Exploring Music through Sound: Sonification of Emotion, Gesture, and Corpora

R. Michael Winters IV

Schulich School of Music
Department of Music Research
McGill University, Montréal

November 2013

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Arts in Music Technology

2013/11/25

Abstract

Contemporary music research is a data-rich domain, integrating a diversity of approaches to data collection, analysis, and display. Though the idea of using sound to perceive scientific information is not new, using it as a tool to study music is a special case, unfortunately lacking proper development. To explore this prospect, this thesis examines sonification of three types of data endemic to music research: emotion, gesture, and corpora. Emotion is a type of data most closely associated with the emergent field of affective computing, though its study in music began much earlier. Gesture is studied quantitatively using motion capture systems designed to accurately record the movements of musicians or dancers in performance. Corpora designates large databases of music itself, constituting for instance, the collection of string quartets by Beethoven, or an individual's music library. Though the motivations for using sonification differ in each case, as this thesis makes clear, added benefits arise from the shared medium of sound. In the case of emotion, sonification first benefits from the robust literature on the structural and acoustic determinants of musical emotion and the new computational tools designed to recognize it. Sonification finds application by offering systematic and theoretically informed mappings, capable of accurately instantiating computational models, and abstracting the emotional elicitors of sound from a specific musical context. In gesture, sound can be used to represent a performer's expressive movements in the same medium as the performed music, making relevant visual cues accessible through simultaneous auditory display. A specially designed tool is evaluated for its ability to meet goals of sonification and expressive movement analysis more generally. In the final case, sonification is applied to the analysis of corpora. Playing through Bach's chorales, Beethoven's string quartets or Monteverdi's madrigals at high speeds (up to 10^4 notes/second) yields characteristically different sounds, and can be applied as a technique for analysis of pitch-transcription algorithms.

Résumé

La recherche actuelle en musique implique la collecte, l'analyse et l'affichage d'un large volume de données, abordées selon différentes approches. Bien que l'idée d'utiliser le son afin de percevoir l'information scientifique ait déjà été explorée, l'utilisation du son comme outil d'étude de la musique constitue un cas particulier encore malheureusement sous-développé. Afin d'explorer cette perspective, trois types de données endémiques en recherche musicale sont examinées dans ce mémoire : émotion, geste et corpus. L'émotion en tant que type de données se retrouve le plus fréquemment au sein du domaine émergent de l'informatique affective, même si la notion fut abordée en musique bien plus tôt. Le geste est étudié de façon quantitative à l'aide de systèmes de capture de mouvement conçus pour enregistrer précisément les mouvements de musiciens ou danseurs lors d'interprétations et performances. Le corpus désigne ici les vastes bases de données sur la musique elle-même que constituent, par exemple, le recueil des quatuors à cordes de Beethoven, ou une collection musicale personnelle. Bien que les motivations pour la sonification diffèrent entre ces trois cas, comme clairement illustré dans ce mémoire, leur relation commune au medium sonore peut engendrer des avantages supplémentaires. Dans le cas de l'émotion, la sonification peut tout d'abord s'appuyer sur les connaissances établie concernant les déterminants acoustiques et structurels de l'émotion musicale, ainsi que sur les nouveaux outils informatiques conçus pour leur identification. La sonification trouve alors son utilité dans les configurations systématiques et théoriquement justifiées qu'elle peut proposer pour précisément instancier un modèle informatique (appliquer un modèle informatique à l'objet d'étude) et extraire d'un contexte musical spécifique les vecteurs d'émotion du son. Pour le geste, le son peut servir à représenter les mouvements expressifs de l'interprète au sein du même medium que la musique interprétée, offrant ainsi un accès auditif simultané correspondant aux indices visuels pertinents. La capacité d'un outil logiciel spécialement conçu à atteindre des objectifs de sonification et d'analyse du mouvement expressif au sens large est évaluée. Enfin, la sonification est appliquée à l'analyse de corpus. La lecture à très haute vitesse (de l'ordre de 10^4 notes par seconde) des chorales de Bach, des quatuors à cordes de Beethoven ou des madrigaux de Monteverdi induit des sons différents et caractéristiques. Cette technique peut être employée pour l'analyse d'algorithmes de transcription de hauteurs.

Acknowledgments

I would like to thank my advisor Marcelo Wanderley for supporting my research and leading this thesis to completion. I am tremendously appreciative of the opportunity to expand the current IDMIL initiative on gesture sonification to include both emotion and music. Darryl Cameron has been helpful in making use of computational resources in the music technology department and overcoming various technical issues in my research. I should also thank members of the IDMIL for offering help and guidance: Dr. Marc Zadel for his help with SuperCollider, Joe Malloch for his help with Max/MSP, and Dr. Steve Sinclair for various programming issues and ideas about the sonification of music. Through lunch time discussions and soccer games, Dr. Bertrand Scherrer of the SPCL has also been helpful in the process. The excellent French in the abstract is due to the kind labours of Dr. Michel Bernays, post-doctoral researcher in the SPCL and IDMIL. Cryptographer Saining Li has been instrumental in the research, helping in the creation of python scripts in the sonification of music experiment, creating Figure 2.3, and in general being a delightful partner and companion in my development over the course of my masters thesis.

Ian Hattwick was my primary collaborator in the early stages of my research in emotion. Some of his ideas and corrections are present in the typology presented in Chapter 2 (Winters, Hattwick, & Wanderley, 2013), and our weekly meetings in the summer of 2012 in no doubt informed my thinking about sonification strategies for emotion in the fall. Dr. Stephen McAdams was also very helpful, directing me to two articles on emotion that became seminal to the present work (Juslin & Västfjäll, 2008; Scherer, 2004). He also permitted me to experiment with the sonification of some of his collected data on musical emotion in 2011, my first ever experience of working with the data-type. Alexandre Savard was another collaborator in the current research, offering through his Sonification Toolbox (Savard, 2009) a context for framing issues in the sonification of expressive gesture. I am tremendously grateful for the time he put in towards updating the Sonification Toolbox in the Fall of 2011, leading ultimately to our shared publication in December 2012 (Winters, Savard, Verfaille, & Wanderley, 2012). The ideas and friendship of Dr. Florian Grond have been influential on in the work on gesture, listening aesthetics, and sonification in general.

I am also very grateful to my friends and classmates that participated in the user test presented in Chapter 7. Along with them, I would specifically like to mention my colleagues Greg Burlet, Hannah Matthews, Gautum Bhattacharya, Chuck Bronson, Aaron Krajeski,

Yinan Cao, Ying Qin, Ye Lu, and Indiana Wollman, friends and companions in the thesis process. DDMAL members Alastair Porter and Greg Burlet proofread an early edition of Chapter 7 and I am thankful for their suggestions. Sound Engineer Scott Levine, and other members of the sound recording department have been formative in orienting thoughts on sonification and its use as music technology.

I would finally like to thank friends and family. The sculptor Kevin Reiswig, songwriter David Yontz, and cermaicist Claire Miller have been incredibly supportive and influential people in the past few years. Their creativity and accomplishments have been an inspirational force to me in this thesis writing process. I would also like to thank my family for their support throughout the process, helping me move to Montréal and supporting my piano lessons.

Preface

Data sonification is a process whereby numbers are turned into sound so that they can be understood by listening. There are many reasons why a person might want to sonify data, not the least being artistic and musical—giving voice to inaudible fluctuations in the world around us. When I first began doing research in sonification as an undergraduate physicist at the College of Wooster, my goal was to do just that: turn three-dimensional chaotic trajectories into something musically meaningful (Winters, 2009). However, through time, I became aware that sonification was much more than a musical endeavour, that it could be applied towards representing processes that would otherwise be completely in-experiencable (Winters, Blaikie, & O'Neil, 2011), and perhaps unlock mysteries of listening aesthetics (Winters, 2011a), all the while offering a fundamentally different tool for scientific data analysis (Winters, 2010).

As a pianist and musician, the experience of music and sound has been something I have held in great esteem. To this inspiration, I have brought previous experience in physics research advanced through a liberal arts education and independent study (Funk, O'Neil, & Winters, 2012; Lehman et al., 2012). Upon coming to McGill, I confess that despite a degree in music, I had very little knowledge of contemporary music research, much less the applied field of music technology. Since then I have been inspired by ideas and research I had never before dreamed as possible, the experience of music and sound always at the core.

In this thesis, I have tried to capture some of the potential of sound as a tool and technology for music research. It would appear that though sound can be used to represent any kind of data whatsoever, when it is used with music, domain specific benefits arise, mostly due to the shared medium of sound. This benefit is most pronounced with emotion, where the design of sonification itself benefits from the robust literature on musical emotion that has already been established. A sonification can be made to "sound-like" an emotion, advance theories of musical emotion, and be applied for emotional communication. In gesture, sound can be made to represent the expressive movements of musicians, adding an additional sonic layer to the music originating from the musician's body. For corpora, where sound is used to explore huge databases of music, I think that sonification perhaps has the greatest potential, though at present, only marginal results, such as the fact that Bach, Beethoven and Monteverdi sound different at high speeds, can be offered.

Contribution of Authors

In all of these manuscripts presented, my advisor Marcelo Wanderley is listed as co-author. Besides for providing guidance, directing the research, and challenging me to answer the really important questions in each project, he has proofread all of the manuscripts, at times leading to important revisions. The contributions of other collaborators (i.e. Ian Hattwick, Alexandre Savard, and Vincent Verfaille) have been addressed in-text where appropriate. Ian contributed towards Chapter 2 by creating a vocal effects processor for musical performance, which as featured in the text, offers a complementary approach to affective music generation that was motivational for the typology presented in Section 2.2. In preparation for publication, Ian also contributed to written material presented in Sections 2.1 and 2.3. Alexandre Savard made the Sonification Tool that is presented in Section 6.3, and similarly to Ian, wrote the material for that section, which is presented here as proofread and organized by me prior to publication. Vincent Verfaille's ideas and research were foundational to the present work on gesture and is listed as co-author on the corresponding publication (Winters et al., 2012).

Contents

List of Figures

List of Tables

List of Acronyms

AMG	Affective Music Generation
AV	Arousal/Valence
BVP	Blood Volume Pulse
CIRMMT	Center for Interdisciplinary Research in Music Media and Technology
CSEMP	Computer System for Expressive Music Performance
DDMAL	Distributed Digital Music Archives and Libraries Lab
EEG	Electroencephalography
EKG	Electrocardiograph
EMG	Electromyography
EIC	Emotional Imaging Composer
EII	Emotional Imaging Incorporated
GSR	Galvanic Skin Response
GUI	Graphical User Interface
ICAD	International Community for Auditory Display
IDMIL	Input Devices and Music Interaction Laboratory
MER	Music Emotion Recognition
MIR	Music Information Retrieval
PCA	Principal Component Analysis
SID	Sonic Interaction Design
SPCL	Sound Processing and Control Laboratory

Chapter 1

Introduction

1.1 Context & Motivation

Sonification is an interdisciplinary field of research broadly interested in the use of sound to convey information (Kramer et al., 1999). A simple example is the geiger counter, a device that emits audible clicks that vary in frequency, representing the amount of radiation in the nearby environment. Although sonification has been applied to a diversity of tasks and is becoming increasingly established as a scientific practice (Hermann, Hunt, & Neuhoff, 2011), its application to music research is new and undeveloped.

There are, nonetheless, several reasons why sonification might be expected to be useful. Perhaps the most obvious reason is that contemporary music research is a data-rich domain, one in which data often unfolds temporally and patterns emerge through the analysis of multiple parallel data streams. An experiment interested in the bio-physiological or emotional impact of music for instance may continuously record EEG, EMG, heart-rate and skin conductance measurements while a listener or performer engages with music. On the other hand, experiments interested in expressive or functional movements of musicians may use high-resolution motion capture systems to record the three-dimensional position of points on the body during performance. Yet another source of data are the increasingly large databases of music itself, most often encountered in the field of music information retrieval (MIR). Such a database might constitute an individual's music library or perhaps the entire corpus of string quartets by Beethoven. As with other sonification contexts, the faculties of the auditory system can be directed towards perceiving this data, identifying complex temporal patterns in parallel data streams, exploring large databases, and garnering information that may not be immediately obvious through visual techniques (Walker & Nees, 2011).

Unexpected and unique benefits arise from the shared medium of sound. In the case of gesture for instance, sonification can be used to help to identify subtle, non-obvious changes in a performer's movement, but when listening in parallel to the performance audio, the musical meaning of a movement is better understood. Although the same data can be displayed bimodaly through a dynamic visualization, audition renders the data in the same medium as the music itself, a difference in representation that brings characteristically different insights. In the case of large databases of music as well, sound can be used to explore the corpora under investigation, helping the user to garner characteristic information. In this special case, however, the database under investigation is composed of sound, and through sonification, data and data representation can at times be coupled, again bringing different insights to analysis. Collectively, gesture and corpora are alike in that sonification is applied as a technique for communicating and perceiving data. Though visualization also has many benefits and is certainly more widespread, for a field comprising of musicians and music researchers, listening to garner information and meaning is already a well founded, definitive practice.

This benefit of shared medium can be applied in a different direction when sonification is used to instantiate specific structural or acoustic cues that lead to an emotional response in music. For over three-quarters of a century, research has been directed towards determining the structural and acoustic elicitors responsible for musical emotion (Gabrielsson & Lindström, 2010). Although musical emotion is a multifaceted cognitive experience in which culture and learning are fundamental, this branch has been directed towards underlying acoustic details. Sonification finds application here, where it can be used to provide systematic and theoretically informed manipulation of these cues, which according to Juslin and Västfjäll (2008, p. 574), would be a "significant advance" to stimuli selection. Indeed, by choosing wisely, the emotional effect of sound can be isolated from a musical context and applied towards emotional communication. Thus, in addition to being used for data analysis and display, sonification provides a tool for perceptual investigation, furthering the understanding of a complex phenomenon: emotional response to music.

This thesis introduces sonification as a tool for music research, targeting three areas of application: emotion, gesture, and corpora. The utility of sonification is unique in each case, but is always connected to the shared medium of sound. In the first, research originating in the study of musical emotion is applied to sonification of emotion for the purpose of communication and display, as might be useful in affective computing. By evaluating these sonifications using a tool for music emotion recognition, the benefits and limitations of these computational tools are addressed, as well of the capacity of sonification to accurately instantiate such a model. In the second application, sound is used to represent the movements made by musicians while performing. Unlike movements made for the purposes of note production, this study focuses on *expressive* movements, considerably more varied, and requiring flexible tools for analysis. One such sonification tool, first introduced in Savard (2009), is evaluated for its ability to meet goals of sonification and expressive movement analysis. In the last application, sound is studied as a potentially valuable tool for exploration of large databases of symbolic music. In these databases (i.e. Beethoven's String Quartets, Bach's Chorales, Monteverdi's Madrigals), sonification can be applied as a tool for pitch-based error analysis and differentiation of corpora, even at speeds of 10,000 notes per second. For each tool, approach, and application, evaluation is presented as a pivotal element for the discussion.

1.2 Thesis Structure & Connecting Text

The thesis is separated into three parts, representing emotion, gesture, and corpora respectively. Further, each chapter represents a publication that has either been accepted, submitted, or is in preparation for submission. The work on emotion in Part I marks the most substantial contribution, marked by Chapters 2-4. Together, Part II and III provide three more chapters: Chapters 5 and 6 being dedicated to gesture, and Chapter 7 dedicated to corpora. As discussed in Section 1.1, these three topics can also be characterized by the benefit of sonification for each case. In emotion, benefit comes through providing systematic and theoretically informed mappings capable of being used in perceptual and theoretical work. In gesture and corpora, benefits are directed towards the utility of sonification for data analysis and display with the added benefit of shared medium. After presenting these six chapters, Chapter 8 draws out conclusions and future work, discussing the outlook for sonification in music research more generally.

1.2.1 Emotion

Chapter 2 (Winters et al., 2013) begins Part I by describing two approaches to using real-time arousal and valence coordinates in musical performance, one of which is a sonification, the other, a vocal effects processor. A typology is presented to differentiate these two systems and organize systems for affective music generation more generally. The typology abstracts the technology used for emotional input from the characteristics of the generation algorithm, identifying relevant design criteria in each. The sonification mapping is presented in detail, along with a specially designed graphical user interface for multimodal emotion data analysis. In addition to the sonification, the interface includes a realtime arousal/valence visualizer, video player, and an interface for testing and altering sonification mappings. The vocal effects processor is presented as well, and the typology is applied, noting the fundamental difference in output schema in each algorithm. In the sonification, all content is generated from the system, whereas in the vocal effects processor, content is derived from the performer's voice. The most convincing application of sonification is provided by watching the performance while listening simultaneously to her corresponding AV coordinates. Listening to the sonification communicated emotional content not apparent from her visual cues, modulating and amplifying the perception of her emotional state.

Although the sonification was based upon principles from the study of musical emotion and was received positively in public demonstrations, there had not yet been formal investigation into the best strategies for sonification of emotion, and where sonification of emotion would likely find real-world application. Thus, in Chapter 3 (Winters & Wanderley, 2013), a formal literature review is presented discussing sources for auditory emotion induction and applications in affective computing. Acknowledging the dominance of visual and verbal social displays of affect, applications are targeted in which visual or verbal displays are *unavailable*, *misleading*, or *inappropriate*. For these contexts, the use of an underlying arousal/valence space in a *peripheral* auditory display are cited as the most advantageous framework for sonification development. For the purpose of mapping, environmental sound and music are presented as two potential sources for emotion elicitation, each with available cues and determinants. Following a comprehensive review of emotion induced by environmental sounds, reasons are provided why musical emotion provides a more robust framework for future development. However, instead of haphazardly choosing from the available structural and acoustic cues, the paper argues that it is better to first choose mechanisms for emotion induction based upon desirable psychological characteristics. These mechanisms can then be used to determine which emotional cues are desirable for implementation in sonification. To this end, the mechanisms of 'brain stem reflex' and 'emotional contagion' are chosen for the desired psychological properties of low cultural impact/learning, high induction speed, low degree of volitional influence, and finally, their dependence on musical structure. Cues are then presented that trigger these mechanisms, and the sonification design presented in Chapter 2 is evaluated based upon these principles. It is noted that the use of major/minor mode, while not accounted for by either of the two mechanisms, was a very strong communicator of valence.

Following the discussion of mapping strategies and applications that formed the basis of Chapter 3, Chapter 4 (Winters & Wanderley, 2014) introduces the use of computational tools for evaluation and design. The application of such tools, originating in the field of music emotion recognition (MER), draws further attention to the ways that sonification of emotion can benefit from music research, and consequently, how sonification can advance or challenge computational approaches to research on musical emotion. To this end, a second sonification mapping strategy was created that would accurately cover the activity/valence space prescribed by the MIREmotion function (Eerola, Lartillot, & Toiviainen, 2009) using a minimal number of acoustic cues. This "computational" design was then compared and contrasted with the "ecological" mapping strategy presented in Chapter 2 (Winters et al., 2013). Though the computational design performed many times better computationally, the performance of the ecological design was not random and weakly preserved the desired AV space though offset towards higher V and A for every point. Aside from these computational results, both models were considered for their utility for emotional communication and display. In spite of the disparity in computational performance, the ecological design was argued to be more useful in emotional communication due to the greater number of structural and acoustic cues used for display, and more 'naturalistic' synthesis of the fundamental sound. Instead of discounting computational evaluation altogether, these results clarify and address certain computational limitations, which if accounted for can improve mapping, while still maintaining computational accuracy. In spite of these issues, the benefits of computational design and evaluation strongly support their application to sonification of emotion in future research, largely due to the reciprocal relationship with musical emotion.

1.2.2 Gesture

Part II extends the discussion of sonification in music research to movement analysis, specifically the "expressive" or "ancillary" gestures made by musicians while performing. Chapter 5 functions in a similar vein to Chapter 3 by providing a review of the literature, strategies for design and evaluation, and potential applications. The important difference between the two chapters would be that at the time of the publication (Winters & Wanderley, 2012b), the topic of gesture sonification had developed considerably in its own right (Verfaille, Quek, & Wanderley, 2006), while the topic of emotion sonification was still not well defined (Schubert, Ferguson, Farrar, & McPherson, 2011). In this light Chapter 6 has a similar function to Chapter 4, providing a thorough background and motivation to the subject, a specific approaches to sonification, and a large section dedicated to discussion and evaluation.

Chapter 5 (Winters & Wanderley, 2012b) begins Part II by presenting new criteria for design and evaluation of sonifications based upon a review of relevant literature (Winters, 2011c). The new design strategies focus upon conveying higher-level features rather than low-level marker positions and angles. A good sonification for instance, should be able to differentiate gestures characterized by the instrument, the music, and the performer's interpretation. It should also convey structural and emotive cues that indicate emotion or expression in performance, such as amplitude of motion, regularity, fluency, and speed. By taking this higher level approach, the paper argues, relevant visual information is transformed into the auditory domain, enabling a fuller acoustic display of expression than the music alone, one that is potentially closer to the performer's internal representation of the piece. To conclude, it is argued that sonification of these higher-level features provides the best possible avenue for simultaneous auditory display of music and sonification, an application that can make visual performance accessible to the blind or those that cannot see.

Building on the criteria presented in Chapter 5, Chapter 6 applies it to evaluate a tool for expressive movement analysis (Savard, 2009). The chapter begins by first distinguishing expressive movements from "effective"/"goal-oriented" movements (as found in sports). Expressive movements are considerably more varied, and though researchers may use sonification for similar reasons in data analysis, a tool for expressive movement analysis must be more flexible, capable of quickly adjusting itself to new performers, comparing across performers, and conveying the higher-level cues presented in Chapter 5. A background section presents developments since the foundational paper of (Verfaille, Quek, & Wanderley, 2006) and a defence of Principle Component Analysis (PCA) for data preprocessing. The tool is then presented with all of its current functionalities, including the GUI, the data preprocessing and synthesis options. Following this presentation, the tool is then evaluated based upon identified goals of sonification in movement analysis, and the goals of expressive movement analysis more generally. The discussion highlights the benefit of 10 simultaneous synthesis "channels," interactive mapping, integration with performance video, and the data preprocessing options useful to expressive gesture (e.g. body curvature, circular motion, velocity, PCA). The chapter concludes with a fuller discussion on the simultaneous auditory display of music and sonification begun in Winters (2011c) and Winters and Wanderley (2012b), citing areas in *effective* gesture where the benefits of shared medium have already been identified. These discussions are extended to the domain expressive gesture, where as in Chapter 5, it is argued that expressive gesture sonification should be explored as a means to making visual performance accessible to the blind or those that cannot see. The paper concludes by drawing distinctions between sonification of expressive gesture and the mapping of gesture in music performance.

1.2.3 Corpora

Part III presents the third application to music research, sonification of corpora—garnering information about large databases of music through sound. Part III is unlike the others in that it is comparatively brief, condensing all discussion of mapping strategies, applications, background, implementation and evaluation into a single chapter. It is also unlike the others in that it features results from a user test in which participants performed analysis with a specific sonification technique.

Chapter 7 (Winters & Wanderley, 2012a) begins by discussing the current state of sonification in MIR, bringing attention to the fact that it is often used tacitly to display final results rather than as an integrated research tool. Some researchers have begun working towards displaying audio and audio features using sound, but there has not yet been application towards symbolic music—a score-based music representation frequently found in MIR. Such an application is proposed, namely analysis of transcription errors in pitch transcription algorithms. In such a case, high-speed, pitch-based sonification can be used to quickly compare ground truth and transcription by representing each pitch as a short audible sinusoid, and playing through all pitches, each version in a separate stereo channel. When the two pitches diverge, the central location of the auditory stream splits into left and right stereo channels. Two enhance this effect and further distinguish errors, these divergences were made slightly louder than the other notes. When tested on a group of 11 participants in a sorting task, results suggested the technique could be accurately used across three speeds of presentation (10^2, 10^3, and 10^4 notes per second) and three corpora (Monteverdi's Madrigals, Bach's Chorales, and Beethoven's String Quartets).

Although the technique was successful in rapidly conveying the quantity of transcription errors, it is not clear why sonification would be considered useful given a host of other data analysis tools, many of which could perform the same task faster and provide a more accurate numeric analysis. The chapter therefore draws attention to other benefits of listening that were not addressed in the present study. For instance, the dynamic presentation offered by rapid pitch-based mapping allows one to identify relatively short events by their temporal location and pitch. The technique can also be applied towards corpora differentiation: each corpora could be quickly identified at all speeds due to their characteristic sound. Though the reason for this is not known, insights such as these indicate that more information than just transcription errors is afforded through the auditory representation. Furthermore, these insights can be unsuspected, leading to new directions in research.

Part I

Sonification of Emotion

Chapter 2

A Sonification System for Realtime Emotion Display

Winters, R. M., Hattwick, I., & Wanderley, M. M. (2013, June). Integrating emotional data into music performance: Two audio environments for the emotional imaging composer. In *Proceedings of the 3rd international conference on music and emotion.* Jyväskylä, Finland.

Abstract

Technologies capable of automatically sensing and recognizing emotion are becoming increasingly prevalent in performance and compositional practice. Though these technologies are complex and diverse, we present a typology that draws on similarities with computational systems for expressive music performance. This typology provides a framework to present results from the development of two audio environments for the Emotional Imaging Composer, a commercial product for realtime arousal/valence recognition that uses signals from the autonomic nervous system. In the first environment, a spectral delay processor for live vocal performance uses the performer's emotional state to interpolate between subspaces of the arousal/valence plane. For the second, a sonification mapping communicates continuous arousal and valence measurements using tempo, loudness, decay, mode, and roughness. Both were informed by empirical research on musical emotion, though differences in desired output schemas manifested in different mapping strategies.

2.1 Introduction

Ian Hattwick contributed original material to Section 2.1 that was subsequently edited and reorganized by R. Michael Winters.

Emotions form an important part of traditional music performance and expression. It is therefore not surprising that new technologies designed to sense emotion are finding their way into performance practice. Facial expression, physical gesture, and bio-physiological process provide just a sampling of the data streams available. A special class of algorithm abstracts from this information an actual emotion, making the emotion itself (rather than low-level data features) a driving force in the performance.

In this paper, two audio environments are presented that use a performer's emotional state to control audio processing and synthesis. Using a collection of physiological markers representing relevant biological processes, an algorithm outputs continuous arousal and valence coordinates representing the performer's emotional state at each instance of performance. In the first audio environment, these two coordinates drive a sonification model to accurately communicate the emotional information. In the second, the two coordinates control an algorithm for realtime audio processing of the musician's performance.

