advantage of its interaction with the environment to go beyond the classic sense-think-act paradigm (Rowe, 1993). It combines cognitively relevant representations with perceptually-grounded sound synthesis techniques (Cf. Part II) and is based on modern data-flow audio programming practices (Puckette,
3. The SMuSe: A Synthetic, Perceptual, Emotive and Cognitive Music System

1996; Puckette et al., 1998). This provides an intuitive, flexible and distributed control environment that can easily generate complex musical structure in real-time. SMuSe can sense its environment via a variety of sensors, notably physiology-based sensors. The analysis and extraction of relevant information from sensor data allows to re-inject emotion-based feedback to the system based on the responses of the human participant. The SMuSe proposes a set of “pre-wired” emotional mappings from emotions to musical parameters grounded on the literature on music and emotion, as well as a reinforcement learning agent that performs online adaptive mapping. The system design and functionalities have been constantly tested and improved, to adapt to different real-world contexts which led to the presentation of several large scale artistic performances. Central to all these examples of externalist aesthetics has been our new paradigm of situated interactive music composition implemented in the SMuSe. It provides a well grounded approach towards the development of advanced synthetic aesthetic systems and a further understanding of the fundamental psychological processes on which it relies.
Part II.

Composing Sounds: Perceptually Grounded Sound Synthesis
4. From Sound Description to Sound Synthesis

In Chapter 3, we described a set of perceptually grounded parameters and cognitively relevant structures used by the SMuSe to control music generation at the macro levels of pitch, scales and rhythms. While these macro-level perceptual parameters can be manipulated and generated by relatively simple computational models, the micro-level control of sound is more problematic. How can we model and control such a highly dimensional and complex concept as timbre? In this chapter we first provide a short overview of the literature on the two complementary domains of sound description and sound generation (Sections 4.2 and 4.3), and then describe several strategies that can be used to bridge the gap between high-level perceptual sound features and low-level synthesis parameters (Section 4.4). More precisely, we present four main approaches to timbre description that originated in music information retrieval research, psychoacoustics, sound synthesis models and music composition. We then introduce the most common types of sound synthesis algorithms. Finally, we analyze four approaches to the high-level control of sound synthesis based on machine learning, concatenative synthesis, spectral synthesis and physically-inspired models. This background information informs the design of the perceptual synthesizers described in detail in Chapters 5 and 6.

4.1. Introduction

Nowadays, computers are already able to synthesize high quality sounds, and sound synthesis software has become largely accessible. Yet, the use of these tools can be quite intimidating and even counter-intuitive for non-technically oriented users. Most of the time, the relation between a change of low-level synthesis parameter and its effect on perception of the synthesized sound is not predictable. Building new interesting sounds with actual synthesizers often requires a high level of expertise (Cf. Figure 4.1). A key issue in order to make rich sound modeling more widely available and easily usable is the control of sound synthesis in a natural, intuitive way. In order to synthesize a sound, one has to precisely define the time-varying trajectories of a (usually large) set of low-level parameters, which due to the high dimensionality of the parameter space (e.g. in additive synthesis) is usually not practical.

We propose to build a model where the desired high-level, perceptually meaningful acoustic properties are used to control the system. This approach is somewhat analog
4. From Sound Description to Sound Synthesis

Figure 4.1.: A) The Buchla 200e analog modular synthesizer, designed by electronic music pioneer Don Buchla allows to generate new timbres by patching different modules together (waveform generator, oscillator, filters, ADSR envelopes, ...). B) The Alchemy Additive Synthesizer software, by Camel Audio, allows to synthesize new sounds by manually editing the parameters of an additive synthesis algorithm (amplitude and frequency of partials). Both synthesizers require a high level of technical expertise to generate new interesting sounds.

to VRML or X3D\(^1\) in the visual domain, where one can describe visual scenes with e.g., shape descriptors, position descriptors or texture descriptors. Our objective is to achieve fine continuous control over sound generation with a synthesis system based on auditory perception. The first step is to define a set relevant controllers. To this purpose, we give an overview of the different schemes that have been used to describe sounds such as those used in physical model synthesis, defined in psychoacoustics or extracted from signal processing techniques.

4.2. Sound Description

4.2.1. Music Information Retrieval: An Engineering Approach

Music Information Retrieval (MIR\(^2\)) is a discipline at the intersection of computer science, information science and music that emerged relatively recently and is driven by a need from the music industry. Its goal is to improve the access, searching and distribution of music from large databases (Downie, 2003). MIR takes an approach to sound description that is motivated by retrieval and classification of audio files (songs, sound effects, ...). The standard methodology is to extract a large set of features from the

---

\(^1\)http://www.web3d.org/

\(^2\)http://www.ismir.net/
audio files and to compute some similarity measures between each feature vector representing an audio file (Wold et al., 2002; Foote, 1997). The initial features are chosen to describe as many aspects of an audio signal as possible (e.g. the temporal shape, the spectral shape, the harmonic content) (Peeters, 2004). More or less elaborated statistics (mean, variance, higher order statistics, ...) are then used to calculate distances between sounds or groups of sounds. The best features are then defined as the ones that allow a classifier to distinguish the best between classes of labelled sounds (Peeters, 2004). Some sound features such as spectral flatness, loudness, Mel frequency cepstrum coefficients (MFCC) and temporal derivatives of log-loudness stood out (Herre et al., 2004), but there is still no general agreement on what are the best MIR sound descriptors. In fact, most of the MIR systems relying on a standard engineering approach have been found to reach a ceiling on performance that comes from a lack of more refined understanding of the perceptual, cognitive and cultural aspects of timbre perception (Aucourtier and Pachet, 2004).

4.2.2. Psychoacoustics

Sounds can be described by four main perceptual components, namely pitch, loudness, duration and timbre (Rasch and Plomp, 1999). To a first degree of approximation, pitch, loudness and duration are unidimensional parameters that relate to fundamental frequency, sound level, and time respectively, and are relatively easy to model and manipulate (Cf. Section 3.1 for a description of the musical representations and manipulation implemented in the SMuSiC music system). The perceptual control of timbre (at the micro-level of sound) poses a more difficult problem as there exist no standard computational models of timbre. Timbre is notoriously difficult to describe (Sandell, 2008) and is usually defined by the negative as the perceptual attribute that allows a listener to distinguish among sounds that are otherwise equivalent in pitch, loudness and subjective duration (American Standards Association, 1960). Yet, the manipulation of the timbre of classical or electronics instruments is a rich area of exploration for composers (Kaltenecker, 2001; Boulez, 1987) even if traditionally ignored by music theorists.

Over the years, several experimental studies of timbre have been taken, leading to disparate but related results. The principal method of investigation consisted in asking the listeners to group pairs of sound by dissimilarity. The judgements were then analyzed with multidimensional scaling techniques that generated a so-called “timbre space”. Multidimensional analysis techniques allow to extract a set of “principal axes” or dimensions that are commonly assumed to model the criteria used by participants to estimate the dissimilarity. The next step is to find the acoustical descriptors that correlate with the dimensions derived from multidimensional analysis. Although various dimensions have been put forward historically (Cf. Figure 4.2), most of the studies on instrumental sounds currently agree on the following three descriptors: spectral centroid or brightness (Grey, 1977), the log-attack time and the spectral flux (McAdams et al., 1995a; Grey, 1975; Iverson and Krumhansl, 1993; Lakatos, 2000; Wessel, 1979). The use of timbre space space for musical composition is discussed in Wessel 1979.
4. From Sound Description to Sound Synthesis

These psychoacoustics dimensions of timbre are interesting as they are general (they can describe any musical sound) and are specifically defined to correspond to a listener’s perception. Yet, it is still not obvious how to control sound synthesis algorithms directly from psychoacoustics descriptors.

Figure 4.2.: The timbre space of Grey and Gordon (Grey, 1975). Dimension I is the spectral energy distribution. Dimension II is timing of the attack and decay. Dimension III is the amount of inharmonic sound in the attack. (BN - Bassoon C1 - E flat Clarinet C2 - B flat Bass Clarinet EH - English Horn FH - French Horn FL - Flute O1 - Oboe O2 - Oboe (different instrument and player) S1 - Cello, muted sul ponticello S2 - Cello S3 - Cello, muted sul tasto TM - Muted Trombone TP - B flat Trumpet X1 - Saxophone, played mf X2 - Saxophone, played p X3 - Soprano Saxophone)

4.2.3. Schaeffer’s Typology: a Composer’s Approach

Schaeffer’s approach is to find sound descriptions that could be used for composing “musique concrete” (Palombini, 1993). Schaeffer’s taxonomy or “solfege of musical objects” (Schaeffer, 1977) is an intuitive description of synthesized or natural sounds, and is virtually applicable to any class of sound. This taxonomy is based on the pair shape/matter. The matter criteria correspond to what we would hear if we could freeze the sound, while the shape is related to the variation of matter in time.

Within the matter criteria category, the mass is related to the perception of pitchiness of a sound (and indeed its spectral distribution). For classification, this criterion can be
4.3. Sound Synthesis

further divided into three distinct subclasses, namely noisy (unpitched), pitched (single pitch), complex (noise and pitch or multiple pitch).

*Harmonic Timbre* is a finer characterization of the mass associated with brightness (or spectral centroid).

Finally, a *grain* is related to the microstructure of the sound. Again it can be subdivided in subtypes: resonant, for non sustained sounds (like a cymbal), rubbing for sustained sounds (bow, breath), and iterative grain (drum roll).

Within the *shape criteria* category, *dynamics* defines the evolution of intensity or amplitude envelope (We can associate this criterion to evolution of loudness). For further classification purposes, several classes such as varying, varying-impulse, varying iterative (several transients), varying-delta, varying-crescendo, varying-decrescendo or other can be defined. *Allure* is related to Amplitude or frequency modulation.

We now come to the definition of the components of the *variation criteria*. The *melodic profile* is the type of pitch variation (for pitched sounds) or frequency-range variation (for noisy sounds). Unvarying, varying-continuous (siren), varying-stepped (piano), and varying chaotic are possible sub-classification of this criterion.

The *mass profile* is the type of mass variation, i.e. varying from noisy to pitched or unvarying.

Along the same line, a more experimental approach proposed by Bismarck 1974 looked for the best semantic field that would describe sounds accurately. In his study, Bismarck asked the subjects to rate sounds on 30 verbal attributes. Based on a multi-dimensional scaling analysis of the semantic-differential data, he found four axes that were statistically significant, namely the dull-sharp, compact-scattered, full-empty and colorful-colorless pairs.

One of the advantages of Schaeffer’s taxonomy is that it is general enough to be able to describe almost any natural or synthetic sound, it is made by and for music composers, and doesn’t rely on excessively technical terms. Nevertheless, it has not been validated by large scale experiments so far, and in fact most of the terms in the taxonomy can actually be associated to psychoacoustics descriptors (harmonic timbre and centroid, allure and loudness, mass and pitch). This suggests that the sound descriptors coming from psychoacoustics studies are the most appropriate both from a perceptual and compositional point of view.

4.3. Sound Synthesis

In our proposal for perceptual control of sound synthesis, the synthesizer has to recognize a group of perceptually relevant controls and take care of the details of the synthesis algorithm parameters. Which sound synthesis technique can we use? In this section, we follow Smith’s taxonomy (Smith, 1991) and review four main synthesis methods, namely abstract algorithms, processed recordings, spectral models, and physical models (Table 4.1).
4. From Sound Description to Sound Synthesis

<table>
<thead>
<tr>
<th>Processed Recording</th>
<th>Spectral Model</th>
<th>Physical Model</th>
<th>Abstract Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concrète</td>
<td>Wavetable F Additive</td>
<td>Ruiz Strings</td>
<td>VCO, VCA, VCF</td>
</tr>
<tr>
<td>Wavetable T Sampling</td>
<td>Phase Vocoder PARSHL</td>
<td>Karplus-Strong Ext.</td>
<td>Some Music V</td>
</tr>
<tr>
<td>Vector Granular Prin. Comp.</td>
<td>Sines+Noise (Serra)</td>
<td>Waveguide Modal</td>
<td>Original FM</td>
</tr>
<tr>
<td>T Wavelet T</td>
<td>Prin. Comp. F Chant VOSIM Risset FM</td>
<td>Cordis-Anima Mosaic</td>
<td>Feedback FM</td>
</tr>
<tr>
<td></td>
<td>VOSIM Risset FM</td>
<td>Brass Chowning FM Waveshaping Phase</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Voice Subtractive LPC Distortion</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inverse FFT Xenakis Line Clusters Karplus-Strong</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1.: Sound Synthesis Taxonomy (adapted from Smith 1991)

4.3.1. Abstract Models

The first group of synthesis methods is called abstract algorithm models. They are usually simple and easy to implement. Nevertheless, they often sound artificial if compared to other more complex synthesis techniques. Abstract models include synthesis schemes such as FM synthesis (Chowning and Bristow, 1987), particularly efficient for the synthesis of bell-like and metallic sounds, wave-shaping (or non-linear distortion) synthesis, and Karplus-Strong algorithm (Jaffe and Smith, 1983) typically used for the synthesis of plucked string or percussive sounds. Abstract algorithms create interesting sounds but using techniques that are not related to the real world sound production mechanisms. They allow for the creation of new arbitrary sounds, and are computationally efficient for average quality synthesis of existing musical instruments.

4.3.2. Processed Recording

The second sound synthesis group is based on processed recordings. It includes simple sampling synthesis, consisting in the direct play-back of the recording of sounds. Based on this principle, wave-table synthesis stores only typical portions of an instrument sound in tables to later loop them and play them back when appropriate (Bristow-Johnson, 1996). Based on similar principles, granular synthesis represents the sound by elementary units or grains, which shape and temporal distribution modulate the synthesized sound (Roads, 1988; Truax, 1988).

Processed recording techniques allow for the reproduction of existing recorded sounds. It is also possible to merge, and morph between recorded sounds. Moreover, short grain of sounds, or recordings can be used to produce new sounds. Sampling techniques are particularly useful for applications demanding high-quality audio rendering, but lack the flexibility of model based techniques.
4.3. Sound Synthesis

4.3.3. Spectral Models

In principle, spectral synthesis techniques can be used to model any kind of sound, as they rely on a general analysis/synthesis framework that can be applied to any audio file. Techniques based on the spectral models include additive synthesis, that models the signal by a sum of weighted sinusoids, the phase vocoder (that can be viewed either as a bank of filters or as a short term Fourier transform analyzer) (Dolson, 1986; Flanagan et al., 1965; Moorer, 1978), source-filter synthesis (Cook, 2002), McAulay-Quatieri algorithm (McAulay and Quatieri, 1986) and Spectral Modeling Synthesis (SMS) (Serra and Smith, 1990), that decomposes any signal into a deterministic part and a stochastic part. SMS is well suited for the simulation and analysis of existing sounds and is adapted to good pitch-shifting and time-scaling modifications. It is able to synthesize a large range of sounds while sounding quite natural (which unfortunately is often not the case for abstract models). Spectral modeling allows to directly modulate parameters of the sound spectrum (frequency, amplitude of the partials) that already have some perceptual relevance.

4.3.4. Physical Models

Physical models model the sound production mechanism of a real musical instrument and offer physically realistic control of sound generation. Physical models regroup modeling techniques that include numerical solving of partial differential equations, source filter modeling, vibrating mass-spring network, modal synthesis and waveguide synthesis (Cook, 1997; van den Doel and Pai, 1996; Cook, 2002; Smith, 1991). These type of models are specially interesting for the simulation and analysis of real physical instruments. Applications requiring high fidelity control can take advantage of the fine physical control characteristics of physical models. One major problem though is that for each instrument, one may need an entirely new model and algorithm to produce a realistic sound. In practice, simplified physical models of different instruments (clarinet, trumpet, violin, flute, percussion, etc) can be implemented quite easily (Cook, 2002). The sounds produced by these dynamic systems are very similar to the corresponding instruments, and there is a direct link between physical controls and the timbre of the synthetic sound. Some perceptual characteristics such as the volume of the object, the material, the elasticity are even directly related to the physical parameters of the modes of vibration (McAdams and Bigand, 1993). Other important factors come from the resonating properties of instruments, as well as from different ways of exciting the instrument (hitting it, bowing it, breathing into it, etc.). Control parameters of physical models are at the same time low-level because they are related to a mathematical model, and high-level in the sense that they are strongly correlated to a perceptual physical reality (Arfib et al., 2003). Thus physical models are a particularly interesting way of generating sounds, as the synthesis parameters are already high-level and intuitive by definition.
4. From Sound Description to Sound Synthesis

4.4. Perceptual Mappings: From Sound Description to Sound Generation

Several mapping techniques have already been proposed to bridge the gap between sound description and sound generation models. They provide different complementary views on how to frame the problem of mapping high-level controls to synthesis. We can distinguish three main trends: the machine learning view, where a model of the timbre is learned automatically from audio data, the concatenative view, where the timbre is “constructed” by the juxtaposition of pre-existing sonic grains, and the signal processing view, where the transformations on timbre are model-dependent and direct applications of the signal model affordances.

One of the first illustrations of the machine learning point of view is found in Lee and Wessel 1992, where a detailed description of a connectionist model for real time control of synthesis is articulated by the authors. The study focused on the transformation of performance gestures into control parameters and emphasized the importance of both the pre-processing of gestural data and the perceptual representation of sound. This initial approach was later compared to a neural network model including a memory based machine learning technique (Wessel et al., 1998). The network described in this second study received fundamental frequency and loudness in its input layer, and was trained to estimate amplitudes and frequencies of the sound partials. A related study using Radial Basis Function neural networks to model speech and music signal was presented in Roebel 1997. While the neural network was particularly adapted for the prediction of time series, the system lacked variety in the choice of controllers. In Jehan and Schoner 2001, the authors replaced the artificial neural network component by a Cluster Weighted Model (Gershenfeld et al., 1999) (similar to a hierarchical mixture of experts (Jordan and Jacobs, 1994)) and added a model of noise analysis/synthesis. More recent works include spectral envelope morphing based on evolutionary computing (Caetano and Rodet, 2009) and a feature-based synthesis framework based on our approach (Hoffman and Cook, 2006b).

In the concatenative synthesis approach, a sound is defined as a combination (concatenation) of pre-existing samples in a database (Schwarz, 2004). These samples (or units) are already analyzed and classified in the database, and can be retrieved by their audio characteristics. High-level control of synthesis is achieved by selecting the units that correspond to a desired value of a sound attribute. For a more detailed review of concatenative synthesis, the reader is reported to Schwarz 2007.

Serra and Bonada 1998 have proposed yet another approach based on spectral modeling synthesis (SMS) (Serra and Smith, 1990). It takes advantage of the fact that the SMS model, which is a framework to analyze and synthesize the time-varying spectrum of a sound, directly allows for transformation such as translation and distortion of the harmonic spectrum, vibrato, tremolo and gender transformation of a voice.

Finally, an hybrid class of models inspired by physical modeling techniques and using
4.5. Conclusions

Spectral synthesis techniques has recently led to interesting results (Aramaki et al., 2006). These models take advantage of a simple physical model of impact sounds to obtain both a parameterization of the sound source (via physics-based parameters of object and excitation) and the control of sound timbre features (inharmonicity, roughness, brightness, attack-time) that can be explicitly mapped to modal synthesis parameters (Aramaki et al., 2006; McAdams et al., 2004).

4.5. Conclusions

On the one hand, we reviewed a variety of audio descriptors selected from the separate domains of MIR, signal processing, psychoacoustics and music composition. We emphasized the importance of fundamental perceptual attributes of timbre such as spectral centroid, spectral flux, attack time, inharmonicity and roughness. On the other hand, we found that synthesis techniques such as spectral/additive synthesis and physical models were the most flexible and closer to perception. Based on the results from previous works, we propose two main methods for achieving high-level control of sound synthesis. The first implicit approach is based on a machine learning algorithm that learns to automatically map perceptual features to the parameters of an additive synthesizer. The second explicit approach is based on a physical model that provides access to perceptual features via the physical parameters. Both methods are explored in detail in the chapters 5 and 6.
5. A Machine Learning Approach to Perceptual Synthesis

Nowadays, powerful synthesis algorithms that run on standard home computers are able to generate almost any kind of sound. The main issue at stake is now that of intuitive control and transformation of these synthetic sounds. This problem is hard as the synthesis algorithm’s parameters are usually numerous, low-level and relate to perception in a complex non-linear way. This chapter presents a new system based on machine learning that allows to control an additive sound synthesis model from perceptually relevant high-level sound features. We propose a general framework for the extraction, abstraction, reproduction and transformation of timbral characteristics of a sound analyzed from recordings and introduce a method to train, tune and evaluate our system in an automatic, consistent and reproducible fashion (Le Groux and Verschure, 2008). We start by a description of the additive model used for sound analysis and synthesis (Section 5.1). The problem of the high-dimensionality of the parameter space of this model is then tackled using Principal Component Analysis (Section 5.1.2). Finally, the mapping from perceptual controls to synthesis parameters is achieve thanks to a Support Vector machine learning algorithm that can be automatically parameterized (Sections 5.3 and 5.4).

5.1. Sound Analysis and Synthesis Framework

We chose an additive synthesis model for its ability to synthesize a large range of sounds. Unlike physical models, it is based on original recorded sounds. It doesn’t require having a theoretical representation of the physical properties of each different instrument can be used for interesting sound manipulations as the frequency, amplitude and phase of every sound partial can be controlled independently. One drawback of additive synthesis is that a large number of synthesis parameters is needed to obtain good quality results. This is solved by using a PCA-based synthesis model (Section 5.1.2).

5.1.1. Harmonic Additive Model

The sound is represented as a sum of weighted sinusoids at frequencies multiple of the fundamental frequency. This model presupposes the sound is quasi harmonic and the
5. A Machine Learning Approach to Perceptual Synthesis

frequencies, amplitudes and phases are varying slowly; nevertheless, the results obtained with this type of analysis prove satisfactory for our research objectives.

A quasi-periodic tone can be decomposed in a sum of sine waves with time-varying amplitudes and frequencies as such (McAulay and Quatieri, 1986):

\[
y(n) \simeq \sum_{k=0}^{N-1} a_k(n) \sin(2\pi f_k(n) + \phi_k)
\]

(5.1)

where \( n \) is the time index, \( N \) is the number of harmonics in the synthetic signal, \( a_k(n) \) is the time-varying amplitude of the \( k \)th partial, \( f_k(n) \) is the time-varying frequency of the \( k \)th partial, and \( \phi_k \) is the corresponding phase. with \( f_k = k.f_0 \) where \( f_0 \) is the fundamental frequency (harmonicity hypothesis) and \( a_k(t) \) and \( f_k(t) \) are slowly varying.

For the resynthesis, the phase information, which is not perceptually relevant, is discarded. We discuss this choice in the following paragraph. We assume the sample \( y(n) \) can be reconstructed from the fundamental frequency estimation vector \( \mathbf{f} \) (dimension number of frames) and the matrix of partial amplitudes \( \mathbf{A}_{i,j} \) (dimension number of harmonics by number of frames).

\[
\begin{bmatrix}
a_{1,1} & \cdots & a_{1,M} \\
\vdots & & \vdots \\
a_{N,1} & \cdots & a_{N,M}
\end{bmatrix}
\]

(5.2)

where \( N \) is the number of partials, and \( M \) is the number of analysis frames.

For the re-synthesis, the samples \( y(n) \) are reconstituted at the proper sample rate \( F_s \) by linear interpolation between each frame of the analysis matrix \( \mathbf{A}_{i,j} \) to give the coefficients \( a_k(n) \). The procedure is similar for obtaining the coefficients \( f_k(n) \). The analysis hop size gives the number of samples represented by each column of the matrix: \( hop = round(F_s.frameDuration) \) where \( F_s \) is the sample rate in Hertz and \( frameDuration \) is in seconds.

The interpolation scheme allows the calculation of the time-varying coefficients between each frame:

\[
a_k^j(n) = \left( \frac{a_{k,j+1} - a_{k,j}}{hop} \right) n + a_{k,j}
\]

(5.3)

where \( a_k^j(n) \) is the time-varying amplitude of the partial \( k \) between the frame \( j \) and \( j+1 \), \( a_k \) is the element of matrix \( \mathbf{A} \) corresponding to the partial \( k \) and to the frame \( j \).

The same procedure is repeated for the interpolation of the fundamental frequency values between each frame and the synthesized sound is computed from equation 5.1. The analysis information provides a reliable abstract representation of the original audio signal.

We previously assumed that the influence of the phase was not primordial for re-synthesis purposes. This assumption deserves an explanation. The phase for each partial (or each
5.1. Sound Analysis and Synthesis Framework

Sinusoidal oscillators can be observed as a difference in initial amplitude and direction. We based our decision to neglect the effects of the phase on an article by (Risset and Wessel, 1982), where the authors discuss Ohm’s acoustic law and state that “if the Fourier representations of two sounds have the same pattern of harmonic amplitudes but have different patterns of phase relationships, a listener will be unable to perceive a difference between the two sounds, even though the two sounds may have very different waveforms.” Notice that this insensitivity to phase is only true for periodic tones, and doesn’t hold anymore in a dispersive medium environments where frequencies propagate at different speeds. Nevertheless, for audible sounds, air can be considered as a non-dispersive medium (it is not true anymore for ultrasonic frequencies superior to 30kHz). We favor expressive control of synthesis over high quality of sound generation.

We based our sound analysis on the SMS model which we won’t detail here since it has been extensively studied and presented elsewhere. We report the reader to (Serra, 1997) for a comprehensive explanation of the spectral analysis/synthesis paradigm, and for the exact methods employed to extract the frequencies, amplitude an phase parameters of the additive model we previously introduced. The main steps of the analysis process are summarized in Figure 5.1. Because we chose to use a simplified additive model for re-synthesis as a first approach, we just need the sine magnitude and phase parameters. The practical choice of the analysis parameters such as the size of the short-term Fourier transform, the hop-size and the smoothing windows are given in the Section 5.3 that deals with feature extraction. At this stage, we have obtained a parametrization of a sound represented by the matrix of time-varying amplitudes, and a matrix of the time-varying partial frequencies inferred from the time-varying fundamental frequency. Typically, this parametrization is of high dimension (hundreds of partials). In the next section we will explain how we managed to do synthesis from a reduced set of parameters while preserving the original quality of the synthesis.

5.1.2. Principal Component Synthesis

We want to reduce the dimensionality of the parameter space by finding a compact representation of the time-varying amplitude matrix. For this purpose, we use the Principal Component Analysis (PCA) (Jackson, 2003). PCA is one of the most commonly used techniques in modern data analysis. The goal of PCA is to compute the most meaningful basis to re-express a noisy data set. It has proved to be a powerful tool for many applications (Computer vision, Bio-informatics, Statistics, ...). In the context of musical timbre research, several studies have used PCA to analyze physical parameters of sound, searching for relationships to perceptive characteristics (Hourdin et al., 1997b,a). In this chapter, we use PCA as a data reduction tool for additive analysis parameters. The perceptual validity of this approach has been extensively studied in (Sandell and Martens, 1995) in the case of phase vocoder parameters. Here we propose a method to synthesize a quasi-harmonic sound from the low-dimensional principal component decomposition of the harmonic amplitude matrix.
Figure 5.1: The SMS analysis provides a robust methodology for the extraction of synthesis parameters such as amplitude, frequency and phase from an audio file (Serra and Smith, 1990).

If we take a statistics formalism, one row of the partial amplitude matrix $\mathbf{A}_{i,j}$ is a variable (a particular partial amplitude trajectory), and one column is a measurement (a particular spectrum) for the $j$-th analysis frame (Cf. Figure 5.2).

First, we subtract the mean in each dimension (amplitude trajectory) and then, the covariance of the partial amplitude matrix is calculated.

$$C_A = \frac{1}{n-1} \mathbf{A} \mathbf{A}^T$$ (5.4)

The $ij^{th}$ element of $C_A$ is the dot product between the $i^{th}$ and $j^{th}$ partial amplitude trajectories. $C_A$ is a square symmetric matrix, which diagonal terms are the variance of the amplitude partials and the off-diagonal terms are their covariances.

This correlation matrix captures the correlations between all possible pairs, and reflects the noise (small diagonal values) and redundancy (large off-diagonal values) in the additive analysis. The purpose of this principal component analysis is to reduce the redundancy (off-diagonal) that might be encountered in the additive analysis parameter space, and maximize the variance (diagonal). In other words, it comes down to diagonalizing $C_A$.

$C_A$ is square, and an eigenvector decomposition is performed in order to find an orthonormal basis of eigenvectors. The eigenvectors of the covariance matrix associated with the largest eigen-values lie in the direction of the largest data variance. We choose a small set of eigen (“feature”) vectors with the largest eigenvalues so that only the
5.1. Sound Analysis and Synthesis Framework

Figure 5.2: The time-varying partial amplitudes from matrix $A$ represent the first 50 partials of a 5 notes solo saxophone. Each color represents a different partial (one row of the matrix).

The smallest amount of information is lost in the mean-square sense. They form the matrix $F$ of dimension number of partials by number of eigen-vector retained, or Principal Components (PC).

The data in the new bases is easily computed, by a projection of the original data on the principal components (eigen-vectors):

$$D = F^T A$$  \hspace{1cm} (5.5)

where $D$ is the new data matrix (dimension number of principal components by number of frames). $F$ is the feature matrix of eigen-vectors (dimension number of partials by number of principal components) and $A$ is the partial amplitude matrix as described in 5.2(of dimension number of partials by number of frames).

The time-varying envelopes $D_t$ (Cf. Figure 5.3 in dimension three) are trajectories in the new Principal Component (PC) basis $F$ (Figure 5.4 and 5.5). The new PC bases represent “principal spectra”, and a time-varying partial amplitude is a time-varying linear combination of those basis spectra.