The audio environments were developed in collaboration with Emotional Imaging Incorporated (EII), a company specializing in media products infused with technologies for emotion recognition. In the current project, development was directed towards the Emotional Imaging Composer (EIC), described as "a multimedia instrument that translates biosignals into [emotionally] responsive environments in realtime."[1] Previously, the responsive environment had taken the form of an abstract, fluid computer visualization. For the present research, a platform for responsive audio was designed.

Our systems are framed in the context of interactive affective music generation. Given the numerous systems that have thus far been implemented, we introduce our system through analogy to a typology introduced for computer systems for expressive music performance (CSEMP) (Kirke & Miranda, 2013a). The typology abstracts the algorithm for music generation from the tool for realtime interaction. For our purposes, the tools themselves are then distinguished by the degree to which the high-level emotional data stream is the driving force of performance, and how easily the tool can be controlled.

[1] Emotional Imaging Composer [Online]: `http://www.emotionalimaging.com/products.html`

2.2 A Typology of Affective Music Generation Systems

Affective Music Generation (AMG) encompasses a diversity of computational practices directed towards communicating or expressing affect through music. New technologies enable realtime data streams to guide the algorithm and consequently the emotional progression of the piece.

Closely related to affective music generation are so called computer systems for expressive music performance (CSEMP) (Kirke & Miranda, 2013b, p. 2). The goal of these systems is to create expressive performances or compositions, which are in some way more realistic or humanistic than the more "robotic" performances that might otherwise characterize computer generated music. For CSEMPs, it is not uncommon to design a system to compose music to match a particular emotion or "mood," though this feature is certainly not dominant (Kirke & Miranda, 2013b, Table 1.1).

A point of distinction is evident, namely that musical expression is not necessarily synonymous with musical emotion. Having music express an emotion might contribute to its expressivity more generally (Juslin, 2003), but a performance might be expressive without having the direct goal of conveying an emotion to its audience (Davies, 1994). It is also the case that non-speech sound can communicate an emotion without being in any way musically expressive. The emotional space occupied by environmental sounds is a strong example (Bradley & Lang, 2000), and continuous auditory display of arousal and valence variation is another (Winters & Wanderley, 2013).

2.2.1 Systems for Algorithmic Generation

Nevertheless, computer systems for expressive music performance and affective music generation share common questions for design and implementation. The first question concerns content generation, including the type of input, the algorithm itself, and the sound output. With regards to input, a CSEMP has been classified as either "automatic" or "semi-automatic" depending on whether it accepts realtime input (Kirke & Miranda, 2013b). This distinction also applies to AMGs, but of equal or more importance is the type of emotional data driving the algorithm. This data might take the form of a high-level emotional model, which might be discrete or dimensional, or can be mapped from control data output if using a technology for realtime input.

Also similar to CSEMPs, a commonly used strategy in affective music generation is to translate empirically derived results from psychological studies into defined rules for the generation algorithm. However, the desired output schema closely guides this translation. An output schema might include manipulation of symbolic music or audio recordings, realtime sound synthesis/processing, or other techniques for content generation, but for the purposes of AMG, output schema is characterized by the degree to which the emotional data is responsible for content generation. A system that requires input of another type, whether it be symbolic music, audio recordings or live audio input, has less influence over content generation than a system in which sound or music comes directly from the algorithm. In the latter case, the system determines all content, in the former, a portion of the content has been generated independently from the system. Categorizing output in this way abstracts the AMG from a performance context, where a system might as a whole be relegated to a more or less prominent role depending upon aesthetic choices of the musicians involved.

The algorithm is the third part of the AMG that needs to be considered, but in principle sits between the input data and the output schema. Its importance is evident from the fact that it is possible, given the same input data and output schema, to have remarkably different acoustic results. In order to generate affective music, the algorithm must implement acoustic, structural, or performative features to express or communicate the desired emotion. It is natural to direct these choices from the large literature on features that convey or induce musical emotion, but their implementation will change depending upon choices made by the system designer. The designer might favor certain features over others, or include features that do not directly contribute to emotional communication or expression. By including a graphical user interface, mapping decisions might be provided to the user, contributing to flexibility and usability without changing the input data or fundamental output schema.

2.2.2 Technologies for Realtime Emotional Data in Music Performance

However, the question of algorithm for music generation only addresses part of the overall aesthetic of a performance. As with CSEMPs, one must additionally consider the possible technologies for realtime interactive control (Fabiani, Friberg, & Bresin, 2013). These technologies can be assimilated into a music performance, adding a “performer” or “performers” that in some way determine the emotional input data. For AMG, these technologies can be classified by the degree to which emotion is recognized and the amount of control provided to the user.

For this typology, a technology is capable of “emotion recognition” if it generates realtime emotional coordinates from an auxiliary data stream (e.g. biosignals, motion sensors). The output model might be discrete or dimensional, but in either case, the technology in some way “recognizes” an emotion from low-level control data input. In the context of CSEMP, these realtime emotional coordinates provide high-level, “semiotic” control (Fabiani et al., 2013).

By contrast to technologies for emotion recognition, this typology adopts the term “emotion sensing” to describe technologies used in AMG that do not include an algorithmic model for extracting emotional coordinates. Instead, data features from the input device (e.g. biosignals, motion sensors) are mapped directly to the generation algorithm. These data features may correlate with emotions—for instance, amount of motion correlating with arousal in a motion capture system—but the translation from these signals to an emotion-space is lacking. One could map input from a gestural controller (Miranda & Wanderley, 2006) to a set of emotionally salient parameters (e.g. tempo, loudness, etc.) and express an emotion like sadness (Bresin & Friberg, 2011), but if the output of the controller is mapped directly into the acoustic feature space, side-passing an emotion-model, it is classified in this typology as emotion sensing. Only if the gesture itself is first classified as an emotion (e.g. sadness, or AV coordinates) does it become a technology for emotion recognition.

The issue of emotion sensing versus recognition should be separated from a parallel consideration: the degree to which a user can directly control the input to the AMG. For example, the computer mouse has a high degree of control, and might be applied to realtime movement through an arousal-valence space. By moving this way, a performer can directly control the emotional input to the system, and the mouse would qualifying as a tool for emotion recognition. The term “recognition” suffices to distinguish it from the possible direct control of emotionally salient low-level parameters such as tempo and loudness. In that case, the mouse no longer outputs arousal and valence coordinates, and is thus classified as a tool for emotion sensing.

Other systems provide less control to a user. In the present case, physiological measures such as galvanic skin response, heart rate, and phalange temperature are the input to the system. These inputs are relatively more difficult to control than the computer mouse, but still might be applied to emotion sensing or recognition. Presently, realtime arousal and valence are derived from the measures, and used to drive the generation algorithm. In other cases, low-level data features (e.g. heart rate, temperature) might pass directly to sound generation parameters without being recognized as an emotion.

It is important to note that for interactive affective music generation, a high-degree of control is not always desirable. Technologies that are difficult to control (such as biosignals) allow less room for mediation, and might be considered to provide more "genuine" emotional data stream as input. A high-degree of control might be the best for conveying a performer's subjective feeling of emotion, but in performance, requires both honesty and attention on the part of the performer.

2.2.3 Summary

As in CSEMP, the tool for realtime interaction can be separated from the algorithm for music generation. The algorithm for generation is determined by its input, the generation algorithm, and output schema. Input data can come from either a "high-level" emotional model or low-level control input. The portion of performance content that is generated directly from algorithm categorizes the output schema. The generation algorithm implements structural, acoustic or performative cues determined by the system designer to communicate or express emotion given the input data and desired output schema.

Technologies for realtime control are determined by degree of emotion recognition and control. If the technology makes a translation from low-level data features to emotional coordinates (e.g. sadness, activity, valence), it is called "emotion recognition," otherwise, it is termed "emotion sensing." Degree of control is determined by the degree to which a performer can consciously manipulate input data, a feature that is not always desirable. In light of the above typology, the two audio environments introduced presently use a tool for emotion recognition with a low degree of control. They feature two different output schemas: the audio-processing environment uses additional input from a performer's voice and the sonification environment generates content independently. The two translation algorithms implement cues based upon psychological results from music emotion, but are not directly comparable due to the difference in output schemas.

2.3 Details Regarding Test Case

Section 2.3 was written by Ian Hattwick and edited prior to publication by R. Michael Winters.

In this section we present details about the test case scenario used for the development of the audio environments. We discuss the biosensors used to collect physiological data, the emotion recognition engine in the Emotional Imaging Composer, and the musical and aesthetic aspects of the performance.

2.3.1 Biosensors

The performer's physiological data was recorded at 64Hz using Thought Technologies' ProComp Infiniti[2] biofeedback system. The specific biosignals recorded were galvanic skin response (GSR), blood volume pulse (BVP), phalange temperature, heart electrical activity using an electrocardiograph (EKG), and respiration.

2.3.2 The Emotional Imaging Composer

The Emotional Imaging Composer takes the raw physiological data and processes it using four steps in order to produce arousal and valence data (Benovoy, Cooperstock, & Deitcher, 2008). The four steps are:

1. Preprocessing: raw signals are processed to reduce motion artifacts and high frequency noise.

2. Feature Extraction: 225 features are extracted from the noise-filtered biosignals and their first and second derivatives. Examples of features include heart rate mean, acceleration and deceleration, and respiration power spectrum at different frequency bands.

3. Feature Selection: Redundant and irrelevant data is removed from the feature set using a greedy sequential forward selection algorithm.

4. Feature Space Reduction: the remaining features are projected onto a 2-dimensional arousal/valence space using Fisher discriminant analysis.

2.3.3 Performance Details

As Emotional Imaging's primary goal for the EIC is "to investigate the mapping of [emotional] states to expressive control over virtual environments and multimedia instruments"

[2] ProComp Infiniti [Online]: `http://www.thoughttechnology.com/proinf.htm`

(Benovoy et al., 2008), a test case scenario was presented to guide the development of the audio environments. This scenario involved method-trained actress Laurence Dauphinais interacting closely with a small audience while performing “You Put A Spell On Me” (by Screamin’ Jay Hawkins and made famous by Nina Simone) along with the corresponding audio, biosignal, arousal, and valence data. Since the EIC uses data regarding physiological processes over which performers have little conscious control, the intention of EII is for it to produce output that transparently reflects the inner emotional state of the performer. Though challenging, Dauphinais had previously demonstrated the ability to use her method acting training to reliably reach certain emotional states.

The video recording used to test the audio environments during development contains a single audio track that consists of both vocals and piano. Dauphinais improvised variations on the basic song, and used her method acting training to help her move through various emotional states. Her performance and the piano accompaniment were in the jazz vocal tradition. Since the video was recorded before the development of the audio environments, her performance does not take into consideration any additional digital processing or accompaniment. While this presented a challenge, it also reflects the desire of Emotional Imaging for the EIC to function in a wide variety of performance aesthetics. In order for the EIC to meet this goal, it has to be able to work in parallel to a previously existing performance tradition.

The research performed during the creation of the audio environments, therefore, centered on the effective mapping of emotional data to audio processing and synthesis in realtime musical performance. Additional goals were for the sonification environment to clearly present the data, and for the performance environment to use the data to augment the musician’s acoustic sound production.

2.4 Sonification System

The sonification system was written in SuperCollider,[3] an environment and programming language for realtime audio synthesis (McCartney, 1996; Wilson, Cottle, & Collins, 2011). The characteristic sound of the system was a resonant object excited by impulse in alternating stereo channels. This sound was created using the DynKlank UGen, which instantiates a bank of frequency resonators with independent control of center frequency, amplitude,

[3] SuperCollider [Freely Available Online]: `http://supercollider.sourceforge.net/`

and decay time (T60s) for each resonant mode.

Though initialized with resonant modes at 400, 800, 1200, and 1600 Hz, amplitudes of 0.3, 0.1, 0.1, and 0.2, and decay times of 1 second respectively, the GUI allows the user to create new sounds randomly by resetting center frequency, amplitude and decay of the four nodes. The new center frequency was between ±200Hz of the original, amplitude was randomly set between $(0.1, 0.5)$, and decay time between $(0.5, 1.5)$ seconds. This action was implemented by pressing a button, and also randomly generating a new visual ball representing the position in the arousal and valence space.

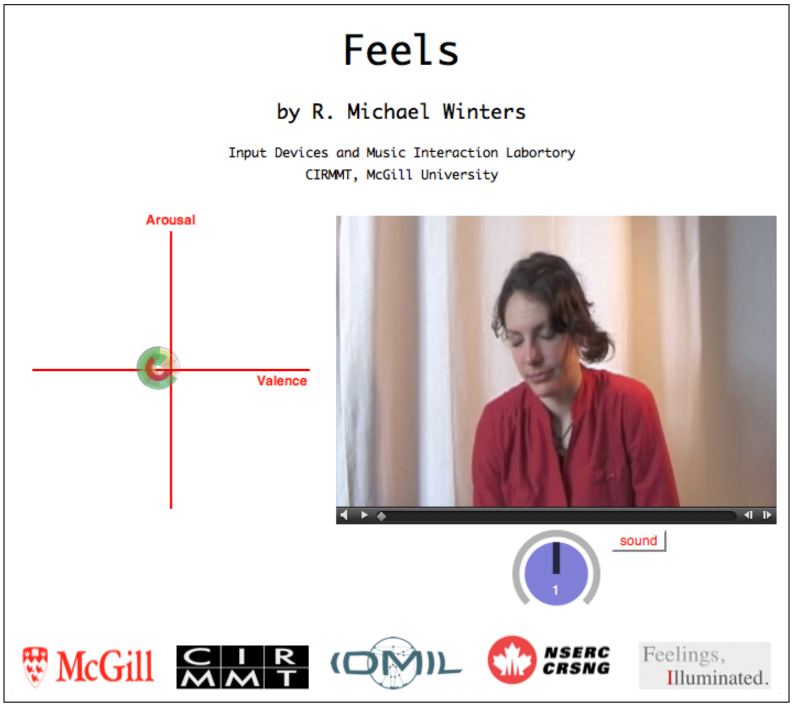
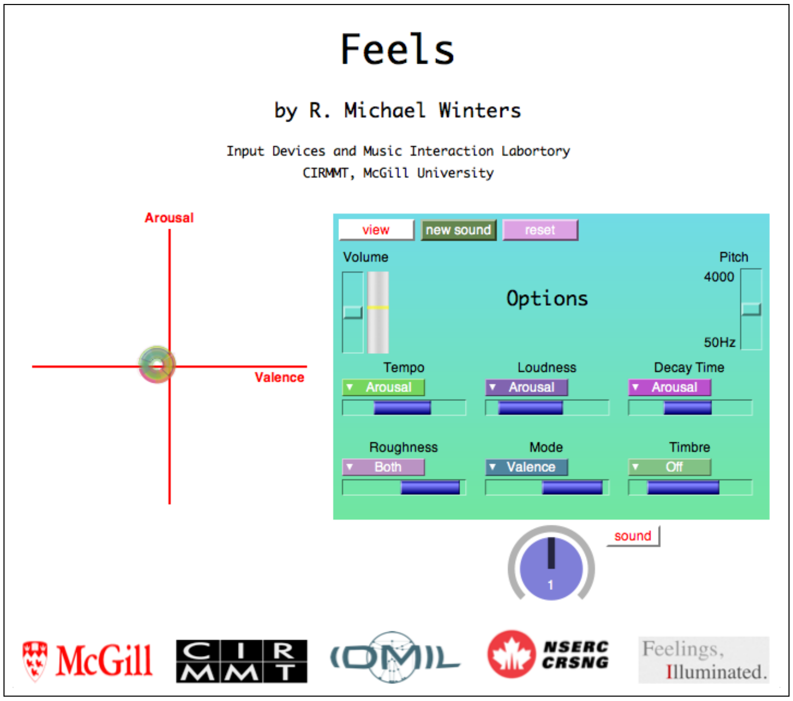


Fig. 2.1: The primary user interface. On the left, an arousal and valence (AV) graph contains a multicolored ball centered on the initial AV coordinate. On the right, a movie player displays method actress Laurence Dauphinais, for whom the AV trajectory corresponds. Pressing play in the movie player starts the movie and the time-aligned AV data. The blue knob below the video player controls play-through speed. Clicking on the AV graph un-mutes the sonification. The user can freely move the ball in the space if desired to learn the mappings.

The front end of the GUI (Fig. 2.1) displays an arousal and valence coordinate system



Feels

by R. Michael Winters

Input Devices and Music Interaction Labortory
CIRMMT, McGill University

Options

Tempo: Arousal
Loudness: Arousal
Decay Time: Arousal
Roughness: Both
Mode: Valence
Timbre: Off

Fig. 2.2: The mapping interface. By double-clicking on the AV graph, the video is replaced by the mapping interface, which allows the user to control aspects of the mapping. In this figure, the ranges of tempo, loudness, decay time, roughness, mode, and timbre can be changed, as well as to which dimension they are mapped. These functions were not implemented currently, but are kept in place for future development.

with a small multicolored ball representing the current arousal and valence coordinate. By clicking once on the arousal/valence (AV) graph, the sonification begins to play using the current AV position of the ball. By holding down the mouse-button, the user can drag the ball through the entire AV space hearing all of the possible sounds. Letting go of the mouse button snaps the ball back to its "true" coordinate, which is either the origin if there is no data, or elsewhere if the data is playing through. Pressing the graph again turns off the sound of the sonification, and double clicking exposes the back end, which is located behind the video player, and allows more user control of the sonification mapping.

Adjacent to the AV graph is a video player, which can be used to display corresponding live video if it is available. In the current context, a method actress sings through a song while her different emotions are identified as AV coordinates by the emotional imaging composer using physiological markers. When pressing play in the video begins to play and the AV data begins to drive the ball in the adjacent graph. The data is time-aligned with the video, so speeding through the video, skipping to particular points, all creates a change in the AV graph that reflects the coordinate of that instant in time. Just below the video player, a knob allows the user to control the speed the video plays through. Speed could be set anywhere between $e^{-1.5} \approx 0.2$ and $e^{1.5} \approx 4.5$ times the normal speed.

2.4.1 Mapping

A summary of the mapping decisions is provided in Figure 2.3. As discussed previously, the fundamental sound is a resonant object that is excited through impulse, with impulses alternating between left and right stereo channel. Tempo was conveyed by the rate at which impulses were presented. Arousal values were mapped exponentially from 0.75 to 5 impulses per second in each channel, creating between 1.5 to 10 impulses per second together. Loudness was also mapped to arousal, with the lowest arousal being 1/10th the amplitude of the highest arousal. Articulation was the third and final cue used for arousal, implemented by uniformly increasing or decreasing the decay times (T60s) of all resonant modes. At the lowest arousal, decay time was 2 seconds, at highest arousal, decay time was 0.5 seconds. These choices meant that each new excitation of the resonator occurred before the sound fully decayed.

Globally, valence was controlled by increasing "majorness" or "minorness" of the resonator as valence became more positive or negative respectively. Although at neutral valence there was only one sound, moving either positively or negatively in valence introduced three additional notes from either a major or minor triad. For example, given the initial fundamental of 400Hz with partials at 800Hz, 1200Hz, and 1600Hz, the neutrally valenced sound was most nearly G4. If increasing in valence however, B4, D5 and G5 would slowly increase in amplitude. The fifth, would reach maximum loudness at ±0.5 valence. The third would reach maximum loudness at ±0.75 valence, though it would be a major third (B4) for positive valence, and a minor third (B$^\flat$4) for negative valence. Finally, the octave (G5) reached maximum loudness at ±1 valence.

Sensory dissonance was used to convey the second quadrant (negative valence, high arousal), and was implemented by creating an identical copy of the sound (including third, fifth, and octave), and pitch shifting. The amplitude of the copy increased with radial proximity to $3\pi/4$, being 0 at both $\pi/2$ and π. Within the second quadrant, sensory dissonance increased with radial distance from the origin. At maximum distance, the copy

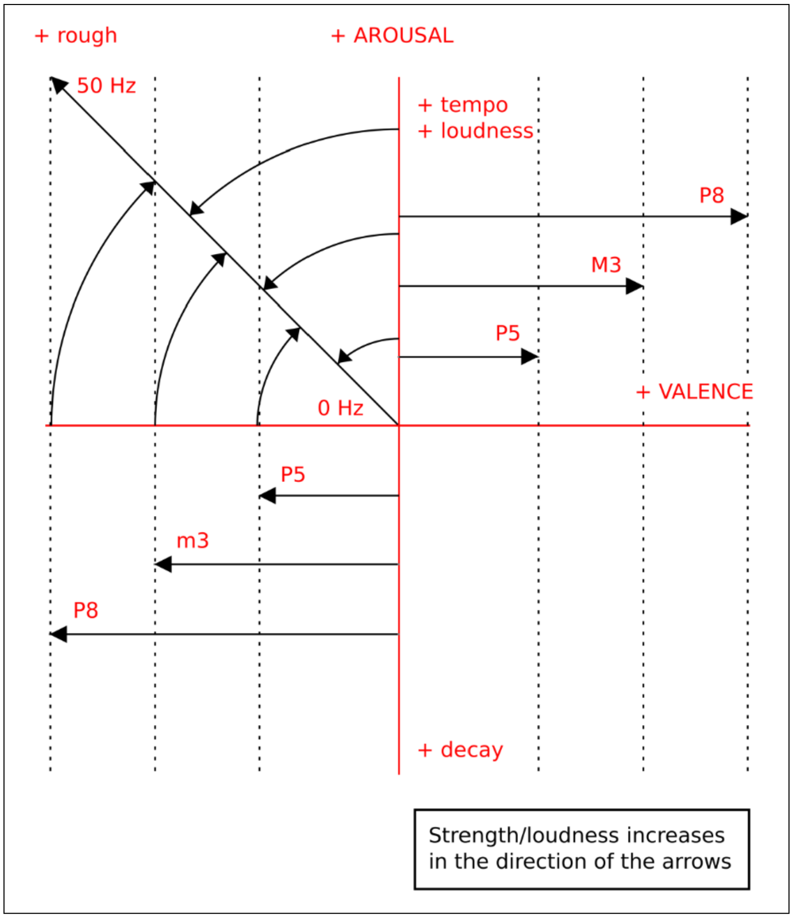


Fig. 2.3: A summary of the mapping decisions on a two-dimensional arousal arousal/valence plot. Arrow direction indicates increasing strength.

was pitch shifted by 50Hz—at the origin, there was no pitch shifting.

2.4.2 Evaluation

The system was created with the express goal that emotional communication through audio should be as clear as possible. Informal evaluations from public demonstrations have been affirmative of the strategy. Holding the ball fixed in different regions of the AV space could convey markedly different emotions that expressed categorical emotions like sad, happy, boredom, anger, and fear. Using sensory dissonance in the second quadrant was particularly salient for listeners. Though the major-happy/minor-sad cue is culturally specific, remarks from listeners at public demonstrations supported its viability as a cue for conveying the difference between positive and negative emotions of similar arousals. Listeners also liked the ability to generate new sounds by clicking a button. It was hypothesized that new sounds could refresh the listener's attention, which could otherwise diminish when using the same sound for long periods of time.

Interesting results were provided through listening to the sound in the background while watching the method actress. The auditory display of her emotions provided information that was not obvious through visual cues alone. For example, the sonification could be "nervous sounding" or "happy sounding" even when the cues from the actresses facial expression and gesture suggested otherwise. Because the sound was assumed to be the emotional representation that was "felt" by the actress, the added sound contributed to a deeper understanding of the actress' emotional experience. Further, this auditory representation allowed visual attention to be directed towards the actor's expression rather the visual AV graph.

2.4.3 Future Work

Although the decisions implemented in this model were informed by research on the structural and acoustic cues of musical emotion, a more rigorous framework has been provided in (Winters & Wanderley, 2013), which considers possible environmental sources of auditory emotion induction, and additional structural and acoustic cues guided by a more psychologically grounded approach to feature selection. The additional psychoacoustic features of sharpness, attack, tonalness, and regularity for instance have not yet been implemented, but should be in future work.

2.5 Performance Environment

> *Section 2.5 was written Ian Hattwick representing the system he developed for the collaboration with EIC. The text has since then been edited for consistency by R. Michael Winters. It is kept in place as a reference for the typology in Section 2.2, which is applied towards comparing the two systems.*

The test case scenario presented by Emotional Imaging presents different constraints from other approaches incorporating emotion data into music performance such as affective music generation systems (Wallis, Ingalls, & Campana, 2008) or performances in which all of the musical characteristics are generated in response to emotional data (Clay et al., 2012). In the chosen test case, the structure of the performance environment was heavily driven by the fact that the song determined the harmony, form, and rhythm of the singer's performance. In addition, it was desirable for the effects of the singer's emotion to be seen as part of the singer's performance rather than as an accompaniment. Due to these considerations we chose to also implement a performance system that processed the singer's voice rather than generating an autonomous additional audio source.

The fact that the source material was a human voice raised other issues relating to performance practice. Juslin and Laukka (2003) note that the human voice is a primary factor in the development of emotional audio cues. We quickly identified that drastic alterations of vocal timbre through distortion, pitch shifting, and filtering not only sounded unnatural within the context of the song but also served to obscure the emotional cues already present within the voice. For this reason we chose to implement a spectral delay algorithm that enables the creation of virtual spaces representing different emotional states.

2.5.1 Spectral Delay

A spectral delay system divides an incoming audio stream into a discrete number of audio bands, and each band is then individually stored in an audio buffer. The buffer containing each band is then played back with its own delay, feedback, gain, and panning settings. We also implemented an optional additional amplitude envelope stage. This stage occurs after the gain stage, and a 32-step sequencer whose parameters are controlled by the output of the EIC triggers the envelopes. The spectral delay implemented for this project was developed in Max/MSP and draws upon prior work by John Gibson's work on spectral delays (Gibson, 2009) and Jean-Francois Charles' use of jitter matrixes to store frequency domain audio data (Charles, 2008).

2.5.2 Graphic Programming Interface and Preset Management

Two separate graphic user interfaces were developed for easy programming of the spectral delay as well as the mapping strategies. A two stage preset management system was also implemented, of which the first stage allows for the user to save presets containing spectral delay and sequencer parameters.

The second preset stage contains parameters pertaining to the mapping of different spectral delay presets to the two-dimensional *AV* space. Five different delay presets are assigned to separate nodes. Each node consists of a central point and a radius within which the delay preset is activated. When the radii of multiple nodes overlap the parameters for the presets they refer to are interpolated. Parameters stored in this stage include the preset assigned to each node, the location and radii of each node, and the color assigned to each node. Five nodes were initially implemented in order to allow for one node for each quadrant of the emotional space as well as one node for a neutral "in-between" state. In practice, it was found that the performer navigated within a relatively small terrain within the emotional space and therefore an irregular assignment of nodes was more musically effective.