It is possible to get an approximation of the original matrix from the low dimensional principal component space by multiplication of the PC bases matrix and the time-varying envelope matrix (Figure 5.6).

$$\hat{A} \approx FD$$  \hspace{1cm} (5.6)

We are now able to synthesize a sound from a reduced set of PC controls. The next step is to map perceptual controls to those PC controls. This step is taken care of by the support vector machine learning algorithm, which will briefly explain in the following
5. A Machine Learning Approach to Perceptual Synthesis

![Figure 5.3: The time-varying envelopes represent trajectories in the Principal Component bases.](image)

section.

5.2. Introduction to Machine Learning

Machine Learning evolved from the field of Artificial Intelligence which aims to reproduce intelligent human behaviors in machines. The Machine Learning field studies the possibilities of making machines able to learn. Learning here is understood as inductive inference, or the capacity to predict based on observations. Learning implies a lot of different processes but basically consists in gaining knowledge, or understanding by experience. It can thus be said that a machine has learned when it changes its program, its data, or its structure, based on the inputs it receive, and in such a way that the future performance is expected to improve. There are two main kind of learning. **Unsupervised learning** tries to uncover hidden regularities such as clusters or to detect anomalies in the data. In **supervised learning**, we know in advance the label, or target values, of a set of examples we train the machine on. When the target is discrete, the task is a classification problem. If the target is real-valued, the task is a regression problem.

One might wonder why we would use a machine that learns, instead of directly designing a machine that executes the task we want. If the study of human cognition is an argument in favor of exploring machine learning models, there are also some important engineering tasks that require this type of model. For instance, some tasks cannot be defined well except by example. Some important relationships or correlation information might be hidden in the data and extracted by machine learning. Some applications
need on-the-job improvements of what is already existing. Finally, one might need an application that adapts to time-varying environments.

In our research, we want to be able to map continuous perceptual controllers to sound synthesis parameters. This mapping is based on the analysis of a recorded sound and its descriptors. We can have access to both the input and the desired target for a training set, thus a good way to define our task is by example. We can specify the input/output pairs (perceptual controls/synthesis parameters) from the audio analysis, but not a precise relationship between inputs and desired outputs. In consequence, supervised machine learning algorithms are well-suited to the kind of task we wish to tackle.

A mapping function associates an input vector to an output vector. Our problem is to learn this function $f$. The hypothesis about the function is denoted $h$. $f$ and $h$ are functions of a real-valued input $x = (x_1, ..., x_i, ..., x_n)$ with $n$ components. The function $h$ is selected based on a training set of input vector examples $S$. For supervised learning, we assume that if we can find a $h$ that closely agree with $f$ on the training data set, then this hypothesis $h$ will be a good guess for $f$ when the dataset is large enough.

Supervised machine learning relies on three main steps. In the first step, we extract the relevant inputs (perceptual control) and outputs (synthesis parameters) from the data.
5. A Machine Learning Approach to Perceptual Synthesis

Figure 5.5.: 3D representation of the trajectories in the PC bases space.

set. In a second step, we define the training parameters and run the training algorithm. Finally, the training session is validated by using the system to predict outputs (synthesis parameters) based on new inputs (perceptual controls).

5.3. Feature extraction

The features are obtained from short-term additive analysis based on the SMS model (Serra and Smith, 1990). The data is organized in SDIF time-tagged frames (Wright et al., 1998). Each frame is composed of the frequencies, amplitudes and phases of the quasi-harmonic components. The SMS analysis window (BlackmanHarris92) size is 1024 samples. Hop size is 256 samples. The sampling rate 44100Hz. We kept 80 harmonics trajectories \( \dim(A) = 80 \times N\text{Frames} \).

5.3.1. Controls: Perceptual Features

5.3.1.1. Pitch Estimate

Fundamental frequency is crucial for the quality of the re-synthesis. We chose to use the Yin algorithm to extract an accurate fundamental frequency from the audio signal. The Yin method is inspired by auditory models of pitch perception and has been shown to be reliable and accurate on a large range of speech and musical sounds (de Cheveigné and Kawahara, 2002). We had to carry out three main post-processing tasks on the raw data output by the Yin algorithm. First, based on the aperiodicity, and energy factor (also provided by Yin), we could clean up the noisy data in the pitch track. If the product of the normalized periodicity, power and coherence windows was inferior to 0.5, then
the corresponding data points were removed (de Cheveigné and Kawahara, 2002). In a second step, we used a median filter and interpolated pitch during unvoiced regions to produce a smooth pitch curve.

Finally, since the size of the analysis window for the Yin algorithm can be different from the SMS harmonic analysis window, we resample the result to match the difference in frameshifts. In practice, if we want a good estimate of the pitch, the algorithm needs to be fed with the range of variation of fundamental frequency.

![Yin smoothed analysis frame](image)

**Figure 5.7.:** Yin fundamental frequency estimation, with noisy data points removal and smoothing.

### 5.3.1.2. Loudness Estimate

We extracted the loudness estimate by applying a simple loudness summation model proposed by Zwicker and Scharf 1965:

\[
I(n) = \sum_{k=1}^{N} a_k(n) 
\]  

\[5.7\]

### 5.3.1.3. Brightness Estimate

The exploration of fine timbral control is also possible. Our system is designed in such a way that it allows to “shape” the synthesis of a sound using any kind of continuous feature extracted from audio as a controller. For instance, we can obtain a measure of brightness by computing the centroid for each spectral frame (Schubert and Tarnopolsky, 2004):

\[
SC(n) = \frac{\sum_{k=1}^{N} a_k \cdot f_k(n)}{\sum_{k=1}^{N} a_k(n)}
\]

\[5.8\]
5. A Machine Learning Approach to Perceptual Synthesis

Our high-level input is of dimension two (loudness and fundamental frequency) by NFrames.

Finally, we have a high-level input vector of dimension three (loudness summation, fundamental frequency, and spectral centroid) by number of frames (Figure 5.8). The number of input analysis frame is synchronized with that of the output.

Figure 5.8.: The three-dimensional control vector includes estimations of pitch (fundamental frequency), loudness (loudness summation) and brightness (spectral centroid).

5.3.2. Targets: Time-Varying Principal Component Trajectories

We have found a way to reduce the dimension of the target space with PCA analysis, and are able to re-synthesize an estimate of the original sound from those parameters without any significant perceptual loss. It is possible to choose the number of Principal Components (NPCs) by entering the percentage of variance of the data we wish to take into account (Cf. Figure 5.9). In this test case, we can see that most of the variance of the data is explained by only 3 PC. In the rest of this chapter, we will assume that the parameter space is of dimension 3 (Figure 5.10), knowing that we can reconstruct a good approximation of the original data (Figure 5.11).
5.3. Feature extraction

Figure 5.9: **Pareto chart** of the number of Principal Components (PC) against the percentage of variance expressed. For this database, most of the variance in the data can be expressed by only 3 PCs.

Figure 5.10: **Test set trajectory** in a three-dimensional PC space.
The dimension of the target vector is now NPCs by NFrames. In addition to dimension reduction, PCA allows the de-correlation of the new target vector, which is a good preprocessing practice for the machine learning algorithm used for mapping. A problem of PCA decomposition is that the most significant PC parameters don’t necessarily have an obvious relation to human perception. We need another layer to map perceptual,
intuitive, musically relevant controls to this lower dimensional synthesis parameter space. This task is handled by the Support Vector Regression (SVR) algorithm that maps sonic percepts to sound generation parameters using our PCA synthesis model.

5.4. Training and results

5.4.1. Support Vector Mapping Machine

Support Vector Machines (SVMs) and kernel methods have become increasingly popular tools. We chose to apply this technique to the perceptual mapping problem based on good results the SVM algorithms obtained in various surveys from the machine learning community (Bennett and Campbell, 2000), and on the improvement it could bring to previous related works. Compared to other approaches that have been used for perceptual mapping such as Artificial Neural Networks or Cluster-Weighted Modeling (Wessel et al., 1998; Schoner, 2000b), SVMs exhibit a lot of interesting properties. Namely, we can build highly non-linear regression function without getting stuck in a local minima. Moreover, there’s only a few model parameters to pick, and the final results are stable and reproducible, unlike neural networks models where the result depends on the initial starting point. Finally, SVMs have been proved to be robust to noise in many applications (Bennett and Campbell, 2000).

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression (Scholkopf and Smola, 2002). They use a technique known as "kernel trick" to apply linear techniques to non-linear problems.

The ε-SVR algorithm is a generalization of the better known Support Vector for Classification algorithm. For regression, the goal is to estimate an unknown continuous-valued function based on a finite set of training samples. Given $n$ training vectors $\mathbf{x}_i$ and a vector $y$: $(\mathbf{x}_i, y_i), (i = 1, ..., n)$ where $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$, we optimally estimate the function $y = f(\mathbf{x})$, in the Structural Risk Minimization sense, by:

$$f(x) = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) k(x_i, x) + b \quad (5.9)$$

where $b$ is a bias, $k(x_i, x)$ is the kernel function and $\alpha_i^*, \alpha_i$ the solution of the quadratic problem:

Maximize

$$\frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) k(x_i, x_j) - \varepsilon \sum_{i=1}^{l} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \quad (5.10)$$

Subject to

$$\sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0$$

$$\alpha_i - \alpha_i^* \in [0, C]$$
5. A Machine Learning Approach to Perceptual Synthesis

Where C and ε are the chosen by the user.

The “kernel trick” is a means to convert a linear regression learning into a non-linear one by mapping the original observation samples into a higher-dimensional space, in order to account for non-linearities in the estimate function (Figure 5.12). \[ k(x, y) = \langle \psi(x), \psi(y) \rangle, \text{ where } \psi(x) : \mathbb{R}^d \rightarrow H \text{ and where } H \text{ is a high-dimensional feature space with dot product.} \] A commonly used kernel is the Radial Basis Function (RBF) kernel:

\[ k(x, y) = \exp(-\gamma \|x - y\|^2) \tag{5.11} \]

where the width parameter \( \gamma \) is selected by the user.

![Kernel methods map the training data non linearly into a higher-dimensional feature space and construct a separating hyperplane with maximum margin.](image)

Hence, the Support Vector Regression algorithm allows to find a non-linear mapping between high-level controls and the PC synthesis parameters.

### 5.4.2. The Dataset

We normalized our dataset in order to have zero mean and unity standard deviation. We ended up with training examples with a three dimensional (fundamental, loudness summation and spectral centroid) real-valued vector as input and a three-dimensional (3 PCs) real-valued vector as output. The example training set was extracted from an audio file of 2 minutes of solo saxophone with as much variability as possible in the attack, dynamics and range. The data was split into a training set and a testing set (roughly 2/3, 1/3).

The support vector regression algorithm described in (?) is designed for only one output. Thus we had to split the problem into \( n \) distinct (and supposedly independent) function estimation problems, considering each time a different “PC trajectory” as output, \( n \) being the number of PC necessary for re-synthesizing a sound without significant perceptual loss.
Figure 5.13.: The continuous perceptual input controls are pitch, loudness and brightness.

5.4.3. Model selection

Resampling approaches, commonly used for SVM, are very expensive in terms of computational costs and data requirements. The approach we used for parameter selection is based on a work by Cherkassky and Ma 2004 who propose a practical analytical approach to the choice of SVM regression parameters based directly on the training data.

We used Cherkassky's prescription for the regularization parameter

\[ C = \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|) \] (5.12)

where \( \bar{y} \) is the mean of the training responses, and \( \sigma_y \) is the standard deviation of the training response.

The value of \( \varepsilon \) was chosen to be proportional to the input noise level and we assumed that the standard deviation of noise \( \sigma \) could be estimated from data. We used the
5. A Machine Learning Approach to Perceptual Synthesis

prescription in Cherkassky and Ma 2004 for noise variance estimation via $k$-nearest neighbor’s method:

$$\hat{\sigma}^2 = 1.5 \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)$$  \hspace{1cm} (5.13)

where $n$ is the number of training data, $y_i$ the training response and $\hat{y}_i$ the fit by the $k$-nn regression.

After empirical comparisons, Cherkassky and Ma 2004 proposed the dependency equation:

$$\varepsilon = \tau \sigma \sqrt{\frac{\ln n}{n}}$$  \hspace{1cm} (5.14)

with $\tau = 3$ giving a good performance. The kernel width was selected to roughly reflect the input range of the training/test data.

5.4.4. Evaluation

For evaluating the performance of our system, we used two functional measures on the test dataset.

- The first one is the RMS error, which gives a measure of the overall distance between two trajectories

$$E_{rms} = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\hat{y}_i - y_i)^2}$$  \hspace{1cm} (5.15)

where $y_i$ is the target original value, $\hat{y}_i$ the estimation, and $n_{test}$ the number of sample in the test dataset.

- The second measure is the correlation score, that indicates similarity of shape and synchrony between two trajectories.

$$r = \frac{\sum_i (\hat{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_i (\hat{y}_i - \bar{y})^2 \sum_i (y_i - \bar{y})^2}}$$  \hspace{1cm} (5.16)

Finally, by re-synthesis from the estimated output, we obtain a perceptual, subjective equivalent of those functional measures.
5.4. Training and results

5.4.5. Results

Results on the data described in Section 5.3 give good correlation scores between the target partials and their estimate for the three principal components (Cf. Table 5.1). Visualization of the training data and test data for the first trajectory show that the SVR-based synthesizer is able to predict a good estimate of the PC trajectory based on the input (5.14). As can be seen, the estimation of the principal component trajectory on the test data (with input controls unknown to the support vector machine) is good (Figure 5.15) and the system generalizes well.

<table>
<thead>
<tr>
<th>PC trajectory</th>
<th>RMS error</th>
<th>Correlation score</th>
</tr>
</thead>
<tbody>
<tr>
<td>First PC</td>
<td>0.009</td>
<td>0.991</td>
</tr>
<tr>
<td>Second PC</td>
<td>0.0159</td>
<td>0.99</td>
</tr>
<tr>
<td>Third PC</td>
<td>0.0534</td>
<td>0.973</td>
</tr>
</tbody>
</table>

Table 5.1.: **Correlation and RMS** results on a 2 min saxophone solo.

![Partial estimation results. A) Target and estimated first PC trajectory of the training set B) Target and estimated first PC trajectories of the test set.](image)
5. A Machine Learning Approach to Perceptual Synthesis

Figure 5.15.: **The correlation between target and estimate** on the test data set shows that the SVR gives a good estimation of the trajectory when presented with unknown inputs.

### 5.5. Applications

One of the most powerful feature of this system is its ability to change the original melody contour, brightness or loudness control while preserving the musical identity of the audio sequence on which the machine learning algorithm has been trained. The spectral properties of the sound are preserved when the control information varies, providing the new controls are rescaled in a similar range as the training data.

Our system proposes a new approach to *pitch shifting/time warping* applications. Because of the good generalization properties of the SVR algorithm, the musician can directly enter the desired time-varying frequency as a control input. The limitation of this technique is the range of the data on which the machine learning algorithm has been trained. The results will be poor if the frequencies given as an input to the system are too far from the frequencies encountered in the training set.

Our model well-suited for *cross-synthesis* applications where the control parameters typical of one instrument can be used in combination with the “support vector timbre model” of another instrument. Another possibility of cross-synthesis is at the level of the PC synthesis, where the PCs, or basis spectra, of one instrument can be replace by those of another instrument.

Due to the characteristics of the PC synthesis, our model has an interesting property of *scalability*. The number of useful PC bases can be chosen depending on the quality of the sound required at the moment. This behavior is particularly interesting in networked applications, where the available bandwidth is variable.
In this paper we have mostly studied the control of a synthesis model from loudness estimation, pitch and brightness parameters, letting the machine learning component take care of the overall timbre generation. This property, that allows direct generation of sound from continuous timbral features such as brightness, is extremely interesting, and suggests an analysis-by-synthesis type of applications. For instance, in the Music Information Retrieval community, the problem of finding relevant descriptors and judging their perceptual influence is still open. Our tool allowing feature-based synthesis would prove useful to investigate the perceptual relevance of abstract descriptors devised by MIR researchers.

5.6. Conclusions

We have proposed a system that allows for flexible and intuitive control of sound generation from high-level sonic percepts. It provides the user with continuous control over a sound directly analyzed from recordings. We devised an analysis/synthesis paradigm well-suited for a machine learning approach to this non-linear mapping problem, and found a way to reduce the dimensionality of the synthesis parameter space, while preserving the auditory quality of the re-synthesis. We described a framework to train, tune and evaluate our system in an automatic, consistent and reproducible fashion. This system paves the way for various original audio and musical applications.

The results of this preliminary work are based on relatively small databases. In the future, we would like to extend our experiments with the actual system on larger datasets including as much variance as possible. We should try our algorithms on many different type of sustained excitation instruments for which the control and interaction is continuous. We’d like to study how the size of the database influence the quality of the re-synthesis when the system is presented to unseen controller values.

The work presented in this chapter did not include a memory mechanism for previous states of an instrument. We know, for example, that in a rapid passage, a given note on a wind instrument will sound different depending on whether it was preceded by a higher or lower pitch. Future work should model and study the influence of these temporal dependencies. Prior state can be easily introduced into the machine learning architecture. A qualitative study of the influence of a memory mechanism on the quality of the synthesis could be pursued.

At the time of this writing, the analysis and synthesis of sounds are realized off-line. Nevertheless, the synthesis models presented are causal and can be implemented in real-time. Future research and musical application will benefit greatly from real-time interaction between the musician and this new kind of instrumental controller.

Another possible future direction of research is the application of some other more advanced techniques for dimension reduction. In this analysis framework, the amplitude matrix is non-negative, so Non-Negative Matrix Factorization (NNMF) and Non-
Negative Independent Component Analysis (NNICA) could also be used to compute basis vectors for low-dimensional models of the amplitude matrix.

Such an implicit and automatic paradigm for timbre modeling is very interesting as it is quite general and can potentially model any kind of sound. Yet, one practical problem is that it requires complex machine learning computation and a series of data pre and post processing stages which can be tedious. These problems can be solved by a more explicit, but also less generic approach to sound synthesis using physically inspired models.
5.6. Conclusions

Figure 5.6.: Estimation of the time-varying amplitudes for the first six partials of the spectrum.
6. Explicit Control of Perceptual Parameters with Physically Inspired Synthesis

Although it is powerful and general, the machine learning approach to perceptual sound synthesis requires complex data processing, and has to be fed with plausible time-varying inputs. In this chapter, we present a simpler alternative based on physically informed synthesis of impact sounds which leads to a straightforward real-time implementation. After a description of the synthesis model and its implementation, we explain how we can directly construct perceptual controls from the synthesis model. This type of synthesis can produce realistic sounds efficiently and intuitively from a three-dimensional timbre space model mapped to a reduced set of physical parameters.

6.1. Physically-Inspired Sound Synthesis

In modal synthesis, a sound is modeled as a combination of modes which oscillate independently (Figure 6.1). While this kind of modeling is in theory only accurate for sounds produced by linear phenomena, it allows for efficient real-time implementations (Cook, 2002; Adrien, 1991; van den Doel and Pai, 1996).

6.1.1. Modal Synthesis

The dynamics of a simple unidimensional mass-spring-damper system (our model of a mode) is given by the second order differential equation:

\[ m\ddot{x} + \mu \dot{x} + kx = 0 \]  

(6.1)

where \( m \) is the mass, \( \mu \) the damping factor, \( k \) the tension and \( x \) the displacement.

The solution is an exponentially decaying oscillator:

\[ x = e^{-\alpha t} A \cos(\omega_0 t + \phi) \]  

(6.2)

where \( \alpha = \frac{\mu}{2m} \), \( \omega_0 = \sqrt{\frac{k}{m} - \left(\frac{\mu}{2m}\right)^2} \), \( A \) is the amplitude and \( \phi \) the phase.
6. Explicit Control of Perceptual Parameters with Physically Inspired Synthesis

The modal approximation consists in modeling the time-varying sound \( s(t) \) by a linear combination of damped oscillators representing the vibrating properties of a structure. In the general case, the damping factor \( \alpha_n \) is frequency-dependent and relates to the physical properties of the material.

\[
s(t) = \sum_{n=1}^{\infty} s_n(t) = \sum_{n=1} A_n e^{2\pi f_n t} e^{-\alpha_n t}
\]

Figure 6.1.: Modal analysis: In the modal analysis paradigm, a complex vibrating structure is modeled as a sum of simpler vibrating modes.

6.1.2. Computer Implementation

By discretizing equation 6.1 with finite difference and using a sampling interval \( T = \frac{T}{n} \) we obtain:

\[
x[n] = x[n-1] \frac{2m + Tr}{m + Tr + T^2k} - x[n-2] \frac{m}{m + Tr + T^2k}
\]

which is the formulation of a standard IIR two-poles resonant filter. Consequently we implemented a real-time synthesizer as a bank of biquad resonant filters in Max5 (Jehan et al., 1999; Zicarelli, 2002) excited by an impulse (Figure 6.2).

6.2. Perceptual Control of the Synthesis Parameters

Previous psychoacoustics studies on timbre have emphasized the perceptual importance of features such as spectral centroid, log-attack time and spectral flux (Cf. Chapter 4). As a means to generate perceptually relevant sounds, we have built a simple and efficient
6.2. Perceptual Control of the Synthesis Parameters

Implementation of the modal Synthesizer: The sounds are produced by an impulse that injects energy in a bank of exponentially-decaying resonant filters with parameters Frequency (F), Amplitude (A) and decay rate (d), the output of the filter bank is then modulated by an ADSR time envelope.

physically-inspired modal synthesizer that produces realistic percussive sounds. We will now show how to map those psychoacoustics features to sound synthesis parameters.

Here we follow this last approach with a simple physical model of impact sound and propose a set of acoustic parameters corresponding to the tridimensional model of timbre proposed by psychoacoustics studies McAdams et al. 1995a; Grey 1975; Iverson and Krumhansl 1993 (Cf. Section 4.2.2).

6.2.1. Tristimulus

We propose to control the spectral centroid of the sound by using the tristimulus parameterization. The stimulus model provides a perceptually relevant control of a sound’s brightness from a two-dimensional surface (Pollard and Jansson, 1982; Riley and Howard, 2004).

The tristimulus analysis of timbre proposes to quantify timbre in terms of three coordinates \((x, y, z)\) associated with band-loudness values. Inspired from the tristimulus theory of color perception, it associates high values of \(x\) to dominant high-frequencies, high values of \(y\) to dominant mid-frequency components and high values of \(z\) to dominant
fundamental frequency. The coordinates are normalized so that \( x + y + z = 1 \).

\[
T_1 = \sum_{h=1}^{H} a_h, T_2 = \sum_{h=1}^{H} a_2 + a_3 + a_4, T_3 = \sum_{h=1}^{H} a_h
\]  

(6.5)

Figure 6.3: **Tristimulus representation of timbre:** The tristimulus timbre space can be represented in 2D as a triangle. The arrow represents a specific time-varying trajectory of a sound in the space (Pollard and Jansson, 1982).

For synthesis, the sound is approximated by a sum of weighted damped oscillators at frequencies multiple of the fundamental. We use an additive model and neglect the effect of phase:

\[
s(t) \approx \sum_{k=0}^{N-1} e^{-\alpha_k t} A_k \cos(2\pi f_k t)
\]

(6.6)

where \( n \) is the time index, \( N \) is the number of harmonics in the synthetic signal, \( A_k \) is the amplitude of the \( k \)th partial, \( f_k \) is the frequency of the \( k \)th partial. with \( f_k = k \cdot f_0 \) where \( f_0 \) is the fundamental frequency (harmonicity hypothesis).

In the synthesis model, the harmonic modes belong to three distinct frequency bands or tristimulus bands: \( f_0 \) belongs to the first low frequency band, frequencies \( f_2, \ldots, f_4 \) belong to the second mid-frequency band, and the remaining partials \( f_5, \ldots, f_N \) belong to the high-frequency band.

The relative intensities in the three bands can be visualized on a tristimulus triangular diagram where each corner represents a specific frequency band. We use this representation as an intuitive spatial interface for timbral control of the synthesized sounds (Cf.
Figure 6.3). The inharmonicity of the sound was set by scaling the values of partials following a piano-like law proposed by (Aramaki et al., 2006) \( f_k = k f_0 \sqrt{1 + \beta k^2} \).

Figure 6.4.: **The modal synthesizer’s graphical interface** allows for graphical control over A) the tristimulus values (left) and automatically updates and displays the amplitudes of the corresponding partials (right). B) the damping parameters \( \alpha_g \) and \( \alpha_r \). C) the piano-like inharmonicity law D) the log-attack time via the Attack Decay Sustain Release (ADSR) envelope E) The amount of noise in the input signal as well as the ration even/odd partials. Additionally, the synthesizers provides a real-time visualization of estimation of pitch, loudness, brightness, noisiness and spectral flux (F). The synthesizer can be controlled by any midi device (G). The configuration of parameters can be saved and retrieved for later use (H).

### 6.2.2. Damping

While the tristimulus is only a static property of the spectrum, the modal synthesis technique described in Section 6.1 also allows for realistic control of the time-varying
attenuation of the spectral components. As a matter of fact, the spectral flux, defined as the mean value of the variation of the spectral components, is known to play an important role in the perception of sounds (McAdams et al., 1995a). We decided, as a first approximation, to indirectly control the spectral flux -or variation of brightness- by modulating the relative damping value $\alpha_r$ of the frequency-dependent damping parameter $\alpha$ (Figure 6.4 and Equation 6.3) proposed by (Aramaki et al., 2006):

$$\alpha(\omega) = \exp(\alpha_g + \alpha_r \omega)$$  \hfill (6.7)

6.2.3. Time Envelope

The log-attack time (LAT) is the logarithm (decimal base) of the time duration between the time the signal starts to the time it reaches its stable part (Peeters et al., 2000). Here, we control the log-attack time manually using our synthesizer’s interface (Figure 6.4). Each time the synthesizer receives a MIDI note-on message, an Attack-Decay-Sustain-Release (ADSR) time envelope corresponding to the desired LAT parameter is triggered (Figure 6.5).

$$LAT = \log_{10}(t_{\text{stop}} - t_{\text{start}})$$  \hfill (6.8)

![Envelope](image)

Figure 6.5.: The ADSR time envelope allows for direct control of the Log-Attack Time (LAT) each time a note is triggered.

6.3. Conclusions

We proposed a physically-inspired model that synthesizes sound samples from a set of well-defined and perceptually-grounded sound parameters. Compared to the previous
6.3. Conclusions

machine learning approach, this model is less demanding in terms of computation, which allows for an easy and efficient real-time implementation. Moreover, there is no need for preprocessing or training the system. It can be used directly. Another interesting aspect of physical modeling is that it also provides access to the physical properties of the sound source such as the material, size and shape of the object or the force, hardness, and position of the excitation. Although we did not explore this aspect in detail, this is certainly a promising idea that allows a three-layer control of synthesis from the sound source to the synthesis parameters via sound descriptors (Aramaki and Kronland-Martinet, 2006; Aramaki et al., 2006). One disadvantage of the physically-based synthesis approach comes from the constraints imposed by the synthesis model itself which is specifically designed to generate “impact-like” sounds only. We gain in efficiency, but lose in generality. Nevertheless, this model is a good candidate for timbre experiments because of its simplicity, ease of use and clear definition of parameters both from a perceptual and physical point of view.
Part III.

Composing Emotions: Audio Kansei Engineering and Healthcare Applications
7. Introduction

Music and its effect on the listener has long been a subject of fascination and scientific exploration. Theories of musical emotions can be traced back as far as the ancient Greeks who believed that “the power of music to arouse the human emotions lay in the representations, by melody, of the human speaking voice when expressing the various emotions” (Kivy, 2002). Nowadays, music has become omnipresent in our daily lives, and emotional reactions to music are increasingly used for utilitarian purposes (e.g. the “soothing” Muzak\(^1\) that invades public spaces such as elevators, shopping malls, parking lots, airport, etc). Music is well known for affecting our emotional states and in fact, most people primarily enjoy music because of the emotions it can evoke. Music, often referred to as the “language of emotions”, appears to deeply affect emotional, cerebral and physiological states (Sacks, 2008; Krumhansl, 1997b; Peretz and Coltheart, 2003; Koelsch and Siebel, 2005). Its effect on stress and anxiety has been established using a variety of self-report, physiological, and observational means (Khalfa et al., 2003; Hanser, 1985). Practically, it is increasingly used as an alternative or complement to other therapies in order to support and encourage physical, mental, social and emotional well-being (Bunt and Pavlicevic, 2001).

Curiously, although emotion seems to be a crucial aspect of music listening and performance, the scientific literature on music and emotion is relatively scarce and recent when compared to the literature on music cognition or perception (Juslin and Sloboda, 2001). The first seminal works came from musicologists and paved the way for more systematic research (Meyer, 1956). A deeper understanding of how music affects the listener needs to be grounded on a solid empirical basis that includes cognitive neuroscience and psychology (e.g. psychoacoustics), guided by a well defined taxonomy of musical content (Meyer, 1956; Gabrielsson and Lindström, 2001; Le Groux et al., 2008b; Krumhansl, 1997a; Bradley and Lang, 2000; Le Groux and Verschure, 2010b).

In this chapter, we introduce key theories of emotion (Section 7.1 and 7.2) and describe the components of music that have been found to influence emotional responses (Section 7.3). We briefly discuss the time-course of emotional responses to music (Section 7.4) and review the potential underlying mechanisms (Section 7.5). Finally, we report the methods that are commonly used to measure emotional responses (Section 7.6).

\(^1\)http://www.muzak.com/
7. Introduction

7.1. Theories of Emotion

The nature and role of emotions and how they relate to reason have been the center of many heated debates and has fascinated many philosophers and scientists. Theories of emotions go back as far as the Greek Stoics, but also Aristotle and Plato. Plato, for instance, thought the soul was composed of a combination of “cognition”, “emotion” and “conation”. Later modern philosophers such as Descartes, Hume or Spinoza also elaborated complex theories of emotion. Descartes, for instance, inspired by his research on the decomposition of white light into different colors, was the first to put forward the concept of primary and secondary emotions, where secondary emotions can be decomposed into a combination of primary emotions. Emotion research nowadays is still heavily influenced by these early conceptualizations. For instance, researchers have found evidence of the existence of primary basic emotions such as fear, disgust, anger, sadness, happiness and surprise (Ekman, 1992). Research on facial expressions, based on cross-cultural data, showed that these six basic emotional expressions were used and understood among very different cultures and communities (Ekman et al., 1982).

Even if emotion appears to be both a unique and universal human phenomenon, it was not taken seriously by most of the scientific community until fairly recently. Although early theories of emotions were mostly based on physiological models, the study of emotion has been largely influenced by the cognitive revolution. This influence has had an interesting impact on the evolution of modern emotion research and has shed a new light on the debated conflict between passion and reason (LeDoux, 1998). Is emotion based on a completely different system than cognition? Is it only a physiological phenomenon? Are those views compatible?

The James-Lange theory, one of the earliest theories of emotion, stresses the crucial importance of physiology. In this model, the perception of a stimulus first leads to a reaction of the autonomic nervous system triggering physiological responses, which results in a specific emotional state (James, 1998). In this view, emotions are inseparable from physiological responses. A physiological change is needed to be aware of the emotion. Consequently, one could think that different physiological states should trigger different emotions. Yet, this is actually not the case, as the same autonomic responses can occur for very different emotional states. Moreover, physiological changes are typically very slow while the feeling of emotion often happens right after the stimulation. Research also showed that animals could display emotional expression even if their internal organs were separated from the nervous system (Cannon, 1927).

These contradictions motivated the creation of an alternative theory called the Cannon-Bard theory that claims that emotion is not only reduced to the perception of physiological changes. It states that when a stimulating event happens, perception of emotion and physiological changes actually occur at the same time but that the stimuli has to be evaluated independently of the physiological responses (Cannon, 1929).

One drawback of the Cannon-Bard approach is that the theory doesn’t propose any mechanism for emotion. This problem is addressed by the Schacter-Singer theory that
models emotion as a two-factor process composed of physiological arousal and cognition. Cognition determines the type of emotion while the physiological arousal determines the intensity of the emotion (Schachter and Singer, 1962). Later works have proposed three components models, based on physiology, cognition and behavior (facial expression/gesture) (Plutchik, 1980, 1962; Lazarus, 1982). A theory proposed by Damasio states that, as we experience the world, emotions associated to specific stimuli are integrated with cognitive processing via somatic markers from a very early stage (Damasio, 2000). The assumption here is that emotions are part of an evolutionary principle that allows us to prepare for significant events (fight, feed, ...).

There exist a number of variations and improvement on these theories. Even if the exact mechanisms underlying emotion are still not completely well understood, there is a clear evidence for both subcortical and cognitive components of emotion, which is also true for music-induced emotions. Unlike many other well-studied emotional responses, music-induced responses may not be considered as useful for survival as say the fight or flight responses of animals reacting to threatening stimuli. Hence, some researcher make the distinction between utilitarian emotions (that need an appropriate reaction) and aesthetics emotions (with no need of appropriate reaction such as in music or the arts in general) (Scherer, 2004). A thorough discussion of the utilitarian aspect of music is beyond the scope of this chapter, but there is some evidence that music has an impact on health (Sacks, 2008) and brain plasticity (Hyde et al., 2009) and that music could after all be beneficial for survival in the long term.

7.2. Music-Induced Emotions

In parallel to the discussions about the general mechanisms of emotion, musicologist and music psychologists have long been debating about the impact of music on emotions (Gabrielsson, 2002). Defendant of the cognitivist view claim that music only expresses or illustrates emotions, whereas emotivists believe that music actually elicit or induce emotion (Kivy, 2002). For cognitivists, the emotional quality of a specific piece of music can be recognized, but it does not mean that the associated emotion is experienced by the listener. The music does not induce any specific emotional state. Nevertheless, to this date, experimental data best supports the emotivist point of view, as several studies have shown that music could reliably induce specific physical responses such as chills, increased heart rate, tears, laughter, and sweating (Sloboda, 1991; Goldstein, 1980).

7.3. Determinants of Musical Emotion

There is strong evidence that music induces specific emotional responses. Yet, the relationship between specific musical parameters and emotional responses is still not clear. This is partly due to the high complexity of music and the difficulty to control for its many dimension in an experimental protocol. Indeed, in most studies, participants are
subjected to pre-existing musical pieces extracted from professional audio recordings. These pieces are composed of many dimensions which vary simultaneously. This raises the problem of finding the right parameterization in order to achieve controlled and reliable stimuli for the experiments. In this section, we review the different musical features that have been shown to significantly influence emotional responses. They can be divided into structural parameters (the musical score), the sound quality (the timber) and the performance parameters (how the music is performed). Although external factors such as environment or the listener’s individual and sociocultural identity also play a role in the emotional effect of music (Scherer and Zentner, 2001), we focus on the parameters that are intrinsically musical.

7.3.1. Musical Structure

Whether it is conscious or not, composers probably tend to use music to express or illustrate the emotional states they are currently experiencing or have experienced (Cook and Dibben, 2001). In fact, there is empirical evidence from psychological research that the structure of a musical piece given by combinations of melody, rhythm and harmony parameters (Cf. Chapter 3 Section 3.1) influences the emotion perceived by the listener (Gabrielsson and Lindström, 2001; Hevner, 1936) (Table 7.1).

Psychophysiological studies have also confirmed the relationship between musical struc-
7.4. Time-course of Emotional Responses to Music

The contribution of individual performers to the expression of emotion during a music piece is an important factor. This can be partly attributed to the gestures and body postures of the performing musician (Vines et al., 2006, 2004; Wanderley et al., 2005), but also to some expressive modulations of the musical material itself. This is this last point that interests us here. In the context of classical Western music, performers can modulate musical parameters such as tempo, sound level, articulation and vibrato (Friberg et al., 2006). Typically, the articulation parameter (that goes from legato to staccato in musical terms) can change the overall character of a piece from sad to happy or angry (Friberg et al., 2006; Gabrielson and Lindström, 2001). The emotional quality of a performance is also greatly enhanced by an appropriate use of phrasing. Crescendo/decrescendo envelopes are constantly applied to the dynamics and tempo parameters to create more or less arousing musical structures. Moreover, since musicians are not completely accurate when producing and perceiving small timing variations, performance noise is a parameter that influences the perceived “naturalness” of musical expression (Juslin et al., 2002). This information is especially relevant for synthetic systems that need to reproduce a believable performance.

7.3.3. Timbre

Because of its high dimensionality, Timbre is notoriously difficult to define. Sandell, 2008. Nevertheless, the ASA (American Standards Association) proposed a definition by the negative in the sixties that has since been widely adopted. Timbre is “that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar” (American Standards Association, 1960). Whereas many studies have focused on emotional responses to performance or structural musical parameters (Friberg et al., 2006; Gabrielson and Juslin, 1996; Gabrielson and Lindström, 2001), little is know about the role of timbre on emotions. One explanation is that this lack of information is partly due to the difficulty for instrumentalists to independently and reliably control the perceptual dimensions of timbre in a reproducible way. In this respect, the use of computer-based synthesizers is of great interest as it allows to directly control the properties of a sound (Le Groux and Verschure, 2010b).

7.4. Time-course of Emotional Responses to Music

Because of its time-varying nature, emotion reflects a variety of complex temporal processes which can be classified in term of their duration. Emotion refers to the shortest
7. Introduction

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
<th>Emotional expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amplitude envelope</td>
<td>Round</td>
<td>Disgust, sadness, fear, boredom, potency (Scherer &amp; Oshinsky, 1977), tenderness, fear, sadness (Juslin, 1997)</td>
</tr>
<tr>
<td></td>
<td>Sharp</td>
<td>Pleasantness, happiness, surprise, activity (Scherer &amp; Oshinsky, 1977), anger (Juslin, 1997)</td>
</tr>
<tr>
<td>Articulation</td>
<td>Staccato</td>
<td>Fear, anger (Juslin, 1997)</td>
</tr>
<tr>
<td></td>
<td>Legato</td>
<td>Tenderness, sadness (Juslin, 1997)</td>
</tr>
<tr>
<td>Harmony</td>
<td>Simple /</td>
<td>Relaxation, tenderness (Lindström, 1997)</td>
</tr>
<tr>
<td></td>
<td>consonant</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complex /</td>
<td>Anger, Tension (Lindström, 1997), tendency, fear (Krumhansl, 1990; Krumhansl, 1997)</td>
</tr>
<tr>
<td></td>
<td>dissonant</td>
<td></td>
</tr>
<tr>
<td>Loudness</td>
<td>Loud</td>
<td>Anger (Juslin, 1997)</td>
</tr>
<tr>
<td></td>
<td>Soft</td>
<td>Fear, tenderness, sadness (Juslin, 1997)</td>
</tr>
<tr>
<td>Variation in loudness</td>
<td>Large</td>
<td>Fear (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>Happiness, pleasantness, activity (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td></td>
<td>Rapid</td>
<td>Fear (Krumhansl, 1997)</td>
</tr>
<tr>
<td></td>
<td>changes</td>
<td></td>
</tr>
<tr>
<td>Melodic range</td>
<td>Wide</td>
<td>Fear (Krumhansl, 1997), joy (Balkwill &amp; Thompson, 1999)</td>
</tr>
<tr>
<td></td>
<td>Narrow</td>
<td>Sadness (Balkwill &amp; Thompson, 1999)</td>
</tr>
<tr>
<td>Mode</td>
<td>Major</td>
<td>Happiness (Scherer &amp; Oshinsky, 1977; Krumhansl, 1990)</td>
</tr>
<tr>
<td></td>
<td>Minor</td>
<td>Sadness (Krumhansl, 1997), disgust, anger (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td>Pitch level</td>
<td>High</td>
<td>Surprise, potency, anger, fear, activity (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>Boredom, pleasantness, sadness (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td>Tempo</td>
<td>Fast</td>
<td>Activity, surprise, happiness, pleasantness, potency, fear, anger (Scherer &amp; Oshinsky, 1977), happiness (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>Sadness, boredom, disgust (Scherer &amp; Oshinsky, 1977), sadness, tenderness (Juslin, 1997), sadness (Krumhansl, 1997)</td>
</tr>
<tr>
<td>Timbre</td>
<td>Few</td>
<td>Pleasantness, boredom, happiness, sadness (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td></td>
<td>harmonies</td>
<td>Potency, anger, disgust, fear, activity, surprise (Scherer &amp; Oshinsky, 1977)</td>
</tr>
<tr>
<td></td>
<td>Many</td>
<td></td>
</tr>
<tr>
<td></td>
<td>harmonics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Soft</td>
<td>Tenderness, sadness (Juslin, 1997)</td>
</tr>
<tr>
<td></td>
<td>Sharp</td>
<td>Anger (Juslin, 1997)</td>
</tr>
</tbody>
</table>

Table 7.1.: Mappings between musical parameters and perceived emotion
(adapted from Gabrielsson and Lindström 2001)
one. In an increasing duration order, these temporal processes are usually labeled as emotion, mood, trait, plot or disorder (Ekman, 1984). Several previous studies have looked at the time response of musical stimuli. For instance, categorization experiments showed that as little as 250 ms of musical exposure were needed to classify the genre of a musical piece (Gjerdingen and Perrott, 2008). This suggests that cognitive decisions regarding musical content can be made relatively fast. Similar experiments focusing on emotional responses have found that 1 second of a musical excerpt was enough to elicit a consistent and reliable emotional reaction in the listener (as compared to listening to the whole excerpt duration). (Bigand et al., 2005). Another study by Peretz et al. 2001 showed that 250 ms of music were enough to distinguish fast from happy excerpts. This suggest that emotional responses to music also rely on fast mechanisms that do not necessarily depend on cortical processes. This contradicts most of conventional music cognition research that traditionally emphasized the role of cognitive mechanisms such as expectation in musical emotions (Jackendoff, 1991; Meyer, 2001, 1956).

7.5. Underlying Mechanisms of Musical Emotions

Studies on musical emotion have been influenced by traditional research on emotion and usually assumed that musical emotion relied on general emotional mechanisms such as cognitive appraisal. Based on seminal works in musicology and psychology of music (Meyer, 1956; Mandler, 1984; Narmour, 1990), the concept of musical expectation has recently re-emerged as a fundamental notion that could unify different mechanisms at stake with music and emotion (Huron, 2006). This idea is developed in Huron's ITPRA theory that proposes five different response systems namely Imagination and Tension (which are pre-outcome responses) and Prediction, Reaction and Appraisal (which are post-outcome responses) (Huron, 2006). Imagination responses correspond to the contemplation of a potential future. Tension responses are a preparation for a forthcoming event and involve physiological changes of arousal. Prediction responses are short term rewards that depend on the validation of expectations. The Reaction response is a fast process that occurs automatically and sub-cortically. Appraisal responses on the contrary are a conscious assessment of the stimulus outcome. One interesting aspect of the ITPRA theory is that the relative amplitude of each specific response for a given stimulus could explain the formation of complex emotions such as awe (Huron, 2006; Thompson, 2009). However, the ITPRA theory is still quite new and there generally is a lack of specific detailed experimental evidence. In fact, on could argue that there are many mechanisms involved in emotional responses to music that do not directly relate to expectation.

Indeed, a recent review article (Juslin and Västfjäll, 2008) attempted to distinguish and describe all those mechanisms. Starting from the observation that most emotional responses to music do not involve implications for goals in life, the article challenges the cognitive appraisal approach, and put forward six additional mechanisms that should be accounted for. Namely, brain stem reflex, evaluative conditioning, emotional con-
7. Introduction

tagion, visual imagery, episodic memory and musical expectancy. Brain stem reflex is a subcortical mechanism responding to basic acoustics features of a sound that may be related to its biological relevance (e.g. loud and sudden sounds could represent an imminent danger). Evaluative conditioning happens when a specific sound stimulus is repeatedly associated with a positive or negative event and hence itself acquires a specific emotional meaning via a phenomenon called classical conditioning (Pavlov, 1927). Emotional contagion describes a phenomenon that occurs when perceiving an emotion. The fact of perceiving the expression of an emotion can lead to actually feel the same emotion. Visual imagery occurs when a sound evokes a powerful imagery, which in turn evokes an emotional response. Episodic memory relates to memories of emotional events that happened while a specific sound or music was heard. Musical expectancies relate to the ITPRA theory and how music can induce emotional responses by playing with the listener's expectation.

In Chapter 9, we describe an experiment with dementia patients that supports the hypothesis of important subcortical processes involved in the mechanisms of emotional responses.

7.6. Measuring Emotional Responses

There exists a number of different methodologies to measure emotional responses. They can be broadly classified into three categories depending on the extent to which they access subjective experience, alterations of behavior or the impact on physiological states (Levensons, 1994). Self-reports of emotional experience measure the participant’s “subjective feeling” of emotions, i.e. “the consciously felt experience of emotions as expressed by the individual” (Stout, 1986). They can be subcategorized into verbal (Watson, 1988) and visual self-reports (Russell, 1980). Behavioral responses include explicit responses such as facial expression and gestures. Finally, physiological responses such as electrodermal activity, heart rate, or respiration correspond to implicit responses from the autonomic nervous system. Generally, the studies involving music and emotion are based on the same paradigm: emotional responses are measured while the subject is presented with a sound sample with specific acoustic characteristics or with an excerpt of music representative of a certain type of emotions (Meyer, 1956; Gabrielsson and Lindström, 2001; Le Groux et al., 2008b; Krumhansl, 1997a; Bradley and Lang, 2000; Le Groux and Verschure, 2010b).

7.6.1. Self-reports

Two main models of emotion have traditionally been emphasized. Each one is associated with specific self-report measures. The theory of basic emotion proposes that emotions are discrete and are based on a a set of innate emotions such as sadness, happiness, fear, anger, disgust and guilt (Ekman, 1992); while the dimensional model put forward
7.6. Measuring Emotional Responses

a representation where emotions are composed of core continuous dimensions such as valence and arousal (Russell, 1980). The most commonly used measures of the effect of music on emotion are free verbal description, affective scales and emotional spaces.

In the free description approach, participants are asked to freely describe a piece of music (Gabrielsson and Wik, 2003). An obvious drawback of this method is the lack of standardization. The adjective scale approach uses a somewhat more standardized methodology based on a number of descriptive scales such as anger, fear, confidence or elation to describe musical extracts (Hevner, 1936). Here, the problem of the choice of the set of basic relevant scales still remains. Moreover, those measures rely on theories of basic emotions which states that an independent neural system subserves every emotion. This seems to be in contradiction with a large number of empirical observations from studies in affective neuroscience (Posner et al., 2005).

On the contrary, emotional space approaches rely on generic, well-established dimensional models of emotion that states that all affective states arise from cognitive interpretations of core neural sensations that are the product of independent neurophysiological systems. First proposed in the nineteenth century by Wundt (Wundt, 1904), the dimensional approach to emotion has received support from many independent researchers (Schlosberg, 1954; Russell, 1980; Plutchik, 1962; Osgood et al., 1965; Niedenthal and Setterlund, 1994; Davitz, 1964; Posner et al., 2005). Originally, the proposed dimensions were pleasantness-unpleasantness (valence), excitement-relaxation (arousal) and tension-release. Yet, the relevance of the third dimension of tension-release has been subject to controversy. The tendency nowadays is to replace the tension-release dimension by the potency (Schlosberg, 1954) or dominance-submissiveness (Russell and Mehrabian, 1977) dimension. One of the most popular measure of emotion is the visual self-report scale used by the International Affective Picture System (IAPS) (Lang et al., 1999) called Self Assessment Manikin (SAM) (Lang, 1980) (Figure 7.2).

Another widely used measure is a bi-dimensional space called the circumplex model of affects (Figure 7.3) that represents emotional responses as a point on the dimension of arousal and valence (Russell, 1980). This model has been found to be consistent with recent findings from behavioral, cognitive neuroscience, neuroimaging and developmental studies of affect (Posner et al., 2005).

One should note that generally these measures are given after the participant has listened to the musical stimulus. This might be seen has a problem in the context of music since it is an idiom that develops in time, and most probably emotional responses to music somehow follow this time-varying evolution. One option is to limit the experiment to short specific musical extracts and collect the ratings post-performance. Another approach is to explicitly look at moment-by-moment variations of the emotional ratings in real-time. This type of real-time assessment is mostly based on the circumplex model of affect (Russell, 1980) and allows continuous capture of the participant’s responses using graphical user interfaces (Schubert, 2004; Cowie et al., 2000; Krumhansl, 1997a). Nevertheless, the improvement brought by online ratings methods over retrospective assessment is not clear, as the high cognitive load associated with the rating of emotional
7. Introduction

Figure 7.2.: The **Self Assessment Manikin (SAM) scale** evaluates emotional responses on the scale of Valence (top), Arousal (middle) and Dominance (bottom) (Lang, 1980)

Figure 7.3.: The **circumplex model of affects** represents emotional responses as a point on the dimension of arousal and valence (Russell, 1980)
dimensions online could interfere with the emotion induction process itself (Rosenberg and Ekman, 2005).

### 7.6.2. Physiology

There has been a great deal of studies looking at the relationship between physiology and emotions in general, but to a lesser extend to music-induced emotion and physiology. Traditionally, most of the researchers interested in the link between musical features and emotion have relied on verbal reports. Nevertheless, as biosignal monitoring became widely available and reliable, the study of the bodily responses to music received an increased attention.

As a matter of fact, there exists a strong relationship between the affective state of humans and somatic changes controlled by the autonomic nervous system such as the skin conductance, heart rate, pupil dilation, respiration, blood volume pulse (Andreassi, 2006), and the usefulness of these measures for the study of human computer interaction has been recognized (Lin et al., 2005; Picard, 2000) (also see Section 3.3.3).

Heart rate variability has been described as reflecting the emotional significance or valence of stimuli. Negative stimuli have been shown to slow down the heart rate while positive stimuli have the opposite effect (Bradley and Lang, 2000; Bradley et al., 1996) and the changes due to negative valence are longer lasting than those induced by positive valence (Brosschet and Thayer, 2003). The conductive changes of the skin due to the activity of the sweat glands often referred to as the Galvanic Skin Response (GSR) or Electro Dermic Response (EDR) has been recognized as a direct measure of stress and arousal (Critchley, 2002). In addition, it has become clear which neuronal systems underlie the changes in EDR. Functional magnetic resonance imaging (fMRI) studies have shown a distinction between a cognitive anticipatory EDR controlled by the ventromedial prefrontal cortex and an associative valence based EDR controlled by the amygdala (Critchley, 2002). Electroencephalography (EEG) is also widely used as a means to observe specific mental states. The study of brainwaves measured by EEG has a rich history, and different brainwave activities have been shown to correlate with different mental states such as concentration, frustration, relaxation. For instance, an increase of energy in the alpha wave frequency typically correlates with states of relaxation (Nunez, 2005).

In many studies involving physiology, the music stimuli, due to its intrinsic complexity and high-dimensionality, was not well-controlled, which did not allow to draw strong conclusions from the experiments (Gomez and Danuser, 2007). One of the first significant study that methodically looked at the impact of musical features on emotion involved measures of the skin conductance, heart rate, vascular and respiratory functions of participants while they were listening to musical extracts (Krumhansl, 1997a). The participants were asked to rate the extracts on emotional scales that provided a measure of happiness, sadness, fear, and tension. The relationship between changes in the physiology and the ratings was then evaluated. The study showed higher changes in
7. Introduction

heart rate, blood pressure and skin conductance for sad excerpts, while happy excerpts produced larger changes in respiration and fear excerpts larger changes in blood pressure (also see Section 3.3.3 and Chapter 12). In a later study, Bradley and Lang (Bradley and Lang, 2000) conducted a similar experiment with environmental sounds but including additional measures such as electromyography (EMG) and startle reflexes. The study revealed that unpleasant sounds provoked bigger startle reflexes, increased EMG activity and lower heart rate, while skin conductance varied more for arousing sounds. Overall the most significant signals were the skin conductance and heart rate, especially on the emotional dimension of arousal.

In fact, while autonomic and physiological recordings have proved to be good indicators of emotion perception in music, it is still not clear how those responses can be mapped to emotional scales (e.g. of valence or arousal). We approach this problem in the pilot study described in Chapter 12.

7.7. Summary

There is ample evidence that music provokes strong emotional responses in the form of physiological arousal, motor expression and subjective feelings. Yet, how this happens still remains unclear. One question concerns the perceptual determinants of emotional responses to music. What are the specific musical parameters that can evoke emotional responses and how can they be mapped to emotional states of valence, arousal and dominance? How does this relate to physiological responses? Parameters such as musical structure, performance and timbre seem to play an important role. In the following chapters, we propose a quantitative and qualitative approach that uses the synthetic composition system called SMuSe to generate controlled synthetic stimuli. Chapter 8 presents an experimental study of the impact of a set of well-defined musical parameters on emotional responses, and Chapter 12 investigates the mappings between musical parameters, emotion and physiology. It is known that sound is processed by both cortical and subcortical areas; yet, the role of subcortical processes in musical emotion has mostly been ignored by traditional musicologists who emphasized the role of expectation and cognitive appraisal. Evidence for the importance of subcortical processes is given in Chapter 9 in the form of a large scale study involving patients who suffer from cognitive deficits of the dementia type. Applications for diagnosis and therapy are proposed.
8. A Synthetic Approach to the Study of Musically-induced Emotions

It is widely acknowledged that music can evoke emotions and synchronized reactions of experiential, expressive and physiological components of emotion have been observed while listening to music (Lundqvist et al., 2009). A key question is how musical parameters can be mapped to emotional states of valence, arousal and dominance. In most of the cases, studies on music and emotion are based on the same paradigm: one measures emotional responses while the participant is presented with an excerpt of recorded music. These recordings are often extracted from well-known pieces of the repertoire and interpreted by human performers who follow specific expressive instructions. One drawback of this methodology is that expressive interpretation can vary quite a lot from one performer to another, which compromises the generality of the results. Moreover, it is difficult, even for a professional musician, to accurately modulate one single expressive dimension independently of the others. Many dimensions of the stimuli might not be controlled for. Besides, pre-made recordings do not provide any control over the musical content and structure.

In this chapter, we propose to tackle these limitations by using the synthetic composition system SMuSe to generate stimuli for the experiment. The SMuSe is able to generate ecologically valid musical pieces and to modulate expressive musical material such as pitch, velocity, articulation, tempo, scale, mode, harmony and timbre. It provides accurate, replicable and independent control over perceptually relevant time-varying dimensions of music (Cf. Chapter 3).

We present a synthetic psychology experiment where the SMuSe generates a set of well-defined musical stimuli that vary in structure (mode, register), expressive performance parameters (tempo, dynamics articulation), and timbre (attack time, brightness, and damping). The emotional responses we obtained from a group of 13 participants on the scale of valence, arousal and dominance were similar to previous studies that used human-produced musical excerpts. This shows that the SMuSe is able to reliably evoke coherent emotional responses in a controlled manner.
8. A Synthetic Approach to the Study of Musically-induced Emotions

![SMuSe client's graphical user interface](http://www.dtic.upf.edu/~slegroux/phd/downloads.html)

Figure 8.1. **SMuSe client’s graphical user interface** is used to generate a well-parameterized musical stimulus by modulating macro-level parameters such as articulation, tempo, velocity, density of notes, mode, register and instrumentation. Those parameters are sent to the SMuSe server from the GUI via OSC messages (Cf. Section 3.2.4).

8.1. Methods

8.1.1. Stimuli

This experiment investigates the effects of a set of well-defined musical parameters within the three main musical determinants of emotions, namely structure, performance and timbre. In order to obtain a well-parameterized set of stimuli, all the sound samples were synthetically generated. The composition engine SMuSe described in Chapter 3 allowed the modulation of macro-level musical parameters via a graphical user interface (Figure 8.1), while the physically-informed synthesizer PhySynth\(^1\) (Cf. Chapter 6) allowed to control micro-level sound parameters (Cf. Figure 6.4 for an description of PhySynth’s GUI). Each parameter was considered at three different levels (Low, Medium, High). All the sound samples were 5s. long and normalized in amplitude with the Peak Pro audio editing and processing software (BIAS\(^2\)).

**Musical Structure:** To look at the influence of musical structure on emotion, we focused on two simple but fundamental structural parameters namely register (Bass, Tenor and Soprano) and mode (Random, Minor, Major). A total of 9 sound samples (3 Register * 3 Mode levels) were generated by SMuSe (Figure 8.2).

\(^{1}\)http://www.dtic.upf.edu/~slegroux/phd/downloads.html

\(^{2}\)http://www.bias-inc.com/
8.1. Methods

Figure 8.2.: **Musical structure samples:** 2 structural parameters (Register and Mode) are modulated over 9 sequences (3*3 combinations).

**Expressivity Parameters:** Our study of the influence of musical performance parameters on emotion relies on three expressive parameters, namely tempo, dynamics, and articulation that are commonly modulated by live musicians during performance. A total of 27 sound samples (3 Tempo * 3 Dynamics * 3 Articulation) were generated by SMuSe (Figure 8.3).

Figure 8.3.: **Musical performance samples:** 3 performance parameters were modulated over 27 musical sequences (3*3*3 combinations of Tempo (BPM), Dynamics (MIDI velocity) and Articulation (duration factor) levels).

**Timbre:** For timbre, we focused on parameters that relate to the three main dimension of timbre namely brightness (controlled by tristimulus value), attack-time and spectral flux (controlled by damping). A total of 27 sound samples (3 Attack Time * 3 Brightness * 3 Damping) were generated by PhySynth (Figure 8.4).
8. A Synthetic Approach to the Study of Musically-induced Emotions

Figure 8.4: **Timbre samples**: 3 timbre parameters are modulated over 27 samples (3*3*3 combinations of Attack (ms), Brightness (tristimulus band), Damping (relative damping $\alpha$)). The other parameters of PhySynth were fixed: decay=300ms, sustain=900ms, release=500ms and global damping $\alpha_g = 0.23$.

### 8.1.2. Procedure

We investigated the influence of different sound features on the emotional state of the patients using a fully automated and computer-based stimulus presentation and response registration system. In our experiment, each subject was seated in front of a PC computer with a 15.4" LCD screen and interacted with custom-made stimulus delivery and data acquisition software called PsyMuse\(^3\) (Figure 8.5) made with the Max-MSP \(^4\) programming language (Zicarelli, 2002). Sound stimuli were presented through headphones (K-66 from AKG).

At the beginning of the experiment, the subject was exposed to a sinusoidal sound generator to calibrate the sound level to a comfortable level and was explained how to use PsyMuse’s interface (Figure 8.5). Subsequently, a number of sound samples with specific sonic characteristics were presented together with the different scales (Figure 8.5) in three experimental blocks (structure, performance, timbre) containing all the sound conditions presented randomly.

For each block, after each sound, the participants rated the sound in terms of its emotional content (valence, arousal, dominance) by clicking on the SAM manikin representing her emotion (Lang, 1980). The participants were given the possibility to repeat the playback of the samples. The SAM graphical scale was converted into an ordinal scale (from 0 to 4) where 0 corresponds to the most dominated, aroused and positive and 4 to the most dominant, calm and negative (Figure 8.5). The data was automatically stored

\(^3\)http://www.dtic.upf.edu/~slegroux/phd/downloads.html

\(^4\)http://cycling74.com/
8.2. Results

Figure 8.5.: The presentation software PsyMuse uses the SAM scales (axes of Dominance, Arousal and Valence) (Lang, 1980) to measure the participant’s emotional responses to a database of sounds.

into a SQLite\textsuperscript{5} database composed of a table for demographics and a table containing the emotional ratings. SPSS\textsuperscript{6} (from IBM) statistical software suite was used to assess the significance of the influence of sound parameters on the affective responses of the subjects.