Several initial delay characteristics pertaining to emotional states were identified, including delay brightness, density, amplitude, stereo width, and length. Emotional cues contained within music performance as detailed by Juslin and Timmers (2010) were found to correlate to these characteristics as well. One useful facet of the spectral delay we implemented is that each characteristic can be realized by a variety of different approaches. For example, lowering the brightness would normally be achieved by lowering the gain of the higher frequency bands; however it can also be achieved by lowering their feedback, delay time, or panning. Many of these settings are consistent with real-world acoustics, such as the attenuation of high frequencies as sound radiates in a room, but the possibility for unnatural acoustic characteristics is retained.

2.5.3 Evaluation

The video of the test case with emotional data from the EIC was used to evaluate the performance environment and mapping strategies. It was quickly found that creating spaces which correlate to emotional states was relatively easy to do; however, by themselves they did not serve to create the desired emotional impact due to the fact that listeners discern the emotional cues contained within the vocal performance as more relevant than those provided by the acoustic space. However, once the performer's emotional signals cause the delay to move from one delay preset to another the sonic change was easily perceived and made a stronger contribution to the perceived emotion of the performer. The importance of moving between delay presets in order to create emotional cues underscores the importance of the location of the nodes within the mapping preset. Since performers will tend to move within a limited number of emotional states, the borders between nodes will need to be located near the junctions of those states in which the performers spend the most time.

2.6 Conclusion

This paper presented two systems for interactive affective music generation. Using a collection of biosignals from the autonomic nervous system, the Emotional Imaging Composer outputs realtime arousal and valence coordinates. In Section 2.2 we presented a typology for affective music generation that drew upon analogies with computational systems for expressive music performance (Fabiani et al., 2013; Kirke & Miranda, 2013a). We distinguish our system as one relying on emotion recognition rather than emotion sensing and being relatively difficult to consciously control.

Though both audio environments use realtime arousal and valence coordinates, and emotionally salient structural and acoustic cues, the difference in desired output schema resulted in markedly different generation algorithms. The sonification environment approached sound generation for the purposes of emotional communication and display, resulting in an autonomous sound that differed at every point in the AV space. The performance environment targeted live input from the human voice for audio processing, thus modified existing performance content.

The sonification was received well in public demonstrations, and users liked the ability to quickly select new sounds with the click of a button. Sensory dissonance and mode were used to convey valence; tempo, loudness and decay were used to convey arousal. The most compelling use context was provided by watching the method actress perform and listening to the auditory display, which provided more information on the performer's emotional state than was available visually.

Chapter 3

Applications & Strategies for Continuous Auditory Display of Emotion

Winters, R. M., & Wanderley, M. M. (2013, June). Sonification of emotion: Strategies for continuous auditory display of arousal and valence. In *Proceedings of the 3rd international conference on music and emotion*. Jyväskylä, Finland.

Abstract

Sonification is an interdisciplinary field of research broadly interested in the use of sound to convey information. A fundamental attribute of sound is its ability to evoke emotion, but the display of emotion as a continuous data type has not yet received adequate attention. This paper motivates the use of sonification for display of emotion in affective computing, and as a means of targeting mechanisms of emotion elicitation in music. Environmental sound and music are presented as two possible sources for non-verbal auditory emotion elicitation, each with specific determinants and available features. The review concludes that the auditory-cognitive mechanisms of brain stem reflex and emotional contagion provide the most advantageous framework for development. A sonification model is presented that implements cues that target these mechanisms. Computationally based strategies for evaluation are presented drawing upon the music emotion recognition literature. Additional aesthetic considerations are discussed that benefit usability and attractiveness of the display.

3.1 Introduction

Sonification is an interdisciplinary field of research broadly interested in the use of sound (usually "non-speech audio") to convey information (Kramer et al., 1999). A classic example of sonification, the Geiger counter, conveys the amount of radiation in the nearby environment using audible clicks. Although sonification has found many applications, this small sample exemplifies two compelling functions. Namely, sound can i) display a stream of information that is not visually obvious and ii) leave the eyes free to direct attention to other tasks. Like radiation, emotion is not always visually accessible, and displaying emotional information through sound does not require visual attention. Unique to emotion however, sonification can recruit resources from a cognitive apparatus that is well-equipped for auditory emotion perception.

In the field of sonification, the subject of continuous emotion display has not yet received adequate attention. Sonification applications have included assistive technologies, bio-acoustic feedback, data exploration, alarms, and process monitoring (Hermann et al., 2011), but the subject of emotion is rare. Though it has been recognized for its role in sound quality and interaction (Serafin et al., 2011), and is relevant to preference and pleasantness in sonification aesthetics (Vickers, 2011), only short, discrete sounds have thus far been applied. Such examples include using auditory icons to communicate emotional associations of the weather (Hermann, Drees, & Ritter, 2003) and using earcons for emotional communication in driver-vehicle interfaces (Larsson, 2010) and robotics (Jee, Jeong, Kim, Kwon, & Kobayahi, 2009), but the display of emotion as a continuous realtime data type is absent. The subject as a whole is much more at home in the realms of contemporary research in affective computing (Picard, 1997) and musical emotion (Juslin & Sloboda, 2010), where emotion expression and communication is considered computationally and music's affective capacity is studied in depth.

Furthermore, affective computing and musical emotion stand to benefit from the development of sonification strategies for emotion. Although embodied conversational agents (Hyniewska, Niewiadomski, Mancini, & Pelachaud, 2010) and emotional speech (Schröder, Burkhardt, & Krstulović, 2010) are the predominantly used modalities for affect display and communication, non-speech audio is an *un-embodied medium*, requiring neither a face or a voice to be understood, and by extension, leaving visual and verbal attention untaxed. When used in combination with other display modalities, this auxiliary channel may contribute to a more meaningful data interpretation.

Sonification of emotion can also be useful to the study of musical emotion. A great number of psychological studies have thus far been applied to determining the acoustic, structural (Gabrielsson & Lindström, 2010), and performative (Juslin & Timmers, 2010) elicitors for musical emotion. However, these results have yet to be applied to creating a "systematic and theoretically informed" manipulation of musical stimuli, which according to Juslin and Västfjäll (2008, p. 574), would be a "significant advance" to stimuli selection. Parallel to psychological studies, music emotion recognition (MER) (Yang & Chen, 2011) has created models for musical emotion using sets of psychoacoustic features, reaching approximately 65% accuracy for categorical emotion recognition in large corpora of music (Kim et al., 2010, p. 261). Sonification offers the possibility of targeting the mechanisms for emotion induction that rely upon the same low-level acoustic cues as these algorithms, increasing (or even decreasing) recognition accuracy, leading to interesting conclusions.

This paper motivates the use of sonification for affective computing and presents strategies for continuous auditory monitoring of arousal and valence. After presenting relevant results from environmental sound, a framework is proposed founded upon two mechanisms for emotion induction in music. A sonification model that implements a select number of these acoustic cues is discussed. Goals and methods for evaluation are presented.

3.2 Background

Affective computing is defined as computing that relates to, arises from or deliberately influences emotion and other affective phenomenon (Picard, 1997). This definition is broad enough to include some uniquely musical pursuits, most of which would not normally be considered as related to affective computing. The first is music emotion recognition (MER), where automated, computational systems for emotion or "mood" recognition based on audio and/or text-based information have received increasing attention (Kim et al., 2010). The second are systems for affective music generation, where music composition is computationally infused with results from psychological studies of music emotion (e.g. Gabrielsson & Lindström, 2010). Within affective computing, music has been recognized as a "socially accepted form of mood manipulation" (Picard, 1997, p. 234), which for example has been applied to noted performance gains in sports (Eliakim, Bodner, Eliakim, Nemet, & Meckel, 2012), gaming (Cassidy & MacDonald, 2009), and driving mood (Zwaag, Fairclough, Spiridon, & Westerink, 2011).

Among these alternatives, sonification of emotion is most closely related to the development of affective music generation systems. Both share emotional data as input and create an "emotional mapping" to sound parameters. Furthermore, sonification can be listened to musically (Vickers & Hogg, 2006) and even integrated into affective music generation systems (Winters et al., 2013). However, they can be distinguished both by the goals of the system designer and the way that they are meant to be listened to. Borrowing from the standard definition of sonification, the goal of a designer is to create a "transformation of data relationships into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation" (Kramer et al., 1999). In this light, the sound resulting from sonification is most comprehensively a *signal* that for the listener communicates or interprets important data relationships. If the data is emotion, than sonification, even when explicitly borrowing acoustic features from music, is simply a signal that communicates or interprets the data for the user.

The definition of sonification in fact, most closely parallels the third of four non-exclusive areas of affective computing: technologies for displaying emotional information or mediating the expression or communication of emotion (Picard, 2009). Although this area most commonly makes use of social signals (Vinciarelli et al., 2012) such as facial, gestural and vocal expressions in embodied conversational agents (Hyniewska et al., 2010), and the task of knowing the social display rules that govern *when* to display *which* affect has been referred to as the "hardest challenge" (Picard, 2009, p. 13), there are contexts in which the relative simplicity of accurate realtime auditory display of emotion would beneficial.

For communication, these contexts arise when social displays of affect are *unavailable*, *misleading*, or *inappropriate*. A social display might be unavailable in cases when an agent is physically removed from or incapable of generating signals recognizable to the receiver. Social displays might be misleading if they are purposely masked, neutralized, or changed in magnitude (Matsumoto, 2009). A social display might be inappropriate if verbal or visual attention needs to be directed elsewhere, like when engaging in other more primary tasks. If paired with a social display, the auditory channel might be likened to the use of music in film, where sound contributes to the emotional expression of a multimodal scene. In visually based analysis tasks, the addition of the auditory channel might draw attention to data relationships not obvious if using visual-only methods.

Sonification of emotion is further motivated by increasingly sophisticated and diverse technologies for realtime emotion measurement and recognition. In these contexts, the subjective experience of emotion is often represented dimensionally (Fontaine, 2009), and the two-dimensional arousal/valence model of Russell is particularly prominent (Russell, 1980). To create a continuous sonification that would be successful in the use-contexts previously described, an objective and systematic mapping of arousal and valence appears most prudent. The content of the next section determines which of the many possible features of non-speech sounds make good candidates for emotion display. Section 3.4 presents an approach for mapping them from realtime arousal and valence coordinates.

3.3 Determining Best Strategies

Potential sources for auditory display of affect come from two broad categories of sound: environmental sound and music. Though speech is another candidate, the stated goal is to create a display that does not conflict with verbal communication. Although some of the cues used in vocal expression of emotion might be shared by the auditory display (as in music; Juslin & Laukka, 2003), the goal here is not to use speech.

Within environmental sounds and music, there are additional requirements imposed by the conditions of realtime data monitoring as a background task in parallel to other more primary tasks. In sonification, this context is most often associated with process monitoring applications, and the present case is most closely a peripheral rather than direct or serendipitous display (Vickers, 2011). As noted by Vickers, common issues raised in process monitoring design are intrusion or distraction, fatigue and annoyance, poor aesthetic or ecological choices, and comprehensibility. These concerns are in turn grounded in the underlying need for appropriate aesthetic and semiotic choices. Through an analysis of acoustic features that communicate emotion in music and environmental sounds, this review shows that ultimately music provides the strongest theoretical framework for development due to the wealth of research and the continuous and malleable nature of its elicitors.

3.3.1 Emotion in the Acoustic Environment

Research on the acoustic elicitors of emotion in the natural environment has been most commonly presented in the psychoacoustic literature or in the pleasantness or annoyance of product sounds. However, recent research has sought a more ecological approach to sound perception in which psychological determinants take prominence to strictly signal characteristics, and the role of emotion becomes more complex. “Emoacoustics” (emotional acoustics) research embodies this trend towards a focus on listener and context, and contributes intriguing new methods and results (Asutay et al., 2012).

Perhaps the most thorough review of emotion in the auditory environment comes from Tajadura-Jimnez who categorizes “auditory-induced emotions” into four determinants (Tajadura-Jiménez, 2008, Ch. 4):

1. Physical Determinants
2. Identification/Psychological Determinants
3. Spatial Determinants
4. Cross-Modal Determinants

Physical determinants are those related to the signal itself and are best studied using “meaningless” sounds (Västfjäll, 2012) like broad-band noise, and amplitude or frequency modulated tones, as is done in the psychoacoustics literature (Fastl & Zwicker, 2007). Factors related to identification enter when a sound has meaning due to the recognition and cognitive associations of the listener. Experiments using the International Affective Digitized Sounds Library (Bradley & Lang, 2007) have targeted this determinant and found similarities with corresponding affective pictures (Bradley & Lang, 2000). Spatial determinants arise when some aspect of the space contributes to the emotion. Issues of proximity, location, room size (Tajadura-Jiménez, Väljamäe, Asutay, & Västfjäll, 2010), and approaching or receding sound sources (Tajadura-Jiménez, Larsson, Väljamäe, Västfjäll, & Kleiner, 2010) have been studied in combination with different sound types (Hagman, 2010). Cross-modality effects occur when emotionally salient information from one modality impacts another. For sound, visual or tactile information might contribute to the emotional meaning of a sound, though this effect has been studied the least.

Although these categories are valid, only the first three pertain to audio-only display. From these, identifiability requires special consideration. As mentioned in the introduction, identifiable sounds (a.k.a. auditory icons; Brazil & Fernström, 2011) have been applied thus far to conveying emotional associations of the weather (Hermann et al., 2003). Although the affective space occupied by these sounds has been shown to convey a variety of emotions (Bradley & Lang, 2000), sounds notably fall upon two motivations, "appetitive" and "defensive," creating a 'V' shape in the AV space. If this trend were to continue for all identifiable sounds, it would leave gaps that could not be well communicated through sound.

Movement is another problem for the use of identifiable sounds. To convey a transition from high arousal, high valence to low arousal, high valence, would require the interpolation through many sounds. If this transition were to occur rapidly, the identifiability of these sounds might be compromised due to their short length. This problem might be avoided by using evolutionary objects (Buxton, Gaver, & Bly, 1994), which, as identified in the auditory icons literature, allow sound properties to be updated while playing (Brazil & Fernström, 2011). If using an evolutionary object, the sound would need to be able to occupy the entire AV space, so it might be best to start with a sound which is more or less emotionally neutral. A promising candidate for this feature is self-referential sounds (Tajadura-Jiménez & Västfjäll, 2008), or sounds related to one's own body and its natural movements (e.g. walking, breathing). The sound of a heartbeat for instance could be changed in tempo or loudness to convey arousal, and perhaps sharpness, roughness, and tonalness to convey valence.

The capacity of using spatial determinants for continuous display is worth mentioning, though is also limited. Increasing room size (reverberation time) creates a systematic decrease in valence and increase in arousal for sounds with neutral emotional connotation (e.g. clarinet, duck quack), but not for negative connotation (e.g. dog growl) (Tajadura-Jiménez, Larsson, et al., 2010). Evidence supporting this effect of neutral sounds is also present in (Västfjäll, Larsson, & Kleiner, 2002), though the effect on arousal was less pronounced. Arousal, in fact, decreased for the condition of highest reverberation, attributed to a decrease in "presence." Presence, though lacking a precise acoustic definition, has been defined as the perceptual illusion of non-mediation (Lombard & Ditton, 1997), and has been strongly connected to the emotion in auditory virtual environments (Västfjäll, 2003), perhaps most analogously correlated with the degree of "realism" (Frija, 1988). Creating the illusion of "approach" is possible by increasing loudness, and creates an increase in emotional intensity, but only for identifiable sounds deemed "unpleasant" (Tajadura-Jiménez, Väljamäe, et al., 2010). Finally, in general, sounds perceived as coming from behind the individual are more emotionally arousing (Tajadura-Jiménez, Larsson, et al., 2010). In general however, one must be aware that the use of spatial effects for emotion display or expression is challenged by incongruent visual information, which can diminish the strength of the desired auditory illusion (Larsson, Västfjäll, Olsson, & Kleiner, 2007).

The results are most clear-cut with psychoacoustic literature using broadband noise, and amplitude or frequency modulated tones (Fastl & Zwicker, 2007). Composite models for sensory pleasantness (p. 245) and psychoacoustic annoyance (p. 328) have been developed using well-defined metrics for roughness, sharpness, loudness, tonality, and fluctuation strength. These have been shown to predictive of ratings of pleasantness and annoyance of product sounds, though they were not designed to be able to predict the position in a full 2-D arousal valence model (Västfjäll, 2012). They make good candidates as features for sonification, though using ecologically valid stimuli should not be abandoned. Results from sonic interaction design (SID) have shown that "naturalness" creates a systematic increase in valence compared to synthesized sounds with similar spectral centroid and tonality (Lemaitre, Houix, Susini, Visell, & Franinović, 2012). However, as in SID, it might be best to consider naturalness as an overall aesthetic property that should be conserved, contributing to the attractiveness of the sound and "usability" of the sonification (Norman, 2004).

This review has assessed different possible features for emotion communication in environmental sounds. If using identifiable sounds, it would be best to use evolutionary sounds, perhaps in some way self-representational. The use of spatial effects can be considered if one is mindful of visual dominance. Psychoacoustic features are the most promising for sonification, but "naturalness" and "realism" are global properties that should be conserved. Overall, it would appear that the strongest emotional determinant of environmental sound—identifiability—is not viable for sonification due to the problem of continuous movement and gaps in the AV space, dramatically diminishing the framework as a whole. The field of Emoacoustics is still developing, and future results might be more favorable. For the time being, a much stronger framework is founded in contemporary research on music and emotion, which will be discussed in the next section.

3.3.2 Mechanisms of Musical Emotion

On the surface, it would seem that the most useful results for sonification of emotion come from the wealth of results linking structural, acoustic, and performative cues in music to defined regions of the arousal/valence space. Instead however, a more rigorous approach first determines which psychological mechanisms are favorable for emotion elicitation given defined properties such as cultural specificity/learning, volitional influence, and induction time. These mechanisms in turn encompass subsets of the available structural/acoustic feature space, making the process of selection easier.

Many psychological studies have been conducted to determine what structural, acoustic, and performative parameters contribute to emotional communication in music (Gabrielsson & Lindström, 2010; Juslin & Timmers, 2010). Additionally, new computational approaches to feature determination have been introduced in the field of music emotion recognition (Yang & Chen, 2011). This literature affirms that there is no dominant single feature, and musical emotion is best predicted using a multiplicity (Kim et al., 2010). The literature on performance cue utilization (Juslin, 2000) has also advanced—recent results have introduced defined ranges for communication of discrete emotions (Bresin & Friberg, 2011).

Collectively, these results offer an abundance of possible features for emotion communication in sonification, but music research offers a more fundamental approach, that of the auditory-cognitive mechanism. In this vein, a collection of six mechanisms for emotion elicitation in music has been proposed (Juslin & Västfjäll, 2008), of which two can be used for continuous auditory display in the contexts thus far mentioned: brain stem reflex and emotional contagion. Both have a low-degree of cultural and volitional influence, high induction speed, and a medium dependence upon musical structure (Juslin & Västfjäll, 2008, Table 4). It is worthy of note that the psychoacoustic features from the environmental sounds literature that are the most viable for sonification are accounted for by these mechanisms, and as noted in (Tajadura-Jiménez, 2008, p. 26), mostly the brain stem reflex.

Acoustic features drawing upon the brain stem reflex recruit innate structures of the brain that bear upon the organism's survival. As noted in (Juslin & Västfjäll, 2008, p. 574), these features are most commonly studied in the psychoacoustics literature and include sharpness, loudness, roughness, tonality, and fluctuation strength. In the music literature, a close relative of sharpness is the height of the spectral centroid. Likened to roughness is sensory dissonance, and tonality (a.k.a "tonalness"; Egmond, 2009, p. 79) refers to how tone-like the timbre is as opposed to broadband. The spatial cues discussed in Section 3.3.1, might be considered in this list in that spatial hearing is also shared and important to an organisms survival, though effects that are dependent upon the sound identification are likely cognitively mediated.

Emotional contagion is a process whereby emotion is induced by perceiving the expression of the stimulus itself and then "mimicking" it internally (Juslin & Västfjäll, 2008). The theory suggests that because music shares many of the acoustic features used in vocal expression of emotion, music becomes like a super-expressive voice (Juslin, 2001). Further, musical features are decoded by an "emotion-perception module" (Juslin & Laukka, 2003, p. 803) of the brain that does not distinguish between music and the voice. Evidence supporting this claim comes from an extensive review of literature in musical and vocal expression showing that a number of prominent features governing expression of five discrete emotions were shared in music and speech (Juslin & Laukka, 2003, Table 7). The cross-modal features relevant to this proposed module are tempo, intensity, intensity variability, high-frequency energy, pitch-level, pitch variability, pitch contour, attack and microstructural regularity (taken at the note-to-note level; Bunt & Pavicevic, 2001).

Implementation of these reflex and contagion features requires two levels of acoustic content, timbral and note-based. For the brain stem reflex and psychoacoustics, spectral content and intensity must be manipulated—the sonification must include a structure that allows malleability of sharpness (amount of high-frequency energy), tonality (amount of noise versus tonal components in the spectra), roughness (including fluctuation strength), and loudness. To use emotional contagion features, a note-based structure must be available for manipulation of tempo, pitch, and attack.

The strength of these features is their low cultural influence, low volitional influence, induction speed and their dependence upon structure. The resulting structure in the sonification is not necessarily "musical" for these mechanisms are on the one hand biological, and on the other, processed by an emotion-perception module that processes speech as well (Juslin & Laukka, 2003, p. 803). Other acoustic cues that rely upon different mechanisms can (and perhaps should) be used in sonification, but they can be expected to be more culturally dependent, with potentially lower induction speed, and more subject to volitional influence. An example of such a cue would be the major-minor mode, which in western classical music may be used to convey positive and negative valence. However, this connotation is not learned until the age of six to eight (Gabrielsson & Lindström, 2010, p. 393), and thus might be accounted for by the mechanism of musical expectancy.

3.3.3 Summary

Having compared mechanisms for emotional elicitation in both environmental sounds and music, it is clear that sonification of emotion finds more substantive support in the mechanisms described in musical emotion research. From environmental sounds, emotion determined through identification and appraisal of the sound was found to be a strong factor influencing emotion. However, the emotional space occupied by these sounds is incomplete, and the problem of movement suggests the use of emotionally neutral evolutionary or self-representational sounds for which acoustic properties can be easily manipulated. Though not well researched, "naturalness" of the sound should be conserved at a global level to maximize pleasantness and usability of the display. Along with these suggestions, the "presence" and "realism" of a virtual auditory environment may be applied towards increasing emotional intensity.

Ultimately however, the results from this literature are much less developed than those from musical emotion, and factors such as cultural dependency, induction speed, degree of volitional influence have not been adequately assessed. From music research, two viable mechanisms for sonification have been proposed, each with well-defined psychological properties. Further, the brain stem reflex accounts for the psychoacoustic and spatial results in the environmental sounds literature that would otherwise be most promising for sonification. The additional mechanism of emotional contagion presents additional musical features for sonification including tempo, attack, pitch information, and regularity.

3.4 Sonification Model

For the purposes of discussion, this section introduces an existing sonification that implements some of these cues for communication of arousal/valence space (Winters et al., 2013). The mapping strategies are then evaluated using the criteria developed in Section 3.3.

A single note forms the basis of the sonification. This note is created as a bank of resonant modes with independent control of center frequencies, amplitudes and decay times. The resonant object is excited through impulse in alternating left-right stereo channels. The choice of this sound was motivated by its "naturalness"—it is capable of generating sounds that resemble materials like glass, wood, metal, etc. For sonification, tempo, and loudness are mapped to increasing arousal, and the decay time increases with decreasing arousal.

Increasing positive or negative valence is conveyed by slowly increasing the loudness of the fifth, third (M3/m3), and octave above the original note. Sensory dissonance is conveyed in the second quadrant by taking an identical copy and pitch-shifting upwards. Loudness of the copy increases with radial proximity to the line 3/4, and the pitch shift increases with distance from the origin. These decisions are summarized in Figure 2.3 presented in Chapter 2.

3.4.1 Evaluation and Future Work

The decisions for tempo, loudness, and roughness are supported by the present discussion. Tempo is a feature from the emotion contagion mechanism, roughness is a feature of the brain stem reflex, and both share loudness. Tempo and loudness increased with increasing arousal as in speech, and increasing roughness and loudness both increased with sensory un-pleasantness. The decisions for major-minor and decay are musical features that are not supported by the present discussion but were found to be useful for conveying valence and decreasing arousal respectively. In fact, these two decisions contributed more to the aesthetic appeal of the display than the decisions of loudness, tempo, and roughness. Although mindful that when using features not accounted for by brain stem reflex and emotion contagion, desired psychological properties (e.g. low cultural specificity) are not guaranteed, the use of cultural associations has been supported in the design of process monitoring sonifications (Vickers, 2011) as well as in aesthetic computing (Fishwick, 2002). Drawing upon a listener's cultural associations can create a convincing display that enhances aesthetic appeal, but the designer should be mindful of its limitations. "Major-happy, minor-sad," for example is culturally learned and may not necessarily be understood by children under six to eight years old.

A yet undeveloped benefit of using strategies from music research, and perhaps most attractive for evaluation, are the growing number of models for music emotion recognition (Yang & Chen, 2011). Using audio-only features, these systems are capable of recognizing emotions categorically or dimensionally, and some systems are designed for time-varying, "second-by-second" emotion detection (Coutinho & Cangelosi, 2011; Schmidt & Kim, 2011). Because these models are sometimes designed for large corpora of music, stretching beyond those of western-classical tradition, the features used for recognition may be less culturally specific. As demonstrated in Chapter 4, these computational models can provide a preliminary metric of the accuracy of communication in the arousal/valence space.

As of yet, several of the features supported in this analysis have not been implemented. From the brain stem reflex, these include sharpness, tonalness, and fluctuation strength. From emotional contagion, these include pitch-level and its variation, contour, intensity variability, and attack. The spatial cues of increasing reverberation time and the auditory-illusion of "behind" might also be investigated. The framework of resonant synthesis creates sounds that are relatively more "natural" than other synthesis techniques. This strategy should be continued in further implementations, although synthesis of self-representational or evolutionary sounds might be assessed as well.

Although presently evaluated using the framework developed in Section 3.3, future evaluation of the design needs to determine how well the underlying AV space is conveyed. With this established, it will be necessary to perform user testing to evaluate sonification in the context of realtime peripheral process monitoring.

3.5 Conclusions

Realtime continuous auditory display of arousal and valence has not yet received adequate attention in the sonification literature, though the pursuit of technologies for realtime emotion recognition makes the data-type eminent. Benefits of sonification include displaying emotional information when visual or verbal cues are unavailable, misleading, or inappropriate, and providing an auxiliary channel for emotional display that can contribute to emotional expression or visual-based data analysis. With these contexts in mind, the present paper targets the development of peripheral display strategies on an underlying AV space.

Determining the best strategies for display requires careful aesthetic and ecological choices, for which research on the emotional impact of environmental sound and music provide two possible categories for the designer. Currently, the most robust foundation for development is presented by research in musical emotion and specifically cues recruiting the mechanisms of brain stem reflex and emotional contagion. These mechanisms account for most of the viable acoustic cues from environmental sound and propose additional cues shared with speech. These cues can be expected to have a low degree of cultural influence, a high induction speed, and a low degree of volitional influence.

The sonification model discussed explicitly uses some of these features, though others are presented for future work. As presented in detail in Chapter 4 to evaluate the model, it may be possible to use models for music emotion recognition as a preliminary design metric. With the accuracy of the mapping strategy assessed, user testing needs to evaluate how well the sonification performs in the defined use contexts of affective computing.

Chapter 4

Benefits & Limits of Computational Evaluation

Winters, R. M., & Wanderley, M. M. (2014). Sonification of Emotion: Strategies and results from the intersection with music. *Organised Sound, 19*(1), Accepted.

Abstract

Emotion is a word not often heard in sonification, though advances in affective computing make the data type imminent. At times contentious due to implied overlap with music, this paper clarifies the relationship between the two, demonstrating how in the case of emotion, this relationship can be mutually beneficial. After identifying contexts favorable to auditory display of emotion, and the utility of its development to research in musical emotion, the current state of the field is addressed, reiterating the necessary conditions for sound to qualify as a sonification of emotion. With this framework, strategies for display are presented that use acoustic and structural cues designed to target select auditory-cognitive mechanisms of musical emotion. Two sonifications are then described using these strategies to convey arousal and valence though differing in design methodology: one designed ecologically, the other computationally. Each model is sampled at 15-second intervals at 49 evenly distributed points on the *AV* space, and evaluated using a publicly available tool for computational music emotion recognition. The computational design performed 65 times better in this test, but the ecological design is argued to be more useful for emotional

communication. Conscious of these limitations, computational design and evaluation is supported for future development.

4.1 Introduction

Sonification is an interdisciplinary field of research broadly interested in the use of sound to convey information (Kramer et al., 1999). Though there are many techniques of sonification and many tasks to which it has been applied, a continual problem is that of definition (Supper, 2012). Always obfuscating, compromising, and testing the mettle of a concise and encompassing delineation are the various artistic and musical practices whereby data is also transformed into sound.

Music has been called a 'language of emotion,' (Crooke, 1957) and with good cause: a vast and expanding literature describes the ways that music comes to convey or induce an emotion in listeners (Juslin & Timmers, 2010). Sonification on the other hand, seems to be anything but a language of emotion. Of the approximate 2.4 million standard words in *The Sonification Handbook* (Hermann et al., 2011), the word 'emotion' appears a mere 78 times. Recent discussions of what might be considered a 'sonification of emotion' have even brought contention in the sonification community, due to potential overlaps with music (Preti & Schubert, 2011; Schubert et al., 2011).

In spite of this difficulty, there are several reasons why the field of sonification should consider the representation and communication of emotion more seriously. The first and perhaps most obvious reason is that emotion as a form of data is becoming increasingly common. In the field of affective computing (Picard, 1997), algorithms have been designed to detect and measure emotion from all manner of possible sources, including but not limited to physiological process, EEG, facial, gestural, and vocal expression (Picard & Daily, 2005). In addition to these indirect measures, technologies for continuous self-report are being used to collect readings of an individual's time-varying emotional experience (Schubert, 2010). Just as with other data types, the facilities of audition can be directed to perceiving this information, identifying patterns, and supporting communication when verbal or visual attention is already occupied (Walker & Nees, 2011).

Another, and perhaps more exciting prospect stems from the utility of the auditory-cognitive system as a non-verbal, non-visual channel for emotional communication. As evidence of the strength of this channel, one need look no further than the importance of music in film, where sound itself brings insurmountable intensity to a scene, even to the point of overriding incongruent visual and verbal emotional cues.

To create a sonification of emotion however, one does not have to create music. As will be discussed presently, many of the most promising applications benefit from the use of sound as a background display. Music, in all of its cognitive complexity, may obscure communication if it does not systematically convey the data, requires too much attention, or uses culturally learned schemas. Instead, by selecting wisely from emotionally salient acoustic cues, many of which are nevertheless used in music, emotion can be conveyed as a background information stream with desirable features such as high induction speed and low volitional influence.

After introducing the benefits and contexts favorable to the auditory display of emotion, the current state of research is presented, reiterating the necessary conditions for sound to be considered a 'sonification of emotion.' Although a number of structural and acoustic cues are used in the expression of musical emotion, a select group is chosen for sonification from desired psychological properties. Two sonification mappings are then presented for conveying arousal and valence but differing in design methodology. The first is designed ecologically using recommendations from the musical emotion literature, while the second is designed computationally using a publicly available model for music emotion recognition. Although the latter performs significantly better on a computational test, the former is argued to be more useful for emotion communication. These results help clarify the relationship between music and sonification, identify areas of mutual benefit, and facilitate future collaboration in emotion display.

4.2 Motivation & Background

The auditory display of emotion is a timely pursuit supported by research agendas originating in affective computing and musical emotion. Applications arise in both, either for emotional communication or model evaluation. Music research in particular offers a robust framework for development, which is applied to the present research. After presenting these relationships in detail, the current status of emotion in auditory display is described, highlighting the requirements for a technique to be appropriately termed a sonification of emotion.

4.2.1 Affective Computing

Affective computing has been defined as computing that relates to, arises from or deliberately influences emotion and other affective phenomenon (Picard, 1997). Though this definition is rather broad, technologies for display, expression, or communication of emotion constitute the third of four major research foci (Picard, 2009). In this context, sonification contrasts and complements existing display modalities, many of which require a face, voice, or body for communication. By contrast, non-speech sound offers an *unembodied* medium for emotional communication that can be ideal in situations when verbal and/or visual attention is already occupied. By extension, sonification of emotion can be added to an existing emotional display, potentially facilitating communication or expression.

The complexities of the rules governing when to display which affect has been described as 'the hardest challenge' of realtime emotion display (Picard, 2009, p.13). However, Winters and Wanderley (2013) list three cases in which the relative simplicity of realtime, accurate auditory display of emotion can be beneficial. These contexts arise when social signals (e.g. facial, vocal, gestural; Vinciarelli et al., 2012) are unavailable, misleading, or inappropriate.

A social display might be *unavailable* when an agent is either physically removed from or incapable of generating the social signals that would be otherwise recognizable to a receiver. In the case of autism for instance, where a person has difficulty utilizing social cues that would allow for their emotional reaction to be recognized, sonification might be used to assist the receiver and cue them into an otherwise hidden emotional experience. A social display might be *misleading* when social signals of emotion are consciously or unconsciously masked, neutralized, or changed in magnitude (Matsumoto, 2009). In this case, verbal and visual attention can remain dedicated to the socially displayed content, but the auditory display once again provides access to a hidden emotional layer, and perhaps a deeper understanding of the agent's state. Finally, a social display may be *inappropriate* when visual and/or verbal attention need to be directed elsewhere, such as when engaged in complex, more primary tasks.

In any of these contexts, the auditory display needs to be clear but also not so complex as to demand unnecessary attentional resources on the part of the user. This function most closely parallels sonification techniques related to process monitoring (Vickers, 2011). Furthermore, because the user's primary attention is directed elsewhere, but the information content of the display is important to the overall goal; the sonification would be classified as peripheral.

4.2.2 Musical Emotion

The auditory display of emotion should not exclusively direct itself towards contexts for realtime emotional communication. To consider this purpose as the exclusive benefit is to miss a potentially advantageous link with a close partner, the study of musical emotion. Musical emotion describes emotions induced or conveyed by music, and while its discussion is old (Budd, 1985), in the past few decades, its scientific axes have expanded, and a variety of psychophysiological, behavioral, and computational methods have been introduced.

Sonification of emotion intersects with musical emotion insofar as the study profits from systematic and theoretically informed mappings of acoustic features. For over three-quarters of a century, research has been directed to determining the structural and acoustic elicitors responsible for musical emotion (Gabrielsson & Lindström, 2010). Although music listening is a multifaceted process in which cultural learning and cognitive associations are fundamental, this branch has directed itself towards the underlying acoustic details. Though beginning with psychological studies, machine-learning approaches have recently gained momentum, offering signal-level correlates of music perception and composite computational models (Yang & Chen, 2011).

Using this background of musical emotion, sonification is afforded a wealth of knowledge on auditory emotion, and can make use of well-developed theories and results. These form the basis for the sonification strategies introduced in Section 4.3.1. However, sonification can also benefit the study of musical emotion by providing ‘systematic and theoretically informed’ approaches, which, according to Juslin and Västfjäll (2008, p. 574) would be a ‘significant advance’ to stimuli selection. In this way, both fields can profit from the other’s research developments.

This benefit is most easily applied to computational models for music emotion recognition, many of which use purely signal/content level attributes for prediction (Kim et al., 2010). These models are complex, using a multiplicity of acoustic features and functions for combination, but can ideally be generalized to large corpora of music, potentially spanning many genres (Ogihara & Kim, 2012). Using sonification, these purely computational models can be acoustically instantiated, satisfying a broad range of model requirements, and potentially isolating these low-level acoustic features from the higher-level cultural and cognitive mechanisms involved in music listening. Both of the sonifications presented in Section 4.3 are measured by such a model, forming the basis for evaluation in Section 4.4.

4.2.3 The Sonification Perspective

The subject of emotion is rare in the sonification literature, and at times even contentious for the definition of sonification (Schubert et al., 2011). To frame the present research, the current state of emotion in the field is addressed, identifying contexts where sonification has thus far been used, its relationship to aesthetics, and the conditions that qualify a technique as a 'sonification of emotion.'

The actual use of sound to communicate or express emotional information has thus far been limited to short, discrete sounds that would either qualify as auditory icons (Brazil & Fernström, 2011) or earcons (McGookin & Brewster, 2011). Hermann et al. (2003), for instance have explored the use of auditory icons to communicate emotional associations in auditory weather reports. These emotive markers (e.g. bird, sigh, scream) were played alongside auditory icons indicating more descriptive information such as temperature, windiness, and humidity. Later, Larsson (2010) introduced two software tools for designing earcons for communication of urgency in auditory-in-vehicle interfaces. As with the weather reports, the emotive content of these sounds were meant to be paired with descriptive identifiers (e.g. seatbelt reminder, collision warning).

Robotics has been another venue for application. Jee et al. (2009) have studied the use of short musical excerpts to express discrete emotional states such as happiness, sadness, or fear. The authors later conducted a review of 275 earcons used for communication of emotion and intention in two popular science-fiction robots (Jee, Jeong, Kim, & Kobayahi, 2010), applying the results to the design of seven musical sounds for expression in an English teacher robot.

These uses of sound to convey emotional information can be contrasted with aesthetic and design studies where the discussion of auditory emotion also appears. In sonic interaction design for example, emotions have been studied in users performing tasks with 'the flops glass,' an acoustically and computationally augmented physical object (Lemaitre et al., 2012). Results suggested that pleasant/positively valenced sounds could make the task seem easier, and provided the user with a stronger sense of control. These results, in combination with similar results from product sound quality suggest that sounds are not only emotionally differentiable, but that emotions can be predictive of product assessment (Västfjäll, Kleiner, & Gärling, 2003). In sonification, where sound can take on any number of forms, 'pleasantness' and 'ecological validity' are championed in design, for the reason that their consideration makes the process of listening easier and increases the ability to perceive the desired information content (Vickers & Hogg, 2006).

It has recently been posited that music might be considered a sonification of emotion: a potential challenge to traditional definitions of sonification (Schubert et al., 2011). The argument stems from the capacity of music (at times) to successfully communicate emotion—the composer or performer encoding an emotion, and the listener decoding. The conditions introduced by (Hermann, 2008) can be applied presently to clarify what qualifies as a sonification of emotion.

According to Hermann (2008), a sonification must be objective, systematic, reproducible, and able to be used with different data. For sonification of emotion, this fundamentally requires an underlying data space that represents emotion, such that the sound can reflect properties and relationships in this space. There must furthermore be a precise definition for how each point in this data space becomes a sound, even to the point that sampling the data multiple times at the same coordinate will create structurally identical resulting sounds. As will be clear in the following sections, the sonification strategies introduced presently satisfy all of these criteria, and the features chosen for communication make the association with music secondary.

4.3 Two Models for Sonification of Emotion

From the previous discussion, the most advantageous avenue for development is the peripheral display of emotion, one that takes advantage of results from musical emotion. Sonification has thus far only made use of auditory-icons and earcons to convey short emotional states, while the realtime continuous display has not yet been sufficiently developed. After discussing strategies for auditory display of emotion, two models are introduced for displaying arousal and valence, two theoretical dimensions of emotion. One of the models was designed to be more ecologically valid and pleasant, the other was designed computationally using a tool for music emotion recognition and specially designed software for analysis.

4.3.1 Strategies

Winters and Wanderley (2013) discuss in detail strategies for auditory display of emotion in a process monitoring setting. Although environmental sounds and music are two broad categories of sound, each capable of emotion induction and communication, music is chosen as the framework for development. It proves advantageous because of the flexibility of acoustic elicitors, the encompassing wealth of knowledge, and problems inherent to using environmental sound for emotion display.

Within music, there are many structural and acoustic cues that correlate with musical emotion and that might be used for communication (Gabrielsson & Lindström, 2010; Juslin & Timmers, 2010). Instead of haphazardly selecting from the available cues, a more psychologically grounded approach first considers psychological properties that would be advantageous to the contexts thus far mentioned. This directs attention to specific auditory-cognitive mechanisms responsible for auditory emotion expression, and the more limited set of acoustic cues to which they respond.

The desired psychological properties for this sonification context include high induction speed, low volitional influence, and importantly, dependence upon structural and acoustic content. Using the framework provided in Juslin and Västfjäll (2008), this narrows the list of potential mechanisms for induction to 'brain stem reflex' and 'emotional contagion.'

The brain stem reflex is a biological mechanism, often triggered by sudden, or loud changes in sound that bear immediate impact upon an organism's survival. Structural and acoustic cues that can trigger this mechanism include loudness, sharpness, roughness, tonality, and fluctuation strength, all of which are studied in detail in the psychoacoustics literature (Fastl & Zwicker, 2007). Emotional contagion is a process whereby a sound triggers an emotion in virtue of having acoustic features that the listener perceives as expressing an emotion, and the listener then 'mimicks' this expression internally. Acoustic features that trigger this mechanism are shared with emotional speech (Juslin & Laukka, 2003), and include tempo, loudness, loudness variability, high frequency energy, pitch-level, pitch variability, pitch contour, attack and irregularity at the event-to-event level.

Using these features, it might be possible to create a systematic and reproducible mapping of an 'emotion,' but to satisfy the objective and different data requirements of Hermann (2008), it is necessary to make a choice of underlying data space. For this purpose, the two-dimensional arousal/valence space is chosen. This so called 'circumplex' (Russell, 1980) model of affect has been prevalent in both affective computing and musical emotion, and can be contrasted with basic or discrete models of emotion and models using more or different dimensions. In addition to its prevalence, other benefits include the continuous nature of its underlying data space and documented correspondence with discrete emotion models (Eerola & Vuoskoski, 2011).

The following two sonifications implement a collection of these cues, differing insofar as they have been designed in two fundamentally different ways. In the first, the desire was to create a mapping strategy that would be pleasant, ecologically-valid, and perceptually clear for all points on the AV space, such that it might even be usable in a concert setting (Winters et al., 2013). By contrast, the second sonification was designed computationally using software for music emotion recognition that uses a linear combination of nine underlying signal characteristics. After briefly discussing the details of the mapping strategies, they are evaluated in Section 4.4.

4.3.2 Ecological Design

The details of this model are presented in (Winters et al., 2013), and are summarized here. The foundation of the sonification is a resonant object created using the DynKlank unit generator in SuperCollider, a programming environment for realtime audio synthesis. By using modal synthesis, DynKlank can produce realistic sounds resembling physical materials (e.g. wood, ceramic, glass) through independent control of resonant modes, their amplitudes, and decay times. As with physical objects, to make sound, the object must be struck (i.e. ‘excited’). In this case, excitation always comes through impulse in alternating left-right stereo channels.

To convey emotion, the sonification uses tempo, loudness, decay, roughness and mode. Increasing arousal increases the speed at which the object is excited as well as the overall loudness of the sound. Decreasing arousal increases the length of decay time, the time at which it takes the amplitude of the sound to decay by 60dB. To convey valence, the original sound is copied and frequency shifted by major/minor third, perfect fifth, and perfect octave. As valence increases in magnitude, either positively or negatively, the amplitudes of the 3rd, 5th, and octave increase incrementally such that in a normalized AV space, the third reaches maximum amplitude at $V = \pm 0.5$, the fifth reaches maximum amplitude at $V = \pm 0.75$, and the octave at $V = \pm 1$. The third is major or minor depending on whether valence is positive or negative respectively.[1] Finally, the second quadrant of the AV space (i.e. low valence, high arousal) is conveyed using roughness. While within this region of the space, an identical copy of the original sound is pitch shifted up to 50Hz with radial distance from the origin, and is increased in amplitude with radial distance from the line $3\pi/4$.

4.3.3 Computational Design

The second model was designed with the goal of acoustically instantiating a computational model for music emotion recognition. The model chosen for this purpose was the MIREmotion function (Eerola et al., 2009) from the MIRToolbox, a MATLAB toolbox with many useful functions for audio-based music information retrieval. The MIREmotion function can generate emotion scores for each of five categorical emotion concepts (happiness, sadness, tenderness, anger, and fear), and three emotional dimensions (activity, tension, and valence). To determine each score, the model uses a linear combination of four to five audio-based descriptors, determined through a process of multiple linear regression on a database of 110 musical examples and a collection of 29 non-redundant features. Although three dimensions were available, Eerola et al. (2009) demonstrated moderate to high correlation between tension and the other dimensions, while the correlation between activity and valence was marginal. Reasoning that activity and arousal were closely related conceptually, manipulation was directed towards activity and valence.

In the MIREmotion function, activity is determined by the RMS, maximum value of the summarized fluctuation, spectral centroid, spectral entropy, and spectral spread. Valence is determined by the standard deviation of the RMS (σRMS), maximum value of the summarized fluctuation, novelty, mode, and key clarity. From these features, the computational sonification manipulates RMS, σRMS, key clarity and mode. These features were measured using 16-bit, 15-second wave files recorded from the sonification at desired data points in the AV space. To have the greatest degree of control over these features, the fundamental sonification strategy was simplified to a bank of three sinusoidal oscillators, creating a root-position closed major G-chord on G3. To control RMS and σRMS the sound as a whole was periodically amplitude modulated by a strictly determined square wave at 0.4Hz. To control key clarity and mode, the amplitudes of the third and fifth were increased

[1] At this point, it is worth mentioning that coincidentally, Schubert et al. (2011) suggested the same mapping of tempo, loudness, and mode.

or decreased in amplitude. The strategy for conveying valence varied with position in the normalized AV space: from -1 to 0 valence, σRMS was systematically decreased, from 0 to 1 valence, the key clarity, and to a lesser degree, mode were systematically increased. Increasing activity was conveyed by increasing RMS, but at no point was there digital clipping in any of the measured audio files.

4.4 Computational Evaluation

Both of the models in Section 4.3 were designed with the goal of conveying a continuous arousal and valence emotion space. As previously discussed, their mapping strategies vary due to differences in design goals and methods. The first model was designed using acoustic cues suggested by the psychological study of musical emotion, while the second was designed computationally using a publicly available tool for music emotion recognition, and specially designed software for analysis.

After presenting the software and the computational results, both models are evaluated for their expected utility in both emotional communication and musical emotion research. This comparison brings attention to limitations of computational evaluation, but also its benefits, and the ways in which these difficulties can be addressed.

4.4.1 Software for Analysis

For the purpose of evaluation, two GUI frameworks[2] were developed to analyze the output of the MIREmotion function on both individual and groups of soundfiles. Without these tools, the process of designing sounds is tedious: the default visualizations of the MIREmotion output do not indicate the contribution of the five underlying audio features to the emotion score, and do not represent these constitutive features in ways conducive to their systematic analysis and manipulation.

To analyze individual soundfiles, the ‘myemotion’ function visualizes the audio features determining the emotion score under analysis, including the magnitude of their individual contribution and distance from a reference point, usually ±0. A ‘play’ button in the upper left-hand corner allows the user to listen to the analyzed file, which is helpful for identifying distortions in the recording, or understanding the temporal evolution of measured features.

[2] Freely available [Online]: `https://github.com/mikewinters/MIREmotion-Visualizer`

A collection of radio-buttons allows the user to quickly change the emotion dimension or concept under analysis, though only the visualizations for activity and valence have thus far been implemented. To facilitate documentation, if the user creates a title for the graph, it is used to automatically export .eps, .fig, and a .wav file copy into a dated directory. A figure displaying the interface for activity is provided in Figure 4.1 and includes six graphs: one for each of the five constitutive audio features, and a bar-graph summary.

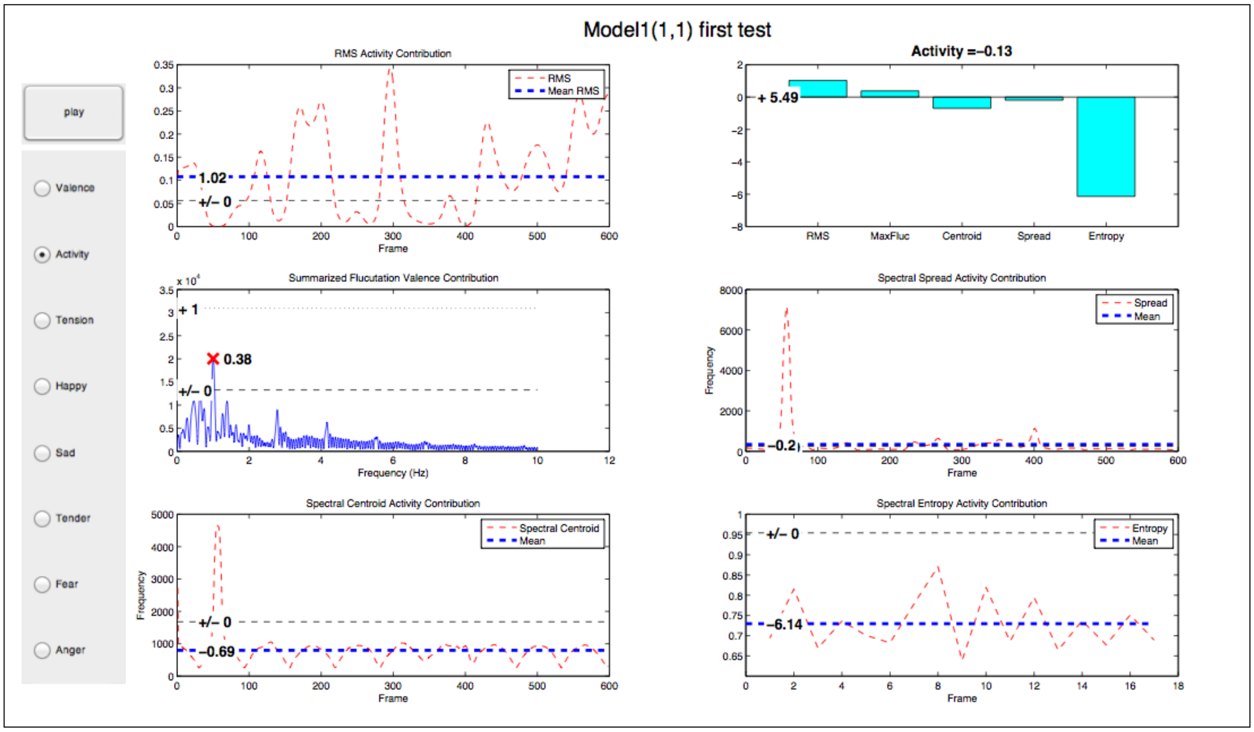


Fig. 4.1: A figure displaying an activity visualization using the myemotion function.

By contrast, the 'avmap' function visualizes the distance of multiple individual wave files to desired points in an AV space, and is designed for analysis of a mapping strategy as a whole. Positioned on a two-dimensional plot are the desired point (accumulated from the name of the wave file), the MIREmotion coordinate, and a line connecting the two points. Colored markers of different shapes help to differentiate the measured points. Adjacent to this plot are two bar graphs displaying in detail the five audio features contributing to each emotion score. Each includes a 'detail' button triggering the myemotion visualization for that dimension. Clicking on points of the graph makes their line-width and marker size bigger for visual feedback and changes the content of corresponding bar graphs. A unique title is generated for the two-dimensional graph indicating the Euclidean distance of all measured sound-files to their desired point on the graph. An example of an avmap visualization for 16 soundfiles is provided in Figure 4.2.

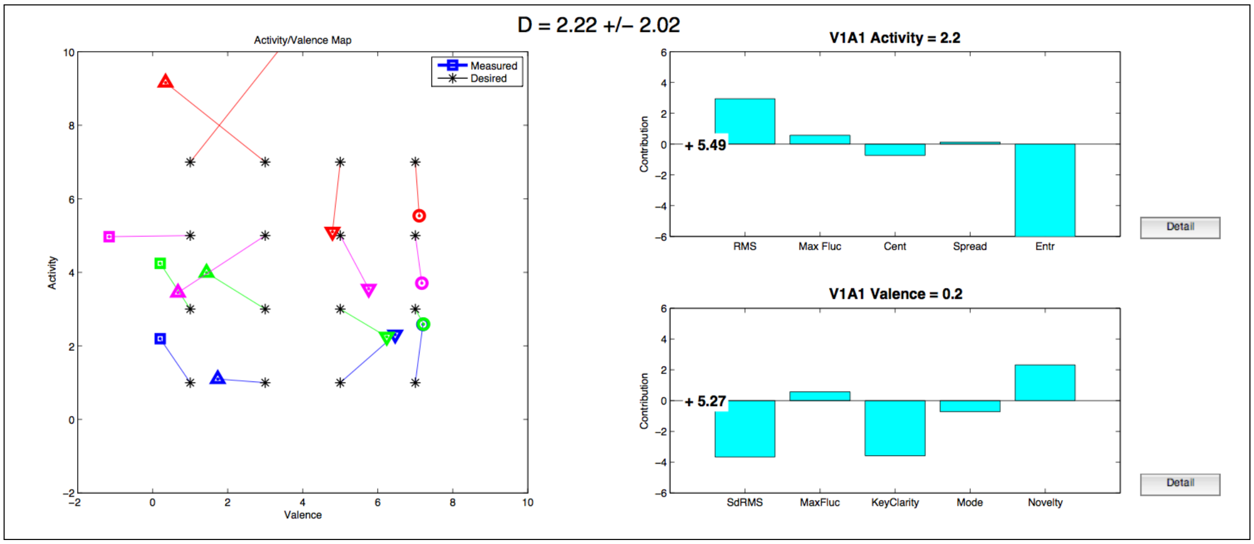


Fig. 4.2: A figure displaying the avmap visualization for a sonification of emotion. The long lines between colored markers and black stars indicates that the sonification does not conform well to the MIREmotion function.