8.1.3. Participants

A total of $N=13$ university students ($5$ women, $M_{age} = 25.8$, range $= 22-31$) with normal hearing took part in the pilot experiment. The experiment was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki\textsuperscript{7}. Six of the subjects had musical background ranging from two to seven years of instrumental practice.

8.2. Results

The experiment followed a within participant design where for each block (structure, performance, timbre) every participant experienced all the conditions.

\textsuperscript{5}http://www.sqlite.org/
\textsuperscript{6}http://www.spss.com/
\textsuperscript{7}http://www.wma.net/en/30publications/10policies/b3/index.html
8. A Synthetic Approach to the Study of Musically-induced Emotions

8.2.1. Musical Structure

To study the emotional effect of the structural aspects of music, we looked at two independent factors (register and mode) with three levels each (soprano, bass, tenor and major, minor, random respectively) and three dependent variables (Arousal, Valence, Dominance). The Kolmogorov-Smirnov test showed that the data is normally distributed. Hence, we carried a Two-Way Repeated Measure Multivariate Analysis of Variance (MANOVA).

The analysis showed a multivariate effect for the mode * register interaction $V(12,144) = 1.92, p < 0.05$. Mauchly tests indicated that assumption of sphericity is met for the main effects of register and mode as well as for the interaction effect. Hence we did not correct the F-ratios for follow-up univariate analysis.

Follow-up univariate analysis revealed an effect of register on arousal $F(2,24) = 2.70, p < 0.05$ and mode on valence $F(2, 24) = 3.08, p < 0.05$ as well as a mode * register interaction effect on arousal $F(4,48) = 2.24, p < 0.05$, dominance $F(4,48) = 2.64, p < 0.05$ and valence $F(4,48) = 2.73, p < 0.05$ (Cf. Table 8.6).

<table>
<thead>
<tr>
<th>A)</th>
<th>Register</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>Arousal</td>
<td>1.97 (.16)</td>
<td>1.87 (.13)</td>
</tr>
<tr>
<td>Valence</td>
<td>1.85 (.20)</td>
<td>1.85 (.20)</td>
</tr>
<tr>
<td>Dominance</td>
<td>2.46 (.19)</td>
<td>2.36 (.22)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B)</th>
<th>ANOVAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register</td>
<td>Mode</td>
</tr>
<tr>
<td>Arousal</td>
<td>$F(2,24)=2.70, *p&lt;0.05$</td>
</tr>
<tr>
<td>Valence</td>
<td>$F(2,24)=3.079, *p&lt;0.05$</td>
</tr>
<tr>
<td>Dominance</td>
<td>$F(2,48)=2.731, *p&lt;0.05$</td>
</tr>
</tbody>
</table>

Figure 8.6.: Effect of mode and register on the emotional scales of arousal, valence and dominance. A) shows mean and standard error values B) shows the statistically significant effects.

A post-hoc pairwise comparison with Bonferroni correction showed a significant mean difference of -0.3 between High and Low register (1 vs. 2) and of -0.18 between High and Medium (1 vs. 3) on the arousal scale (Figure 8.7 B). High register appeared more arousing than medium and low register.

A pairwise comparison with Bonferroni correction showed a significant mean difference of -0.436 between random and major (3 vs. 1) (Figure 8.7 A). Random mode was perceived as more negative than major mode.
8.2. Results

Figure 8.7.: **Influence of structural parameters (register and mode) on arousal and valence.** A) A musical sequence played using random notes (level 3) and using a minor scale (level 2) is perceived as significantly more negative than a sequence played using a major scale (level 1). **B)** A musical sequence played in the soprano range (high register level 1) (respectively bass range, low register level 2) is significantly more (respectively less) arousing than the same sequence played in the tenor range (medium register level 3). Estimated Marginal Means are obtained by taking the average of the means for a given condition.

The interaction effect between mode and register suggests that the random mode (level 3) has a tendency to make a melody with medium register less arousing (Figure 8.8, A). Moreover, the minor mode tended to make high register more positive and low register more negative (Figure 8.8, B). The combination of high register and random mode created a sensation of dominance (Figure 8.8, C).

8.2.2. Expressive Performance Parameters

To study the emotional effect of some expressive aspects of music during performance, we decided to look at three independent factors (Articulation, Tempo, Dynamics) with three levels each (high, low, medium) and three dependent variables (Arousal, Valence, Dominance). The Kolmogorov-Smirnov test showed that the data was normally distributed. We did a Three-Way Repeated Measure Multivariate Analysis of Variance.

The analysis showed a multivariate effect for **Articulation** $V(4.16, 3) < 0.05$, **Tempo** $V(11.6, 3) < 0.01$ and **dynamics** $V(34.9, 3) < 0.01$. No interaction effects were found.

Mauchly tests indicated that the assumption of sphericity was met for the main effects of articulation, tempo and dynamics on arousal and valence but not dominance. Hence we corrected the F-ratios for univariate analysis for dominance with Greenhouse-Geisser.
8. A Synthetic Approach to the Study of Musically-induced Emotions

Figure 8.8: **Structure: interaction between mode and register for arousal, valence and dominance.** A) When using a random scale, a sequence in the tenor range (level 3) becomes less arousing B) When using a minor scale (level 2), a sequence played within the soprano range (level 1) becomes the most positive. C) When using a random scale, bass and soprano sequences are the most dominant whereas tenor becomes the less dominant.
8.2. Results

<table>
<thead>
<tr>
<th></th>
<th>Articulation</th>
<th>Tempo</th>
<th>Dynamics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Arousal</strong></td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>Articulation</td>
<td>2.04 (0.15)</td>
<td>1.73</td>
<td>(0.12)</td>
</tr>
<tr>
<td>Tempo</td>
<td>2.17 (0.18)</td>
<td>2.06</td>
<td>(0.16)</td>
</tr>
<tr>
<td>Dynamics</td>
<td>2.49 (0.16)</td>
<td>1.51</td>
<td>(0.14)</td>
</tr>
<tr>
<td><strong>Valence</strong></td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>Articulation</td>
<td>2.15 (0.17)</td>
<td>1.87</td>
<td>(0.17)</td>
</tr>
<tr>
<td>Tempo</td>
<td>2.20 (0.12)</td>
<td>2.64</td>
<td>(0.1)</td>
</tr>
<tr>
<td>Dynamics</td>
<td>2.23 (0.12)</td>
<td>1.89</td>
<td>(0.16)</td>
</tr>
<tr>
<td><strong>Dominance</strong></td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>Articulation</td>
<td>2.21 (0.15)</td>
<td>1.701</td>
<td>(0.21)</td>
</tr>
<tr>
<td>Tempo</td>
<td>2.13 (0.17)</td>
<td>2.52</td>
<td>(0.2)</td>
</tr>
<tr>
<td>Dynamics</td>
<td>2.23 (0.16)</td>
<td>1.89</td>
<td>(0.16)</td>
</tr>
</tbody>
</table>

Figure 8.9: **Effect of articulation, tempo and dynamics** on self-reported emotional responses on the scale of valence, arousal and dominance. **A)** shows mean and standard error values **B)** shows the statistically significant effects.

### 8.2.2.1. Arousal

Follow-up univariate analysis revealed an effect of **articulation** $F(6.76, 2) < 0.01$, **tempo** $F(27.1, 2) < 0.01$, and **dynamics** $F(45.77, 2) < 0.05$ on arousal (Table 8.9).

A post-hoc pairwise comparison with Bonferroni correction showed a significant mean difference of 0.32 between the **articulation** staccato and legato (level 1 vs. 2) (Figure 8.10 A). The musical sequence played staccato was perceived as more arousing.

A pairwise comparison with Bonferroni correction showed a significant mean difference of -1.316 between high **tempo** and low tempo (1 vs. 2) and -0.89 between high and medium tempo (1 vs. 3) (Figure 8.10 B). This shows that a musical sequence with higher tempi was perceived as more arousing.

A pairwise comparison with Bonferroni correction showed a significant mean difference of -0.8 between forte and piano **dynamics** (1 vs. 2), -0.385 between forte and regular (1 vs. 3) and 0.41 between piano and regular (2 vs. 3) (Figure 8.10 C). This shows that a musical sequence played at higher dynamics was perceived as more arousing.
8. A Synthetic Approach to the Study of Musically-induced Emotions

Figure 8.10.: **Effect of performance parameters (Articulation, Tempo and Dynamics) on Arousal.** A) A sequence played with articulation staccato (level 2) is more arousing than legato (level 1) B) A sequence played with the tempo indication presto (level 1) is more arousing than both moderato (level 3) and lento (level 2). C) A sequence played forte (respectively piano) was more arousing (respectively less arousing) than the same sequence played mezzo forte (level 3).

### 8.2.2.2. Valence

Follow-up univariate analysis revealed an effect of **articulation** $F(7.31, 2) < 0.01$, **tempo** $F(4.3, 2) < 0.01$, and **dynamics** $F(18.9, 2) < 0.01$ on valence (Table 8.9).

A post-hoc pairwise comparison with Bonferroni correction showed a significant mean difference of -0.32 between the **articulation** staccato and legato (level 1 vs. 2) (Figure 8.10 A). The musical sequences played with shorter articulations were perceived as more positive.

A pairwise comparison with Bonferroni correction showed a significant mean difference of 0.48 between high **tempo** and medium tempo (1 vs. 3) (Figure 8.11 B). This shows that sequences with higher tempi tended be perceived as more negatively valenced.
8.2. Results

A pairwise comparison with Bonferroni correction showed a significant mean difference of 0.77 between high and low dynamics (level 1 vs. 2) and -0.513 between low and medium (level 2 vs. 3). (Figure 8.11 C). This shows that musical sequences played with higher dynamics were perceived more negatively.

Figure 8.11.: Effect of performance parameters (Articulation, Tempo and Dynamics) on Valence. A) A musical sequence played staccato (level 2) induce a more negative reaction than when played legato (level 1) B) A musical sequence played presto (level 1) is also inducing a more negative response than played moderato (level 3). C) A musical sequence played forte (respectively piano) is rated as more negative (respectively positive) than a sequence played mezzo forte.

8.2.2.3. Dominance

Follow-up univariate analysis revealed an effect Tempo $F(8, 2) < 0.01$, and dynamics $F(9,7, 2) < 0.01$ on valence (Table 8.9).

A pairwise comparison with Bonferroni correction showed a significant mean difference of -0.821 between high tempo and low tempo (1 vs. 2) and -0.53 between high tempo
8. A Synthetic Approach to the Study of Musically-induced Emotions

and medium tempo (level 1 vs 3) (Figure 8.12 A). This shows that sequences with higher tempi tended to make the listener feel dominated.

A pairwise comparison with Bonferroni correction showed a significant mean difference of -0.55 between high and low dynamics (1 vs. 2) and 0.308 between low and medium (2 vs. 3). (Figure 8.12 B). This shows that when listening to musical sequences played with higher dynamics, the participants felt more dominated.

![Figure 8.12: Effect of performance parameters (Tempo and Dynamics) on Dominance. A) A musical sequence played with a tempo presto (respectively lento) is considered more dominant (respectively less dominant) than played moderato (level 3) B) A musical sequence played forte (respectively piano) is considered more dominant (respectively less dominant) than played mezzo-forte (level 3)](image)

8.2.3. Timbre

To study the emotional effect of the timbral aspects of music, we decided to look at three independent factors known to contribute to the perception of Timbre (McAdams et al., 1995a; Grey, 1977; Lakatos, 2000) (Attack time, Damping and Brightness) with three levels each (high, low, medium) and three dependent variables (Arousal, Valence, Dominance). The Kolmogorov-Smirnov test showed that the data is normally distributed. We did a Three-Way Repeated Measure Multivariate Analysis of Variance.

The analysis showed a multivariate effect for **brightness** \(V(6,34) = 3.76, p < 0.01\), **damping** \(V(6,34) = 3.22, p < 0.05\) and **attack time** \(V(6,34) = 4.19, p < 0.01\) and an interaction effect of **brightness \* damping** \(V(12,108) = 2.8 < 0.01\)

Mauchly tests indicated that assumption of sphericity was met for the main effects of articulation, tempo and dynamics on arousal and valence but not dominance. Hence we corrected the F-ratios for univariate analysis for dominance with Greenhouse-Geisser.
8.2. Results

<table>
<thead>
<tr>
<th>A)</th>
<th>Brightness</th>
<th>Damping</th>
<th>Attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arousal</td>
<td>High (1.96 (.24))</td>
<td>Low (3.13 (.19))</td>
<td>Medium (2.4 (.24))</td>
</tr>
<tr>
<td></td>
<td>High (2.89 (.19))</td>
<td>Low (2.12 (.23))</td>
<td>Medium (2.49 (.23))</td>
</tr>
<tr>
<td></td>
<td>Medium (2.4 (.22))</td>
<td>Low (2.47 (.2))</td>
<td>Medium (2.58 (.2))</td>
</tr>
<tr>
<td>Valence</td>
<td>High (1.6 (.18))</td>
<td>Low (1.97 (.07))</td>
<td>Medium (1.92 (.08))</td>
</tr>
<tr>
<td></td>
<td>Medium (2.1 (.1))</td>
<td>Low (1.94 (.08))</td>
<td>Medium (2.1 (.07))</td>
</tr>
<tr>
<td></td>
<td>Low (1.9 (.08))</td>
<td>Medium (1.95 (.05))</td>
<td></td>
</tr>
<tr>
<td>Dominance</td>
<td>High (2.21 (.01))</td>
<td>Low (2.11 (.014))</td>
<td>Medium (2.13 (.017))</td>
</tr>
<tr>
<td></td>
<td>Medium (2.37 (.285))</td>
<td>Low (2.7 (.26))</td>
<td>Medium (2.62 (.27))</td>
</tr>
<tr>
<td></td>
<td>Low (2.7 (.25))</td>
<td>Medium (2.75 (.24))</td>
<td></td>
</tr>
</tbody>
</table>

B) | Brightness | Damping | Attack | Brightness*Damping |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arousal</td>
<td>F(2,18)=29.09,***p&lt;0.001</td>
<td>F(2,18)=16.03,***p&lt;0.001</td>
<td>F(2,18)=3.54,*p&lt;0.05</td>
<td></td>
</tr>
<tr>
<td>Valence</td>
<td>F(2,18)=5.99,**p&lt;0.01</td>
<td>F(2,18)=7.26,**p&lt;0.01</td>
<td>F(4,36)=5.82,**p&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Dominance</td>
<td>F(1.49,13.45)=6.56,*p&lt;0.05</td>
<td>F(1.05,10.915)=4.7, *p&lt;0.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 8.13: **Effect of brightness, damping and attack** on self-reported emotion on the scales of valence, arousal and dominance. **A)** shows mean and standard error values **B)** shows the statistically significant effects.

8.2.3.1. Arousal

Follow-up univariate analysis revealed the main effects of **Brightness** $F(2, 18) = 29.09 < 0.001$, **Damping** $F(2, 18) = 16.03 < 0.001$, **Attack** $F(2, 18) = 3.54 < 0.05$, and interaction effect **Brightness * Damping** $F(4, 36) = 7.47, p < 0.001$ on Arousal (Figure 8.13).

A post-hoc pairwise comparison with Bonferroni correction showed a significant mean difference between high, low and medium **brightness**. There was a significant difference of -1.18 between high and low brightness (level 1 vs. level 2), -0.450 between high and medium (level 1 and 3) and -0.73 between medium and low (level 3 and 2). The brighter the sounds the more arousing.

Similarly significant mean difference of .780 between high and low (level 1 and 2) **damping** and -0.37 between low and medium (level 2 and 3) damping were found. The more damped, the less arousing.

For the **attack time** parameter, a significant mean difference of -0.11 was found between short and medium attack (level 2 and 3). Shorter attack time were found more arousing.
8. A Synthetic Approach to the Study of Musically-induced Emotions

Figure 8.14.: **Effect of timbre parameters (Brightness, Damping and Attack time) on Arousal.** A) Brighter sounds induced more arousing responses. B) Sounds with more damping were less arousing. C) Sounds with short attack time (level 2) were more arousing than medium attack time (level 3). D) Interaction effects show that less damping and more brightness lead to more arousal.

### 8.2.3.2. Valence

Follow-up univariate analysis revealed main effects of **Brightness** $F(2, 18) = 5.99 < 0.01$ and **Attack** $F(2, 18) = 7.26 < 0.01$, and interaction effect **Brightness * Damping** $F(4, 36) = 5.82, p < 0.01$ on Valence (Figure 8.13).

Follow up pairwise comparisons with Bonferroni correction showed significant mean differences of 0.78 between high and low **brightness** (level 1 and 2) and 0.19 between short and long **attacks** (level 1 and 2) and long and medium attacks (level 1 and 3). Longer attacks and brighter sounds were perceived as more negative (Figure 8.15).
8.2. Results

Figure 8.15.: **Effect of timbre parameters (Brightness, Damping and Attack time) on Valence.** A) Longer attack time are perceived as more negative B) Bright sounds tend to be perceived more negatively than dull sounds C) Interaction effects between damping and brightness show that a sound with high damping attenuates the negative valence due to high brightness.

8.2.3.3. Dominance

Follow-up univariate analysis revealed main effects of **Brightness** $F(1.49, 13.45) = 6.55, p < 0.05$ and **Damping** $F(1.05, 10.915) = 4.7, p < 0.05$ on Dominance (Figure 8.13).

A significant mean difference of -0.743 was found between high and low **brightness** (level 1 and 2). The brighter the more dominant.

A significant mean difference of 0.33 was found between medium and low **damping** factor (level 2 and 3). The more damped the less dominant.
8. A Synthetic Approach to the Study of Musically-induced Emotions

Figure 8.16.: Effect of timbre parameters (Brightness and Damping) on Dominance. A) Bright sounds are perceived as more dominant than dull sounds. B) A sound with medium damping is perceived as less dominant than low damping.

8.3. Conclusions

This study validates the use of the SMuSe as an “affective music engine”. The different levels of musical parameters that were experimentally tested provoked significantly different emotional responses. The tendency of minor mode to increase negative valence and of high register to increase arousal (Figure 8.7) corroborates the results of (Krumhansl, 1997a; Scherer and Oshinsky, 1977), and is complemented by interaction effects (Figure 8.8). The tendency of short articulation to be more arousing and more negative (Figure 8.10 and 8.11) confirms results reported in (Juslin, 1997; Juslin and Sloboda, 2001; Friberg et al., 2006). Similarly, higher tempi have a tendency to increase arousal and decrease valence (Figure 8.10 and 8.11) are also reported in (Juslin, 1997; Juslin and Sloboda, 2001; Krumhansl, 1997a; Scherer and Oshinsky, 1977; Gabrielsson and Lindström, 2001; Friberg et al., 2006). The present study also indicates that higher tempi are perceived as more dominant (Figure 8.12). Musical sequences that were played louder were found more arousing and more negative (Figure 8.10 and 8.11) which is also reported in (Juslin, 1997; Juslin and Sloboda, 2001; Krumhansl, 1997a; Scherer and Oshinsky, 1977; Gabrielsson and Lindström, 2001; Friberg et al., 2006), but also more dominant (Figure 8.12). The fact that higher brightness tends to evoke more arousing and negative responses (Figure 8.14 and 8.15) has been reported (but in terms of number of harmonics in the spectrum) in (Scherer and Oshinsky, 1977). Additionally, brighter sounds are perceived as more dominant (Figure 8.16). Damped sounds are less arousing and dominant (Figure 8.14 and 8.16). Sharp attacks are more arousing and more positive (Figure 8.14 and 8.15). Similar results were also reported by (Juslin, 1997). Additionally, this study revealed interesting interaction effects between damping and brightness (Figure 8.14 and 8.15).
8.3. Conclusions

Most of the studies that investigate the determinants of musical emotion use recordings of musical excerpts as stimuli. In this experiment, we looked at the effect of a well-controlled set of synthetic stimuli (generated by the SMuSe) on the listener’s emotional responses. We developed an automated test procedure that assessed the correlation between a few parameters of musical structure, expressivity and timbre with the self-reported emotional state of the participants. Our results generally corroborated the results of previous meta-analyses (Juslin and Sloboda, 2001), which suggests our synthetic system is able to evoke emotional reactions as well as “real” musical recordings. One advantage of such a system for experimental studies though, is that it allows for precise and independent control over the musical parameter space, which can be difficult to obtain, even from professional musicians. Moreover with this synthetic approach, we can precisely quantify the level of the specific musical parameters that led to emotional responses on the scale of arousal, valence and dominance. These results pave the way for an interactive approach to the study of musical emotion, with potential application to interactive sound-based therapies. In the future, a similar synthetic approach could be developed to further investigate the time-varying characteristics of emotional reactions using continuous two-dimensional scales and physiology (Grewe et al., 2007; Schubert, 2004).
9. Towards Sound-Based Therapy and Diagnosis for Dementia

Although music is widely used in therapeutic settings, the precise relationship between musical parameters and affective states is not clear. To address this issue in the context of dementia and Alzheimer Disease (AD), we developed an automated test procedure that assesses the correlation between the sound source, brightness, sound level, tempo and consonance of a set of parameterized sound samples with the self-reported emotional state of dementia patients on the scale of arousal, valence, dominance and pleasantness.

Sound is processed by both cortical and subcortical areas (Peretz, 2010). One could thus expect that in diseases such as AD, which includes a disconnection syndrome (Delbeuck et al., 2003), the experience of the quality of sound sources will change. Here we explicitly investigate this hypothesis with the goal to develop a rational basis for a music-based approach towards the diagnosis of AD and the alleviation of its symptoms.

We found that the source of a sound had a direct impact on the reported emotional state of the patients. Water sounds were reported to be happier, while water and synthetic sounds were more relaxing. This suggests that a rational approach to the design of soundscapes for sonic therapy is possible.

Moreover, our results show that the control and patient groups had significantly different emotional reactions to the sound samples, suggesting possible applications in automatic diagnosis of dementia. Overall, patients tended to be more radical in their ratings. A detailed analysis showed synthetic and medium loudness sounds were less arousing for patients, medium consonance was less dominant and slow tempo was less pleasant.

We also found that the stage of the disease influenced the ratings of dementia and AD patients. The more the disease progressed, the more the sounds were perceived as happy, relaxing, dominant and pleasant. A detailed analysis showed that voice and synthetic sounds, high loudness and high tempo sounds were reported as significantly happier for the most advanced stage of dementia disease. Likewise, voice and water sounds, high and medium loudness sounds, low brightness and high tempo sounds were considered more pleasant as the disease developed. Sounds with high brightness were considered more angry. This increased sensitivity to sound stimuli with the decrease of cognitive abilities puts into evidence the importance of subcortical processes in the perception of sound-based emotions.
9. Towards Sound-Based Therapy and Diagnosis for Dementia

9.1. Introduction

Dementia can be defined as a progressive loss of mental function which results in a restriction of daily activities and in most cases leads in the long term to the need for care (Kurz, 2007). Given the increased life expectancy in the western world, the total number of dementia patients will increase while also their fraction will increase due to a decreased population growth. Although the most common type of dementia is Alzheimer’s Disease (AD) (60-70% of the cases), there are other known types of dementia like vascular dementia, dementia with Lewy bodies, frontotemporal dementia and Creutzfeldt-Jakob disease. The common symptoms of dementia are cognitive impairments such as memory impairment, aphasia, apraxia, disorientation and mental slowing accompanied with non-cognitive deficits as well as psychiatric and behavioral problems (Alzheimer’s association, 2006). The latter are common in more advanced phases of AD including: depression (60-98% of patients), anxiety and agitation (80% of dementia patients), sleep problems (70% of dementia), and wandering. These behavioral problems appear to be the main cause of carer burden (Coen et al., 1997; Lyketsos and Olin, 2002; Tremont et al., 2006) and currently the use of restraints and medication is one of the most common choices to deal with problem behavior (Testad et al., 2005). These disorders have a great impact on the duration of inpatient treatment, reduction of self-care and the probability of nursing home placement (Wancata et al., 2003).

However, the pharmacological treatment of these non-cognitive symptoms is only of limited effect (Sink et al., 2005) and is considered inappropriate due to an increased risk of mortality. Hence, non-pharmacological interventions to combat the behavioral and psychiatric symptoms of AD are urgently needed. However, currently it is not clear what alternative approaches such as sensory interventions, social contact (real or simulated), behavior therapy, staff training, structured activities, environmental interventions, medical/nursing care interventions, and combination therapies have to offer (Cohen-Mansfield, 2001). Here, we investigate the effect of sound properties on the emotional state of AD patients. Our hypothesis is that a disconnection syndrome characterizing AD and involving cortical and subcortical areas (both involved in the processing of sounds) influences the perception of the quality of sounds in patients.

The effect of music on stress and anxiety has been established using a variety of self-report, physiological, and observational means (Khalfa et al., 2003; Hanser, 1985). Music therapy can have physiological and psychological benefits on patients with a variety of diagnoses and undergoing differing medical treatments (Standley, 1986), while the effect of music on stress induced arousal has shown consistent results using physiological, behavioral and self-report measures (Pelletier, 2004). Interestingly enough this analysis also suggested that specific musical properties including slow tempo, low pitches, primarily based on string composition excluding lyrics and regular rhythmic patterns without extreme changes in dynamics are more effective than the subject’s preferred music. Music and music therapy has been shown to affect, at different time scales, a

\[1\text{http://www.fda.gov/cder/drug/advisory/antipsychotics.htm}\]
number of neuromodulatory and neurohormonal systems that are correlated with depression including 5HT and melatonin (Kumar et al., 1999). Although these results show great promise, a more recent meta-analysis has shown that strong conclusions cannot be drawn from the literature on music therapy due to a number of methodological problems (Sung and Chang, 2005). The same conclusion was drawn from a meta-analysis of the effect of music therapy on agitation treatment in AD patients (Lou, 2001).

In this chapter we focus on the study of a parameterized set of sound samples. We propose to investigate the effect of sound features such as timbre, brightness, speed, consonance and sound level on the subject’s emotional state. We present a quantitative psychoacoustics procedure using self-report measures with well-diagnosed and monitored patient groups to define a foundation on which sonic based therapy can be applied to dementia and AD. Our results show that well defined sound samples can be used as a diagnostic for AD and its progression and that in the construction of music therapy systems, the disease’s stage must be taken into account.

9.2. Methods

9.2.1. Participants

A total of 77 subjects (43 female, 34 males) ranging from 50 to 88 years of age (Mean=72.89, SD=9.483), attending two day-care centers in Barcelona participated in this study. Criteria for inclusion in the study included the following: (a) a neurologist diagnosis of dementia, (b) verbal ability sufficient to answer simple verbal instructions, and (c) signed consent form to participate in the study. The patients were classified into three groups: Alzheimer type (AD, DTA, DSTA, Frontal N=58, 75.3%), Mild Cognitive Impairment (MCI: N=5, 6.5%) and others (Vascular, Mixed, Lewy body, Parkinson, Anoxia: N=14, 18.2%) at three levels of impairment: Level 3 (mild decline: N=7, 9.1%), Level 4 (moderate decline: N=44, 57.1%) and Level 5 (moderately severe decline: N=25, 32.5%).

The age-matched control group consisted of 19 subjects (12 females, 7 males) ranging from 62 to 83 years old (Mean=73.87, SD=6.402).

9.2.2. Stimuli

We parameterized the sound space by distinguishing a number of categories (natural environmental, natural human, synthetic abstract) and perceptually relevant features such as Speed, Consonance, Sound level and Brightness. The sound database comprised 75 different sound samples that covered the full parameter space (Figure 9.1).
9. Towards Sound-Based Therapy and Diagnosis for Dementia

Figure 9.1.: **Parametric tree** describing the sound feature space consisting of sound source, speed, consonance, sound level and brightness.

#### 9.2.2.1. Voice Sample Generation

We recorded a female professional singer who was asked to sing three simple patterns (Figures 9.2a, 9.2b and 9.2c).

![Pattern 1](image1)

(a) Pattern 1

![Pattern 2](image2)

(b) Pattern 2

![Pattern 3](image3)

(c) Pattern 3

Figure 9.2.: **The three voice patterns** were simple melodies sang by a female voice.

All samples for sound level, spectrum and tempo were generated using these three patterns as solo voice.

- **Sound level:** from soft to loud there is a variation of 3 dB between each one.

- **Brightness:** the solo voice was filtered to obtain different brightnesses. The “low brightness voice” was low-pass filtered using 440 Hz as cutting frequency and a slope of 3 dB per octave. The “medium brightness voice” was not filtered, it is the original voice. The “high brightness voice” was high-pass filtered using a cutting frequency of 880 Hz and a slope of 3 dB per octave.
9.2. Methods

- **Tempo:** The singer was conducted to perform the melodic patterns according to the tempo variation. The tempo variation was controlled by a pulse click of 60, 100 and 140 BPM respectively.

- **Consonance:** The three original melodic patterns were superposed over each other in order to produce harmonic intervals, yielding to three harmonic patterns. In each harmonic pattern there are four harmonic intervals. Each harmonic pattern is then classified according to the number of dissonant intervals it contains (Figure 9.3).

![Images of musical patterns](a) pattern 1+2, (b) Pattern 1+3, (c) Pattern 2+3

Figure 9.3.: **Parametric definition of consonance:** Pattern 9.3a was considered the highest consonance, since it has unison, a minor third, unison and octave. Pattern 9.3b was considered the medium consonance, since it has unison, a major second, a minor second and unison. Pattern 9.3c was considered the lowest consonance, since it has unison, a minor second, a minor second and octave.

9.2.2.2. **Water Sample Generation**

- **Sound Level:** a variation of 3 dB between each pattern was applied from soft to loud.

- **Brightness:** the synthetic samples were filtered to obtain different brightnesses. The low brightness samples were low-pass filtered using 440 Hz as cutting frequency and a slop of 3 dB per octave. The medium brightness samples were not filtered (original samples). The high brightness samples were high-pass filtered using a cutting frequency of 880 Hz and a slop of 3 dB per octave.

- **Speed:** the speed of the sample was varied by time stretching the initial sample by factors 0.6 (60/100) and 1.4 (140/100) (so the speed ratios correspond to previous BPM ratios)

- **Consonance:**

9.2.2.3. **Synthetic Sample Generation**

The synthetic samples were produced using computer-generated sinusoidal tones multiplied by a trapezoidal amplitude envelope.
9. Towards Sound-Based Therapy and Diagnosis for Dementia

- **Sound Level**: a variation of 3 dB between each pattern was applied from soft to loud.
- **Brightness**: the synthetic samples were filtered to obtain different brightnesses. The low brightness samples were low-pass filtered using 440 Hz as cutting frequency and a slope of 3 dB per octave. The medium brightness samples were not filtered (original samples). The high brightness samples were high-pass filtered using a cutting frequency of 880 Hz and a slope of 3 dB per octave.
- **Tempo**: the rhythmic patterns were built with quarter notes that lasted 1, 0.6 and 0.42 seconds (60, 100 and 140 BPM respectively).
- **Consonance**: three patterns were generated using superposition of layers or voices with different melodic profiles. The consonance-dissonance balance was related to the harmonic intervals produced by superimposed tones. The highest the number of dissonant intervals the lowest the consonance of the sequence.

9.2.3. Procedure

We have investigated the influence of different sound features on the emotional state of the patients using a fully automated and computer based stimulus presentation and response registration system called Z-blast. In our experiment, each subject was seated in front of a PC with a 15.4” LCD screen (SyncMaster 152T, Samsung) and interacted with the Z-blast stimulus delivery and data acquisition software. Audio was presented via a pair of PC loudspeakers (121, Altec Lansing). At the beginning of the experiment, the subject was exposed to a number of sound snippets to calibrate the sound level to a comfortable level. Subsequently, a number of sound samples with specific sonic characteristics were presented together with a dialog (Figure 9.4).

The subject had to rate each sound sample in terms of their emotional content (valence, arousal, dominance, pleasantness) using a percentage scale by moving a slider on the screen (Figure 9.4). After listening to each sound the subjects were asked to give ratings for four main emotional states: on the valence scale from happy (0%) to sad (100%), on the arousal scale from relaxed (0%) to stressed (100%), on the dominance scale from fearful (0%) to angry (100%), on the pleasantness scale from like (0%) to dislike (100%). The subject was given the possibility to skip samples, and a music therapist was available for subjects who had difficulty manipulating the mouse. Each sample was looped for a maximum of 30 seconds or until the subject had responded to all four dialog boxes. The data was collected and stored on the hard drive, and subsequently processed and formatted with Python\(^2\) scripts and stored in a SQLite\(^3\) database. SPSS\(^4\) statistical software suite was used to assess the significance of the influence of sound parameters on the affective responses of the subjects.

\(^2\)http://www.python.org
\(^3\)http://www.sqlite.org/
\(^4\)http://www.spss.com
Figure 9.4: **Zblast** - The stimulus delivery and data acquisition system’s Graphical User Interface: The user rates the sound samples on a scale from happy to sad based on her emotional state. The user can click the box “no lo sé” (I do not know) or “pasar al siguiente” (pass to the next) to skip to the next dialog. The blue bar at the top indicates the duration that the sample has played and will play.

### 9.3. Results

We conducted four main analyses. Firstly, a global comparison of the ratings between the control and the patients groups on the above-mentioned four emotional scales for all sound samples was done. This global analysis was followed by a more detailed between-group analysis for each sound feature separately. Secondly, we pursued a separate within-subject analysis of the influence of the level of different sound features on the emotional response for both the control group and the dementia group. Finally, we undertook a between-group analysis to look at the global effect of the disease level of dementia patients. This global analysis was followed by a more detailed between-group analysis (for the different disease stages) of the effect of each sound feature separately on the subjective emotional ratings.

#### 9.3.1. Global between-groups analysis (control vs. patients)

At first sight, the distribution of emotional responses to the sound stimuli appears quite different between controls and patients. Patients tend to give more extreme ratings than controls (Figure 9.5).

Since Shapiro-Wilk tests confirmed the non-normality of the responses’ distribution, we used the Mann-Whitney non-parametric test to assess the statistical significance of those
9. Towards Sound-Based Therapy and Diagnosis for Dementia

Figure 9.5.: Global distributions of the emotional responses from the patient and control groups. The histograms on the Valence (A), Pleasantness (B), Dominance (C) and Arousal (D) scales show that dementia patients (in blue) have a tendency to be more extreme in their ratings than the control group (in green).

...differences on the four emotional scales of valence, pleasantness, dominance and arousal.

The analysis shows that controls (Mdn=53) generally feel more happy (ratings closer to 0% on the sadness scale) than patients (Mdn=78) when listening to the sound samples, U=127291.5, p<0.01 while patients (Mdn=21.2) feel significantly more relaxed (arousal ratings closer to 0%) than controls (Mdn=47.29) U=80271.5, p<0.01 (Figure 9.6). Another interesting observation concerns the emotional scales we used. The boxplot of the dominance ratings exhibits many outliers, for the control group (Figure 9.6). This suggests that the dominance axis is perceived as confusing (See Figure 9.5).

To summarize, dementia patients give more extreme ratings than the control group and rate sound samples as being more relaxing and less happy. Moreover, patients seem more sensitive to the dominance scale than the controls.
9.3. Results

Figure 9.6.: **Global between-group analysis.** Overall, patients rate the sound samples as sadder (A) and less arousing (B) than controls. Ratings on the the dominance axis exhibit many outliers (C), especially for the control group. This suggests the dominance axis might not be relevant for musical stimuli.

### 9.3.2. Detailed between-group analysis (control vs. patients)

#### 9.3.2.1. Statistical Analysis

Shapiro-Wilk tests confirmed the non-normality of the responses’ distribution, for the “sound type” and “loudness” features, and confirmed normality for “consonance” and “tempo” features respectively. Consequently, we use Mann-Whitney (respectively independent t-test) to assess the statistical significance of those differences on the four emotional scales of valence, pleasantness, dominance and arousal. We report only significant results.

#### 9.3.2.2. Sound Type

**Arousal**  A Mann-Whitney analysis shows that controls (Mdn=42.8) feel more stressed than patients (Mdn=24.5) when listening to synthetic sound samples $U=348$, $p<.05$ (Figure 9.7).

#### 9.3.2.3. Loudness

**Arousal**  A Mann-Whitney analysis shows that controls (Mdn=51.05) feel more stressed than patients (Mdn=48.5) when listening to sounds with medium loudness $U=192.5$, $p<.05$ (Figure 9.7).

#### 9.3.2.4. Consonance

**Dominance**  An independent t-test analysis shows that on average, patients (M=39.29, SE=4.3) felt less dominant than controls (M=54.8, SE=5.57) when listening to samples with medium consonance $t(29)=-2.032$, $p<0.05$ (Figure 9.7).
9. Towards Sound-Based Therapy and Diagnosis for Dementia

9.3.2.5. Tempo

Unpleasantness On average, controls (M=42.8, SE=4.0) like better slow tempo than patients (M=53.74, SE=10.79) t(50.27)=2.78, p<0.05 (Figure 9.7).

9.3.3. Within-subject analysis for the control group

The different factors we are studying are the sound source (synthetic, natural environmental, natural singing voice), and specific sound features such as consonance, speed, sound level and brightness at different levels (low, medium, high) (Figure ).

9.3.3.1. Statistical analysis

The within-subject analysis for the control group follows a repeated-measure design, where the missing values for each factor are replaced by the series mean. The Shapiro-
Wilk tests show that the distribution are not normal for all the different factors, hence we do non-parametric Friedman ANOVA analysis. Follow-up posthoc Wilcoxon comparison tests with Bonferroni correction are used and all effects are reported at a 0.0167 level of significance.

Figure 9.8.: **Within-subject analysis (control group):** Responses to loudness on the arousal scale (low=1, medium=2, high=3) show that low sound levels are less arousing.

### 9.3.3.2. Sound level and arousal

Our analysis shows that the arousal (relaxed vs. exciting) ratings were influenced by the sound level levels ($\chi^2(2)=7.125, p<0.05$). We found a significant difference between the low and medium levels. Medium levels (Mdn=53.38) were rated as less relaxing than low levels (Mdn=45.58), $T=26$, $z=-2.172$, $p<0.0167$ (Figure 9.8). Other comparisons were not significant.

### 9.3.4. Within-subject analysis for the dementia group

**9.3.4.1. Statistical Analysis**

The within-subject analysis for the dementia group follows a repeated-measure design. For most features, the Shapiro-Wilk tests show that the distributions are not normal for the different factors, hence we compute a non-parametric Friedman ANOVA analysis. Follow-up posthoc Wilcoxon comparison tests with Bonferroni correction are accomplished (effects are reported at a 0.0167 level of significance).
9. Towards Sound-Based Therapy and Diagnosis for Dementia

![Figure 9.9: Within-subject analysis (patient group): water sound samples are perceived as happier (A) and more relaxing (B). Voice samples are evaluated as being sadder (A), more arousing (B) and more dominant (C).](image)

9.3.4.2. Influence of the origin of the sound

**Happiness**  Our analysis showed that the “Happiness” ratings were influenced by the origin of the sound ($\chi^2(2)=10.292$, $p<0.01$). We found a significant difference for the synth/water and the voice/water comparison. Synthetic samples (Mdn=75.3) were rated as less happy than water samples (Mdn=50.41) $T=289.5$, $z=-4.854$, $p<0.0167$. Voice samples (Mdn=74) were significantly less happy than water samples (Mdn=50.41), $T=366.5$, $z=-4.242$, $p<0.0167$. Other comparisons were not significant. (cf Figure 9.9).

**Arousal**  Analysis showed that the “Arousal” ratings were influenced by the origin of the sound ($\chi^2(2)=6.986$, $p<0.05$). We found a significant difference for the synth/voice and the voice/water comparison. Synthetic samples (Mdn=24.5) were rated as more relaxing than voice samples (Mdn=43.76) $T=551$, $z=-2.917$, $p<0.0167$. Voice samples (Mdn=43.76) were significantly less relaxing than water samples (Mdn=26.4), $T=545$, $z=-2.516$, $p<0.0167$. Other comparisons were not significant. (cf Figure 9.9).

9.3.4.3. Influence of the loudness of the sound

**Dominance**  The “Dominance” ratings were influenced by the loudness ($\chi^2(2)=6.64$, $p<0.05$). We found significant differences for the low/medium comparison. Medium loudness (Mdn=45.0) was more fearful than low sound level (Mdn=48.78), $T=43.0$, $z=-2.318$, $p<0.0167$. (cf Figure 9.9)
9.3.5. Between group (disease stage) analysis for dementia patients

9.3.5.1. Statistical Analysis

The Shapiro-Wilk test showed that ratings were not normally distributed, hence we computed a Kruskall-Wallis analysis. Follow-up posthoc Mann-Whitney comparison tests with Bonferroni correction were accomplished, so all effects are reported at a 0.0167 level of significance.

![Box plots showing the distribution of ratings by disease stage.](image)

Figure 9.10.: Global between-group analysis (dementia disease stages): The more the disease progresses, the more the patients tend to rate the sound samples as being happier (A) and more pleasant (B).

9.3.5.2. Influence of the disease stage

**Sadness** Analysis showed that the “Happiness” ratings were influenced by the level of the dementia $H(2)=15.246$, p<0.01. We found a significant difference for the levels 4/5. At stage 5 of the disease samples were perceived as more happy than at stage 4, $U=247.0$ (Figure 9.10).

**Unpleasantness** The “Pleasantness” ratings were influenced by the level of the dementia $H(2)=8.164$, p<0.05. We found significant differences for the comparisons between level 4 and 5 where sounds were rated as increasingly more pleasant when going from level 4 to level 5 ($U=346.5$) (Figure 9.10).
9. Towards Sound-Based Therapy and Diagnosis for Dementia

9.3.6. Detailed between group analysis (disease stage)

9.3.6.1. Statistical Analysis

The Shapiro-Wilk test showed that ratings were not normally distributed, hence we computed a Kruskall-Wallis analysis. Follow-up posthoc Mann-Whitney comparison tests with Bonferroni correction were accomplished, so all effects are reported at a 0.0167 level of significance.

9.3.6.2. Sound Type

![Box plots of sound samples rated for sadness and unpleasantness across disease stages](image)

**Figure 9.11.:** Detailed between-group analysis (dementia disease stages) for sound types: the more the dementia progresses (from stage 3 to 5), the more water (A) and voice (B) sound samples are rated as happier and synthetic (C) and water (D) sounds are rated as more pleasant.

**Sadness** The rating of synthetic sounds on the sadness scale were significantly different depending on the dementia level $H(2)=16.69$, $p<0.01$. At more advanced stage
of the disease, patients rated synthetic sounds as happier. Pairwise comparisons were significant for levels 4/5 (U=197) and levels 3/5 (U=25.5) (Figure 9.11).

The rating of voice samples on the sadness scale were significantly different depending on the dementia level H(2)=9.736, p<0.01. At more advanced stage of the disease, patients rated voice sounds as happier. Pairwise comparisons were significant for levels 4/5 (U=301.5) (Figure 9.11).

**Unpleasantness** The rating of voice samples on the unpleasantness scale were significantly different depending on the dementia level H(2)=7.708, p<0.05. At more advanced stage of the disease, patients rated synthetic sounds as more pleasant. Pairwise comparisons were significant for levels 4/5 (U=321.5) (Figure 9.11).

The rating of water samples on the unpleasantness scale were significantly different depending on the dementia level H(2)=6.573, p<0.05. At more advanced stage of the disease, patients rated water sounds as more pleasant. Pairwise comparisons were significant for levels 3/5 (U=20.0) (Figure 9.11).

**9.3.6.3. Loudness**

**Sadness** The rating of high loudness sounds on the sadness scale were significantly different depending on the dementia level H(2)=8.727, p<0.05. At more advanced stage of the disease, patients rated loud sounds as happier. Pairwise comparisons were significant for levels 4/5 (U=148.5) (Figure 9.12).

The rating of medium loudness sounds on the sadness scale were significantly different depending on the dementia level H(2)=7.356, p<0.05. At more advanced stage of the disease, patients rated medium-loud sounds as happier. Pairwise comparisons were significant for levels 4/5 (U=173.5) (Figure 9.12).

**Unpleasantness** The rating of samples with high loudness on the unpleasantness scale were significantly different depending on the dementia level H(2)=9.225, p<0.05. At more advanced stage of the disease, patients rated synthetic sounds with high loudness as more pleasant. Pairwise comparisons were significant for levels 4/5 (U=158) and 3/5 (U=9.0) (Figure 9.12).

The rating of samples with low loudness on the unpleasantness scale were significantly different depending on the dementia level H(2)=6.17, p<0.05. At more advanced stage of the disease, patients rated sounds with low loudness as more pleasant. Pairwise comparisons were significant for levels 4/5 (U=139.5) (Figure 9.12).

**9.3.6.4. Brightness**

**Dominance** The rating of samples with low brightness on the dominance scale were significantly different depending on the dementia level H(2)=9.791 p<0.01. At more ad-
9. Towards Sound-Based Therapy and Diagnosis for Dementia

Figure 9.12: **Detailed between-group analysis (dementia disease stages) for loudness:** When the dementia progresses (from stage 3 to 5), samples with high loudness (A) and medium loudness (B) are rated as happier and more pleasant respectively (C, D).

Advanced stage of the disease, patients rated sounds with low brightness as more dominant. Pairwise comparisons were significant for levels 4/5 (U=103.5) (Figure 9.13).

**Unpleasantness**  The rating of samples with low brightness on the unpleasantness scale were significantly different depending on the dementia level H(2)=6.614, p<0.05. At more advanced stage of the disease, patients rated sounds with low brightness as more pleasant. Pairwise comparisons were significant for levels 4/5 (U=139.5) (Figure 9.13).

9.3.6.5. **Consonance**

**Sadness**  The rating of sounds with medium consonance on the sadness scale were significantly different depending on the dementia level H(2)=6.39, p<0.05. At more advanced stage of the disease, patients rated medium consonant sounds as happier. Pairwise comparisons were significant for levels 4/5 (U=72.0) (Figure 9.14).
9.4. Discussion

9.4.1. Methodological considerations

9.4.1.1. Self-report and subjective feeling

As of today, there is still no consensual definition of what an emotion is (Izard, 2007). Nevertheless it is generally admitted that emotion is a complex phenomenon involving physiological arousal, subjective feeling, cognitive appraisal, action tendencies and expressive motor behavior (Niedenthal et al., 2006). In the context of this study we used self-report measures which goal is to provide information about the subjective feeling of the subjects.
9. Towards Sound-Based Therapy and Diagnosis for Dementia

9.4.1.2. Emotional scales

Two main models of emotion have traditionally been emphasized. Each one is associated with specific self-report measures. The theory of basic emotions proposes that emotions are discrete and are based on a set of innate emotions such as sadness, happiness, fear, anger, disgust and guilt (Ekman, 1992); while the dimensional models put forward a representation where emotions are composed of core continuous dimensions such as valence and arousal (Russell, 1980). Here, we proposed to combine both models into a representation of musical emotions on four continuous scales of valence, arousal, dominance and pleasantness represented by a slider that ranged from 0 to 100%.

9.4.1.3. Online vs. retrospective measures

The subjects rated the sound samples retrospectively since it has been shown that online ratings could interfere with the emotion induction process itself (Rosenberg and Ekman, 2005).

Figure 9.14.: Detailed between-group analysis (dementia disease stages) for consonance: medium consonance was perceived as happier at higher stage of the dementia.

By asking subjects to rate the sounds they heard in terms of the emotion they felt, we implicitly postulated that there is a coherence across experiential and behavioral response systems. Although this has been shown to not be completely obvious for generic emotions such as surprise and anxiety (Mauss et al., 2004; Reisenzein, 2000), it appears mostly true for musical emotions (Lundqvist et al., 2009).
9.4. Discussion

Figure 9.15.: Detailed between-group analysis (dementia disease stages) for tempo: When the dementia progresses (from stage 3 to 5), samples with high tempo are perceived as happier (A) and more pleasant (B).

9.4.1.4. Duration of the stimuli

Previous work showed that emotional responses to music are very stable within and between participants and do not depend on musical expertise nor, to a certain extent, excerpt duration (Bigand et al., 2005; Peretz et al., 1998). This justifies the use of a database containing relatively short affective sound samples (with a duration of 5 s.).

9.4.2. Towards affective sound design for music therapy

A within analysis for the control group showed that sounds with medium loudness were more arousing than sounds with low loudness which is consistent with previous studies on healthy subjects Staum and Brotons 2000; Gabrielson and Lindström 2001; Gomez and Danuser 2007.

The within analysis for the dementia group showed that water sounds were reported to be happier while water and synthetic sounds were more relaxing. Previous studies have stressed the influence of different musical instruments on the responses of healthy subjects. It appeared that the choice of a specific musical instrument or combination of instruments is highly relevant to the expression of emotion (Balkwill and Thompson, 1999; Behrens and Green, 1993; Juslin, 1997; Gabrielson and Juslin, 1996). This shows that different sound timbre can elicit significant emotional responses in a similar fashion than more commonly studied (non timbral) features such as pitch, harmony or sound level Hailstone et al. 2009a. As a matter of fact, it is rather common to use environmental sounds as well as musical sounds in music therapy protocols Götell et al. 2003; Gomez and Danuser 2004.

Depression and agitation decreases the quality of life of dementia sufferers, may reduce the duration of survival and would therefore appear to be important to treat. Our study
show that a sound-based therapy with a well-defined set of sound samples with specific acoustic properties can have a soothing effect on the affective state of the patients.

The classification of the different emotional responses to specific and well-defined sound features in patients with dementia can serve as a basis for designing soundscapes for music therapy. For this purpose, we have been developing an interactive music technology called the SMuSe (Situated Music Server) (Le Groux and Verschure, 2009b). The SMuSe offers an original perceptually-grounded synthesis model, and the possibility to parameterize and generate musical structures and soundscapes in real-time. By using a mapping scheme based on the present study and psycho-physiological evidence (Le Groux et al., 2008b; Gabrielsson and Lindström, 2001; Le Groux and Verschure, 2010b), the SMuSe will allow for the generation of personalized interactive affective music.

9.4.3. Subcortical processing of sonic emotions in dementia

One intriguing finding of the present study is that the more the disease progresses, the more dementia patients tend to give happy and pleased ratings.

This seems to be in contradiction with the fact that depression is one of the most frequent neuropsychiatric comorbidities of Alzheimer dementia (AD) (Chinello et al., 2007). These depressive symptoms tend to occur early, are persistent, and are becoming increasingly common as dementia progresses (Amore et al., 2007).

We propose that, as the cognitive abilities of the subjects degrade with the level of the disease, the subjects are in fact getting more and more sensitive to subcortical emotional processes as compared to higher cognitive processes such as evaluative conditioning, emotional contagion, episodic memory, or musical expectancy (Juslin and Västfjäll, 2008). Our results reinforce the view of the subcortical nature of emotional processing of perceptual features of sound (Peretz et al., 1998; Kurylo et al., 1993; Blood and Zatorre, 2001). We can hypothesize that as dementia, which is a disconnection syndrome, progresses, a deterioration of the connections and interactions between subcortical and cortical musical processing areas occurs.

The gradual differences in emotional responses to sound stimuli can be put in relation with the theoretical framework of the Progressively Lowered Stress Threshold (PLST) . This model states that, as the dementia progresses and neurological damage increases, the person is less able to receive and process information from the external environment and their stress threshold gradually decreases. (Hall and Buckwalter, 1987; Sherratt et al., 2004). In our case, we can suggest a similar reversed mechanism, where the more the disease progresses, the more the patients are sensitive to positive effects of sound stimuli.

9.4.4. Towards a sonic-based diagnosis of dementia

A global analysis of the differences between the control and patient groups indicated a general tendency of dementia patients to give more radical ratings than the control
group. Overall the patients found the sound samples sadder and more relaxing than the control group. The fact that responses on the dominance scale were not clear for the control group concords with previous literature on emotional responses to sound in healthy subjects (Resnicow et al., 2004; Robazza et al., 1994; Hailstone et al., 2009a).

A detailed between-group analysis showed that the responses to synthetic and medium loudness sounds on the arousal scale, the responses to medium consonance on the dominance scale, and the responses to slow tempo on the unpleasantness scale can be used to distinguish between controls and patients groups.

A detailed analysis revealed significant differences as the dementia condition progresses from level 3 to 5 on the sadness scale for synthetic and voice samples, medium and high loudness samples, medium consonance and high tempo. On the unpleasantness scale, responses to voice and water samples, high and low loudness samples, low brightness and high tempo samples were significantly different for different stages of dementia. Similarly, low brightness sounds evoked different responses on the dominance scale for different stages of dementia.

These between-group analysis confirm the assumption that there are systematic relationships between reported subjective states and the sonic properties of the audio samples and show that dementia patients have significantly different responses to audio stimuli than healthy patients. Moreover, those responses increase with the level of the dementia. These results suggest that a sonic-based diagnosis for dementia and the disease state is possible.

9.5. Conclusions

We developed an automated test procedure that assesses the correlation between the origin, brightness, sound level, tempo and consonance of a set of parameterized sound samples with the self-reported emotional state of dementia patients. Our results show that the control and patient groups had significantly different emotional reactions to the sound samples, suggesting possible applications in automatic dementia diagnosis. We also found that the stage of the disease influenced the ratings of dementia patients and the more the disease was advanced, the more the sounds were perceived as happy, relaxing dominant and pleasant.
Part IV.

The SMuSe in the flesh: Real-World Applications
10. XiM-SMuSE: Sonification of Spatial Behaviors in the Mixed-Reality Environment XiM

It is generally admitted that music is a powerful carrier of emotions (Gabrielson and Juslin, 1996; Juslin and Sloboda, 2001), and that audition can play an important role in enhancing the sensation of presence in Virtual Environments (Gilkey et al., 1995). In mixed-reality environments and interactive multi-media systems such as Massively Multiplayer Online Games (MMORPG), the improvement of the user’s perception of immersion is crucial. Nonetheless, the sonification of those environments is often reduced to its simplest expression, namely a set of prerecorded sound tracks. Background music many times relies on repetitive, predetermined and somewhat predictable musical material. Hence, there is a need for a sonification scheme that can generate context sensitive, adaptive, rich and consistent music in real-time. In this chapter we introduce an application of the SMuSe (Cf. Chapter 3 for a description of the system) to the sonification of spatial behavior of multiple human and synthetic characters in a Mixed-Reality environment. We propose a semantic layer that maps sensor data onto intuitive parameters for the control of music generation, and show that the musical events are directly influenced by the spatial behavior of human and synthetic characters in the space, thus creating a behavior-dependent sonification that enhance the user’s perception of immersion.

10.1. Introduction

Sonification is a relatively new field of research that has developed rapidly over the last years. It is now commonly used in different scientific contexts and standard techniques and applications have recently emerged (Hermann and Hunt, 2005). For instance, Auditory Icons and Earcons models are used to signal specific events to the users. Audification models convert data series into sound samples for scientific data mining (Hermann and Hunt, 2005). Furthermore, model-based sonification allows a better understanding of the structure of data by using plausible physical sound synthesis algorithms driven by the data itself. Our work relates to yet another type of sonification, we call behavioral mapping, where data from the environment and the behavior of agents (human and/or machines) are mapped to acoustic and musical attributes. If scientific sonification of abstract data is usually intended to give an insight in the structure of the data
for analysis purposes, our objective is somewhat different. We aim at improving the
social interaction in the environment and create a musical world that has to be both
informative and aesthetically appealing. We want to enhance the comprehension of the
environment, embed the user in a rich and consistent audio-visual environment while
producing a musical experience. It is widely acknowledged that music is a powerful
carrier of emotions, and that auditory cues play an important role in enhancing the
sensation of presence in virtual environments (Gabrielson and Juslin, 1996; Juslin and
Laukka, 2003; Wassermann et al., 2003). Mixed-Reality spaces are becoming more and
more common and various labs are building their own environments (e.g. the Allosphere
at UCSB (Amatriain et al., 2007), the intelligent House at MIT, the Nanohouse at UTS
(Muir et al., 2007) and the Sentient Lab in the Faculty of Architecture, University of
Sydney (Beilharz, 2005)…). Until now music generation and sonification for large-scale
mixed-reality environments has not been extensively studied. However, we can extract
basic design guidelines from previous studies. For instance, Freeman and Lessiter 2001
showed that multichannel systems enhance the sense of presence. It is also assumed
that moving sound sources increases the dynamics of the environment, and that inter-
active sonification makes it more interesting for the user (Serafin and Serafin, 2004). In
addition, repetitive loops are likely to be detected by the user and perceived as artifi-
cial (Serafin and Serafin, 2004), which means our system should be able to produce a
reasonable variety of sounds. In this chapter we present an original framework for adap-
tive sonification of the spatial behavior of entities in a mixed-reality space based on the
aforementioned guidelines. In contrast with many data sonification approaches, we do
not sonify directly the sensory data, but we rely on higher-level behavioral information
that is extracted from sensor data. We propose to extract this high-level spatial/social
behavior from sensor data via a neuromorphic processes designed with a real-time the
neural network simulator IQR (Bernardet and Verschure, 2010). This spatial behavioral
information is then used to modulate a real-time real-world composition engine. The fi-
nal correlation between auditory representation and understanding of the space is based
on the study of perceptual effects of musical and sound generation parameters. The
same paradigm could also be used in virtual rehabilitation therapy systems as a way to
promote coordinated spatial movements (Cameirão et al., 2008; Jack et al., 2002).

10.2. XIM: a mixed-reality space

To better understand the context of our work, we give a brief description of the mixed-
reality space called XIM we have built in our lab (Bernardet et al., 2007). This mixed-
reality environment is the continuation of the Ada project, an interactive space exhibited
for the Swiss national Expo in 2002 (Eng et al., 2003). It consists of two main parts.
On the one hand we have a physical interactive space of 7x7 m. called XIM (eXperi-
ence Induction Machine) and on the other hand the virtual environment called PVC
(for Persistent Virtual Community). The PVC is divided into four nested regions: the
Clubhouse is the virtual counterpart of the XIM physical space, and is itself surrounded
by the P-Garden and the Avatar-Heaven, a retreat area for avatars (see Figure 10.1).

![Diagram of XIM: a mixed-reality space](image)

Figure 10.1.: **The Persistent Virtual Community** consists of the real-world physical installation (called the Experience Induction Machine), the ClubHouse, the P-Garden and Avatar Heaven. Users can visit the physical installation XIM, connect over the Internet (remote users) or visit a CAVE installation.

The community is interfaced to this mixed-reality environment via several communication channels. For instance, interactions can take place between visitors of the physical space, and/or between users that connect over the Internet (Remote Users), but it can also take place between fully synthetic characters or the virtual environment itself. The Persistent Virtual Community consists of the physical installation (the Experience Induction Machine), the ClubHouse, the P-Garden and Avatar Heaven. Users can visit the physical installation XIM, connect over the Internet (Remote Users) or visit a CAVE installation.

From the general structure of the PVC, we distinguish four main areas where sound should contribute to the overall interactive experience. First, we would like to generate sound in the physical space itself (the XIM) to accompany, illustrate, transform and influence the interaction of physical and virtual users within the physical space. Secondly we wish to find a coherent sonification scheme where each avatar’s personal contribution is taken into account and added to the global sonic output. Finally, the virtual environment at both the client and server side has a role to play in the global sonification scheme, involving remote control, sound streaming, client/server sound synthesis. The variety of the possible modes of interaction and communication paradigms is apparent, and the sonification of such a complex mixed-reality system is an intricate task. Hence, there are three main tasks for the sonification process: 1) soundscapes: here the concept is to embed the physical and virtual environment within a dynamic soundscape. The sonic result can be related to a naturalist approach in which the sounds are used...
to correlate natural or biological-like structures with sonic events. Concepts such as Sonic Ecologic can be related to this approach, 2) synthetic voices: here the concept is to bring to the acoustic space the agent’s voices and from that to construct structures such as dialogues and narratives and 3) musical realm: here the concept is to use the SMuSe to create a dynamic musical discourse in which the agent’s behavior is related to expression and emotional experience. In this chapter we will more specifically focus on the sonification of the physical space XIM in relation to the behavior of humans and avatars that populate this mixed-reality environment.