4.4.2 Results

For computational evaluation, the MIREmotion function was applied to a collection of 49 15-second wav files recorded from evenly distributed points on each underlying AV space. Because the function was trained using a seven-point Likert Scale on the interval from $[1, 7]$, the collection represents all possible integer combinations of activity and valence. The time scale of 15 seconds was chosen to closely match the average duration of the Soundtrack110 data set used to train the function (Eerola et al., 2009).

Figure 4.3 shows the comparison of the two sonifications side by side. For the non-computationally designed model, the average distance d from the measured point (V_m, A_m) to the desired point (V_d, A_d) is $d = 7.13 \pm 1.02$. For the computationally designed model, $d = 0.11 \pm 0.10$, a difference factor of approximately 65.

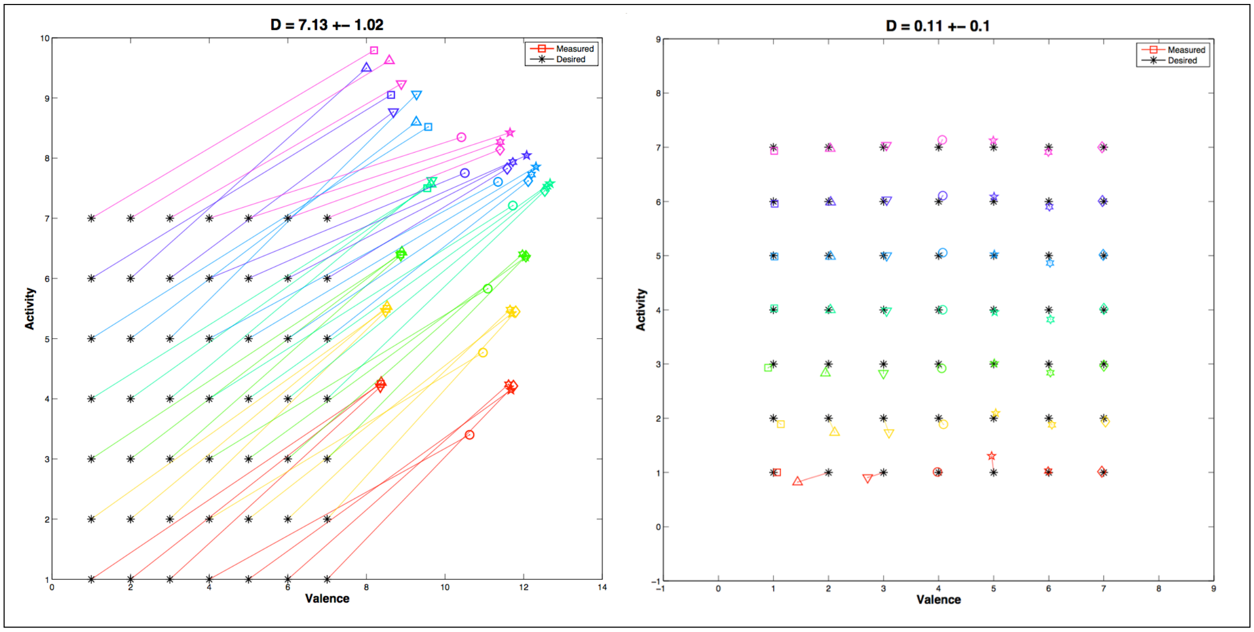


Fig. 4.3: A comparison of the two sonifications analyzed by the MIREmotion function using the avmap function described in Section 4.4.1. The ecological design (left) has larger distances between desired and measured AV coordinate than the computational design (right).

From visual analysis, it is clear that the computationally designed sonification closely matches most of the desired points in the MIREmotion function. The worst scoring point on the sonification corresponds to $(V_d, A_d) = (2, 1)$ with $d = 0.57$. In general, points (V_m, A_m) of poor performance are found in regions of low activity and valence. This issue stems from the inherent difficulty of creating points in this region for the MIREmotion function. Due to constraints of the model, the solution of a single sinusoid with strict control of both RMS and σRMS is one of few possibilities. These two audio features however are implicitly connected, making the systematic variation of V and A in this quadrant more challenging.

It is also apparent that the ecological sonification does not conform well to the MIREmotion function. There is a systematic offset of all measured coordinates to a space between $A_m \approx (4, 10)$ and $V_m \approx (8, 13)$, and for all points $(V_m, A_m) > (V_d, A_d)$. Points of equivalent A_d cluster together for $V_d = [1, 3]$ and $V_d = [5, 7]$, though the latter are systematically higher in valence than the former. Similarly, for every line of equivalent valence, activity incrementally increases from $A_d = [1, 4]$, and to a lesser extent from $A_d = [5, 7]$. In this light, the worst performance is in the region $A_d = [4, 7]$, $V_d = [4, 7]$, which clusters into a very small region between $V_m \approx (10, 12)$ and $A_m \approx (7, 8)$. In spite of these problems, it is interesting to note that the AV structure is more or less preserved. The distribution of points in Figure 4.2 for instance, is considerably more random.

4.4.3 Analysis

To compare these two models computationally is one method for evaluation. As evident, benefits include visual graphs (lending itself to visual analysis) and rapid evaluation. Computational models can also be used to direct mapping strategies, and increase the 'accuracy' of the sonification with respect to it. However, there are reasons why in the present case, it would be unwise to base evaluation exclusively upon this method.

In this section it is argued that in spite of its performance in the computational test, the ecological design would still fare better in the contexts of emotional communication thus far mentioned. The reasons for this include the abundance and type of acoustic cues, and the more 'natural' sound created by the synthesis. Mindful of these limitations, reasons are provided why the computational approach should continue to be applied in evaluation and design.

Limits of Computational Design

As demonstrated here, it is possible to design a sonification of emotion to almost perfectly match a computational model of musical emotion using a small number of acoustic cues. At this limit, changes in the mapping may no longer increase computational accuracy, though may still benefit emotional communication. Further, to attain the highest degree of accuracy, it might even be advantageous to use simple sounds (such as sinusoids or noise) to provide greater systematic control of the constitutive audio features.

Thus, though each model represents an underlying arousal/valence space using structural and acoustic cues shared with musical emotion, it is instructive to highlight reasons why in the present comparison, the ecological design would likely still be more useful for the communication contexts listed in Section 4.2.1. The first reason stems from the number and type of cues used in each model. Whereas the ecological design uses three cues to convey arousal (tempo, loudness, and decay), the computational design used exclusively RMS (loudness) for this dimension and maintained a constant speed of amplitude modulation (tempo) for the entire AV space. As for valence, similar strategies were used to convey high V (key clarity/majorness), but the two differed in their approaches to low V. The computational design used σRMS, and the ecological design used minorness and roughness, a difference not only in number but also in type. Though the use of minor mode was desirable for low V, the use of σRMS of a single sinusoid was dictated by model constraints discussed in Section 4.4.2. In either case, an abundance of cues is likely to have a greater emotional salience and/or magnitude than a singular cue. Using many cues also provides a degree of redundancy, which might be useful to users that attend to different qualities in the sound.

Besides for the cues, the ecological design also makes use of modal synthesis to create the underlying sound. This type of synthesis lends itself to creating 'naturalistic' sounds, which might resemble struck materials such as wood, metal, or glass for instance. On the other hand, the computational design uses a collection of three sinusoids, and for half of the space is limited to just one, centered on G3. The computational model has no mechanism for recognizing something like 'naturalness', yet from the environmental sounds discussion in Winters and Wanderley (2013), it is a feature that should be preserved, having demonstrated emotional salience and behavioral impact in sonic interaction design (Lemaitre et al., 2012). Similarly, the naturalness in the ecological design might be expected to be preferred to the sinusoids of the computational design, in turn benefiting the utility of the display for communication.

Benefits of Computational Design

Although in this case, the ecological design is predicted to perform better in contexts of emotion communication, there are many reasons why the use of computational tools for evaluation and design should continue. Beyond rapid evaluation and graphs, they provide a framework for design, one that is already systematically informed by listeners' emotional ratings. They are also valuable tools for music emotion research, acoustically instantiating an otherwise abstract mathematical model. The issues encountered in the present case originate in part from restrictions inherent to the model being used (i.e. constraints for low V, low A) and in part from the desire to clarify and address limitations of the computational approach.

That being said, more cues could be applied in the present computational design—specifically in areas not as restricted as the low V, low A quadrant. From the previous discussion, contributing more cues would be beneficial to emotional communication and computational accuracy may still be maintained. Though the problem of computationally recognizing 'naturalness' may persist, other computational models might be expected to be more sensitive to this feature, especially if the model was trained on listening tests including 'natural' and non-'natural' (i.e. sinusoids/noise) test sounds.

Further, neglecting these tools in sonification stymies collaboration with the field of musical emotion, an exchange this paper hopes to demonstrate as mutually valuable. As noted in Section 4.2.2, sonification offers musical emotion systematic and theoretically informed manipulations of acoustic cues. Although sonification by definition provides a systematic manipulation, and both models are theoretically informed, the computational model goes much further, acoustically instantiating an otherwise exclusively mathematical model of musical emotion and accurately covering a two-dimensional space. Though the ecological design uses suggestions from psychological studies, it follows no theoretical rules for their combination or implementation on an underlying AV space.

By providing this acoustic instantiation, results from listening tests can also be directly applied towards refining the model and extending its predictive power. Although in music emotion recognition the highest scoring classifiers can reach accuracy levels of $\approx 65\%$ (Kim et al., 2010), it is possible that future performance would increase if cognitive factors due to recognition or genre preference are minimized. By using sonifications rather than music, these models would also become more predictive of the success of a sonification design than if trained using strictly musical examples. Better tools lead to better sonification designs, and can further contribute to the understanding of musical and auditory-induced emotion more generally.

4.5 Conclusion

In this paper, the subject of sonification of emotion was addressed in detail. Contexts favorable to realtime accurate auditory display were identified and the benefit to musical emotion research was highlighted. To frame this research, the current state of emotion in sonification was presented including a reiteration of the necessary qualifications for a sound to qualify as a sonification of emotion. Strategies for display were presented that draw heavily upon research in musical emotion and target the auditory cognitive mechanisms of brain stem reflex and emotional contagion. Two sonification mapping strategies were then presented that use these cues to display arousal and valence, two underlying dimensions of emotion. Both were evaluated computationally using the MIREmotion function and custom software for analysis. The significant difference in the performance in this test reflected fundamental differences in their method of design. Though the computational design performed better, the 'naturalness' and the number and type of cues used in the ecological design called to question whether this accuracy would equate to better performance in emotion communication. Mindful of these limitations in the computational approach, its application in sonification of emotion was supported for future research.

In total, this research demonstrates how tools and research in musical emotion can be applied to research in sonification of emotion, and also how sonification might be beneficial to music research. In this reciprocal relationship, computational tools can be applied as a design metric, but listening remains of utmost importance. It is hoped that this research can help to establish the display of emotion as a worthwhile pursuit in sonification, a pursuit that can make use of the wealth of resources from music rather than be confounded by them.

Part II

Sonification of Expressive Gesture

Chapter 5

New Directions for Expressive Movement Sonification

Winters, R. M., & Wanderley, M. M. (2012, June). New directions for sonification of expressive movement in music. In *Proceedings of the 18th international conference on auditory display.* Atlanta, Georgia.

Abstract

Expert musical performance is rich with movements that facilitate performance accuracy and expressive communication. As in sports or rehabilitation, these movements can be sonified for analysis or to provide realtime feedback to the performer. Expressive movement is different however in that movements are not strictly goal-oriented and highly idiosyncratic. Drawing upon insights from the literature, this paper argues that for expressive movement in music, sonifications should be evaluated based upon their capacity to convey information that is relevant to visual perception and the relationship of movement, performer and music. Benefits of the synchronous presentation of sonification and music are identified, and examples of this display type are provided.

5.1 Sonification of Expressive Movement

Recent developments in auditory display have infused human motion with sound for the purpose of analysis, motor learning, and adapted physical activity (Höner, 2011). However, human motion is not limited to goal oriented movements like those frequently found in sports. In music for example, expressive (Delalande, 1988) or ancillary (Wanderley, 1999, 2002) gestures refer to movements that are not responsible for sound production, but nevertheless common in performance. Though complex and diverse—varying with the instrument, performer, and musical piece—these movements are otherwise highly consistent over time and reflect musical structure and expressive intention (Wanderley, Vines, Middleton, McKay, & Hatch, 2005).

The use of high-resolution motion capture systems has enabled the quantitative study of these movements. In a typical setting, a performer wears reflective markers that are tracked over time in three spatial dimensions using an array of calibrated infrared cameras. Due to the size and complexity of the data sets, sonification can be used to quickly browse through the data, make non-obvious relationships more apparent, and facilitate the process of data analysis.

5.1.1 Previous Work

The use of sonification for studying expressive gesture in performance began with a study of four clarinetists (Verfaille, Quek, & Wanderley, 2006) who were asked to play the same piece of music with exaggerated, normal, and immobilized playing modes. Though mapping choices were discernible and could be used to expose data relationships that were not visually obvious, the mapping was not easily extendible to other performers due to the high variability in the movement patterns between subjects.

A more recent work (Grond, Hermann, Verfaille, & Wanderley, 2010) has compared Principle Component Analysis (PCA) and velocity of markers as preprocessing steps for sonification in a bimodal context using a stickman visualization. Using an open task, they found that sonification would work well in directing the attention of the user to aspects of the visual display in the velocity based mapping, but not in the PCA.

5.2 A New Methodology

Gesture in music performance is a rich field for sonification, but the expressive nature of these movements warrants special consideration that is distinct from goal-oriented movements that are common in sports. What is more important than the exact positions or velocities of points and angles on the body are the higher-level structural and emotional information they carry. This information can be organized around the relationship of movement performer and music, and what the movements convey to the viewer.

5.2.1 The relationship of movement, performer, and music

Building upon a foundational work in the study of expressive movement (Wanderley, 2002), there are three levels of gestures that need to be conveyed in sonification, the material, structural, and interpretive. Material gestures are those that are defined by the instrument being played. For example, the cello is more limited in possible expressive movements than the clarinet, resulting in different movement patterns. For a good sonification, a listener should be able to identify this type of difference.

The structural level of gesture concerns the relationship to the underlying music. For instance, highly difficult passages of music often impede mobility while easy passages and phrase boundaries see an increase in movement (Vines, Krumhansl, Wanderley, & Levitin, 2006). Though each performer moves differently, these sorts of structural cues are important and should be clear in sonification.

Finally, the interpretive gestures concern the performer's unique interpretation of the piece and convey their structural and emotional representation. For a good sonification, a listener should be able to identify two takes of the same performer playing a piece of music and likewise perceive that a different performer has played.

5.2.2 The perception of movement in musical performance

In the perception of music, the visual context provides cues that can modulate the emotional and structural perception of a piece. For instance, simply viewing a performer can extend the perceived length of phrases and reduce or augment ratings of tension (Vines et al., 2006). In another study, Dahl and Friberg (2007) showed that the visual perception of regularity, fluency, speed, and amount of motion could predict the emotional ratings of happiness, sadness, and anger.

Results of (Dahl & Friberg, 2007) supported a possible invariance between viewing conditions, instrument, and musician. This invariance was supported by (Nusseck & Wanderley, 2009), who modified stickman avatars derived from motion capture data of real performers. Completely immobilizing the arms or torso, or even playing the avatar in reverse did not significantly effect judgements of tension, intensity, fluency, or professionalism. Increasing the amplitude of motion of the whole body was important however, implying this factor was more important than the movement of individual body regions.

If factors such as amplitude of motion are indeed more important to visual perception than the exact part of the body being moved, than it is wise that sonification of performers prioritize this cue. Additionally, if the regularity, fluency, and speed are important cues for conveyed emotion, likewise sonifications should focus on the ability to correctly display this information.

5.3 Sonification for Music-Data Analysis

New music research abounds with large, complex, time-varying data sets. For this data, sonification as a tool for analysis or display benefits from the shared medium of music and sonification. For gesture in particular, some of these benefits have already been identified by researchers using interactive sonification to teach bowing technique of the violin.

The first benefit, identified by Larkin, Koerselman, Ong, and Ng (2008), stressed that the shared temporal nature of music and the data could be used to understand data events as they occur temporally relative to the music. Later, Grosshauser and Hermann (2009) identified that for sonification and music research, listening is a familiar and widely used medium. Also, the shared acoustic medium could provide a more direct access to relationship of data and performance audio. For expressive gesture, this may provide a fuller display of the performer's expressive intension than the music alone, and may be closer to the performer's internal representation of the structural and emotional content of the piece (Winters, 2011c).

A benefit that has not yet been identified is that through sonification, the visual aspect of musical performance is made accessible to the blind (or those who cannot see). If a sonification design is able to convey the structural and emotional cues discussed in Section 2, then it is a display medium that can be used to make expressive gesture accessible through sound.

Videos hosted on the IDMIL website and Vimeo provide examples of this display type. In the first example, a performer's expressive gestures are sonified and presented with performance audio and video.[1] In the second example, sonification of the eigenmodes of a subject dancing to music (Toiviainen, Luck, & Thompson, 2010) displays four metrical layers that can be compared to the metrical layers of the music itself.[2] In both of these examples, sonification provides a dynamic display that conveys non-obvious information as well as the performer's unique representation of the piece.

5.4 Conclusions & Future Work

This article has argued that for sonification, expressive movement should be treated differently than goal-oriented movement. Evaluation should be based upon the ability to convey movement cues that are relevant to visual perception and that highlight the relationship of instrument, music, and performer. Pairing music and sonification has benefits for analysis and display that are unique to their shared medium. In this way, a successful sonification can make expressive gesture accessible and provides a more complete display of a performer's expressive intentions in the same medium as the performed music.

[1] IDMIL Sonification Project [Online]: http://www.idmil.org/projects/sonification_project

[2] Movement Sonification 2 [Online]: http://vimeo.com/42395861

Chapter 6

A Sonification Tool for Expressive Gesture

Winters, R. M., Savard, A., Verfaille, V., & Wanderley, M. M. (2012). A sonification tool for the analysis of large databases of expressive gesture. *International Journal of Multimedia and Its Applications, 4*(6), 13-26.

Abstract

Expert musical performance is rich with movements that facilitate performance accuracy and expressive communication. Studying these movements quantitatively using high-resolution motion capture systems has been fruitful, but analysis is arduous due to the size of the data sets and performance idiosyncrasies. Compared to visual-only methods, sonification provides an interesting alternative that can ease the process of data analysis and provide additional insights. To this end, a sonification tool was designed in Max/MSP that provides interactive access to synthesis mappings and data preprocessing functions that are specific to expressive movement. The tool is evaluated in terms of its ability to fulfil the goals of sonification in this domain and the goals of expressive movement analysis more generally. Additional benefits of sonification are discussed in light of the expressive and musical context.

6.1 Introduction & Motivation

In its most concise form, sonification is defined as the use of non-speech audio to convey information (Kramer et al., 1999; Hermann et al., 2011). Since it began as an international field of research in 1992 (Kramer, 1994), it has found continual application in many areas due to its highly interdisciplinary nature. New developments in sonification seek to display human movement patters in order to augment human performance (as in sports) or to provide a complement to visual-only analysis methods (as in rehabilitation, Höner, 2011). By comparison, it is rare that sonification is used to display expressive movement patterns such as those made in the performance of music.

Several important features of quantitative research in "expressive" (J. W. Davidson, 1993) or "ancillary" (Wanderley, 2002) gestures in music performance make analysis difficult. Primarily, motion capture systems generate extremely large amounts of data over relatively short amounts of time. Five minutes worth of data capture can take months to analyze. Further, unlike "effective" (Delalande, 1988) gestures, which are required for sound production, expressive movements can be highly idiosyncratic, dependent upon many factors including the instrument, the performer, and the piece of music. Techniques for analysis therefore benefit from flexibility. A technique that was fruitful for one performer often needs to be revised and translated in order to work for another. Movement patterns also vary across instrument and musical piece—each instrument has different ranges for potential expressive movement, and each piece of music has a unique structural and emotional character.

Though the visual display of expressive movement is intuitive and our visual facilities are well prepared for motion perception (Johansson, 1975), recent research has explored the auditory channel as an alternative or complement. By transforming the motion capture data into sound, researchers hope to benefit from capacities sound as a data-bearing medium. Proposed benefits include the enhanced perception of periodicity, regularity, and speed (Kapur, Tzanetakis, Virji-Babul, Wang, & Cook, 2005), non-obvious visual features and fast-transient movements (Verfaille, Quek, & Wanderley, 2006), abstracted orientation and attention (Grond, Hermann, et al., 2010), and similarities and differences between repetitive movements (Grond, Bouënard, Hermann, & Wanderley, 2010).

In this paper, a tool is presented for researchers interested in the potential of sonification as a complement to visual-only methods for the analysis of expressive gesture in music. A GUI provides the front-end of a synthesis and computation engine written in Max/MSP. The interface allows up to 10 synthesis channels to run simultaneously for any number of performers, all the while providing interactive access to several synthesis mappings and their modifiable parameters. After explaining the inner workings of the tool, it is evaluated in terms of the goals of sonification in this domain and the goals of expressive movement analysis more generally. New uses for the sonification of expressive movements are also presented.

6.2 Previous Work

Sonification as a tool for quantitative analysis of musicians' ancillary or expressive gestures was first demonstrated in Verfaille, Quek, and Wanderley (2006). The researchers used motion capture data from clarinetists as they played through Stravinsky's *Three Pieces for Solo Clarinet* in three expressive manners: normal, immobilized, and exaggerated (Wanderley et al., 2005). Four gestures were chosen for sonification: circular movements of the clarinet bell were mapped to pitch, body weight transfer to tremolo rate, and body curvature and knee bending controlled timbral attributes. Additionally, the velocity of each parameter modulated sound amplitude, and the weight transfer and circular bell movements were mapped to panning.

Although the mapping choices were discernable, they were not extendible to other performers due to the high degree of variability in movement patterns. The group suggested that interactive, realtime sonification would avoid this problem by allowing the user to adapt mapping and data conditioning settings for multiple display varieties. It was also clear that the addition of audio enhanced the perception of certain gestures (i.e. the clarinet bell) that were hidden in the corresponding video.

A later work has compared two different sonification techniques in a bimodal display with "stickman" visualizations (Grond, Hermann, et al., 2010). Gestures were sonified using either direct velocity sonification or Principle Component Analysis (PCA) as a preprocessing step. Data parameters were mapped using a resonant filter with logarithmically separated center frequencies between 150-4000Hz (velocity) and 300-2000Hz (PCA). Data fluctuations modulated the amplitude, center frequency, and bandwidth ratio of each frequency. To test the displays, an open task was created in which participants marked by mouse-click "events" that they encountered in the display. The visualization was presented with each sonification separately with and without audio. From their results, the authors concluded that velocity sonification was more efficient at exposing non-obvious visual features and was generally what users preferred for the context. They hypothesized that because the PCA did not have an obvious correspondence to the display, it was difficult to "connect" the two displays.

6.2.1 Current Trends

The analysis task chosen for evaluation in Grond, Hermann, et al. (2010) is valid, but is ultimately one of many possible use contexts. It is furthermore a context that benefits from bimodal integration, a feature that is best realized by redundant audio-visual information (Spence & Soto-Faraco, 2010). While not optimal for bimodal display, the PCA remains a useful preprocessing tool for expressive movement in light of its generalizability. Researchers in Toiviainen et al. (2010) for instance used PCA to compare "eigenmovements" across a pool of 18 participants as they danced to four pieces of music of different tempi. The PCA offered a way to abstract from each subject's individual movement patterns and thereby study global characteristics. In the design of a sonification tool, we posit that all useful tools should be made available, and thus both PCA and velocity sonifications are present.

New developments (Winters & Wanderley, 2012b) have reconsidered how a sonification system for expressive gesture should be evaluated. Instead of focusing on the perception of events, the authors argued, the sonification should be evaluated on its ability to convey higher level, more abstract features, specifically those that are important for the perception of emotion and structure. The present tool is evaluated in terms of this theory.

6.3 The Tool

Section 6.3 was written originally by Alexandre Savard, representing his thesis work. The text was expanded, edited, reorganized, and proofread prior to publication by R. Michael Winters.

6.3.1 Introduction

The tool was designed first and foremost to provide an accessible interface for researchers who do not necessarily have experience in signal processing, sound synthesis, or mapping. Following an offline preprocessing step in MATLAB, it can be used with any data format from any motion capture system, and can quickly be re-calibrated to each individual performer or data run. It allows six synthesis mapping options and up to ten independent mappings for each performer and playing condition running simultaneously. Six data preprocessing functions, including principal component analysis on individual body regions, provide features for analysis. The resulting sonifications can be integrated with video from the capture and easily recorded onto the desktop.

The GUI is displayed in Figure 6.1. For each data preprocessing option and synthesis parameter, there is a subinterface that allows the user to make fine adjustments in realtime. The output sonification of each channel is sent to the sonification mixer (bottom of Figure 6.1) that gives users the flexibility to balance the gain of different mappings on a global level and stress specific gestures.