10.3. System Architecture

The real-time/real-world system we designed is made up of three main interconnected modules (Cf. Figure 10.2) that can be described in terms of a sensor interface, a central-processing stage, and a sound synthesis stage (Wanderley et al., 2002).

![Diagram of system architecture](image)

Figure 10.2.: The architecture the mixed-reality music system couples sensing with neuromorphic processing and musical responses.

The first module consists of the mixed-reality environment populated by “behaving” human and synthetic characters. The second module consists of a neural model that processes the sensor data and interprets it as behavioral data. Finally, the third module based on the SMuSe (Cf. Chapter 3) deals with the interactive composition of musical structure and the generation of sounds. The musical output is modulated by the behavior of the human and synthetic users. The different modules are all interconnected via TCP/IP connections via Ethernet on a local network.
10.4. Overview

10.4.1. Sensing

Accurate and robust tracking of movement and people interacting with the XIM is a complex task carried on by a team of colleagues in our lab. Their approach focuses on multimodal information fusion to improve tracking, automatic assessment of sensor data quality and analysis of methods for intelligent recruitment of sensors to resolve conflicting situations (Mathews et al., 2007). XIM is equipped with a number of sensors that are used to determine its internal states and to map the physical XIM onto its virtual representation. In addition, XIM can use its effectors to influence the behavior of its as well as to regulate human-avatar interactions and to provide the best operating conditions for all sub-systems. The sensors that are currently used are 3 overhead cameras and frame grabbers, 4 gazers, 3 microphones, and the pressure sensitive floor. They provide information for the accurate tracking of the visitors and can influence the emotions and behaviors of XIM. The effectors are 8 speakers, sound synthesizers, 6 video projectors, 8 lightfingers and the floor (Cf. Figure 10.3). With our technology, we can offer a reliable tracking of movement and people in the XIM, which, combined with the trajectories of the avatars, form an accurate map of each individual trajectory.

Figure 10.3.: XIM is provided with a set of sensor and effectors. Sensors allows for the tracking of people in the space. Effectors are used to influence the emotions and behavior of the visitors.
10.4.2. Neuromorphic analysis of user behavior

The neurom simulator environment IQR\(^1\) handles the real-time processing of the tracking system outputs. IQR is a flexible tool for creating and running simulations of large-scale neural models and provides a user-friendly graphical interface to design multilevel neural models (see Figure 10.4) (Bernardet and Verschure, 2010).

![Diagram of IQR interface](image)

Figure 10.4.: IQR provides a user-friendly interface to design multilevel neural models. Models in IQR are organized at different levels. The top level is the system containing an arbitrary number of processes, and connections. This example illustrates an IQR module where two processes are interconnected. Each process contain a neuronal circuit made up of three neuronal groups.

In this project, we used IQR to transform the tracking data from the real and synthetic characters into higher-level relevant control parameters for an interactive real-world music generator. IQR allows us to model a biologically plausible neural network for spatial/behavioral analysis. The spatial behavior of each individual is mapped to a neuron group called Floor Layout of size 56 * 53 neurons representing the spatial topology of XIM (Figure 10.5). When individuals move from one point in space to another, the corresponding neurons are activated. With this representation, we can design neuromorphic processes performing an analysis of the behavior of the population in the space. For instance, motion is defined as the difference of neural activity in the cell group representing people’s position at two time intervals. We also defined a neuronal group that counts the number of people populating the space by summing up the total activity of the occupancy group. We then virtually divided the space in four main activity areas (up/down/left/right corners), and created the corresponding four neuronal groups. The synapse connectivity from the floor layout to those four groups can be defined so that each group receives the activity of only a subset of the Floor Layout group, thus allowing us to know which of the four areas of the space is the most active. The different modes of interaction of XIM are based on such analysis (if one individual is in the space, the

\(^1\)http://iqr.sourceforge.net/
XIM will behave differently than if there are two people, more, or none. Cf. Figure 10.5).

Figure 10.5.: **Behavior analysis:** IQR allows to graphically design a module that extracts behavioral information from the tracking system. The cell activity of the different interconnected neuronal groups is related to the activity of avatars and human interacting in the mixed-reality environment.

We now have a high-level representation of the spatial behavior of people in the space and are able to define musical parameter and specific neuronal groups, which activity will control the musical output generation.

### 10.5. Music Generation

In order to synchronize musical composition to the behavior of real and synthetic characters in the mixed-reality environment, we represent the music as a parameterized structure, which values are dynamically controlled over time. We classify the musical parameters into three main categories. On the one hand are the macro-level structural parameters such as tonality, rhythm, tempo, etc. On the other hand are the micro-level parameters that relate to the synthesis of sound (instrumentation, brightness, harmonicity, spectral complexity, etc.). Finally are the parameters that relate to the localization of the sound source (spatialization).

We chose a set of standard musical parameters for controlling the generation of music based on the requirement that their modulation should have a clear perceptual effect. We kept a list of parameters that has been extensively studied, and which effect on emotional expression is widely acknowledged, as described in Chapter 8 and reviewed in
Gabrielsson and Lindström 2001. Those parameters are tempo, mode, volume, register, tonality and rhythm at the macro level and articulation, brightness and inharmonicity at the micro level.

The generation of music, based on a real-world situated composition paradigm, used the SMuSe (Cf. Part I). Prepared musical material was dynamically modulated as the users interacted with the mixed-reality space. A group of various related musical cells defined a certain style or atmosphere and corresponded to a specific mode of interaction in the XIM (see Figure 10.6). When the interaction between people and XIM took place, these basic musical events were dynamically modified. The musical material was modulated following a carefully designed emotional mapping (Cf. Section 10.6). The initial musical material was amplified, transformed, nuanced, as the interaction between the XIM and the users evolved.

Figure 10.6.: **Musical Cells:** A group of related musical cells defines a certain style file which specific structural and musical parameters are associated with a mode of interaction of the XIM. When a style file is enabled, we cycle through the different cells whose properties are modulated in real-time via OSC control messages.

To provide our system with fine timbral control over the generation of sound, we implemented a simple FM synthesizer. We chose the FM synthesis algorithm because of its simplicity, efficiency, and ability to synthesize a wide range of timbre with only a few control parameters Chowning and Bristow 1987. We used the MIDI instruments for most of the melodic “discrete” material while the FM synthesizer added continuous layers of sound with slowly varying timbre characteristics. Finally, we designed a soundscape generator that could trigger samples of environmental natural sounds (such as forest, sea, footsteps, etc.) chosen from the Freesound database\(^2\).

\(^2\)http://www.freesound.org
Spatialization is a crucial parameter to take into account for improving the sensation of presence in a VR environment (Gilkey et al., 1995). It creates the perceptual illusion of a physically realistic sound field. We wanted to be able to virtually place various sound sources in the environment and move them around the listener in real-time in XIM. For this purpose, we used the Vector Based Amplitude Panning (VBAP) algorithm, a flexible and efficient method for positioning virtual sources using multiple speakers (Pulkki, 1997).

10.6. Mapping Spatial Behavior to Emotion and Music Generation

In this project, we relied on a three-layer mapping from spatial behavior to emotion expression to music generation (Figure 10.7).

Figure 10.7.: **Three layers mapping** from the spatial behavior of virtual and human characters in the mixed reality space to the semantics of emotional states we want to express, and finally to the musical parameters illustrating such emotional states.

First, we defined a set of relations between musical structural parameters and musical expression based on a review of literature on music and emotion (Gabrielson and Juslin, 1996; Juslin and Laukka, 2003; Gabrielsson and Lindström, 2001) (See Table 10.1).

Then, we defined mappings from the perceptual domain to FM synthesis technical parameters (Chowning and Bristow, 1987) (Cf. Table 10.2).

Finally, in order to design the interaction between spatial behavior and music generation, we made use of the semantic mappings previously introduced. We described the musical effects we wanted to express, and associated them with a corresponding set of music and sound parameters using tables 10.1 and 10.2. The choice of the set of musical parameters...
Table 10.1.: Relationship between a set of standard musical parameters and semantics of musical expression (adapted from Gabrielsson and Lindström 2001)

<table>
<thead>
<tr>
<th>Musical Parameter</th>
<th>Level</th>
<th>Semantics of Musical Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tempo</td>
<td>Slow</td>
<td>Sadness, Calmness, Dignity, Boredom</td>
</tr>
<tr>
<td></td>
<td>Fast</td>
<td>Happiness, Activity, Surprise, Anger</td>
</tr>
<tr>
<td>Mode</td>
<td>Minor</td>
<td>Sadness, Dreamy, Anger</td>
</tr>
<tr>
<td></td>
<td>Major</td>
<td>Happiness, Grace, Serenity</td>
</tr>
<tr>
<td>Volume</td>
<td>Loud</td>
<td>Joy, Intensity, Power, Anger</td>
</tr>
<tr>
<td></td>
<td>Soft</td>
<td>Sadness, Tenderness, Solemnity, Fear</td>
</tr>
<tr>
<td>Register</td>
<td>High</td>
<td>Happiness, Grace, Excitement, Anger, Activity</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>Sadness, Dignity, Solemnity, Boredom</td>
</tr>
<tr>
<td>Tonality</td>
<td>Tonal</td>
<td>Joyful, Dull</td>
</tr>
<tr>
<td></td>
<td>Atonal</td>
<td>Angry</td>
</tr>
<tr>
<td>Rhythm</td>
<td>Regular</td>
<td>Happiness, Dignity, Peace</td>
</tr>
<tr>
<td></td>
<td>Irregular</td>
<td>Amusement, Uneasiness, Anger</td>
</tr>
<tr>
<td>Articulation</td>
<td>Staccato (short)</td>
<td>Gaiety, energy, fear</td>
</tr>
<tr>
<td></td>
<td>Legato (long)</td>
<td>Sadness, Tenderness, solemnity</td>
</tr>
</tbody>
</table>

in relation to behavioral/spatial information was the result of a trial and error design process.

We found that our music generation system offered a good representation of the spatial behavior of people in the space. The combination of soundscapes, event-based samples, midi melodies, and slowly varying FM synthesis layers provided a rich and lively sonic environment. Different set of parameters produced significantly distinct sounding results. Figure 10.8 and 10.9 illustrate the spectral content of the audio signal generated by the system for two different interaction scenarios where augmentation of tempo and of density of events were correlated to the sensation of increased activity in the space. The users (6 university students) perceived this auditory feedback as natural and playful. They often started playing with the space by running around, moving slowly, staying still, thus trying to provoke various musical outputs. In the future we wish to perform a series of experiments to quantitatively study the relation between the emotional state
10.7. Conclusions

<table>
<thead>
<tr>
<th>Sound Parameter</th>
<th>Level</th>
<th>Semantics of musical expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brightness</td>
<td>High</td>
<td>Potency, Anger, Disgust, Fear, Activity, Surprise</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>Tenderness, Sadness</td>
</tr>
<tr>
<td>Harmonicity</td>
<td>High</td>
<td>Calmness, tenderness, Light</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>Energy, Anger</td>
</tr>
</tbody>
</table>

Table 10.2.: Relationship between a reduced set of sound synthesis parameters and musical expression. (adapted from Gabrielson and Lindström 2001)

<table>
<thead>
<tr>
<th>Spatial Parameter</th>
<th>Semantics</th>
<th>Sonification Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial location</td>
<td>Activity, Surprise</td>
<td>Trigger specific samples, Modulate brightness.</td>
</tr>
<tr>
<td>Motion</td>
<td>Happiness, Activity, Excitement</td>
<td>Modulate temp, velocity, register</td>
</tr>
<tr>
<td>Number of users</td>
<td>Activity, Excitement</td>
<td>Density of events, rhythmic regularity, harmonicity, articulation</td>
</tr>
<tr>
<td>Interaction scenario</td>
<td>Surprise, Change, Anger</td>
<td>Change of style file, scale, mode</td>
</tr>
</tbody>
</table>

Table 10.3.: Relation between spatial behaviors of human and avatars in the XIM and sonification parameters.

of the users and the different parameters of our system.

10.7. Conclusions

In this chapter, we have presented a complete system that generates a complex spatialized sonification of the behavior of multiple humans and avatars in a Mixed-Reality environment based on their spatial/social behavior. We proposed a system architecture that relies on interaction in real-time between a mixed reality space, a neural network simulator for neuromorphic sensor processing and a sound and music generator. We introduced a semantic mapping layer to help the interaction design and ease the choice of relevant musical and sonic parameters.

In the future, we would like to operate on a higher level of integration between the neural control and the musical structure. To achieve this we want to use a more elaborate and adaptive neuromorphic model of sensor processing combined with a more advanced cognitive architecture called Distributed Adaptive Control (DAC) (Verschure et al., 2003), for musical learning and decision-making generating a synthetic behavioral
music composition system. DAC distinguishes three levels of control in a cognitive behaving system: reactive, adaptive and contextual. In the current system, the mappings are made directly from the spatial sensor data. By relying on the DAC paradigm to infer more refined behavioral and emotional information, we can use this higher-level information as a composition parameter.

In the sound synthesis domain, we used basic mappings from perceptual properties to FM synthesis parameters; this could be extended to other synthesis algorithm and parameters, notably our physically-based synthesizer. Further experiments involving self-reported as well as physiological measurements of users’ emotional response to the different set of parameters of the system will have to be done to gain more insights into the efficiency of this mixed-reality space sonification.

Here, we limited the scope of this chapter to the sonification of the XIM, and didn’t address the problem of sonification at the client side for remote users. The simple solution we propose to solve this problem is to have one centralized music server based on the SMuSe and real-time audio streaming of the SMuSe’s musical output to each client computer.
Figure 10.9.: **Six users interaction:** This spectrogram (x is time in seconds and y frequency in kHz) illustrates an interactive session where six users are moving around in the space. We observe that the density of event is high. There is also a large amount of energy in the high frequency spectrum.
11. Neuro-SMuSe: Neurofeedback Through Musical Interaction

Human aural system is arguably one of the most refined sensor we possess. It is sensitive to such highly complex stimuli as conversations or musical pieces. Be it all speaking voice or a band playing live, we are able to easily perceive relaxed or agitated states in an auditory stream. In turn, our own state of agitation can now be detected via electroencephalography technologies. In this chapter we propose to explore both ideas in the form of a framework for conscious learning of relaxation through sonic feedback. After presenting the general paradigm of neurofeedback, we describe a set of tools to analyze electroencephalogram (EEG) data in real-time and we introduce a carefully designed, perceptually-grounded interactive music feedback system that helps the listener to keep track of and modulate her agitation state measured by EEG.

11.1. Introduction

Music is generally acknowledged to be a powerful carrier of emotions and mood regulator, and as such has become omnipresent in our day to day life. Many stores now use energetic dance music to instill a festive atmosphere during our shopping hours. Moreover, with the advent of new human-computer interaction technologies, it has now become possible to derive some information about our emotional or mental state from physiological data. As a matter of fact, certain specific activity patterns of our brain, measurable by electroencephalography, correlate with different state of anxiety, tranquility or concentration. In parallel, the developments of music technology and psychoacoustics now allow for the generation of “affective” parametric and synthetic sounds in real-time. This paves the way for a new type of application that takes advantage of real-time physiology analysis and musical feedback and interaction to understand and learn to control better our mental states. With this work, we propose a framework for training our brain to learn to reach a relaxing state by relying on a perceptually-grounded interactive music system.
11.2. Background

11.2.1. Electroencephalography

Electroencephalography (EEG) devices measure the summed activity of post-synaptic currents in the brain. The electrical voltage of an individual neuron can’t be detected by an EEG electrode placed on the scalp, but a surface EEG reading is the summation of the synchronous activity of thousands of neurons. If a group of neurons fire in synchrony, the activity will result in the measurement of a large signal whereas asynchronous firing will trigger a smaller irregular signal. Scalp EEG activity can oscillate at different frequencies representing specific rhythmic, synchronized activity: the brain waves (Nunez, 2005).

11.2.2. Brain Waves

The rhythmic activity describing the EEG is divided into bands by frequency, and most of the cerebral signal observed in the scalp EEG falls into the range 1-20 Hz. Brainwaves are usually categorized into the bands known as delta, theta, alpha and beta which have been shown to correlate with different mental states (Nunez, 2005) (Cf. Table 11.1).

Figure 11.1.: Brainwave patterns

**Delta** wave (0-4 Hz) is the slowest and is associated with deep sleep and can be measured frontally. It is the dominant wave rhythm in infants.

**Theta** wave (4-8 Hz) is associated with dream sleep, meditation and creative inspiration and is strong in children with attention deficit disorders.
11.2. Background

**Alpha** wave (8-12 Hz) can be measured from the posterior regions of the head and is associated with a relaxed state. Only closing one’s eyes increases the generation of alpha waves.

**Beta** wave (12-30 Hz) is most evident frontally and is associated with an alert state of mind, anxiety, concentration and mental activity.

**Gamma** waves (30-80Hz) correlate with higher mental processes, perception and consciousness. High frequency waves (alpha, beta and gamma) dominate during wakefulness.

<table>
<thead>
<tr>
<th>Rhythm</th>
<th>Frequency Range</th>
<th>Location</th>
<th>Mental state</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delta</td>
<td>(0-4) Hz</td>
<td>Frontal lobe</td>
<td>Deep sleep</td>
</tr>
<tr>
<td>Theta</td>
<td>(4-8) Hz</td>
<td>Midline, temporal</td>
<td>Drowsiness, meditation, problem solving</td>
</tr>
<tr>
<td>Alpha</td>
<td>(8-13) Hz</td>
<td>Frontal, Occipital</td>
<td>Relaxing, closed eyes</td>
</tr>
<tr>
<td>Mu</td>
<td>(8-13) Hz</td>
<td>Central</td>
<td>Motor</td>
</tr>
<tr>
<td>Beta</td>
<td>(13-30) Hz</td>
<td>Frontal, Central</td>
<td>Concentration, alertness, mental activity</td>
</tr>
</tbody>
</table>

Table 11.1.: **Brainwaves** and their associated mental states

Since different brain-waves activities correlate with different states, it is possible to imagine various protocols to enhance the activity of brainwaves that are related to soothing states. This is the main goal of a type of biofeedback called neurofeedback.

### 11.2.3. Neurofeedback

Neurofeedback is a technique that makes use of real-time feedback on brainwave activity. The goal is to teach people how to control their brainwave activity and limit it to a certain frequency range, representative of a characteristic mental state. A typical application is to control stress-related conditions by learning to increase alpha wave activity, which correspond to a relaxed state (Nunez, 2005). The feedback information allows people being monitored to better understand the process and gain some conscious control over the generation of different brain waves. Interestingly enough EEG studies of Yogis and Zen Masters, showed that high levels of alpha could be observed during meditation (Anand et al., 1969). It is generally admitted that this increased alpha activity leads to less anxiety, improved attention and enhanced cognitive functioning (Norris and Currieri, 1999).
11.2.4. EEG-based Music Composition

One of the first attempts to use brainwaves to generate music was the piece “Music for solo performer” composed by Alvin Lucier in 1965 (Wikipedia, 2008a) where he used brainwaves as a generative source for the whole piece. In this piece the EEG signal from the performer was amplified and relayed to a set of loudspeakers coupled with percussion instruments.

In the seventies, the composer David Rosenboom, started to systematically use EEG output as a means to create or enhance performance art and music. He used biofeedback devices such as EEG to allow performers to create sounds and music using their own brainwaves (Rosenboom, 1989).

More recent research has attempted to create complex musical interaction between particular brainwaves and corresponding sound events where the listener EEG control a music generator imitating the style of a previously listened sample (Miranda et al., 2003).

Data sonification in general and EEG sonification in particular has been the subject of recent studies (Hinterberger and Baier, 2005) showing the ability of the human auditory system to deal with and understand highly complex sonic representation of data.

In the present work, we describe a unified framework where the user can learn to actively control her alpha-wave activity through interactive musical feedback. The musical parameters the user can control via her brainwave activity are based on psychoacoustics studies. They are designed to be perceptually salient in order to enhance the understanding of the interaction.

11.3. The EEG System

11.3.1. Hardware

11.3.1.1. ENOBIO: a dry electrode wireless device

Nowadays most of the scalp EEG use a conductive gel or paste applied on the scalp to reduce impedance and obtain clean recordings of the signal from the electrodes. With the advent of new technologies such as carbon nanotubes, it has now become possible to penetrate the outer layer of the skin and have an improved electrical contact with dry electrodes.

We used such a dry electrode wireless system called ENOBIO from Starlab1, Barcelona (Ruffini et al., 2006). ENOBIO is a wearable, wireless, 4-channel, all-digital electrophysiology recording system that has been optimised for dry electrodes (Figure 11.2). The cables allow any combination of EEG (Electroencephalogram - brain activity), EOG

1http://starlab.es/
11.3. The EEG System

(Electrooculogram - eye movement) and ECG (Electrocardiogram - heart activity) in a single wearable system without loss of signal quality.

![Image of EEG setup](image)

Figure 11.2.: The ENOBIO is a wearable wireless dry electrode electrophysiology recording system.

For the purpose of this chapter we focused on EEG signals only.

### 11.3.1.2. Electrode Placement

Following the international 10-20 system for EEG scalp electrode location, the signal from FP1 and FP2 (Cf. Figure 11.3) are usually used for measuring beta channel whereas O1 and O2 give higher amplitudes in the alpha range (Nunez, 2005). Nevertheless, hair reduces considerably the sensitivity of ENOBIO’s dry electrodes, which excludes the possibility to use O1 and O2 locations in our setup. The advantage of the ENOBIO system though is that it can be easily usable in many different contexts (such as performances, relaxation sessions, etc.) since (dis)connecting is as easy as putting the cap on or off (no gel to put or specific preparation). We opted for the FP1, FP2 only and the ground reference is clipped to the left earlobe. Even if the overall amplitude of alpha waves in the FP area is not optimal, it is still possible to measure the relative increase of energy in the alpha range compared to the beta range.

### 11.3.2. Software

#### 11.3.2.1. A Client-Server Architecture

The ENOBIO device is not provided with any standard data analysis tools yet, but it transfers binary-encoded stream of data via a TCP-IP server called JENOBIOS. We implemented a set of TCP client/decoder threaded externals in C for Max-MSP (Zicarelli, 2002) and PD (Puckette, 1996) to easily communicate with the JENOBIOS server and decode the stream of bytes into the 4 different channels in real-time.²

²[http://www.dtic.upf.edu/~slegroux/phd/downloads](http://www.dtic.upf.edu/~slegroux/phd/downloads)
11. Neuro-SMuSe: Neurofeedback Through Musical Interaction

Figure 11.3.: **The 10-20 system:** Standard location and nomenclature of the electrodes for EEG recording according to the 10-20 system (American Electroencephalographic Society)

11.3.2.2. Filtering

For each channel we implemented a suite of standard signal processing Max-MSP modules to extract each brainwave information as accurately as possible. The raw signal is first high-passed with a cutoff frequency of 0 Hz and low-passed with a cutoff frequency of 35 Hz to avoid ambient noise from the system. We then designed a set of bandpass butterworth filters with center frequency of 10 Hz and 22 Hz and bandwidth of 4Hz and 8Hz respectively for the alpha and beta brainwave specification.

11.3.2.3. Power Band Spectrum

For each channel the powerband spectrum was computed and averaged using a moving average filter of length 20 in order to obtain a relatively smooth representation of the alpha and beta bands that can be fed to the adaptive musical system. We also implemented a set of graphical interfaces and displays in Max-MSP that allow us to check in real-time the evolution of the signal for each channel (Cf. Figure 11.4).
11.4. Interactive Music Generation

11.4.1. Macro Structure: I-Ching

The generation of a musical structure for this project was intentionally fairly simple. We based the real-time algorithmic composition process on the I-Ching principle used in works such as “Music of Changes” by John Cage where precomposed musical cells are subjected to modulations.

We used the SMuSe music engine (Cf. Chapter 3) to define basic musical material such as Rhythm, Pitch, Velocity, Duration and deterministic and probabilistic techniques to modulate this material. For this project we wanted to put the emphasis on the timbre modulation of the musical material, to obtain a minimalist, meditative musical atmosphere. The pitch sequence is based on the following indian scale That Todi: \{C Db Eb F# G Ab Bb\} that was chosen for its meditative mood. A 10 beat primitive rhythmic cell Jhaptal was defined as the following division of beat: \{2 + 3 + 2 + 3\}. The selection principle of each element (note or rhythmic entity) followed a serial music paradigm: in one run each element was played only once in a random order. When all the elements of the list had been played once, the procedure started again. For the velocity and duration parameters, we chose a random walk process with a velocity range between 10 and 120 on the midi scale and a note duration between 40 and 400 ms. This gave a distinct character to the piece\(^3\) since the scale was the same, but enough randomness was injected so that the piece was not completely predictable.

\(^3\)http://www.dtic.upf.edu/~slegroux/phd/neurosmuse.html
11.4.2. Micro Structure: Perceptually Grounded Sound Generation Based on the Tristimulus Timbre Space

Here, we put the emphasis on studying a tristimulus synthesizer that allows the modulation of subtle timbral features that are perceptually relevant (Gabrielson and Juslin, 1996).

We implemented a polyphonic synthesizer with a GUI-based interface (Figure 11.5) in Max-MSP which relied on the tristimulus model of timbre (Pollard and Jansson, 1982; Riley and Howard, 2004). It provided us with a simple and intuitive interface to modulate relevant perceptual features such as brightness, noisiness, harmonicity, or odd/even ratio of an additive plus noise synthesis model.

Figure 11.5.: The tristimulus synthesizer allows control over tristimulus parameters, ADSR envelope, noisiness, loudness, inharmonicity and vibrato.

The tristimulus analysis of timbre proposes to quantify timbre in terms of three coordinates (x, y, z) associated with band-loudness values. Inspired from the tristimulus theory of colour perception, it associates high values of x to dominant high-frequencies, high values of y to dominant mid-frequency components and high values of z to dominant fundamental frequency.

The sound is represented as a sum of weighted sinusoids at frequencies multiple of the fundamental frequency plus a noise factor (McAulay and Quatieri, 1986).

In the synthesis model, the harmonics partials belong to three distinct frequency bands or tristimulus bands: \( f_0(n) \) belongs to the first low frequency band, frequencies \( f_{2\ldots4}(n) \) belong to the second mid-frequency band, and the remaining partials \( f_{5\ldots N}(n) \) belong to the high-frequency band.

The relative intensities in the three bands can be visualized on a tristimulus triangular diagram where each corner represents a specific frequency band (Figure 11.6). With this model the brightness can be modulated by the position of a cursor in the tristimulus timbre space (cf. Figure 11.5), where the lower left corner of the triangle corresponds to the darkest class of sounds whereas lower right corner correspond to the brightest.

The noisy part of the sound was generated following the subtractive synthesis paradigm. We filtered a random noise generator with a bank of three passband filters centered at \( f_0 \), \( 3f_0 \) and \( 9f_0 \) respectively as suggested in (Riley and Howard, 2004) so that the noisy portion of the sound follows a tristimulus spectral distribution.
11.5. Neurofeedback Training with Music

Noisiness is defined as the relative amplitude of the filtered noise generator. Inharmonicity relates to the factor by which successive partials deviate from the harmonic spectrum. Finally, we define the even partial attenuation factor as the relative amplitude level of even partials in the spectrum of the signal.

Figure 11.6: **Real-time spectral analysis and visualization of the tristimulus synthesizer**: the upper display represents the weighting of the harmonics as defined by the tristimulus controller. The middle display represents the time varying harmonics after synthesis. The bottom display represents the audio signal as well as the spectrogram and spectral density.

The synthesizer allows the control of perceptually relevant sound parameters, which variation has been shown to correlate with different emotional states (Gabrielson and Juslin, 1996). Our hope is that the differences in parameters are clear enough so that the listener will easily learn to make the distinction between different musical parameters that lead to different musical “moods”.

11.5. Neurofeedback Training with Music

In our interactive paradigm, the user’s goal should be to consciously control her brainwaves to stay in the alpha band. The music acts as an auditory feedback that helps the listener understand the process. The musical parameters are predefined and should clearly represent positive feedback if the EEG detects more energy in the alpha band, or negative feedback if the energy is principally in the beta band.

For this purpose we chose to use a set of parameters based on psychoacoustical studies that investigated the relation between musical feature and valence (Gabrielson and Juslin, 1996). Noisy, inharmonic, bright (high tristimulus component) and loud sounds were associated to negative feedback whereas harmonic, dark (low tristimulus) component and mellow sounds were associated to positive feedback.
During a set of informal experiments done at the university, it appeared that listeners could practice and learn to gain control over their brainwave emission thanks to the auditory feedback component and that the experience was pleasant. Nevertheless the process was not completely obvious, and a more detailed and systematic study of the parametrization of the system is needed. First impressions were encouraging, but we will need further investigation and quantitative assessment of the neurofeedback paradigm.

11.6. Conclusions

We presented a unified framework for learning to control EEG signals through a perceptually-grounded interactive sonic feedback system. The technology works in real-time, and the hardware is mobile, wireless, and easy to setup, making the system suitable for diverse environments. We believe the sound parameters we chose were good at representing different state of agitation and helped the listener to understand better the interaction. However, the whole system needs to be robustly evaluated. In the near future we plan a series of controlled experiments and statistical analysis to assess the validity of the system. Different interactive scenario are also possible. We can imagine a machine learning algorithm searching the sound synthesis parameter space for the set of parameters that would maximize the power in the alpha band. In this scenario the listener is not consciously adapting anymore. The system does the learning and adaptation itself in order to induce a relaxed state in the listener.

Music is well known for affecting human emotional states, yet the relationship between specific musical parameters and emotional responses is still not clear. With the advent of new human-computer interaction (HCI) technologies, it is now possible to derive emotion-related information from physiological data and use it as an input to interactive music systems. This raises the question of how musical parameters are mapped to emotional states. In this chapter, we assess this question using both verbal and physiological responses. While most of the work on musical interfaces is based on explicit HCI (e.g. involving gestures), here, we focus on the potential of implicit interaction based on human emotional states. Our results show that a significant correlation exists between electrodermal activity, heart rate, heart rate variability and the subjective evaluation of well-defined sound generation parameters. This demonstrates the feasibility of adaptive and interactive music composition based on physiological feedback. Providing implicit musical HCI will be highly relevant for a number of applications including music therapy, automatic generation of music for interactive virtual story telling and games, music for video games and physiologically-based musical instruments.

12.1. Introduction

It is generally acknowledged that music is a powerful carrier of emotions and the effect of music on emotional states has been established using many different self-report, physiological and observational means (Juslin and Sloboda, 2001; Meyer, 1956). Nevertheless, the precise relationship between musical parameters and emotional response is not clear. In the context of a mixed-reality environment called the eXperience Induction Machine (XIM) (Bernardet et al., 2007), we developed a real-world interactive composition and performance system which can produce musical structures and sonic textures in real-time, as a result of the interaction between the system and its human and non-human environment (Cf. Chapter 10). The musical output of an interactive multimedia system is thought as a communication channel that can reflect, express in some way, the sensory, behavioral and internal state of the interactive system itself.

In order to generate original affective music, we investigate the mapping between emotions and the musical output of our real-time composition and performance system. We
want to study the relationship between the musical parameters used to generate music and the listener’s emotional states. Although several methods are available to assess the emotional state of listeners, time-varying physiological measurements seems to be particularly adequate for real-time interactive applications. Here, we focus on the correlations between the implicit, physiological measure of the emotional state of the listener and the musical parameters used to generate music.

12.2. Background

Recent advances in human computer interaction have provided researchers and musicians with easy access to physiology sensing technologies. Although the idea is not new (Knapp and Lusted, 1990; Rosenboom, 1989), the past few years have witnessed a growing interest from the computer music community in using physiological data to generate or transform sound and music. In the literature, we distinguish three main trends: the use of physiology to modulate pre-recorded samples, to directly map physiological data to synthesis parameters (sonification), or to control higher level musical structures with parameters extracted from the physiology. A popular example of the first category is the Fraunhofer StepMan sensing and music playback device (Bieber and Diener, 2005) that adapts the tempo of the music to the speed and rhythm of joggers’ step, calculated from biosensoric data. While this approach appears efficient and successful, the creative possibilities are somewhat limited. In other work (Arslan et al., 2006), the emphasis is put on the signal processing chain for analyzing the physiological data, which in turn is sonified, using ad-hoc experimental mappings. Although raw data sonification can lead to engaging artistic results, these approaches do not use higher-level interpretation of the data to control musical parameters. Finally, musicians and researchers have used physiological data to modulate the activity of groups of predefined musical cells (Hamilton, 2006). This approach allows for interesting and original musical results, but the relation between the emotional information contained in the physiological data and the composer’s intention is usually not explicit.

In this chapter we assess the relationship between physiological response and music generation parameters. If specific musical parameters produce specific physiological responses (thus certain affective states), then those sound parameters can be used as a compositional tool to induce emotional states in the listener.

12.3. Adaptive Music Generator

The music generation, interaction and composition tools are based on the SMuSe (Cf. Chapter 3) and the psychoacoustics of emotional sonic expression (Cf. Part III) (Le Groux et al., 2007c; Gabrielsson and Lindström, 2001). Here we propose to use sensory data provided by the listener’s physiology to generate musical structures and validate the choice of adequate parameters for inducing specific emotional states. First, we give an
overview of our musical generation system (a more detailed description can be found in (?)). We have chosen a set of control musical parameters based on the requirement that their modulation should have a clear perceptual effect. We kept a list of parameters that has been extensively studied, and whose effect on emotional expression is widely acknowledged, as described in Juslin and Sloboda 2001 for instance. Those parameters are tempo, mode, register, tonality, consonance and rhythm for musical structure, articulation, tempo and volume modulation for expressivity and brightness and harmonicity for sound generation.

The musical material consists of the real-time modulation of precomposed musical cells. The SMuSe’s GUI represents one musical cell (rhythmic values and chords) for two voices that communicate with OSC (Cf. Chapter 3 for more details).

Figure 12.1.: The musical material consists of the real-time modulation of precomposed musical cells. The SMuSe’s GUI represents one musical cell (rhythmic values and chords) for two voices that communicate with OSC (Cf. Chapter 3 for more details).

The generation of music is based on the real-world composition paradigm, where prepared musical material is dynamically modulated as the users interact with the system. When the interaction between people and the system takes place, these basic musical events are dynamically modified. The initial musical material is amplified, transformed, nuanced, as the interaction between the system and the users evolves. The relation between structural levels of music generation and emotional have been extensively studied elsewhere ((Juslin and Sloboda, 2001), Chapter 8) , and we decided to limit the scope of this chapter to the quantitative study of a small set of only specific sound features. For fine timbre control, we used the tristimulus synthesizer (Cf. Figure 12.2) which provides a simple and intuitive interface for controlling spectral properties of an additive synthesis model such as spectral content, harmonicity, or odd/even ration (Pollard and Jansson 1982; Riley and Howard 2004, Chapter 11 and 6).

12.4. Emotional Mappings

A comparative review of literature on music and emotion provided us with a set of different musical parameters that have been reported to elicit specific emotional responses (Juslin and Sloboda, 2001). We decided to focus on a subset of parameters such as Loudness, Brightness, Harmonicity, Noisiness and Odd/Even ratio that can be easily produced by our synthesis engine. We also followed the well established bi-polar dimensional theory of emotions: hedonic valence or pleasantness and intensity of activation or arousal (Russell, 1980). Emotions can then be placed in a two-dimensional emotional space, where the valence scale ranges from pleasantness (happy, pleased, hopeful, positive, etc.) to unpleasantness (unhappy, annoyed, despairing, negative, etc.), and the activation scale extends from calmness (relaxed, sleepy or peaceful) to high arousal (excited, stimulated, energized or alert).

12.5. Method

12.5.1. Participants

We performed a pilot study using self-report and physiological measures. A total of 13 students from the university (1 female) ranging from 25 to 30 years of age (mean age 27 years) participated in the experiment. The experiment was conducted in accordance with the ethical standards laid down in the 1964 declaration of Helsinki.
12.5. Method

12.5.2. Apparatus, Materials and Data Collection

The experiment was conducted in a dark room. Each subject was seated in front of a computer, wired to the physiological equipment. Sound was delivered to the participants through circumaural headphones (K66 from AKG\(^1\)).

The physiological data was recorded using the G.tec Mobilab\(^2\) equipment. Electro Dermal Activity (EDA), Heart Rate (HR) and Heart Rate Variability (HRV) were collected using 256 Hz sampling rate. Physiological responses were measured for 3 s before the sound presentation, during the 10 s presentation of each sound and 2 s after sound offset. Processing of physiological data was done in Simulink and Matlab by Mathworks\(^3\).

12.5.3. Stimuli

The stimuli consisted of a set of 8 synthesized sounds from the tristimulus model of timbre with 4 different sound parameters.

- Brightness: low and high brightness corresponded to bottom left corner and bottom right corner of the tristimulus diagram (Cf. Figure 6.3).
- Loudness: the sound level difference between two samples was of 12 dB.
- Noisiness: low (respectively high) noisiness corresponded to a stimulus with harmonic gain 1 and inharmonic gain 0 (respectively harmonic gain 0 and inharmonic gain 1) (Cf. Figure 12.2).
- Inharmonicity: low inharmonicity corresponded to a harmonic spectrum, high inharmonicity to a multiplicative deviation factor from the harmonic spectrum of 4.

12.5.4. Experimental Design

Each of eight samples was presented to participants twice. The samples were presented in the randomized order using the following presentation sequence: 0.5 s. signaling tone, 7 s. pre-stimulus silence period, 10 s. stimulus presentation, 10.5 s. post-stimulus silence period. Participants were instructed to close their eyes once hearing the signaling tone and concentrate on the sound sample. After a sound sample offset, participants opened their eyes and had to rate their emotional state during the post-stimulus silence before the next signaling tone using the 9-point valence and arousal pictorial scales of the Self-Assessment Manikin (SAM) (Lang, 1980)

\(^1\)http://www.akg.com/
\(^2\)http://www.gtec.at/Products/Hardware-and-Accessories/g.MOBIIab-Specs-Features
\(^3\)http://www.mathworks.com/
12.6. Results

Self-reported valence and arousal were analyzed separately using repeated paired-sample t-tests for each sound parameter (averaged over the repetitions). **Loudness** level reached significance on the *arousal* scale, $t(7) = -3.11$, $p < 0.01$ with mean ratings of 4.4 (SE = 0.6) vs. 6.1 (SE = 0.4) for louder sample. In addition, **brightness** level reached significance on the *arousal* scale, $t(7) = 2.6$, $p < 0.03$ with mean ratings of 5.9 (SE = 0.4) vs. 4.5 (SE = 0.5) for the high brightness level. **Noisiness** level reached significance on the *valence* scale, $t(7) = -2.14$, $p < 0.05$ with mean ratings of 4.2 (SE = 0.5) vs. 4.9 (SE = 0.27) for less noisy samples. Other ratings did not reach significant differences (Cf. Table 12.1). These results corroborate previous studies on emotional responses to musical parameters (Cf. Chapter ?? and a review in (Gabrielsson and Lindström, 2001)).

<table>
<thead>
<tr>
<th>Noisiness</th>
<th>Valence (Std Error)</th>
<th>Arousal (Std Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>4.9 (.27)*</td>
<td>5.46 (.4)</td>
</tr>
<tr>
<td>high</td>
<td>4.2 (.5)*</td>
<td>5.5 (.52)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Brightness</th>
<th>Valence (Std Error)</th>
<th>Arousal (Std Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>4.2 (.57)</td>
<td>4.5 (.33)*</td>
</tr>
<tr>
<td>high</td>
<td>4.1 (.49)</td>
<td>5.8 (.44)*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inharmonicity</th>
<th>Valence (Std Error)</th>
<th>Arousal (Std Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>4.92 (.46)</td>
<td>5.46 (.4)</td>
</tr>
<tr>
<td>high</td>
<td>4.37 (.44)</td>
<td>5.46 (.47)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Loudness</th>
<th>Valence (Std Error)</th>
<th>Arousal (Std Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>4.83 (.51)</td>
<td>4.4 (.63)*</td>
</tr>
<tr>
<td>high</td>
<td>4.33 (.54)</td>
<td>6.12 (.41)*</td>
</tr>
</tbody>
</table>

Table 12.1.: **Influence of sound feature levels** on self-reported Valence and Arousal. High Brightness and Loudness levels significantly increase perceived arousal.

Significant correlations between verbal ratings of valence/arousal and physiological measures (averaged) of heart rate, heart rate variability and electrodermal activity were found (Cf. Table 12.2). Subjective level of *arousal* positively correlated with **EDA** ($p < 0.001$) and negatively with **HRV** data ($p < 0.05$). **Valence** ratings positively correlated with **Heart Rate** ($p < 0.05$). These results are in a good agreement with other findings (Dawson et al., 2000) which show that 1) increase in EDA level can be used for monitoring arousal state; 2) heart rate (long term changes) increase for positive states and decrease for negative stimuli (Cuthbert et al., 1996) and 3) heart rate variability (short term changes) reduces with arousing and attention attracting stimuli (Lang, 1990b).

12.7. Conclusions

In this chapter we investigated the potential of using physiological data to extract information about emotional states. We studied a set of well-defined sound parameters
12.7. Conclusions

Valence  Arousal  HR  HRV  EDA
Valence  Pearson Correlation  1  .009  .421*  .121  -.185
        Sig(2-tailed)  .963  .017  .510  .311
Arousal  Pearson Correlation  .009  1  .090  -.411*  .606**
        Sig(2-tailed)  .963  .626  .019  .000
HR      Pearson Correlation  .421*  .090  1  -.213  -.295
        Sig(2-tailed)  .017  .626  -.241  .101
HRV     Pearson Correlation  .121  -.411*  -.213  1  -.295
        Sig(2-tailed)  .510  .019  .241  .366
EDA      Pearson Correlation  -.185  .606**  -.295  -.165  1
        Sig(2-tailed)  .311  .000  .101  .366

Table 12.2: The correlation table between self reports (valence, arousal) and physiology (heart rate, heart rate variability and electrodermal activity) shows significant correlations between heart rate and valence (p<0.05), heart rate variation and arousal (p<0.05) and arousal and EDA (p<0.001)

and showed that a variation of parameters such as noisiness, loudness and brightness triggered significantly different physiological responses, corresponding to distinct self-reported affective states. We found that EDA and HRV were good indicators of arousal, while HR could be used for valence.

In the future, we will add other physiological measures such as respiration or electromyography (EMG) which is known to be a reliable measure of stimuli valence (Schwartz et al., 1980) and would be a good complement to heart rate data.

The verbal ratings showed that the musical samples had valence and arousal scores that were generally rather neutral. In this chapter we focused on subtle timbral properties of sound, not on higher-level musical structure. In the future, we will investigate more extreme values of the synthesis parameters as well as variations of the structural and expressive parameters of music so that the sound samples represent broader range of emotional responses.

One challenge of real-time music generation with physiological feedback based on HR and EDA is the time-varying properties physiological signals. In this study, we used averaged values of HR and EDA over time to compute correlations. In a real-time context, we want to make a decision on the spot. To investigate in more details the potential of musical parameters to induce specific affective states, we wish to expand our analysis to time-varying parameters, and to co-varying parameters.

Our results however already demonstrate that a rational approach towards the definition of interactive music systems driven by emotion is feasible. We can imagine various applications of this framework in such diverse fields as music therapy, automatic generation of music for interactive story telling, music for video games, physiologically-based musical instruments. In the future studies we plan to refine emotional mappings of musical
parameters using adaptive music synthesis and a response evaluation loop.
13. RL-SMuSe: A Closed-loop Interactive Music System Based on Reinforcement Learning

Although music is often defined as the “language of emotion”, the exact nature of the relationship between musical parameters and the emotional response of the listener remains an open question. Whereas traditional psychological research usually focuses on an analytical approach, involving the rating of static sounds or preexisting musical pieces, we propose a synthetic approach based on a novel adaptive interactive music system controlled by an autonomous reinforcement learning agent. Preliminary results suggest an autonomous mapping from musical parameters (such as tempo, articulation and dynamics) to the perception of tension is possible. This paves the way for interesting applications in music therapy, interactive gaming, and physiologically-based musical instruments.

13.1. Introduction

Music is generally admitted to be a powerful carrier of emotion or mood regulator, and various studies have addressed the effect of specific musical parameters on emotional states (Meyer, 1956; Gabrielsson and Lindström, 2001; Le Groux et al., 2008b; Krumhansl, 1997a; Bradley and Lang, 2000; Le Groux and Verschure, 2010b). Although many different self-report, physiological and observational means have been used, in most of the cases those studies are based on the same paradigm: one measures emotional responses while the subject is presented to a static sound sample with specific acoustic characteristics or an excerpt of music representative of a certain type of emotions.

In this chapter, we take a synthetic and dynamic approach to the exploration of mappings between perceived musical tension (Fredrickson, 2000; Krumhansl, 1996) and a set of musical parameters by using reinforcement learning (RL) (Sutton and Barto, 1998).

Reinforcement learning (as well as agent-based technology) has already been used in various musical systems and most notably for improving real time automatic improvisation (Assayag et al., 2006; Franklin and Manfredi, 2002; Thom, 2000; Collins, 2008). Musical systems that have used reinforcement learning can roughly be divided into three main categories based on the choice of the reward characterizing the quality of musical actions. In one scenario the reward is defined to match internal goals (a set of rules for
13. RL-SMuSe: A Closed-loop Interactive Music System Based on Reinforcement Learning

![Diagram of RL-SMuSe system](image)

Figure 13.1: The Reinforcement Learning framework. The system is composed of three main components, namely the music engine (SMuSe) that generates music, the reinforcement learning agent that modulates musical parameters based on feedback from the participant, and the listener who provides the reward signal based on the musical output generated by the SMuSe.

instance), in another scenario it can be given by the audience (a like/dislike criterion), or else it is based on some notion of musical style imitation (Collins, 2008). Unlike most previous examples where the reward relates to some predefined musical rules or quality of improvisation, we are interested in the emotional feedback from the listener in terms of perceived musical tension (Figure 13.1).

Reinforcement learning is a biologically plausible machine learning technique particularly suited for an explorative and adaptive approach to emotional mapping as it tries to find a sequence of parameter change that optimizes a reward function (in our case musical tension). This approach contrasts with expert systems such as the KTH rule system (Friberg et al., 2006; Friberg, 2006) that can modulate the expressivity of music by applying a set of predefined rules inferred from previous extensive music and performance analysis. Here, we propose a paradigm where the system learns to autonomously tune its own parameters in function of the desired reward function (musical tension) without using any a-priori musical rule.

Interestingly enough, the biological validity of RL is supported by numerous studies in psychology and neuroscience that found various examples of reinforcement learning in animal behavior (e.g. foraging behavior of bees (Montague et al., 1995), the dopamine system in primate brains (Schultz et al., 1997), ...).
13.2. A Hierarchy of Musical Agents For Music Generation

We generated the music with SMuSe, the Situated Music Server, which is composed of a hierarchy of perceptually meaningful musical agents interacting and communicating via the OSC protocol (Chapter 3).

In this project, we studied the modulation of music by three parameters that contribute to the perception of musical tension, namely articulation, tempo and dynamics.

While conceptually fairly simple, the music material generator has been designed to keep the balance between predictability and surprise. The real-time algorithmic composition process is inspired by works from minimalist composers such as Terry Riley (*In C, 1964*) where a set of basic precomposed musical cells are chosen and modulated at the time of performance creating an ever-changing piece.

The choice of base musical material relies on the extended serialism paradigm. We a priori defined sets for every parameter (rhythm, pitch, register, dynamics, articulation). The generation of music from these sets is then using non-deterministic selection principles, as proposed by Gottfried Michael Koenig (Laske, 1981).

For this project we used a simple modal pitch serie \{0, 3, 5, 7, 10\} shared by three different voices (2 monophonic and 1 polyphonic). The first monophonic voice is the lead, the second is the bass line, and the third polyphonic voice is the chord accompaniment. The rhythmic values are coded as 16n for a sixteenth note, 8n for a eighth note, etc. The dynamic values are coded as midi velocity from 0 to 127. The other parameters correspond to standard pitch class set and register notation. The pitch content for all the voices is based on the same mode.

- **Voice1:**
  - Rhythm: \{16n 16n 16n 16n 8n 8n 4n 4n\}
  - Pitch: \{0 3 5 7 10\}
  - Register: \{5 5 6 6 7 7 7\}
  - Dynamics: \{90 90 120 50 80\}

- **Voice2:**
  - Rhythm: \{4n 4n 4n 8n 8n\}
  - Pitch: \{0, 3, 5, 7, 10\}
  - Register: \{3 3 3 4 4 4 4\}
  - Dynamics: \{90 90 120 50 80\}

- **Polyphonic Voice:**
  - Rhythm: \{2n 4n 2n 4n\}
13. RL-SMuSe: A Closed-loop Interactive Music System Based on Reinforcement Learning

- Pitch: \{0\ 3\ 5\ 7\ 10\}
- Register: \{5\}
- Dynamics: \{60\ 80\ 90\ 30\}
- with chord variations on the degrees: \{1\ 4\ 5\}

The selection principle was set to “series” for all the parameters so the piece would not repeat in an obvious way\(^1\). This composition paradigm allowed the generation of constantly varying, yet coherent, musical sequences. Properties of the music generation such as articulation, dynamics modulation and tempo are then modulated by the RL algorithm in function of the reward defined as the musical tension perceived by the listener.

13.3. Musical Tension as a Reward Function

We chose to base the autonomous modulation of the musical parameters on the perception of tension. It has often been said that musical experience may be characterized by an ebb and flow of tension that gives rise to emotional responses (Vines et al., 2006; Chapados and Levitin, 2008). Tension is considered a global attribute of music, and there are many musical factors that can contribute to tension such as pitch range, sound level dynamics, note density, harmonic relations, implicit expectations, ...

The validity and properties of this concept in music have been investigated in various psychological studies. In particular, it has been shown that behavioral judgements of tension are intuitive and consistent across participants (Fredrickson, 2000; Krumhansl, 1996). Tension has also been found to correlate with the judgement of the amount of emotion of a musical piece and relates to changes in physiology (electrodermal activity, heart-rate, respiration) (Krumhansl, 1997b).

Since tension is a well-studied one-dimensional parameter representative of a higher-dimensional affective musical experience, it makes a good candidate for the one-dimensional reinforcer signal of our learning agent.

13.4. Pilot Experiment

As a first proof of concept, we looked at the real-time behaviour of the adaptive music system when responding to the musical tension (reward) provided by a human listener. The tension was measured by a slider GUI controlled by a standard computer mouse. The value of the slider was sampled every 100 ms. The listener was given the following instructions before performing the task: “use the slider to express the tension you experience during the musical performance. Move the slider upwards when tension increases and downward when it decreases”.

\(^1\)Sound examples: [http://www.dtic.upf.edu/~slegroux/phd/rlsmuse.html](http://www.dtic.upf.edu/~slegroux/phd/rlsmuse.html)
The music generation is based on the base material described in section 2. The first monophonic voice controlled the right hand of a piano, the second monophonic voice an upright acoustic bass and the polyphonic voice the left hand of a piano. All the instruments were taken from the EXS 24 sampler from Logic Studio\(^2\) from Apple.

The modulation parameter space was of dimension 3. Dynamics modulation were obtained via a midi velocity gain factor between \([0.0, 2.0]\). Articulation was defined on the interval \([0.0, 2.0]\) (where a value > 1 corresponded to a legato and < 1 a staccato). Tempo was modulated from 10 to 200 BPM. Each dimension was discretized into 8 levels, so each action of the reinforcement algorithm produced an audible difference. The reward values were discretized into three values representing musical tension levels (low=0, medium=1 and high=2).

We empirically setup the sarsa(\(\lambda\)) parameters, to \(\varepsilon = 0.4, \lambda = 0.8, \gamma = 0.1, \alpha = 0.05\) in order to have an interesting musical balance between explorative and exploitative behaviors and some influence of memory on learning. \(\varepsilon\) is the probability of taking a random action. \(\lambda\) is the exponential decay of reward (the higher \(\lambda\), the less the agent remembers). \(\alpha\) is the learning rate (if \(\alpha\) is high, the agent learns faster but can lead to suboptimal solutions).

### 13.4.1. One-Dimensional Independent Adaptive Modulation of Dynamics, Articulation and Tempo

As our first test case we looked at the learning of one parameter at a time. For **dynamics**, we found a significant correlation \((r = 0.9, p < 0.01)\): the tension increased when velocity increased (Figure 13.2). This result is consistent with previous psychological literature on tension and musical form (Krumhansl, 2002). Similar trends were found for **articulation** \((r = 0.25, p < 0.01)\) (Figure 13.3) and **tempo** \((r = 0.64, p < 0.01)\) (Figure 13.4). Whereas literature on tempo supports this trend (Husain et al., 2002; Gabrielson and Lindström, 2001), reports on articulation are more ambiguous (Gabrielson and Lindström, 2001).

### 13.4.2. Two-Dimensional Modulation of Tempo and Dynamics

When testing the algorithm on the 2-dimensional parameter space of Tempo and Dynamics, the convergence was slower. For our example trial, an average reward of medium tension (value of 1) was only achieved after 16 minutes of training (1000 s) (Figure 13.5) compared to 3 minutes (200 s) for dynamics only (Figure 13.2). We observed significant correlations between tempo \((r = 0.9, p < 0.01)\), dynamics \((r = 0.9, p < 0.01)\) and reward in this example, so the method remains useful for the study the relationship between parameters and musical tension. Nevertheless, in this setup, the time taken to converge

\(^2\)http://www.apple.com/logicstudio/
Figure 13.2.: **Learning dynamics:** the RL agent automatically learns to map an increase of perceived tension, provided by the listener as a reward signal, to an increase of the dynamics gain. Dynamics gain level is in green, cumulated mean level is in red/thin, reward is in blue/crossed and cumulated mean reward is in red/thick.
13.4. Pilot Experiment

Figure 13.3.: **Learning articulation:** the RL agent learns to map an increase of perceive tension (reward) to longer articulations.

Figure 13.4.: **Learning Tempo:** the RL agent learns to map an increase of musical tension (reward) to faster tempi.
13. RL-SMuSe: A Closed-loop Interactive Music System Based on Reinforcement Learning

Figure 13.5.: The RL agent learns to map an increase of musical tension (reward in blue/thick) to faster tempi (parameter 1 in green/dashed) and higher dynamics (parameter 2 in red/dashed).

towards a maximum mean reward would be too long for real-world applications such as mood induction or music therapy.

13.4.3. Three Dimensions: Adaptive Modulation of Volume, Tempo and Articulation

When generalizing to three musical parameters (three dimensional state space), the results were less obvious within a comparable interactive session time frame. After a training of 15 minutes, the different parameters values were still fluctuating, although we could extract some trends from the data. It appeared that velocity and tempo were increased for higher tension, but the influence of the articulation parameter was not always clear. In figure 13.6 we show some excerpt where a clear relationship between musical parameter modulation and tension could be observed. The piano roll representative of a moment where the user perceived low tension (center) exhibited sparse rhythmic density due to lower tempi, long notes (long articulation) and low velocity (high velocity is represented as red) whereas a passage where the listener perceived high tension (right) exhibited denser, sharper and louder notes. The left figure representing an early stage of the reinforcement learning (beginning of the session) does not seem to exhibit any special characteristics (we can observe both sharp and long articulation, e.g. the low voice (register C1 to C2) is not very dense compared to the other voices).
13.5. Conclusions

From these trends, we can hypothesize that perception of low tension would relate to sparse density, long articulation and low dynamics which corresponds to both intuition and previous offline systematic studies (Krumhansl, 2002).

These preliminary tests are encouraging and suggest that a reinforcement learning framework can be used to teach an interactive music system (with no prior musical mappings) how to adapt to the perception of the listener. To assess the viability of this model, we plan more extensive experiments in future studies.

13.5. Conclusions

In this chapter we proposed a new synthetic framework for the investigation of the relationship between musical parameters and the perception of musical tension. We
created an original algorithmic music piece that could be modulated by parameters such as articulation, velocity and tempo, assumed to influence the perception of musical tension. The modulation of those parameters was autonomously learned in real-time by a reinforcement learning agent optimizing the reward signal based on the musical tension perceived by the listener. This real-time learning of musical parameters provided an interesting alternative to more traditional research on music and emotion. We observed statistically significant correlations between specific musical parameters and an increase of perceived musical tension. Nevertheless, one limitation of this method for real-time adaptive music was the time taken by the algorithm to converge towards a maximum average reward, especially when the parameter space was of higher dimensions. We will improve several aspects of the experiment in follow-up studies. The influence of the reinforcement learning parameters on the convergence needs to be tested in more details, and other relevant musical parameters will be taken into account. In the future we will also run experiments to assess the coherence and statistical significance of these results over a larger population.
14. The Brain Orchestra

Most new digital musical interfaces have evolved upon the intuitive idea that there is a causality between sonic output and physical actions. Nevertheless, the advent of brain-computer interfaces (BCI) now allows us to directly access subjective mental states and express these in the physical world without bodily actions. In the context of an interactive and collaborative live performance, we propose to exploit novel brain-computer technologies to achieve unmediated brain control over music generation and expression. In this chapter, we introduce a general framework for the generation, synchronization and modulation of musical material from brain signal and describe its use in the realization of Xmotion, a multimodal performance for a “brain quartet”.

14.1. Introduction

The Multimodal Brain Orchestra (MBO) demonstrates interactive, affect-based and self-generated musical content based on novel BCI technology. It is an exploration of the musical creative potential of a collection of unmediated brains directly interfaced to the world, bypassing their bodies.

One of the very first piece to use brainwave for generating music was “Music for solo performer” composed by Alvin Lucier in 1965 (Wikipedia, 2008a). He used brainwaves as a generative source for the whole piece. In this piece, the electroencephalogram (EEG) signal from the performer was amplified and relayed to a set of loudspeakers coupled with percussion instruments. Some years later, the composer David Rosenboom started to use biofeedback devices (especially EEG) to allow performers to create sounds and music using their own brainwaves (Rosenboom, 1989). More recent research has attempted to create complex musical interaction between particular brainwaves and corresponding sound events where the listener EEG control a music generator imitating the style of a previously listened sample (Miranda et al., 2003). Data sonification in general and EEG sonification in particular has been the subject of various studies (Hinterberger and Baier, 2005) showing the ability of the human auditory system to deal with and understand highly complex sonic representation of data.

Although there has been a renewed interest in brain-based music over the recent years, most projects are only based on direct mappings from the EEG spectral content to sound generators. They do not rely on explicit volitional control. The Multimodal Brain Orchestra (MBO) takes a different approach by integrating advanced BCI (Brain Computer Interface) technology that allows the performer complete volitional control.
over the command signals that are generated. MBO preserves the level of control of the instrumentalist by relying on classification of specific stimulus triggered events in the EEG. Another unique aspect of the MBO is that it allows for a multimodal and collaborative performance involving four brain orchestra members, a musical conductor and real-time visualization (Cf. Figure 14.1).