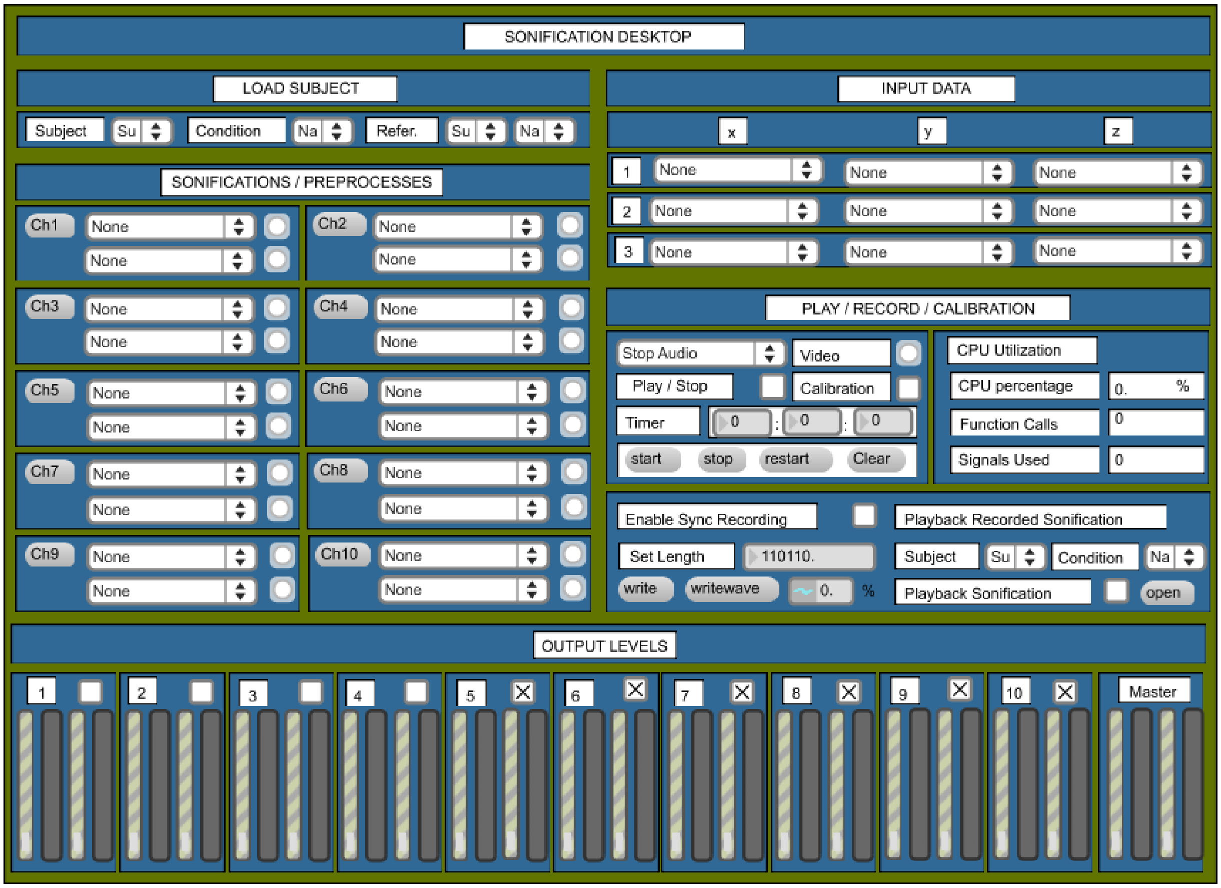
6.3.2 Offline Conversion

For every marker position (x, y, z), a MATLAB script converts the exported motion capture data format into a 100Hz WAVE file using the plug-in-gait biomechanical model. The plug-in-gait model is displayed in Figure 6.2 and provides a general model for marker placement that has been used previously for analysis of movement in performance (Chadefaux, Wanderley, Carrou, Fabre, & Daudet, 2012; Rasamimanana, Bernardin, Wanderley, & Bevilacqua, 2009). The MATLAB script is also used to extract global data parameters such as absolute minimum and maximum values.

The data is then sent to Max/MSP, an audio synthesis graphical programming environment that is designed to process audio data in realtime. In Max/MSP, priority is given to audio processing above data and event processing, so to ensure synchronization between video and sound, the system processes both gesture features and sound synthesis during audio processing callbacks.

6.3.3 Data Preprocessing

For each of the 10 synthesis channels, the user can choose between six data preprocessing functions and the PCA on five individual body regions. Between the six non-PCA options, three are general functions and three are specific to common expressive gestures in music. The three basic options are normalized raw data, Euler distance, and Euler angle. The raw data option uses a marker's x, y, or z position for analysis, the Euler distance provides the distance between two markers' x, y, or z position, and the Euler angle provides the angle between three markers' x, y, or z position.



Timer 0 : 0 : 0

CPU percentage 0. %

Function Calls 0

Signals Used 0

Set Length 110110.

0. %

Fig. 6.1: A screenshot of the sonification desktop. The desktop is the main interface from which users design and manipulate gesture sonification. Usual navigation controls (start, stop, timer) are provided to control playback, and a switch enables the system to recalibrate for different performers. The various data and sound processing techniques are regrouped into several sonification channels. Several menus allow for the selection of data, processing algorithms, sound synthesis and calibration preferences. For a given data or sound process, specific "subinterfaces" can open to modify parameters related to these processes. Sonifications and the control signals that generated them can be saved as PCM audio files (.wav format). Recorded sonifications can be reloaded as well.

The remaining three in the set were designed for gestures that often occur in musical performance (Dahl et al., 2010). These include circular movement, body curvature, and weight transfer, and can be applied to any section of the body. Circular movement describes the degree of circular movement of a marker relative to another marker. In the example of a clarinetist, circular movements of the clarinet bell are often highly indicative of expressive

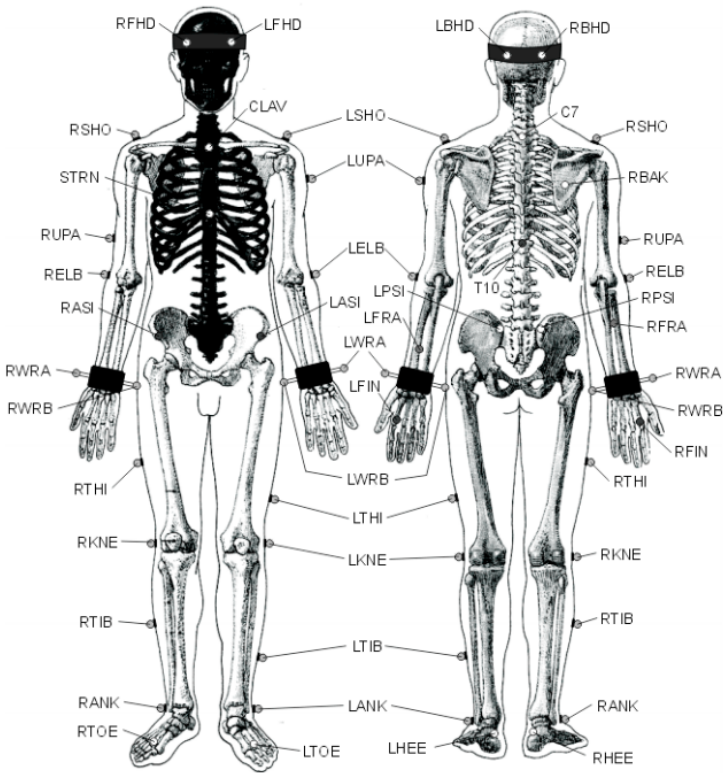


Fig. 6.2: A figure showing the marker placement for the plug-in-gait marker biomechanical model from the IDMIL website.

intention (Wanderley, 2002; Wanderley et al., 2005) and should be measured from the tip of the clarinet bell to a marker located near the mouthpiece. Body curvature is another gesture that is highly expressive in performance. In dance for instance, the extent to which limbs are contracted or expanded with relation to the body has been shown to be predictive of the expression of fear and joy (Camurri, Lagerlöf, & Volpe, 2003). Weight transfer is the last non-PCA preprocessing function available for sonification. It computes the weighted mean position of a set of three markers, and when applied to the torso, can indicate core swaying, fundamental to expression in woodwind and piano performance (J. Davidson, 2012).

6.3.4 Data Reduction

Due to the size of the data set, special care was taken to provide options for data reduction. For this task, principal component analysis (PCA) is used to combine the input signals into the most limited subset that maximizes parameter variance while preserving the most information. A detailed description of the mathematics behind the PCA can be found in Ramsay and Silverman (2005); Daffertshofer, Lamoth, Meijer, and Beek (2004), but the basic idea is to combine information that demonstrates high covariance within the data set in a two-step algorithm that includes the eigenvalue decomposition process and the linear combination reconstruction process. The outputs of the PCA are components that represent a reduction of the data set to its standard basis. Recently, the PCA (Toiviainen et al., 2010) and other projection techniques (Naveda & Leman, 2010) have been used formally for expressive movement analysis. These techniques successfully reduce data quantity but are still able to describe the majority of movement. For instance Toiviainen et al. (2010) found that the first five components of the PCA accounted for an average of $96.7 \pm 2.3\%$ of the variance in eight-beat sections of music-induced movement. In informal investigations involving ancillary gesture motion capture data in clarinet performance, the first three principal components are clearly associated to the motion of the center of mass along the three main axes and are able to describe 85-90% of the total marker movement. The remaining principal components describe less dominant gesture features.

PCA on specific body regions

For the desktop, PCA is available on five local regions of the body independently: the head, the upper trunk, the lower trunk, and both legs. These data sets were augmented to reinforce specific articulations by using derived gesture features such as angles between markers, improving correlations within subgroup markers. From the insights of the PCA on clarinet performers carried out in Savard (2009), it was found that several signals could be discarded as they represent redundant information that do not convey any additional significance of their own. These results are shared presently as they demonstrate the utility of the PCA for data reduction and can potentially be generalized to other instruments. From the plug-in-gait model, the most important parameters were found to be:

- Head mean position

- 7th cervical and 10th thoracic vertebrae (C7 and T10)
- Pelvis mean position
- Left and right knees
- Left and right wrists and angles
- Head orientation
- Spine angle
- Pelvis orientation
- Left and right knee angles

Which would correspond to a reduction from 165 signals to 33 signals. Formal results of a similar nature might be found for other instruments in the future, but for the present tool all marker positions are available.

6.3.5 Normalization

Within the preprocessing functions, normalization can be used to provide proper input for sound synthesis controls, to enhance specific gestural features, and to allow for proper inter-performer comparison. The tool allows the user flexible inter-gesture and inter-performer normalization.

Inter-performance normalization

Gestures range in magnitude of displacement. For example, the motion of the foot as it taps to the beat has a smaller range of motion than the bell of the clarinet. Similarly, the gesture feature extraction algorithms used presently produce several ranges of information. The magnitude difference between gestures is not conserved in the normalization process as all of the principal components have the same maxima and minima. Although this favors the motion of smaller gestures, these are precisely what cannot be conveyed well visually.

Intra-performance normalization

Given a selection of gesture features, both the comparison between different performers and the comparison of different performances require normalization for each gesture type. This step is required for the proper comparison of performances and their relative gestures' velocity. The largest displacement must be the maximum displacement for all data sets, and the relative amplitude of each gesture must be conserved.

6.3.6 Signal Warping

Prior to synthesis mapping, the user can apply warping techniques in order to perceptually enhance or attenuate gestural features to facilitate analysis. Inspired from Verfaille (2003), the following are examples of situations where data modification would be suitable for the sonification of gestures:

1. Attenuate unimportant gestural features that have been amplified or increase important information that has been attenuated through normalization

2. Enhance variation within a signal to emphasize different characteristic positions

3. Warp the signal in order to exploit the full range of a sound synthesis parameter

The normalized input signals $x_i[t] = [0, 1]$ are modified using a transfer function H_i stored in a lookup table $y_i[t] = H_i(x_i[t])$, which can be modified by the user through a subinterface. As in Verfaille, Wanderley, and Depalle (2006), signal warping functions are chosen according to the physical behavior they model into the signals. The warping techniques implemented allow the user to accurately quantify the modification applied to the signals in realtime.

6.3.7 Data Thresholding

One more data modification procedure is acceptable in this context. Continuing the list from Section 6.3.6:

4. Threshold the data in order to filter out undesired information

For each preprocessing feature, velocity information is also extracted. However, in the evaluation of velocity, noise in the signal can obstruct the general motion of the marker making the desired precept less salient. Filtering out this information leads to a better sonification. Threshold could hypothetically be applied to other parameters, but the application to velocity provides an example where this conditioning technique is almost always required. For thresholding, every value of an input signal $x[t]$ that is below a certain threshold is set to zero. To conserve the original range [0, 1], the thresholded signal is stretched to fill the proper range.

Known Issue with Thresholding

Truncation is not a benign procedure, and without properly altering the mean value or range, low-amplitude gestures can be significantly reduced compared to larger gestures. For certain synthesis or mappings, thresholding reduces saliency of these gestures. This situation becomes difficult in the comparison of performers with different expressive intensities. In practice, other data modifications will be necessary to compensate.

6.3.8 Mapping

The basic mapping strategy is based on previous success in Verfaille, Quek, and Wanderley (2006), which presents more detail concerning implementation. Although the user has flexibility in the gesture to sound choices, a reference example based upon the previous work with clarinetists is presented in Table 6.1.

Table 6.1: A table displaying a mapping strategy available in the tool based upon clarinet ancillary gesture. The user can chose what data parameters are mapped to the available sound parameters, and the displayed mapping represents one possibility.

Data Parameter	Sound Parameter
Body Curvature	FM synthesis modulation index
Weight Transfer	Sinusoidal beating effect frequency
Clarinet Circular Motion	Pitch shifting (Risset's infinite loop)
Knee Bending	White noise low-pass filter cutoff frequency
Weight Transfer	Left-right spatialization

For this example, the parameters in Table 6.1 can also be separated by pitch to maximize perceptual segregation. The sinusoidal beating effect could be placed in the highest range, FM synthesis and Risset's infinite loop (Risset, 1969) in the middle range, and the white noise low-pass filter to the lowest range. To further increase segregation, synthesis parameters can be paired with left-right spatialization of the weight transfer. This mapping is provided as suggestion, though others can certainly be implemented.

As suggested in Verfaille, Quek, and Wanderley (2006), gesture velocity, or more exactly the gesture feature derivative, is linked to the sound amplitude. It follows an ecological approach (Gaver, 1993b, 1993a) to the relation between sounds and kinetic events in a bimodal presentation. Loud sounds are produced by high-energy events in the environment and are therefore associated with high velocity. By contrast, absence of motion results in no velocity and zero sound amplitude. This approach was demonstrated successfully in Grond, Hermann, et al. (2010), who found that for a bimodal display, velocity sonification was better than PCA for drawing attention to movement "events." Fundamental to the sonification tool is a degree of flexibility in what data features are mapped to the available sound synthesis parameters. A variety of combinations of one-to-one and one-to-many mappings are available, and the user is able to make decisions that best fit their data set and stage of analysis. Although Table 6.1 presents one successful mapping strategy, the user may find other choices useful. For example, sound amplitude modifies synthesis parameters according to the magnitude of the gesture velocity. To gather information about instantaneous position, this feature should be turned off. At a later stage, by recoupling with sound amplitude, the sound conveniently refers to the gesture velocity again.

6.3.9 Integration with Video

If video was taken during the motion capture session, this can be easily played in the interface. Video and sonification are controlled by the same global play and pause controls, allowing ease in synchronization. The video can be muted to make the sonification as clear as possible, or alternatively, un-muted to listen for the expressive gestures as they align with musical structure. Both of these listening types may bear fruitful results. The availability of video is also meant to guide the user to effective mappings for each performer. For example, choice of markers and preprocessing function might be adjusted for a new performer, and the video can quickly guide the user to the necessary changes. An important benefit of sonification however, is that it can provide a level of detail that video analysis alone cannot.

6.3.10 Examples and Source Distribution

Example sonifications and videos created using this tool can be found on the IDMIL website.[1] The website also features documentation, the most current Max/MSP source-code, an example data set, and instructions for use. Further detail concerning implementation, mapping, and data preprocessing can be found in (Savard, 2009).

6.4 Evaluation

The interface was designed for the use of sonification as a tool for expressive movement analysis. It is presently discussed and evaluated in terms of its ability to fulfill the goals of sonification in this specific domain and its utility for expressive movement analysis more generally.

6.4.1 Goals of Sonification

For sonification as an analysis tool for expressive movement in music, there are three motivating goals that are common in the literature (Verfaille, Quek, & Wanderley, 2006; Grond, Hermann, et al., 2010; Grond, Bouënard, et al., 2010):

1. Sound is ideal for representing patterns in large, dynamic, multivariate data sets with fast, complex, or transient behavior (Barrass & Kramer, 1999).

2. Sound requires neither a particular orientation nor directed attention, making non-obvious visual features more salient (Pauletto & Hunt, 2004).

3. Sound provides a way to reduce the "cognitive load" of a purely visual display and/or allow more information to be perceived (Barrass & Kramer, 1999).

For the first point, the tool offers the ability to quickly "browse" through large databases of motion capture data, determining global information. For example, if a performer was asked to play through a piece multiple times in different expressive styles, the sound of their motion in each condition remains more similar to itself than to other performers. For

[1] IDMIL Sonification Project [Online]: `http://www.idmil.org/projects/sonification_project`

expert performers in particular, “expressive,” “standard,” and “immobilized” performances generate movement patterns that differ primarily in amount of movement while the type of gesture remains mostly similar (Wanderley et al., 2005). Directing attention to the subtle acoustic differences between each style can quickly guide the users to gesture features worthy of further analysis.

For the second point, because the tool allows users to display up to 10 sonification channels for each individual performer or condition, sound can be used to quickly change point of view by altering preprocessing steps and controlling the relative gain of any of the sonification channels. Furthermore, most of the data preprocessing functions offer “views” into the data that are not obvious from the video. For instance, the Euler distance between the knee and the toe can be sonified for both legs and mapped to the left and right stereo channels. This technique highlights these gesture features, re-orienting the user to the degree of their correlation.

In the final point, the tool reduces the “cognitive load” to a degree, but is not meant to be a replacement for visual analysis. By providing flexible access to multiple orientations through the preprocessing functions, gesture features worthy of further analysis can quickly be determined for the whole data set, directing visual analysis to features for further study. As will be discussed more in Section 6.4.2, pairing sonification with the performance audio/video allows the user to listen for important gestures as they occur within the musical structure.

6.4.2 Goals of Expressive Movement Analysis

The sonification tool was designed for analysis of gesture in music. By using the plug-in-gait model, it is also optimized for gross motor as opposed to fine motor analysis. An important distinction in performer gestures are those that are effective and those that are ancillary (Wanderley, 2002). Generally speaking, ancillary gestures are movements not explicitly required for playing the instrument and are usually expressive either to the performer or the viewers. By contrast, effective gestures are required for note generation. Several sonification systems have been designed for analysis or motor learning of effective gesture involving one instrument (Grond, Bouënard, et al., 2010; Larkin et al., 2008; Grosshauser & Hermann, 2009). To our knowledge, this system is the first to provide a general tool specific to the analysis of ancillary gestures across instruments and performers.

Sonification of expressive movement in musical performance bears some similarities to sonification of human movement in general, but with important differences. Fundamentally, for the analysis of expressive movement, there is a high degree of variability in movement between performers, instruments, and musical pieces. The sonification tool presented currently meets these challenges by providing increased flexibility in analysis through interactive mappings (Pauletto & Hunt, 2004), which had originally been suggested for expressive gesture analysis in (Verfaille, Quek, & Wanderley, 2006). With the tool, users can experiment and explore different mappings and preprocessing functions to quickly adjust to different performers. Furthermore, the array of ten mutable channels allows mappings that are meaningful in different sections of the same performance to be dormant until un-muted.

Additionally, while movement can be optimized in sports activities or rehabilitation, leading to measurable performance increase, for expressive movements, optimization is not always well defined and a gesture's expressive quality and context become important points for data analysis. As suggested in (Winters & Wanderley, 2012b), a tool for analysis of expressive movement should be able to convey features important to the perception of structural and emotional content. Expressive movement patterns can be structural when they reflect the properties of the instrument being played, the music itself, or the performer's unique interpretation (Wanderley, 2002). This typology of gesture in music is well-established in the field (Wanderley et al., 2005; Wanderley, 1999), and is useful for categorizing the diversity of movements that can occur in performance. The six non-PCA preprocessing functions convey these structural parameters. For instance, by choosing wisely, a pianist and a violinist can be acoustically distinguished and compared to one another as they play through a sonata. This analysis can be used to determine the subtle gestural cues used in real performance to communicate between performers.

Outside of these structural features, expressive movements carry visual information important to perception of expressive and emotional intention. For instance, gestural differences between staccato and legato notes on a mallet instrument can affect perceived duration (Schutz & Lipscomb, 2007); perceived force of impact can change the perception of loudness; and happiness, sadness, and anger can be characterized by the speed, regularity, fluency, and amount of motion (Dahl & Friberg, 2007). After using the video to optimize the sonification for each performer, the velocity to amplitude mapping and the PCA can be used to convey these features. The velocity can quickly indicate the speed, regularity, and fluency, but the position based preprocessing features can also be useful.

As in Toiviainen et al. (2010), the PCA on the five individual body regions can be used to compare across performers by creating a generalized abstraction.

6.5 Broader Discussion

6.5.1 Listening to Music and Movement

In the present case, sonification is used to convey information about expressive movements made in music performance. Although music can carry structural and emotional information, the movements made by experts during performance can carry structural and emotional content as well. Using sound to convey this type of information provides not only a useful data analysis tool, but also a shared medium for display that can be directly compared to the dynamic character of the underlying music.

The benefits of synchronous presentation of sonification and music were first identified in the mapping of effective gesture for learning the violin. By providing realtime acoustic feedback of the bowing features, Larkin et al. (2008) used sound to help teach bowing technique in string instrument training. Similarly in Grosshauser and Hermann (2009), different sonification approaches were evaluated in terms of their ability to support violinists in learning bowing technique. The authors identified the following benefits of this display type:

1. There is a temporal [relationship] between musical events and data fluctuations (Larkin et al., 2008).

2. Sound provides a medium that is familiar and widely used by musicians (Grosshauser & Hermann, 2009).

3. Sharing the same acoustic medium provides direct access to the relationship between data parameters and the underlying music (Grosshauser & Hermann, 2009).

The three arguments also apply for the analysis of expressive gesture. For expressive gesture, each performer's movements are directly related to their unique structural and emotional representation of the music being performed. Thus, when a performer moves more at phrase boundaries as noted by Vines et al. (2006), this is indicative of their expressive and structural intention. The first point suggests that analysis of expressive gesture becomes most meaningful when the data representation (whether visual or auditory) is temporally matched with the music. Music and sonification are both mediums that evolve temporally, and their temporal relationship is best exposed through synchronous presentation.

The second point posits that sonification is a well-suited data presentation medium for musicians and perhaps music researchers in general. For this community in particular, listening is already strongly associated with research progress and development, and research or performance insights often come through listening. Introducing sonification as a means of data analysis on the movements of performers during performance might find a more hospitable audience here than in other research areas where listening is not as fundamental.

The third point builds upon the temporal matching and listening qualities explained in the first two points. Assuming, as many researchers do, that the emotional and structural content of a musical piece are expressed or somehow mirrored in a performer's movements, the music being performed is not only a point of reference, but necessary for a holistic understanding of the collected data. By listening to sonification and music, a researcher can use the shared medium of sound to integrate a performer's movements in terms of the unique character of the underlying musical piece being performed. Furthermore, considering an expert performer's intimate understanding of sound—fundamental to their practice and performance—the medium of sound may be closer than visualization to the performer's unique cognitive and motoric representation of the piece they perform, contributing to a more meaningful analysis.

6.5.2 Making Visual Performance Accessible

The previous section discussed the benefits of synchronous presentation of expressive gesture with the underlying performance audio. The three arguments for this display type were shared between effective and expressive gesture. However, a fourth benefit of synchronous presentation is specific to expressive gesture (Winters & Wanderley, 2012b):

4. Visual expression in music performance is made accessible to the blind (or those who cannot see).

Although the tool is primarily designed for research, it can also be used to provide a display of a performer's movement for the blind or those that cannot see. As discussed in the third point, the gestures made by performers in performance are important for emotional and structural communication, but are currently only available visually. Sound offers a way to convey this additional content, and the integration of the two mediums may in some cases provide a more profitable listening experience.

As discussed in Section 6.4.2, the sonification tool can be used to make instrumental gestures sound different due to their expressive ranges and be used to target emotional and structural movement cues. Applying this tool to music listening might augment the musical experience by expressing information to the audience that had previously only been accessible visually.

6.5.3 Aesthetic Issues

The kind of listening involved with this display type raises two important issues in the relationship of sonification to music. The first addresses the aesthetic of listening discussed in Grond and Hermann (2011), which identified the types of listening involved in sonification. Though sonification is not music, as the authors argue, it is a scientific and aesthetic practice that can transcend either discipline. By creating a tool designed for analysis of expressive information, it is possible to listen to movements that are inherent to the emotional and structural features of a musical piece. When presented with both music (an expressive medium) and sonification (an objective, data-bearing medium), how do/should we listen? A secondary question, developed by the discussion of listening in the previous two sections is how should a sonification mapping be designed to integrate music as a point of reference or augment the experience of music?

To this end, we provide reference to distinct examples demonstrating the simultaneous presentation of sonification of movement and the corresponding performance audio. An example from previous work in clarinet performance (Verfaille, Quek, & Wanderley, 2006) is provided on the IDMIL website,[2] and another[3] presents a movie of "stickman" avatars dancing to music with PCA sonification as a preprocessing step (Toiviainen et al., 2010). In the latter example, the rhythm and temporal alignment of the movements are acoustically emphasized, allowing the listener to perceive multiple "eigenmodes" or rhythmic layers in the movements. A listener can perceive not only the number and strength of each layer, but also the degree to which each is aligned with the tempo and rhythmic layers of the underlying music.

[2] IDMIL Sonification Project [Online]: `http://www.idmil.org/projects/sonification_project`

[3] Movement Sonification 2 [Online]: `http://vimeo.com/42395861`

The second aesthetic issue deals with sonification's relationship to the musical mapping of gesture. As motion capture technologies have become increasingly available, the uses of human motion in music composition will likely only increase in prevalence. The diversity of such techniques can be clearly seen in the new interfaces for musical expression conference[4] where gestures are commonly used as control parameters in new interfaces. Similar to the movements the sonification tool was designed to convey, these gestures carry expressive and emotional information (Nakra, 2000). However, although sonification can be listened to musically, unlike these musical mappings, the main goal of sonification is not to create music, but to convey data relationships. Some recent works (Fabiani, Dubus, & Bresin, 2010; Goina & Polotti, 2008) have used the term 'sonification' ambiguously, and as the tool presented currently is intended for sonification, Table 6.2 is presented to clarify the differences between the two. Further discussion of these is provided in Chapter 2 of Savard (2009).