14.2. System Architecture

14.2.1. Overview: A client-server Architecture for Multimodal Interaction

The interactive music system of the Multimodal Brain Orchestra is based on a client-server modular architecture, where inter-module communication follows the Open Sound Control (OSC) protocol (Wright, 2005). The MBO consists of three main components (Figure 14.2) namely the orchestra members equipped with BCI, the interactive audiovisual system based on the SMuSe (Cf. Chapter 3), and the conductor. The four members of the “brain quartet” are wired up to two different types of brain-computer interfaces: two P300 and two SSVEP (Steady-State Visual Evoked Potentials) (Cf. Section 14.2.2). The computer-based interactive multimedia system processes inputs from the conductor and the BCIs to generate music and visualization in real-time. This is the core of the system where most of the interaction design choices are made. The interactive multimedia component can itself be decomposed into three subsystems: the EEG signal processing module, the SMuSe music server (Le Groux and Verschure, 2009b) and the real-time visualizer. Finally, the conductor uses a Wii-Baton (Cf. Section 14.2.4) to modulate the
The Multimodal Brain Orchestra is a modular interactive system based on a client-server architecture using the OSC communication protocol. See text for further information.

The tempo of the interactive music generation, trigger different sections of the piece, and cue the orchestra members (Figure 14.2).

14.2.2. Brain Computer Interfaces

The musicians of the orchestra are all connected to brain-computer interfaces that allow them to control sound events and music expressiveness during the performance. These BCIs provide a new communication channel between a brain and a computer. These interfaces are based on the principle that mental activity can lead to observable changes of electrophysiological signals in the brain. These signals can be measured, processed, and later transformed into useful high level messages or commands (Wolpaw, 2007; Felton et al., 2007; Guger et al., 2009).

The MBO is based on two different non-invasive BCI concepts which control the generation and modulation of music and soundscapes, namely the P300 and SSVEP. We
worked with G.tec medical engineering GmbH products\(^1\), providing BCI hardware devices (g.USBamp) and corresponding real-time processing software for MATLAB/Simulink\(^2\). The control commands generated by the classification of the EEG using the, so called, P300 and SSVEP protocols were sent to the music server and visualization module via a Simulink S-function implementing using the OSC protocol for Matlab\(^3\).

14.2.2.1. The P300 Speller

The P300 is an event related potential (ERP) that can be measured with eight electrodes at a latency of approximately 300 ms after an infrequent stimuli occurs. We used the P300 speller paradigm introduced by (Farwell and Donchin, 1988). In our case, two orchestra members were using a 6 by 6 symbol matrix containing alpha-numeric characters (Figure 14.3) in which a row, column or single cell was randomly flashed on. The orchestra member had to focus on the cell containing the symbol to be communicated and to mentally count every time that specific cell flashed (this was to distinguish between common and rare stimuli). This elicits an attention dependent positive deflection of the EEG about 300 ms after stimulus onset, the P300, that can be associated to the specific symbol by the system (Figure 14.3) (Guger et al., 2009). We used this interface to trigger discrete sound events in real-time. Because it is difficult to control the exact time of occurrence of P300 signals, our music server SMuSe (Cf. Section 14.2.3) took care of beat-synchronizing the different P300 events with the rest of the composition (Cf. Chapter ).

A P300 interface is normally trained with 5-40 characters which corresponds to a training time of about 5-45 minutes. A group study with 100 people showed that after a training with 5 characters only, 72% of the users could spell a 5 character word without any mistake (Guger et al., 2009). This motivated the decision to limit the number of symbols used during the performance (Section 14.3.2.2).

14.2.2.2. A SSVEP-based Interface

Another type of interface took advantage of the steady-state visually evoked potentials (SSVEP) that can be observed in the EEG when looking at flickering lights. This method relies on the fact that when the retina is excited by a flickering light with a frequency > 3.5 Hz, the brain generates activity at the same frequency (Allison et al., 2008; Allison and Pineda, 2003). The interface was composed of four different light sources flickering at different frequencies and provided additional brain-controlled music “step controllers” (Figure 14.4).

The SSVEP BCI interface was trained for about 5 minutes during which the user had to look several times at every flickering LED. This session was used to train a classifier

\(^{1}\)http://www.gtec.at/
\(^{2}\)http://www.mathworks.com
\(^{3}\)http://andy.schmeder.net/software/
14.2. System Architecture

Figure 14.3.: **The P300 Speller.**  
**A)** Rare events triggers an ERP (Target X) 300 ms after the onset of the event indicated by the green arrow. **B)** A 6 by 6 symbol matrix is presented to the P300 orchestra member who can potentially trigger 36 specific discrete sound events.

Figure 14.4.: Two members of the orchestra connected to their SSVEP-based BCI interfaces (Prueckl and Guger, 2009)
based on EEG FFT analysis and Linear Discriminant Analysis (LDA). Once the classifier was trained, the interface could be used in real-time to trigger four distinct controllers. In contrast to the P300, the SSVEP BCI gave a continuous control signal that switched from one state to another within about 2-3 seconds. The SSVEP software provided by g.tec solved also the zero-class problem and when the performers were not looking at one of the LEDs then no musical decision was made (Prueckl and Guger, 2009).

During the performance, the two SSVEPs were used to control changes in articulation and dynamics of the music generated by the SMuSe.

### 14.2.2.3. Visual Feedback

We designed a module in openFrameworks\(^4\) that displayed a real-time visualization of the BCI output. More precisely, the different possible control messages detected by g.tec analysis software from the brain signal were sent to the visualizer via OSC and illustrated with simple color coded icons (Cf. Figure 14.5). From the two members of the orchestra using the P300 BCI interfaces we could receive 36 distinct control messages. Each of these 36 symbols was represented using a combination of six geometrical shapes and six different colors. The two other members of the orchestra who used the SSVEP BCI interfaces were able to trigger four possible events corresponding to the four different states (or in other words, four brain activity frequencies), but continuously. Each line of the visualizer corresponded to a member of the orchestra: the first two using P300 and the last two SSEVP. When a P300 member triggered an event, the associated geometrical shape appeared in the left side and moved from left to right as time passed by. For the SSVEP events, the current state was shown in green and the past changes could be seen as they moved from left to right.

The real-time visualization played the role of a real time score. It provided feedback to the audience and was used by the conductor to monitor the different events that had been triggered. The conductor could indicate to the orchestra member when to trigger an event (P300 or SSVEP) but the confirmation of its triggering was indicated by the real time visual score as well as by its musical consequences.

### 14.2.3. The Situated Music Server

Once brain activity was transformed into high-level commands by g.tec EEG software analysis suite, specific OSC message were sent to the SMuSe music server (Le Groux and Verschure, 2009b) and to the visualization module. SSVEP and P300 interfaces provided us with a set of discrete commands that could be transformed into control messages to the SMuSe (Cf. Chapter 3 for more details on the music server).

\(^4\)http://www.openframeworks.cc/
14.2. System Architecture

Figure 14.5.: **The real-time visualizer** displays a visualization of the two P300 systems output (the two upper rows show combinations of shapes and colors) and the two SSVEP system output (the two lower rows show the transitions from the 4 different states).

Figure 14.6.: **The wii-baton module** analyzed 3D acceleration data from the wii-mote so the conductor could use it to modulate the tempo and to trigger specific sections of the piece.

14.2.4. Wii-mote Conductor Baton

We provided the orchestra conductor with additional control over the musical output using the Wii-mote (Nintendo) as a baton. Different sections of the quartet could be triggered by pressing a specific button, and the gestures of the conductor were recorded and analyzed. A processing module filtered the accelerometers and the time-varying accelerations were interpreted in terms of beat pulse and mapped to tempo modulation of the SMuSe playing rate (Figure 14.6).
14. The Brain Orchestra

14.3. XMotion: a Brain-based Musical Performance

14.3.1. Emotion, Cognition and Musical Composition

One of the original motivations of the MBO project was to explore the potential creativity of BCIs as they allow to access subjective mental states and express these in the physical world without bodily actions. The name XMotion designate those states that can be generated and experienced by the unmediated brain when it is both immersed and in charge of the multimodal experience in which it finds itself.

The XMotion performance is based on the assumption that mental states can be organized along the three-dimensional space of valence, arousal and representational content (Mura et al., 2009). Usually emotion is described as decoupled from cognition in a low dimensional space such as the circumplex model of Russell (Russell, 1980). This is a very effective description of emotional states in terms of their valence and arousal. However, these emotional dimensions are not independent of other dimensions such as the representational capacity of consciousness which allows us to evaluate and alter the emotional dimensions (Laureys, 2005). The musical piece composed for XMotion proposes to combine both models into a framework where the emotional dimensions of arousal and valence are expressed by the music, while the conductor evaluates its representational dimension.

Basing our ideas on previous emotion research studies (Gabrielsson and Lindström, 2001; Le Groux et al., 2008b), we decided to control the modulation of music from Russell’s bidimensional model of emotions (Russell, 1980). The higher the values of the dynamics, the higher the expressed arousal and similarly, the longer the articulation, the higher the valence. In addition, a database of sound samples was created where each sample was classified according to the Arousal and Valence taxonomy (Table 14.1).

14.3.2. Musical Material

The musical composition by Jonatas Manzolli consisted of three layers, namely the virtual string quartet, a fixed electroacoustic tape and live triggering of sound events. The four voices of a traditional string quartet setup up were precomposed offline and stored as MIDI events to be modulated (articulation and accentuation) by the MBO members connected to the SSVEP interfaces. The sound rendering was done using state of the art orchestral string sampling technology (using the London Solo Strings library with Kontakt sampler). The second layer consisted of a fixed audio tape soundtrack synchronized with the string quartet material with Ableton Live audio time stretching algorithms. Additionally, we used discrete sound events triggered by the P300 brain

---

5http://www.bigfishaudio.com/
6http://www.native-instruments.com/
7http://www.ableton.com
orchestra members. The orchestra members were coordinated by the musical conductor standing in front of them.

14.3.2.1. String Quartet

The basic composition strategy was to associate different melodic and rhythmic patterns of musical textures to variations in dynamics and articulation producing textural changes in the composition. The inspiration for this music architecture was the so called net-structure technique created by Ligeti using pattern-mecanico material (Clendinning, 1993). The second aspect of the composition was to produce transposition of beats producing an effect of phase-shifting(Cohn, 1992). These two aspects produced a two-dimension gradual transformation in the string quartet textures. In one direction the melodic profile was gradually transformed by the articulation changes. On the other, the shift of accentuation and gradual tempo changes produced phase-shifts. In the first movement a chromatic pattern was repeated and the use of legato articulation progressively increased the superposition of notes. The second and fourth movements worked with a constant chord modulation chain and the third with a canonical structure.

One member of the orchestra used the SSVEP to modulate the articulation of the string quartet (four levels from legato to staccato corresponding to the four light sources frequencies) while the other member modulated the accentuation (from piano to forte) of the quartet.

213
14. The Brain Orchestra

<table>
<thead>
<tr>
<th>Sound Quality</th>
<th>State</th>
<th>Arousal</th>
<th>Valence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharp Strong</td>
<td>A</td>
<td>High</td>
<td>Negative</td>
</tr>
<tr>
<td>Short Percussive</td>
<td>B</td>
<td>High</td>
<td>Negative</td>
</tr>
<tr>
<td>Water Flow</td>
<td>C</td>
<td>Low</td>
<td>Positive</td>
</tr>
<tr>
<td>Harmonic Spectrum</td>
<td>D</td>
<td>Low</td>
<td>Positive</td>
</tr>
</tbody>
</table>

Table 14.1.: **Relationship between sound quality and emotion expression** on the scales of valence and arousal.

14.3.2.2. Soundscape

The soundscape was made of a fixed tape piece composition and discrete sound events triggered according to affective content. The sound events are driven by the conductor’s cues and relate to the visual realm. The tape was created using four primitive sound qualities. The idea was to associate mental states with changes of sound material. “P300 performers” produced discrete events related to four letters: A (sharp strong), B (short percussive), C (water flow) and D (harmonic spectrum). The sounds samples were organized in four different groups corresponding to different emotional levels of valence and arousal (Table 14.1). On the conductor’s cue, the performers concentrated on a specific column and row and triggered the desired sound. Two members of the orchestra were using P300 hundred and concentrated on 4 symbols each. Each symbol triggered a sound sample from the “emotional database” corresponding to the affective taxonomy associated with the symbol (for each symbol or sound quality we had a set of 4 possible sound samples).

14.4. Conclusions

We presented a disembodied interactive system designed for the generation and modulation of musical material from brain signal, and described XMotion, an interactive “brain quartet” piece based on novel brain computer interface technologies. The MBO shows how novel BCI technologies can be used in a multimodal collaborative context where the performers have volitional control over their mental state and the music generation process. Considering that the response time delays of the SSVEP and P300 interfaces are well above audio rate, we do not claim that these interfaces provide the level of subtlety and intimate control more traditional instruments can afford. Nevertheless, it is a promising first step towards the exploration of the creative potential of collaborative brain-based interaction for audio-visual content generation. It is part of a larger effort to include physiological feedback in the interactive generation of music. We can envision many applications of this brain-based systems beyond the area of performance including music therapy (this system fosters musical collaboration and would allow disable people to play music together), neurofeedback (Le Groux and Verschure, 2009a; Egner and Gruzelier, 2003; Mindlin and Rozelle, 2006) and motor rehabilitation (e.g. the use
of musical feedback for neurofeedback training might be a good alternative to visual feedback for people with visual impairment (Nijboer et al., 2008; Pham et al., 2005). We are further exploring both these artistic and practical applications of the MBO.
15. Conclusions and Outlook

The first part of this thesis introduced a conceptual framework for interactive music composition based on situated cognition. We showed the limitations of outdated classic cognition approaches focusing on internal state, centralized control and disembodiment and emphasized the importance of concepts such as parallelism, emergence, embodiment and emotional feedback (Chapter 2). Following a synthetic psychology approach, we built a situated music system called the SMuSe (Situated Music Server) based on our understanding of the psychology and neuroscience of music (Chapter 3). The SMuSe’s architecture is based on a hierarchy of musical agents that allows for flexible and distributed control of musical processes and includes a learning mechanism based on reinforcement. The SMuSe is able to sense its environment via a wide range of sensors which provide feedback about the emotional reactions of the participants interacting with the system. The interaction dynamics that occur between the SMuSe and its environment give rise to complex musical structures from relatively scarce initial musical material. This situated and synthetic approach provides a well-grounded paradigm for the development of advanced synthetic aesthetic systems. At the same time, it can be used as a tool to improve our understanding of the psychological processes on which the system is based.

The second part of the thesis focused on the perceptual control of the micro-level of musical sounds. It proposed two complementary approaches to achieve perceptual control over sound synthesis algorithms. The first approach was based on a support vector regression procedure that extracted perceptual features from audio data and learned to automatically map them to low-level additive synthesis parameters (Chapter 5). This method, albeit general and promising, required extensive data processing and computing. To deal with these problems, we introduced a second approach, more direct but less general. This explicit method, based on a physically inspired model of impact sounds allowed to explore a three-dimensional timbre space in real-time by manipulation of the parameters of a modal synthesizer (Chapter 6). These models bridged the gap between perceptually relevant sound description and sound generation algorithms, thus extending our psychologically-grounded approach to the micro level of music generation.

The third part of the thesis moved from the modeling, control and generation of music towards the study of the musical experience induced by the music generated by the SMuSe. Based on ample empirical evidence that emotions play a fundamental role in musical experience, we used the SMuSe to study the effect of musical parameters such as structure, performance and timbre on emotional responses (Chapter 8). We found that the results were consistent with the existing literature, which validated the use
of the SMuSe for generating controlled experimental stimuli. One big advantage of such a synthetic system is that it provides specific, detailed and independent control of the stimuli’s parameters, which is difficult to obtain using musical recordings. To further investigate the effect of sound on emotional experience, we ran a large scale study involving patients with dementia, a type of illness known to induce affective and cognitive deficits (Chapter 9). It appeared that specific sound features provoked significantly different emotional responses within the patients. We also found that the responses to specific sound parameters differed from the dementia patients to healthy age-matched controls. Finally, the stage of the disease clearly modulated the patients’ emotional experience to sounds, thus suggesting possible applications of our research to sound-based diagnostic tools.

The last part of the thesis presented a selection of applications demonstrating the use of the SMuSe in diverse real-world contexts. Namely, we described the generation of interactive music in a a mixed-reality space (Chapter 10), we introduced a novel music-based neurofeedback, looked at the physiological responses specific musical parameters (Chapter 11 and 12), we studied a closed-loop system for optimization of musical tension (Chapter 13) and we finally presented a large-scale multimedia performance based on brain computer interfaces. These applications demonstrated the robustness and maturity of the system.

The conceptual framework we introduced in this thesis provides a new perspective on the notion of musical creativity. We shifted the focus from the traditional modeling of complex rule-based internal musical processes towards taking advantage of the complexity inherent to real-world interaction with the environment. This change of frame of reference allows to create dynamic musical outputs from relatively simple cognitive structures embedded in a situated system. One unique aspect of the SMuSe is that it is conceived as both a real-time music composition tool and a synthetic psychology experiment that aims at exploring the composition process itself.

We can foresee several interesting artistic and scientific extensions of this work. For instance, we want to further explore the concept of stigmergic music composition, where standard rational musical composition techniques are replaced by distributed collective intelligence. What kind of musical structure and esthetics would emerge if the control of all the musical processes were heavily distributed? To answer these questions, the SMuSe stands out as an ideal platform since a key feature of the system is the distributed on-the-fly control of musical parameters and processes.

Along the same line, we will investigate the musical potential of the interaction between several autonomous instances of the SMuSe (as opposed to one instance with distributed control) “playing” together in the same environment. We will have to extend the capacity of the music analysis modules of the SMuSe so that each instance can extract usable feedback from the collective musical output, to drive its own modulation of musical parameters. One possibility for instance is to use an analysis based on consonance and dissonance to drive the collective musical interaction.

For practical reasons, we chose to implement our models at an abstract level, but the
different modules of the system can be modified to fit to more refined models of auditory perception and/or auditory physiology. At this stage we did not use any detailed neural models. Yet, the system’s modular architecture based on musical agents provides a lot of flexibility for further developments. In the future, specific modules could be replaced by more realistic neuromorphic models (e.g. a cerebellar mode of pulse generation). A promising direction would be to integrate the generic Distributed Adaptive Control (DAC) framework (Verschure et al., 2003) within SMuSe. DAC proposes a biomimetic detailed neural model of classical and operand conditioning. It allows for the learning and generation of sequences of perceptions/actions based on feedback from the environment and internal states, which is particularly relevant to the domain of music generation.

In its current version, the SMuSe only relies on internal goals that are modulated by the environment (e.g. the system chooses a sequence of specific musical parameters to optimize the level of arousal perceived from the environment). One interesting addition to the system would be the incorporation of an internal model of emotion that could both influence the current perception of the environment and modulate the internal goals of the system (e.g. if the SMuSe is “depressed”, the musical output is biased towards negative valence and low arousal).

One particularly difficult problem encountered in the course of this thesis related to the extraction of significant external emotional feedback. It is not easy to obtain clear, stable and robust real-time emotional information from physiology that can mirror the subtlety and variety of emotions felt when listening to music. Physiologically-based interfaces have improved greatly over the last years, but it is still difficult to assess the emotional state of a participant beyond simple arousal levels. For musical stimuli, getting robust information on the valence scale from physiology is still a challenge. One option would be to integrate as many emotional cues as possible at the same time, both implicit and explicit (including facial expression, body posture, EDA, HR, EMG, EEG). The analysis of this data will require the use of advanced processing, analysis and data fusion algorithms.

Our experiments on the influence of auditory features on emotional responses showed the important role of subcortical processes on musically-induced emotions. Music theorists and emotion researchers, influenced by conventional music theory, have traditionally focused on more syntactic aspects of musical experience. We do not minimize the importance of these cognitive processes, but emphasize the importance of the complementary subcortical processes that take place when listening to music. An interesting continuation of this work would be to use our synthetic psychology methodology to investigate how cognitive musical processes such as expectation also influence musical experience (for example using computational models of the Implication Realization paradigm).

This thesis adopted an integrated and multi-disciplinary approach to the study of music perception, cognition, emotion and performance. The results of the experimental studies are encouraging and pave the way for a new generation of interactive training and rehabilitation technologies. Based on this preliminary work, we can envision a wide range of music technology applications that will enhance human experience and quality
15. Conclusions and Outlook

of life such as music-based neurofeedback training, anxiety therapy, diagnostic of cognitive and affective deficits, or self expression and diagnostic tools for patients with deficit of consciousness (coma, locked in syndrome) using brain computer interfaces.
Bibliography


Behrens, G., Green, S., 1993. The ability to identify emotional content of solo improvisations performed vocally and on three different instruments. Psychology of music 21 (1), 20.


URL http://portal.acm.org/citation.cfm?doid=380995.380999


15. Conclusions and Outlook


Cohn, R., 1992. Transpositional Combination of Beat-Class Sets in Steve Reich’s Phase-Shifting Music. Perspectives of New Music 30 (2), 146–177.


Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M., Schr 
"oder, M., 2000. ‘feeltrace’: An instrument for recording perceived emotion in real 
time. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. 
Citeseer, pp. 19–24.

The Neuroscientist 8 (2), 132.

Harper Perennial.

Csikszentmihalyi, M., Hunter, J., 2003. Happiness in everyday life: The uses of experi-

Cuthbert, B., Bradley, M., Lang, P., 1996. Probing picture perception: Activation and 

York:.

Davitz, J., 1964. The communication of emotional meaning.


psychophysiology.

de Cheveigné, A., Kawahara, H., 2002. Yin, a fundamental frequency estimator for 


Reviews Neuroscience 7 (3), 242–249.

Delbeuck, X., Van der Linden, M., Collette, F., 2003. Alzheimer’disease as a disconnec-


Downie, J., 2003. Music information retrieval. Annual review of information science and 
technology 37 (1), 295–340.

wave EEG enhances musical performance. Neuroreport 14 (9), 1221.

344.

15. Conclusions and Outlook


Fletcher, H., Munson, W., 1933. Loudness, its definition, measurement and calculation./. acoust. Soc. Amer 5, 82–108.


Grewe, O., Nagel, F., Kopiez, R., Altenm


15. Conclusions and Outlook


Krumhansl, C., Louhivuori, J., Toiviainen, P., J


230


15. Conclusions and Outlook

URL http://www.cs.ucl.ac.uk/research/vr/Projects/PRESENCIA/Public/presenccia_pub/sharedDocuments/presenccia_publications/Publications/app1/21_LeGroux-ICMC08.pdf


URL http://www.cs.ucl.ac.uk/research/vr/Projects/PRESENCIA/Public/presenccia_pub/sharedDocuments/presenccia_publications/Publications/app1/20_LeGroux-ICMC09.pdf


15. Conclusions and Outlook


URL http://mediatheque.ircam.fr/articles/textes/McAdams95a


15. Conclusions and Outlook


Puckette, M., Ucsd, M., Apel, T., et al., 1998. Real-time audio analysis tools for pd and
msp.

Journal of the Audio Engineering Society 45 (6), 456–466.


Rasch, R., Plomp, R., 1999. The perception of musical tones. d. deutsch (ed.). the
psychology of music (pp. 89-112).

Reisenzein, R., 2000. Exploring the strength of association between the components of


versity of York.

Risset, J., Wessel, D., 1982. Exploration of timbre by analysis and synthesis. The psy-
chology of music, 26–58.

11–13.

Robazza, C., Macaluso, C., D’urso, V., 1994. Emotional reactions to music by gender,
age, and expertise. Perceptual and motor skills 79 (2), 939.


with high technology.

Rosenberg, E., Ekman, P., 2005. Coherence between expressive and experiential systems
in emotion. What the face reveals: Basic and applied studies of spontaneous expression
using the Facial Action Coding System (FACS), 63.

Rosenboom, D., 1989. Biofeedback and the arts: Results of early experiments. In: Com-

Rowe, R., 1993. Interactive music systems: machine listening and composing. MIT Press,
Cambridge, MA, USA.

Ruffini, G., Dunne, S., Farres, E., Watts, P. C. P., Mendoza, E., Silva, R., Grau, C.,
Marco-Pallares, J., Fuentemilla, L., Vandecasteele, B., 2006. Enobio - first tests of
a dry electrophysiology electrode using carbon nanotubes. In: IEEE Engineering in

15. Conclusions and Outlook


Sandell, G., 2008. Definitions of the word" timbre.


Simon, H., 1981. The science of the artificial. Cambridge, USA, MA.


15. Conclusions and Outlook


15. Conclusions and Outlook


Index

A
Ableton Live, 210
ADSR, 102
aesthetics emotions, 109
Algorithmic Composition, 8
Alosphere, 162
Allure, 69
alpha, 176
Aristotle, 107
articulation, 112, 168, 197
attack time, 31
Auditory Icons, 161
Aulay-Quatieri, 71

B
BCI, 203, 206
beta, 176
Biosignals, 19
Brain Opera, 18
Brain Orchestra, 203
Brain stem reflex, 113
brain-computer interfaces, 203
brainwave, 177, 179
brainwaves, 20
brightness, 31, 83, 93, 168
butterworth, 180

C
Cannon-Bard, 108
centralized control, 15
centroid, 31
chroma, 28
classification, 80
cognition, 1
cognitivist, 110
concatenative synthesis, 72
connectionist, 72
control, 13
Creativity, 17
cross-synthesis, 92
Curvasom, 22, 23

d
DAC, 22, 23, 171
Damping, 101
damping, 102
delay, 44
delta, 176
Descartes, 108
Distributed Adaptive Control, 22, 171
duration, 27, 30, 67
dynamics, 197

E
echoic memory, 27, 32
EEG, 176, 203, 204
electroencephalogram, 203
embodied cognitive science, 15
embodiment, 7
emergence, 7, 25
emergent phenomena, 15
EMI, 9
emotion, 2, 169
Emotional contagion, 113
Emotional Feedback, 21
emotional feedback, 7, 25
emotional space, 115
emotions, 107
emotivists, 110
EOG, 178
Episodic memory, 114
ERP, 206

243
Index

Evaluate conditioning, 113
event fusion, 26
event related potential, 206
expressivity, 32
Externalism, 7

F
feature extraction, 27
feedback, 13
FFT, 207
Flow, 17
FM, 70
FM synthesis, 169
form level, 26
forward engineering, 22
Fourier, 71
free description, 114

G
gestures, 112
Giambattista Vico, 22
granular synthesis, 70
G.tec, 206

H
harmonic, 75
harmonicity, 100
harmony, 34, 110
Healthcare, 105
heart rate, 19
Human in the Loop, 17
Hume, 108
Huron, 113

I
IanniX, 18
IAPS, 115
impact sounds, 72
In C, 195
inharmonicity, 168
Interactive Music, 9
internal representations, 15
interpolation, 76
IQR, 22, 24, 166
ITPRA, 113
IXI, 18

J
James-Lange theory, 108

K
Kansei, 105
Karplus-Strong, 70
Kontakt, 210

L
LDA, 207
learning, 80
Ligeti, 211
log-attack time, 67
long-term memory, 35, 45
loudness, 27, 29, 67, 83
LTM, 45

M
machine learning, 31
Markov chain, 9, 36
mass, 68
mass profile, 69
MATLAB, 206
Matlab, 206
matter, 68
Max-MSP, 122
MBO, 204, 206, 212
melodic grouping, 33
melodic profile, 69
melody, 34, 110
MIDI, 24, 210
MIR, 66
mixed-reality, 162
Modal Synthesis, 97
mode, 168
Modes of Interaction, 12
multidimensional scaling, 67
Multimodal Brain Orchestra, 203
Musak, 21, 107
Music Information Retrieval, 66
Music of Changes, 16
Musical expectancies, 114

244
Index

N
neural network, 72
neurofeedback, 177, 212
neuromorphic, 162, 166, 171

O
Open Sound Control, 204
OSC, 194, 204, 206, 208

P
P300, 206, 208
parallel computation., 15
parallelism, 7, 25
PCA, 75, 77, 84, 87
perception, 1
perceptual binding, 27
performance, 1
phase vocoder, 71
Physical models, 71
physiology, 20, 108
pitch, 27, 28, 67
pitch shifting, 92
Plato, 107
Playing Modes, 12
Principal Component Analysis, 77
Psychoacoustics, 67
psychoacoustics, 25, 31
pulse, 33

R
Reactable, 18
register, 168
regression, 80
reinforcement learning, 193
rhythm, 33, 110, 168
rhythmic grouping, 26, 33
Roboser, 22, 23

S
SAM, 115, 122
scalability, 92
Schacter-Singer, 108
Schaeffer, 68, 69
score, 25
SDIF, 82
selection principle, 196
Selection principles, 45
Self Assessment Manikin, 115
disreport, 114
Sense-Think-Act, 14
sense-think-act, 15
sequencer, 8
Sequencing, 8
shape, 69
short-term memory, 35
Simulink, 206
Situated Cognition, 5
Situated Music Server, 2
situated robotics, 7, 24
situatedness, 7, 25
SMS, 71, 72, 77, 82
SMuSe, 2, 24, 25, 164, 168, 204
Sonification, 161
sound level, 112
Spatialization, 169
spectral centroid, 67, 99
spectral flux, 31, 67
Spectral Modeling Synthesis, 71
Spinoza, 108
SSVEP, 206, 207, 210
STM, 45
String Quartet, 211
subharmonics, 25
supervised, 80
Support Vector Machine, 87
Support Vector Regression, 87
SVM, 87, 89
SVR, 92

tempo, 33, 44, 112, 168, 197
tension, 196
The Hands, 18
The Very Nervous System, 18
therapy, 1
theta, 176
Timbre, 30, 112
timbre, 27, 31, 67
timbre space, 67

245
Index

Time Envelope, 102
time warping, 92
tristimulus, 99

U
Unsupervised, 80
utilitarian emotions, 109

V
variation criteria, 69
VBAP, 169
vibrato, 112
Virtual Environments, 161
Visual imagery, 113
volume, 168

W
Wii, 204

X
XIM, 49, 60, 162, 166, 168

Y
Yin, 82