Table 6.2: A table displaying distinctions between musical mapping of gesture and the sonification of gesture.

	Musical Mapping of Gesture	**Sonification of Gesture**
Input Data	Body Movements	Body Movements
Motivation for Mapping	Create Music	Convey Information and Perform a task
Is it interactive?	Yes	Yes if it facilitates the task. Otherwise, no.
Is there an interface?	Yes	Yes
How do/should we listen?	Musically	For data relationships
What increases with practice?	Expression	Ability to determine data relationships
Is there a performer?	Yes	No

6.6 Conclusion

For the analysis of expressive gesture in music, the high degree of variability created by different performers, instruments and music makes data analysis challenging. Sonification provides a complement to visual display methods that can be optimized to quickly browse through these large and complex data sets and expose data relationships that were not visually obvious, facilitating the task of data analysis. A tool was presented for researchers working with motion capture data that are interested in using sonification, but without a

[4] NIME [Online]: `http://www.nime.org`

specific knowledge of programming, signal processing, or sound synthesis. Its main features include:

- Preprocessing features specific to expressive gesture
- A simple recalibration process
- Capacity to easily switch between performers
- Ability to play sonifications at different speeds
- Flexible, interactive mapping options
- Simple integration with video and performance audio

The tool was evaluated in terms of the goals of sonification for movement analysis and goals specific to the analysis of expressive gesture. Example contexts were presented in which the tool could be used to address these desired functions. The integration with performance audio and video that is provided by the tool emphasizes sonification's complementary nature, and optimizes the use of sonification by directing the user to appropriate preprocessing and synthesis mappings for each performer.

As contemporary music research is a quantitatively rich field, sonification in this domain will no doubt continue to develop. When sonification seeks to convey information that is expressive and inherently connected to music—as in the case of expressive gesture—synchronous presentation of sonification and music provides additional benefits for analysis and display. Designing sonifications that can use music as a reference or augment the experience of music is an interesting challenge for future work.

Part III

Sonification of Symbolic Music

Chapter 7

High-Speed Sonification of Pitch in Large Corpora of Music

Winters, R. M., & Wanderley, M. M. (2012a, April). *High speed sonification of pitch in large corpora of music.* Input Devices and Music Interaction Lab. (In Preparation)

Abstract

Sonification is defined as the use of sound to convey information. While it has been used in many fields for a variety of tasks, its use in Music Information Retrieval (MIR) is often tacit and has little formal development. In this paper, a technique for high-speed pitch-based sonification is introduced as a way of exploring large corpora of classical music. The technique is applied to the analysis of pitch transcription algorithms by playing ground-truth and transcription in separate stereo channels and augmenting divergences with slight amplification. Using a group of 11 participants, the technique was tested using artificial "pitch-transcriptions" in which notes in the original corpora were deliberately altered according to experimentally contrived probability distributions. Results from the test showed that the technique could be quickly learned and used to order sets of nine four-second sound-files by number of transcription errors in three corpora (Monteverdi's Madrigals, Bach's Chorales, Beethoven's String Quartets) and three speeds ($10^2, 10^3, 10^4$ notes/second). Additional benefits of listening are discussed that transcend simple error counting.

7.1 Introduction

The use of sound as an information bearing medium is fundamental to human interaction with the world (Hunt & Hermann, 2011). Sonification as a field of research explores ways in which sound can be used to transform data relationships into perceived relationships for the purpose of analysis or display (Kramer et al., 1999; Hermann et al., 2011). Common examples include the geiger counter and submarine sonar systems, but recent examples have grown in diversity and complexity including helicopter flight analysis (Pauletto & Hunt, 2004) and adapted physical activity (Höner, 2011). Sonification has been most successful in situations in which listening to data can provide insights that are difficult to see, or where it can provide a useful alternative to visual display (Barrass & Kramer, 1999).

Central to MIR is the link between data and musical reference. Unlike any other field, data sets commonly refer to notes, durations, chords, instruments, or other musical attributes. As some have argued (Ferguson & Cabrera, 2009), this relationship would seem to make sonification a clear choice for researchers. However, visual displays are by far the most commonly used and sonifications, when they are mentioned (e.g. Ewert, Müller, & Grosche, 2009), tend to be presented for the purpose of display of final results rather than as an integrated research tool. Although the benefits of visual display are many and should not be dismissed, sound is a rich medium for information transfer that can bring the user closer to their data, provide unexpected insights, and be efficient in data analysis.

Working towards this goal, researchers have introduced a sonification technique for display of spectral information and other audio features (Ferguson & Cabrera, 2009; Cabrera & Ferguson, 2006, 2007; Ferguson, 2009). By contrast, this paper offers a technique for high-speed note-based sonification with application to analysis of the performance of pitch transcription algorithms. In the technique, ground truth and transcription are compared at 10^2, 10^3, and 10^4 notes per second by playing through each in a separate stereo channel simultaneously and supplementing transcription errors with loudness cues. After detailing the methods for generation of test data and sonification, the results of a user-test involving 11 participants are presented. The paper concludes with a discussion of the benefits of sonification and other avenues for research in the use of sonification in MIR.

7.2 Data Generation

7.2.1 Extraction from music21

Using the built-in corpora of music21 (Cuthbert & Ariza, 2010),[1] data for sonification was drawn from Bach's Chorales, Beethoven's String Quartets, and Monteverdi's Madrigals using methods native to music21. For each of the three sets of resulting MusicXML files, a Python script was written to parse and extract data.

The script first parsed the MusicXML file, translating it into a music21 stream, the internal score representation in music21. Each score was then transformed into a "flat" representation using the .flat method, which translates any vertical sonority into a horizontal stream with the lowest sounding note played first (e.g., a root position C major chord in four parts becomes the stream [C3, G3, E4, C5]).

Each note within the stream was then converted to its MIDI value and appended to a separate list that held all notes extracted from the corpus in order. This list was exported as a CSV file that was imported into SuperCollider. Using this method, the Monteverdi Madrigals recorded 42,190 notes, Beethoven String Quartets had 167,941 notes, and the Bach Chorales had 125,301 notes.

7.2.2 Generating Transcription Errors

For this study, the "pitch transcription algorithm" is a copy of ground truth in which randomly chosen pitches are deliberately altered using probabilistic error distributions. As opposed to using actual pitch transcription algorithms, this method allowed transcription errors to be distributed arbitrarily, increasing experimental control.

Within each copy of the three corpora, a coin-flip method in SuperCollider was used to select notes from the ground-truth for modification. Once modified, the new note replaced the old note in the copy. The probability of note discrepancy between ground truth and "transcription" was fixed to represent the range of probabilities $p(n)$

$$p(n) = \frac{1}{2^n} \text{ where } n \in [0, 1, ..., 14], \tag{7.1}$$

where n is an index that is varied to produce a desired probability of transcription error.

[1] Music 21 [Freely Available Online]: http://mit.edu/music21/

For instance, a given note in the Bach chorales at $p(4)$ had a 1 in 16 ($1/2^4$) chance of being selected for modification.

When a note was chosen for modification using this probability scheme, it was transposed from the original using a gaussian distribution centered around the chosen note and rounded to the nearest integer. The gaussian distribution used in the perceptual test was fixed to have a standard deviation of $\sigma^2 = 6$ notes. Admittedly this error distribution is unlikely in real-algorithms, which are more likely to generate octave errors for example. However, the method was reasoned to be extendible in virtue of the small frequency separation, which due to perceptual grouping principles (Bergman, 1990) may be less salient than larger frequency separations. By consequence, human performance using the sonification technique on real-transcription algorithms might be expected to be better than for the artificial transcriptions described presently.

Transcription errors were created using the method discussed in Equation 7.1, but for the perceptual experiment, a subset of nine n values were chosen for each of the three sonification speeds:

- For 10^2 notes/second, $n \in [0, 1, .., 8]$
- For 10^3 notes/second, $n \in [3, 4, .., 11]$
- For 10^4 notes/second, $n \in [6, 7, .., 14]$

Practically speaking, if a transcription algorithm is expected to have low probability of transcription error (e.g., $p(10)$ notes misclassified) playing through at 10^2 notes/second would not expose transcription errors as quickly as a higher sonification speed. Likewise, for high error probability (e.g., $p(3)$ notes misclassified), the user would be expected to use a slower speed to capture local detail.

Through sonification, this method of error generation resulted in nine sound-files for each speed and corpus with a range of transcription errors that was approximately the same. The length of the sound file was chosen to be four seconds, starting at a random point in the corpus, resulting in $\approx$ 0-400, $\approx$ 0-500, and $\approx$ 0-600 notes misclassified in each sound file at 10^2, 10^3, and 10^4 notes/second respectively. Though created probabilistically, for sound-files with low error probability (1-10 note transcription errors per sound file), sound-files were selected to be well ordered, so that the lower probability had approximately half transcription errors of the next highest probability.

7.3 Sonification

A sonification technique was written in SuperCollider (McCartney, 1996; Wilson et al., 2011)[2] for playing through the each corpora at a broad range of speeds. In the technique, ground truth and transcription (in the present case a modified version of ground truth) are played synchronously through the left and right stereo channels at 10^2, 10^3, and 10^4 notes/second and enhancing divergences through a small amplification. When the pitch is identical in both versions, the transcription algorithm has performed well, and the pitch is perceived to come from the center of the head. When the pitch is not identical, the transcription algorithm has made an error, and the stream breaks into a pair of two slightly louder, non-identical notes coming from the left and right ear simultaneously. To compensate for changes in global and relative loudness differences created by the difference in speed, a loudness-compensation function was introduced.

7.3.1 Pitch Mapping

From the CSV file exported from music21, each MIDI value was transposed up an octave and a half and converted to the cycles per second of a sinusoid using the .midicps method. The change of an octave and a half increased audibility of lower notes, which tended to be muddled together when not transposed. The change made the frequency of the high notes higher, but the shift of an octave and a half rarely resulted in notes outside the upper limit of the modern piano.

Well known results from auditory perception experiments are the Fletcher-Munson curves (Fletcher & Munson, 1933), or equal-loudness contours (Epstein & Marozeau, 2010). Essentially, pitches in the range of 1.5-7kHz will sound louder than other pitches, increasing their salience relative to the other notes. To compensate for this inequality, the AmpCompA unit generator in SuperCollider was used with the MIDI note 48 (C3) as the reference. The algorithm is designed such that all other notes would be of roughly the same loudness as the reference note. Simple listening reveals that the compensation is not perfect but is significantly better than no compensation.

[2] SuperCollider [Freely Available Online]: `http://supercollider.sourceforge.net/`

7.3.2 Presentation Speed

The speeds used for testing—10^2, 10^3, and 10^4 notes/second—were chosen to represent a broad range of possible speeds a researcher might use. Although the technique could potentially run at higher speeds, testing at 10^5 notes/second was found to crash SuperCollider on a 2×2.66 GHz Dual-Core Intel Xeon processor despite relatively simple synthesis. If necessary, higher speeds may be attainable using a lower-level programming language or a faster computer in the future.

As discussed in Section 7.2 the notes from each corpora were arranged in a "flat" representation, and duration, loudness, and instrument data were not recorded. Consequently, there was no speed which could closely recreate the original music.

7.3.3 Duration and Amplitude Envelope

Given the great speeds of playing through the corpora, notes were designed to be played at the shortest possible duration that could communicate the fundamental frequency of each note. Through informal testing, a reasonable duration was found to be 50ms mediated by a sinusoidal envelope, shown in Figure 7.1. For shorter durations, pitches became noise-like and pitch could not be perceived. For other envelopes, "clipping" and other artifacts were audible. The overlap between notes was therefore minimized though not negligible, varying tremendously with speed: 5 note overlap at 10^2 notes/second, 50 note overlap at 10^3 notes/second, and 500 note overlap at 10^4 notes/second. To automatically compensate for the change in global and relative loudness between "correct" and "error" notes due to overlap, it was necessary to create a loudness compensation function.

7.3.4 Loudness-Compensation Function

The note-based sonification approach created changes in global loudness due to increasing overlap between notes at increasing speeds. Decreasing the loudness of all notes at high speeds could render transcription errors too faint to be heard (especially in noisy environments). To compensate for these differences, a loudness-compensation function was introduced to preserve the relative balance between correct and error notes while maintaining the same global loudness level across speeds.

The amplitude of each note was modified depending both upon the presentation speed

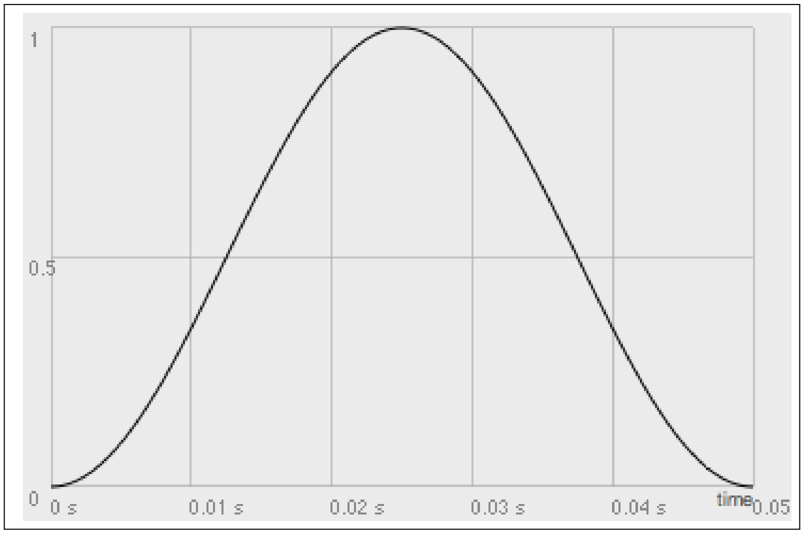


Fig. 7.1: A plot of the amplitude envelope used for sonification, generated using the Env.sine envelope generator in SuperCollider. Each note lasted 50ms and the relative amplitude of error vs. correct notes was dependent upon speed as discussed in Section 7.3.4.

and whether or not it was correctly transcribed. Introducing this foreknowledge into the algorithm allows the machine to help the user without being explicit. Though this will be discussed more later, the value of this approach is in the user's cognitive experience of the data set as a whole and not just in the ability to count transcription errors.

The equation for amplitude $A(s)$ of each note became

$$A(s) = \frac{1 + \alpha_s \gamma}{1 + \alpha_s}, \tag{7.2}$$

where α_s is the control of relative gain between transcription error and correct classification that varies with sonification speed s, and γ is a gate that is 1 when there is an transcription error and 0 when the transcription is correct. The chosen values for α_s were 60, 15, and 4 for $s = 10^4, 10^3$, and 10^2 notes/second respectively. The right level of relative gain gave the impression of an auditory stream that was further away (the correct notes) and a second stream that was much closer (the transcription errors).

7.4 Perceptual Experiment

A perceptual experiment sought to determine how well users could learn and use the sonification technique discussed in Section 7.3 to order sets of nine four-second samples by number of transcription errors. Each of the three corpora were represented once at each speed (10^2, 10^3, 10^4 notes/second) making a total of nine folders for ordering. An informal questionnaire following the experiment sought aesthetic responses to the sonification technique and ordering task. From a pilot study, it was hypothesized that users could quickly learn the approach, there would be no effect of corpus, but speed might influence performance ability.

7.4.1 Methods and Materials

The method of error generation and sonification mentioned in Sections 7.2 and 7.3 was repeated for each corpora and each speed resulting in nine total folders. Although the files within each folder were randomized, the set of nine folders as a whole was not randomized so that for each participant, folders 1, 4 and 7 were the chorales, 2, 5 and 8 were the string quartets and 3, 6, and 9 were the madrigals. Likewise, folders 1, 2 and 3 were 10^2 notes/second, folders 4, 5 and 6 were 10^3 notes/second, and folders 7, 8, and 9 were 10^4 notes/second. To better study learning effects, the folders should be randomized for all participants in the future.

Participants listened to the recorded AIFF 16bit sound-files on Sennheiser HD 800 headphones in the Audiovisual Editing Lab at CIRMMT. Subjects were instructed to use "Finder," the default file manager used in MacOS X to preview and order sound-files within the folder. Sound-files were previewed by pressing the spacebar on a standard Apple keyboard, and were dragged and dropped using an Apple Mouse. An example of such a folder containing nine sound-files is shown in Figure 7.2 and an ordered folder is shown in Figure 7.3.

After explaining to each subject what an transcription error sounded like, the participants were asked to place on the headphones and sound-files from the first folder were played as examples. The experimenter also showed how files could be easily played and paused with the spacebar and how to use the Apple Mouse to arrange files. With the

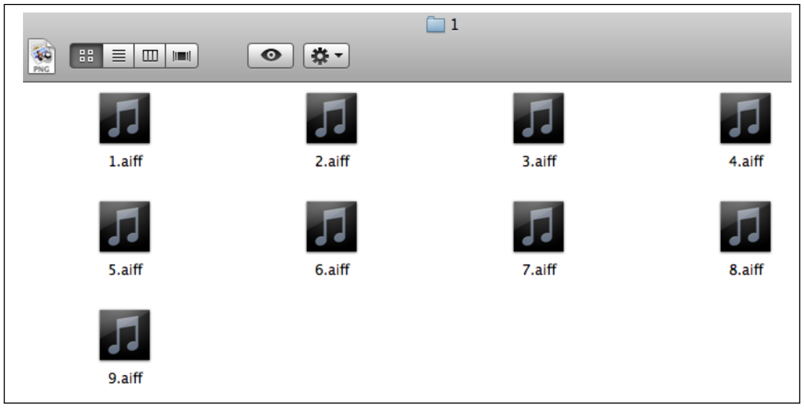


Fig. 7.2: An example folder containing nine unordered four-second sound-files with varying numbers of transcription errors. Participants were asked to order nine of these folders. An example ordering is shown in Figure 7.3.

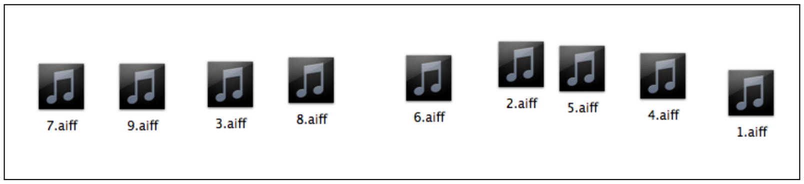


Fig. 7.3: An example of the folder containing the four-second sound-files from Figure 7.2 ordered from most note discrepancies to least note discrepancies as determined by the participant.

first folder partially complete, the participants were asked to start with the second folder and complete the first folder later. Within each folder, once the participant had found an ordering they were happy with, they recorded their answer in written form, which was collected at the end of the experiment and used for data analysis.

7.4.2 Participants

The experiment involved 11 volunteer, unpaid graduate (9) and undergraduate (2) students (4 female, 7 male) studying either music technology (8), information science (1), computer science (1) or psychology (1). All but three had more than 5 years of private music lessons. Participants were told that the experiment would last 20–30 minutes and most finished within this time frame.

Six of the participants had heard examples of the sounds before the experiment. Five participants had heard brief samples when it was demonstrated in a graduate level seminar, and the other participant had heard them several times during development of the technique, been involved with discussions of the technique, and had participated in a pilot experiment. This participant attained the highest score of any of the participants in the experiment.

7.4.3 Results

A plot of the results from the experiment is displayed in Figure 7.4. The greatest deviation occurs in folder 2, the first folder that the participants were asked to order. As can be seen, in general participants did very well and parts of folders were ordered perfectly for all participants. The ordering errors that did occur tended to be greatest for sound-files with a mid-range of note discrepancies.

Overall, there were very few ordering errors made by the participants. Nine out of eleven participants got at least one set perfectly correct. Among this subset of participants, the mean number of sets ordered perfectly was 4.4, the worst performance was three perfect sets ($n = 1$) and the best performance was seven perfect sets ($n = 1$). By speed, the best performance was for the 10^3 notes/second group (folders 4–6), where the total number of prefect orderings was 19 ($mean = 6.33$) and the highest performance was folder four (Bach Chorales, 10^3 notes/second), which had nine correct orderings.

The high accuracy in the Bach Chorales at 10^3 notes/second (9 perfect orderings) did not continue in the 10^4 notes/second folder (1 perfect ordering). For the three folders at 10^4 notes/second, the total number of perfect orderings was 13 ($mean = 4.33$). The other two folders (8 and 9) at this speed had six correct orderings each which was close to the mean of the 10^3 notes/second group.

The worst performance was in the 10^2 notes/second group which had a total of six perfect orderings ($mean = 2$). Seven participants returned to the first folder after finishing folder nine to complete the partial ordering, and out of them, three ordered it correctly. For the same group of seven, there were no perfect orderings on folder 2 and two perfect orderings on folder 3. The number of and type of errors did not differ between folders 1 and 3 in this subgroup, indicating that most learning happened in folder 2.

Between the two participants that did not order any correctly, analysis of number of ordering errors and ordering error type showed increased performance accuracy over time. Their difference in performance could not be attributed to musical experience as one had

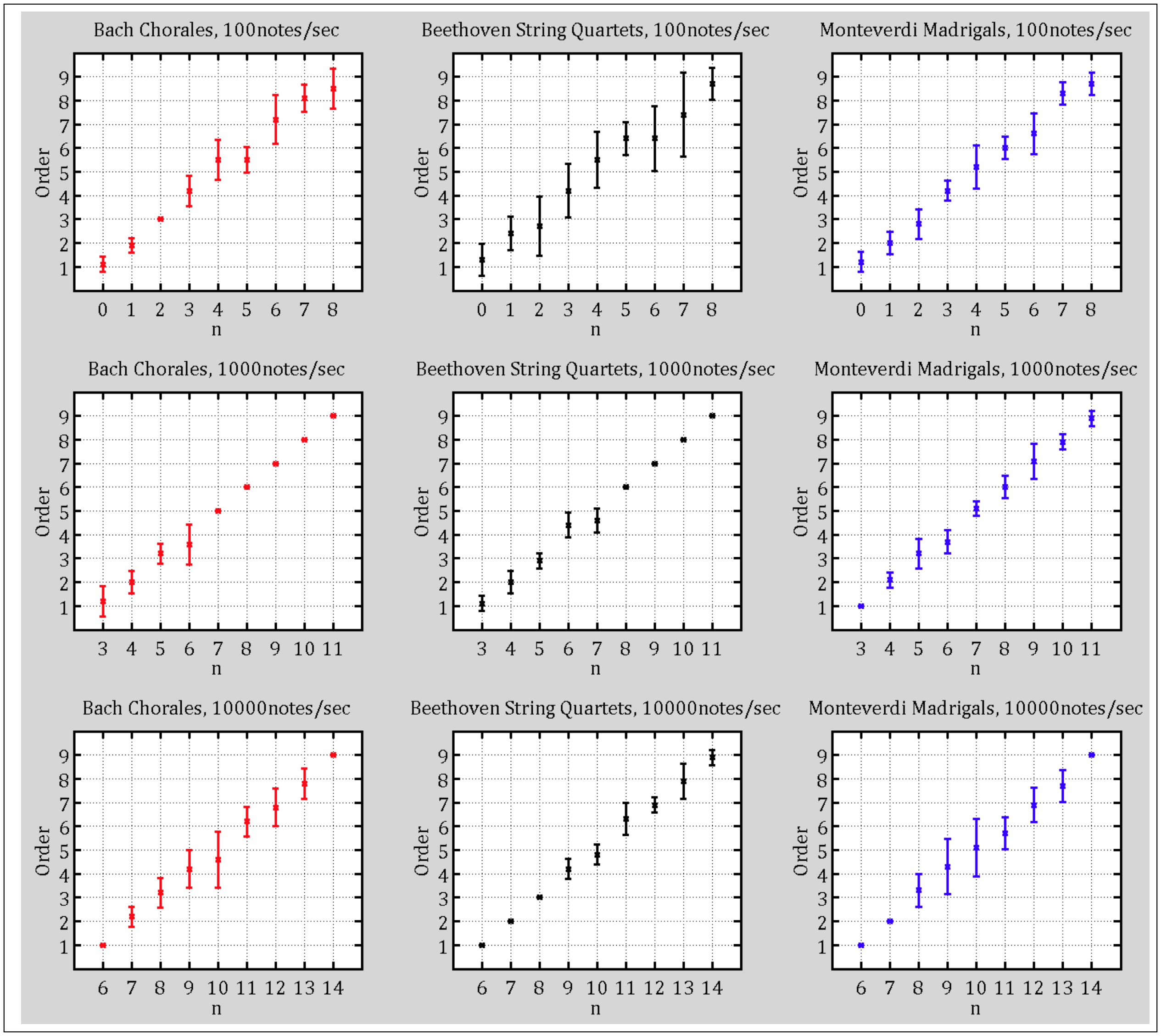


Fig. 7.4: Nine plots showing the performance of all participants on each of the sets. The plots are arranged by number and corpus representing the nine folders the participants were asked to order. The ordinate is the order as arranged by the participants from most errors (1) to least errors (9). The abscissa is the value for n in the probability of transcription error$p(n)$ in Equation 7.2. The error bars represent the standard deviation from the mean order number for each n value.

almost no musical training and the other had many years of training. Other factors such as sleep, attention, understanding of instructions might have influenced their performance.

7.4.4 Feedback

Following the test, an informal verbal questionnaire recorded the impressions of the task in terms of difficulty and enjoyment. Overall, the subjects found the task easy, though some found specific folders harder than others. For some, impressions of “melody” or “structure” in the background made the task more challenging.

When asked what they used to accomplish the task, participants reported that it changed depending upon number of errors. For large number of errors, loudness provided the clearest indication of number of errors. For small number of errors (below 40–50 according to one participant) counting provided a reliable means of ordering. Four participants remarked that they felt the task had become easier with time or that they felt that they were learning. Some wished that they had more time or realized they had made mistakes. One found it difficult and irritating at the end and remarked they would need breaks if doing the task for long periods of time.

For the most part, subjects remarked that the sounds were interesting and not-unpleasant, but would not listen to them for leisure. Among those that liked it best ($n = 3$), descriptors like “nice,” “electronic,” “industrial,” and “nine-inch nails” were used to describe the sound. For those that didn’t like it ($n = 2$), descriptors such as “irritating,” “disgusting,” or “like noise” were used.

Questions of musical experience and music listening patterns were also recorded, but the between subject differences were not found to be significant.

7.4.5 Discussion

The results show that the technique was quite effective overall. The best performance was for the 10^3 notes/second group. The 10^2 notes/second group had the worst performance of the three speeds, which may be due in part to learning effects. Two out of the eleven participants did significantly worse than the others, but error analysis revealed that their ordering accuracy was increasing over time.

The difference between corpus was not found to be significant as the effect of speed. Because the loudness cues scale with speed, increasing the value of α_{100} from Equation 7.2 might result in better performance in the future. Participants found the available cues most useful for categorizing large (> 200) and small (< 50) errors and performance tended to be worse for error numbers in the middle range.

The balance between localization and loudness cues warrants further study. The loudness cues were incorporated to increase performance as they could amplify the distinction between correct and incorrectly classified notes. However, in this experiment, the loudness cues became at times so strong that the spatial cues took a secondary role. Equalizing the loudness between incorrect and correctly classified notes would reveal a threshold for distinction that might be useful to the scientific study of auditory perception.

The pitch modifications used in the experiment had standard distribution of $\sigma^2 = 6$ meaning that the modifications tended to be small. Although frequency resolution is high for the pitch range used presently, as mentioned previously, it is foreseeable that larger pitch differences (as found in real pitch-transcription algorithms) would result in higher performance accuracy due to auditory grouping principles.

7.5 General Discussion

Overall, the results show that the sonification technique could be used to communicate number of errors made by pitch transcription algorithms across a broad range of presentation speeds, corpora, and error probabilities. In general, the technique was easy to learn, not-unpleasant, and could be used at speeds far above those of the actual music. Granted that a machine can quickly judge the performance of a transcription algorithm through comparison to ground truth and quickly present this information to the user, the full benefits of sonification as a research tool in this case and in others needs to be addressed.

7.5.1 Why Sonification?

The fundamental benefit of sonification is that for transcription analysis and MIR in general, sonification provides a medium for data representation that is shared with the underlying data. This representation can provide researchers with a cognitive experience of their data and algorithm that is closer to the data's source. This experience can enrich the data analysis process, providing information that is not visually obvious, leading to further insights and research directions.

As an example, one such insight was provided early in the development of this technique through the high-speed sonification of the extracted pitches of each corpora individually. Informal listening revealed that for all speeds (even 10^4 notes/second) the three corpora could be identified quite easily based upon their characteristic sounds. The string quartets were characterized by more jumping between high and low extremities and a feeling of agitation. The chorales were fixed around a certain pitch element with occasional leaps to higher pitch regions. The madrigals had a mellow timbre that was very well localized in a certain pitch region. Examples can be found at the author's website.[3]

Examples such as this stress the richness of the auditory representation of data in MIR. When used in this way, sound can carry information that quickly communicates features of the data that result from the data's relationship to music. For the sonification technique offered presently, sound offers two levels of information that complement simple error counting:

1. Where the errors occurred in time
2. Type of Error

The first point refers to perception of misclassified notes relative to other notes in the sound-file. Users can use auditory cues such as the pitch of the note and its temporal position relative to other (perhaps correctly transcribed) notes to quickly gauge if a transcription algorithm has made the same or different errors as another algorithm, or if any new notes have been misclassified or correctly classified.

For the second point, pitch can be used to identify the type of error that occurs. For instance, if the transcription system confuses a pitch with its octave equivalent or a tone a minor second away, this pitch discrepancy can be acoustically identified and used to modify or judge the subsequent performance of the algorithm. Identification of this type might be limited to situations of very few transcription errors however, as it might take more cognitive effort for the listener to encode.

7.5.2 Sonification for Music Analysis

The idea of using sonification as a means to present information about music or audio is not new, and has been presented formally by (Ferguson & Cabrera, 2009; Cabrera & Ferguson, 2006, 2007; Ferguson, 2009). The reasons why the pitches of Bach's Chorales, Beethoven's String Quartets, and Monteverdi's Madrigals could be differentiated at 10^4 notes per second deserves further investigation as do the many smaller, local events that

[3] http://www.music.mcgill.ca/~raymond/Sonification_for_Symbolic_Music_Analysis

could be distinguished in each corpora. Why, for instance, in Bach's Chorales are there few and brief moments at 10^4 notes/sec where the pitch "cluster" center suddenly moves up by major second? Is this phenomenon repeated if all pieces are transposed to C? How does the pitch-class content of a composer like Stravinsky change over the course of a career? Due to this exploratory capacity, sonification can be expected to generate new directions and unsuspected insights for future study.

7.5.3 Future Application of the Technique

An additional benefit of the sonification technique is its accessibility. The algorithm for synthesis is very simple, making it easy to implement in many languages. Using MatLab for instance, data can be quickly mapped to a time-varying signal which could be played in stereo using the *soundsc* function. The code used for generation in SuperCollider is available for download on the author's webpage.[4] Making sonification a tool that is accessible to researchers in MIR is an important goal as additional benefits of this analysis and display type are highly likely.

It should be noted that although the current sonification technique was designed for pitch transcription algorithms, it is not necessary that the data be pitch values at all. The technique can work for any situation in which ground-truth needs to be rapidly compared to machine-generated classification or transcription. It is only necessary that the data parameters (whatever they might be) be mapped to pitch values.

7.6 Conclusions

In this paper, a technique for high-speed note-based sonification was introduced and applied to the analysis of pitch transcription algorithms. The technique was tested using modified versions of ground truth with experimentally contrived probability and distribution of note transcription errors. A user-test involving 11 participants showed that the technique could be quickly learned and used across a large range of speeds, transcription error probabilities, and three corpora of classical music. Additional benefits of the technique beyond simple error counting were provided, as well as directions for future application of the technique.

[4] http://www.music.mcgill.ca/~raymond/Sonification_for_Symbolic_Music_Analysis

Chapter 8

Conclusion

In this thesis, sonification was applied to three types of data endemic to contemporary music research: emotion, gesture, and corpora. Each brought attention to the benefits of sonification, specific applications, and means of evaluation. Though sonification can be used to represent any data, when the data is somehow related to music, either through the data being used, or the mapping strategies employed, domain specific benefits arise.

8.1 Summary

In the case of emotional communication, sound can be applied to contexts of affective computing when social displays of emotion are *unavailable*, *misleading*, or *inappropriate*. Though results from emotion elicitation in environmental sounds or music can be used to direct mapping decisions, ultimately the continuous and flexible nature of the cues musical emotion contribute to its place as the strongest framework for development. Environmental sounds, though quite capable of emotional elicitation, cannot rely on identifiability as the main emotional determinant in continuous display, though features such as "naturalness" and "realism" should be preserved, and emotionally-neutral "evolutionary," "self-referential" sounds might be useful in choosing the fundamental sound for display. In the context of musical emotion, one should not haphazardly choose from the documented structural and acoustic cues from musical emotion, but should instead consider first what *mechanisms* of emotion induction would be best for the use context. Emotional contagion and brain stem reflex are two such mechanisms with desirable qualities such as high-induction speed, low degree of volitional influence, low cultural specificity, and dependence on musical structure. The use of major-minor mode for conveying valence might be a powerful conveyor of valence for listeners aware of the cultural connotation, but should be applied with caution as the desired psychological properties supplied by the aforementioned mechanisms are not guaranteed. It is possible, and even wise, to develop sonifications to match the emerging number of computational tools for music emotion recognition, but one must be aware that accurately matching such a model does not guarantee success in emotional communication. The number and type of cues must also be considered, as well as the nature of the sounds being used. By applying and accurately matching a computational model, sonification strips the model of a musical context allowing the emotional elicitors of the sound to be studied in isolation.

In the case of gesture, sound offers the capacity to display additional expressive information in the same channel as the performance audio, enabling a fuller understanding of a performer's expressive intention than the audio alone. When designing a sonification to be capable of this, one has to make careful decisions to not mask or interfere with the performance audio. When designing a tool to analyze expressive gesture using sound, flexibility is paramount. Unlike effective gestures, where movements are more "goal-oriented," expressive gestures are considerably more varied, differing primarily between performer, instrument, and musical piece. Similarly, a sonification tool for analysis should be able to differentiate according to these three factors. It is also necessary for the tool to display features relevant to emotional expression, some of which have already been identified as speed, amplitude, fluency, and regularity. Although the use of PCA as a pre-processing tool in sonification has been criticized in the past, it has since been effectively demonstrated as a tool to compare between different performers, abstracting more general movement characteristics. In this light, the tool provided by Savard (2009) has many useful functions for data analysis, including 10 synthesis channels, interactive mapping, and data-preprocessing functions specific to expressive gesture (including PCA), in addition to more functional tools such as a movie viewer, recording, and playback buttons.

In the subject of corpora, it has been demonstrated that sonification can be used to display note discrepancies between "ground truth" and artificially generated pitch-transcription at high speeds (up to 10^4 notes/second) and for three corpora. Although computers can analyze this sort of algorithm faster and provide a numerical analysis that is more accurate, through sonification, it is evident that more information was presented to users than simply errors in the algorithm's transcription. For example, Monteverdi's madrigals, Bach's chorals, and Beethoven's string quartets all manifested characteristically different sounds at high speeds such that they were acoustically differentiable. Though not available presently, it is likely that extending this technique might become valuable for determining characteristic differences in large corpora of symbolic music, and might be applied to audio-based libraries in the future as well.

In conclusion, sound has been demonstrated as a useful tool for music analysis. All three of the data types presented have demonstrated avenues for future development, all of which benefit in some way from the shared medium of sound. Though sonification as of yet is not a commonly used technology in music research, it will likely continue to be applied and perhaps gain in popularity as more successful applications arise. One should not forget that unlike other contexts of sonification, in music, listening is a fundamental, definitive practice, further supporting the potential of sonification in this domain.

8.2 Contributions

This thesis has made several contributions for future work which should be noted. In the study of emotion, this thesis has first contributed a typology for organizing systems for affective music generation (AMG), and has successfully applied it to differentiate the two systems presented in Chapter 2. It has also contributed a framework for choosing cues for sonification of emotion that is based upon psychological mechanisms for emotional induction. It has also contributed results from the first ever computational evaluation of a sonification of emotion, identifying benefits and limits of the approach, and a means of overcoming computational obstacles. In the realm of gesture, the contribution is most comprehensively a framework for evaluating sonifications of expressive gesture, one which is notably different from effective or "goal-oriented" movements, focusing on relevant emotional and expressive visual cues and structural elements. Along with this, the thesis contributed motivations for designing sonifications for synchronous presentation with the underlying music. This thesis has also contributed preliminary results on the sonification of corpora, an area, which like emotion, had little formal development. Results indicated that sound can be used to rapidly analyze and differentiate large databases of music. Although computers can perform some tasks faster and more accurately, listening can present hidden structures and patterns the were previously unknown or not visually obvious.

The thesis has also contributed a collection of three software tools. The sonification of emotion GUI presented in Chapter 2 is optimized for emotion analysis with corresponding video, and includes an arousal/valence visualizer, a mapping interface, control of speed and playback, and a method for quickly generating fresh sounds. By allowing the eyes to be focused on the video rather than the arousal/valence visualizer, the additional emotional information was communicated to the user without masking socially communicated visual cues. In Chapter 4, two GUI frameworks were created to assist in the design of sonifications of emotion using the MIREmotion function (Eerola et al., 2009). The first framework—the "myemotion" function—allows the user to analyze a single soundfile according to five emotion categories and three dimensions (though only activity and valence were implemented), and analyze the soundfile based upon the features relevant to the emotion score. A second framework—the "avmap" function—analyzes a collection of soundfiles, connecting the measured AV coordinate to the desired AV coordinate, and generating a euclidean error metric to judge the sonification's adherence to the model. In Chapter 6, "The Sonification Desktop" was presented, originally designed for the purpose of expressive/ancillary gesture analysis in Savard (2009). For the present thesis, the desktop was updated, documented (Winters, 2011b), demonstration videos created, and the Max/MSP source code was moved to a GitHub repository.[1]

The thesis has also contributed to the discussion of the relationship between music and sonification. In Chapter 2, it was demonstrated that a sonification of emotion could be integrated into music performance, echoing the idea presented in Vickers and Hogg (2006), that it may not always be instructive to distinguish between the two. However, Chapters 2, 4, and 6 all add to the discussion of how the two can be differentiated. For example, in Chapter 2 (p. 29), it was noted that while a sonification of emotion can be integrated into music performance, in sonification, the sound is most comprehensively a *signal* that communicates or interprets data for the user. In turn, this creates differences in both the goals of the sonification designer and the way that it is meant to be listened to. In Chapter 4, this discussion was expanded, including a formulation of the necessary conditions for a technique to be considered a sonification of emotion based upon the criteria presented in Hermann (2008). Previously, the definition of sonification of emotion had been contentious due to possible overlaps with music (Schubert et al., 2011), but this

[1] Freely available [Online]: `https://github.com/mikewinters/SonificationDesktop`

thesis reiterates necessary features: an underlying data space representing emotion, and a systematic, reproducible mapping for every point in the space. Finally, in the discussion of expressive gesture in Chapter 6, Table 6.2 presented similarities and differences between sonification and musical mapping of gesture, as might be found at the annual New Interfaces for Musical Expression conference.[2] As in Chapter 2, they can be distinguished by the motivation for mapping, and how they are meant to be listened to. In the case of gesture, they are also differentiated by the presence of a performer, and the skill that increases with use (i.e. for music: expression; for sonification: the ability to determine data relationships).

8.3 Limitations & Future Work

In Chapter 2, due to constraints in the project timeline, the Emotional Imaging Composer was not used in a live concert setting, and only a single sample video and the corresponding analyzed emotion trajectory were available for development. In the future, the GUI and sonification technique should be tested with more data, and with different videos and performances. Getting feedback from a live performance scenario would be helpful as well. The ability to interactively change mappings was presented in the GUI of Fig. 2.2, but was not implemented due to time constraints. In the future, if the GUI were to be further developed, access to mappings should be available to the user, including the ability to turn certain cues on and off, and change how they are mapped from the emotion dimensions.

In Chapter 3, musical emotion was chosen as a more robust framework for development than environmental sounds. However, the field of Emoacoustics (Asutay et al., 2012) is quickly emerging, presenting new results, and perhaps a more encompassing view on auditory-induced emotion than provided by music alone. In the future, results from this field may find firmer footing in the subject of emotion sonification, especially in contexts where non-musical emotional communication and display are profitable (i.e. affective computing). The framework presented in Chapter 3 also focused on two theoretical mechanisms musical emotion induction: ‘brain stem reflex’ and ‘emotional contagion.’ However, determining the mechanisms for emotion induction in music is still an active research question (e.g. Scherer & Coutinho, 2013). In the future, theoretical accounts for musical emotion may change, but the sonification strategy should stay the same: Mechanisms should be chosen based upon the desired psychological properties for the context. These in turn, lead

[2] NIME [Online]:`http://www.nime.org`

a sonification designer to a subset of corresponding structural and acoustic cues that can be used for communication.

Chapter 4 presented the first ever computational evaluation using a tool for music emotion recognition (MER). Although the chosen tool was useful in demonstrating the general benefits and limitations of the approach, in the future, the chosen MER tool should be trained on larger corpora of music (potentially spanning many genres) and also using time-varying as opposed to static arousal/valence judgements. The work of Coutinho and Cangelosi (2011), and Schmidt, Scott, and Kim (2012), provide examples where such time-varying models are beginning to be developed. When using models for evaluation, it is important to remember that these models have thus far demonstrated an apparent limit to prediction accuracy at approximately 65% (Kim et al., 2010). As discussed in Section 4.4.3, if the model were trained using music, the prediction accuracy for sonification may be different. In either case, for future work, such computational evaluations should be complemented with listener studies.

In both Chapters 5 and 6 it was posited that adding a well-designed sonification of expressive gesture to performance audio may in some ways enhance the expression of the music by conveying the expression embodied in the performer. Future listening studies should asses this capacity with a variety of music types and performers, determining what types of mappings work best for each case, and choosing cues for sonification relevant to visual perception of emotion and expression. In the future, the Max/MSP tool presented in Chapter 6 should be studied with a larger database of movement, spanning many performers, genres, and instruments. Although it has been evaluated positively for its ability to meet the goals of sonification and expressive movement analysis, it still needs to be tested with actual users.

The study of sonification of corpora in Chapter 7 was thus far limited to the study of only three corpora of music. Large and complete corpora of symbolic music, though rare, are becoming increasingly available, and the sonification technique presented here may bring additional results with more data. Determining how to apply sonification with music, and the features a sonification should display in a music database are difficult questions, which for the moment, limit sonification to a more exploratory nature. In the future, the presented technique might also be extended to large audio-based libraries (e.g. one's entire personal music collection). As with symbolic music, determining which features should be displayed and the most advantageous techniques for sonification are yet to be determined.

References

Asutay, E., Västfjäll, D., Tajadura-Jiménez, A., Genell, A., Bergman, P., & Kleiner, M. (2012). Emoacoustics: A study of the psychoacoustical and psychological dimensions of emotional sound design. *Journal of the Audio Engineering Society*, *60*(1/2), 21-8.

Barrass, S., & Kramer, G. (1999). Using sonification. *Multimedia Systems*, *7*(1), 23-31.

Benovoy, M., Cooperstock, J. R., & Deitcher, J. (2008, January). Biosignals analysis and its application in a performance setting: Towards the development of an emotional-imaging generator. In *Proceedings of the 1st international conference on biomedical electronics and devices* (p. 253-8). Funchal, Madeira.

Bergman, A. (1990). *Auditory scene analysis.* Cambridge, MA: MIT Press.

Bradley, M. M., & Lang, P. J. (2000). Affective reactions to acoustic stimuli. *Psychophysiology*, *37*, 204-15.

Bradley, M. M., & Lang, P. J. (2007). *The international affective digitized sounds (2nd edition; IADS-2): Affective ratings of sounds and instruction manual* (Tech. Rep.). Gainesville, FL: University of Florida.

Brazil, E., & Fernström, M. (2011). Auditory icons. In T. Hermann, A. Hunt, & J. G. Neuhoff (Eds.), *The sonification handbook* (p. 325-38). Berlin, Germany: Logos Verlag.

Bresin, R., & Friberg, A. (2011). Emotion rendering in music: Range and characteristic values of seven musical variables. *Cortex*, *47*, 1068-81.

Budd, M. (1985). *Music and the emotions: The philosophical theories.* London, UK: Routledge & Kegan Paul.

Bunt, L., & Pavicevic, M. (2001). Music and emotion: Perspectives from music therapy. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (p. 181-201). New York, NY: Oxford University Press.

Buxton, W., Gaver, W., & Bly, S. (1994). *Auditory interfaces: The use of non-speech* *audio at the interface.* (Ch. 2: Acoustics and Psychoacoustics)

Cabrera, D., & Ferguson, S. (2006). Auditory display of audio. In *120th audio engineering society convention.* Paris, France.

Cabrera, D., & Ferguson, S. (2007). Sonification of sound: Tools for teaching acoustics and audio. In *Proceedings of the 13th international conference on auditory display* (p. 483-90). Montréal, Canada.

Camurri, A., Lagerlöf, I., & Volpe, G. (2003). Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, *59*(1), 213-25.

Cassidy, G., & MacDonald, R. (2009). The effects of music choice on task performance: A study of the impact of self-selected andd experimenter-selected music on driving game performance and experience. *Musicae Scientiae*, *13*(2), 357-86.

Chadefaux, D., Wanderley, M. M., Carrou, J. L. L., Fabre, B., & Daudet, L. (2012, April). Experimental study of the musician/instrument interaction in the case of the concert harp. In *Proceedings of acoustics 2012.* Nantes, France.

Charles, J.-F. (2008). A Tutorial on Spectral Sound Processing Using Max / MSP and Jitter. *Computer Music Journal*, *32*(3), 87-102.

Clay, A., Couture, N., Decarsin, E., Desainte-Catherine, M., Vulliard, P.-H., & Larralde, J. (2012, May). Movement to emotions to music: Using whole body emotional expression as an interaction for electronic music generation. In *Proceedings of the 12th international conference on new interfaces for musical expression.* Ann Arbor, MI.

Coutinho, E., & Cangelosi, A. (2011). Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion*, *11*(4), 921-37.

Crooke, D. (1957). *The language of music.* London, UK: Oxford University Press.

Cuthbert, M. S., & Ariza, C. (2010). music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the 11th international society for music information retrieval conference* (p. 637-42). Utrecht, Netherlands.

Daffertshofer, A., Lamoth, C. J. C., Meijer, O. G., & Beek, P. J. (2004). PCA in studying coordination and variability: A tutorial. *Clinical Biomechanics*, *19*(4), 415-28.

Dahl, S., Bevilacqua, F., Bresin, R., Clayton, M., Leante, L., Poggi, I., et al. (2010). Gestures in performance. In R. I. Godøy & M. Leman (Eds.), *Musical gestures:*

Sound, movement, and meaning (p. 36-68). New York, NY: Routledge.

Dahl, S., & Friberg, A. (2007). Visual perception of expressiveness in musicians' body movements. *Music Perception: An Interdisciplinary Journal*, *24*(5), 433-54.

Davidson, J. (2012). Bodily movement and facial actions in expressive musical performance by solo and duo instrumentalists: Two distinctive case studies. *Psychology of Music*, *40*(5), 595-633.

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, *21*(2), 103-13.

Davies, S. (1994). *Musical meaning and expression.* Ithaca, NY: Cornell University Press.

Delalande, F. (1988). Glenn gould pluriel. In G. Guertin (Ed.), (p. 85-111). Verdun, Québec: Louise Courteau.

Eerola, T., Lartillot, O., & Toiviainen, P. (2009, October). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *Proceedings of the 10th international society for music information retrieval conference* (p. 621-6). Kobe, Japan.

Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, *39*(1), 18-49.

Egmond, R. V. (2009). The experience of product sounds. In H. N. J. Schifferstein & P. Hekkert (Eds.), *Product experience* (p. 69-89). San Diego, CA: Elsevier.

Eliakim, M., Bodner, E., Eliakim, A., Nemet, D., & Meckel, Y. (2012). Effect of motivational music on lactate levels during recovery from intense exercise. *Journal of Strength and Conditioning Research*, *26*(1), 80-6.

Epstein, M., & Marozeau, J. (2010). Loudness and intensity coding. In C. Plack (Ed.), *The oxford handbook of auditory science: Hearing* (Vol. 3, chap. 3). New York, NY: Oxford University Press.

Ewert, S., Müller, M., & Grosche, P. (2009). High resolution audio synchronization using chroma onset features. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing* (p. 1869-72). Taipei, Taiwan.

Fabiani, M., Dubus, G., & Bresin, R. (2010, April). Interactive sonification of emotionally expressive gestures by means of musical performance. In *Proceedings of the 3rd interactive sonification workshop* (p. 113-16). Stockholm, Sweden.

Fabiani, M., Friberg, A., & Bresin, R. (2013). Systems for interactive control of computer generated music performance. In A. Kirke & E. R. Miranda (Eds.), *Guide to* *computing for expressive music performance* (chap. 2). London, UK: Springer-Verlag.

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics: Facts and models* (3rd ed.). Berlin, Germany: Springer-Verlag.

Ferguson, S. (2009). *Statistical sonifications for the investigation of sound.* Unpublished doctoral dissertation, University of Sydney, Sydney, Australia.

Ferguson, S., & Cabrera, D. (2009). Auditory spectral summarisation for audio signals with musical applications. In *Proceedings of the 10th international society for music information retrieval conference* (p. 567-72). Kobe, Japan.

Fishwick, P. A. (2002). Aesthetic programming: Crafting personalized software. *Leonardo*, *35*(4), 383-90.

Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*, *5*(2), 82-108.

Fontaine, J. R. (2009). Dimensional emotion models. In D. Sander & K. R. Scherer (Eds.), *Oxford companion to emotion and the affective sciences* (p. 119-20). New York, NY: Oxford University Press.

Frija, N. H. (1988). The laws of emotion. *American Psychologist*, *43*(5), 349-58.

Funk, G., O'Neil, D., & Winters, R. M. (2012). What the oblique parameters S, T, and U and their extensions reveal about the 2HDM: A numerical analysis. *International Journal of Modern Physics A*, *27*(5), 1250021 (21 Pages).

Gabrielsson, A., & Lindström, E. (2010). The role of structure in the musical expression of emotions. In P. N. Juslin & J. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (p. 367-400). New York, NY: Oxford University Press.

Gaver, W. W. (1993a). How do we hear in the world? explorations in ecological acoustics. *Ecological Psychology*, *5*(4), 285-313.

Gaver, W. W. (1993b). What in the world do we hear? an ecological approach to auditory event perception. *Ecological Psychology*, *5*(4), 1-29.

Gibson, J. (2009). Spectral Delay as a Compositional Resource. *The Electronic Journal of Electroacoustics*, *11*(4), 9-12.

Goina, M., & Polotti, P. (2008, June). Elementary gestalts for gesture sonification. In *Proceedings of the 8th international conference on new interfaces for musical expression.* Genova, Italy.

Grond, F., Bouënard, A., Hermann, T., & Wanderley, M. M. (2010, September). Virtual Auditory Myography of Timpani-Playing Avatars. In *Proceedings of the 13th* *international conference on digital audio effects* (p. 1-8). Graz, Austria.

Grond, F., & Hermann, T. (2011). Aesthetic strategies in sonification. *AI & Society*, *27*(2), 213-22.

Grond, F., Hermann, T., Verfaille, V., & Wanderley, M. M. (2010). Methods for effective sonification of clarinetists' ancillary gestures. In S. Kopp & I. Wachsmuth (Eds.), *Gesture in embodied communication and human-computer interaction* (p. 171-81). Berlin, Germany: Springer-Verlag.

Grosshauser, T., & Hermann, T. (2009, July). The sonified music stand - an interactive sonification system for musicians. In *Proceedings of the 6th sound and music computing conference* (p. 233-8). Porto, Portugal.

Hagman, F. (2010). *Emotional response to sound: Influence of spatial determinants.* Unpublished master's thesis, Chalmers University of Technology, Göteborg, Sweden.

Hermann, T. (2008, June). Taxonomy and definitions for sonification and auditory display. In *Proceedings of the 14th international conference on auditory display.* Paris, France.

Hermann, T., Drees, J. M., & Ritter, H. (2003, July). Broadcasting auditory weather reports - a pilot project. In *Proceedings of the 9th international conference on auditory display* (p. 208-11). Boston, MA.

Hermann, T., Hunt, A., & Neuhoff, J. G. (Eds.). (2011). *The sonification handbook.* Berlin, Germany: Logos Verlag.

Höner, O. (2011). Multidisciplinary applications of sonification in the field of "exercise, play and sport". In T. Hermann, A. Hunt, & J. G. Neuhoff (Eds.), *The sonification handbook* (p. 525-54). Berlin, Germany: Logos Publishing House.

Hunt, A., & Hermann, T. (2011). Interactive sonification. In T. Hermann, A. Hunt, & J. G. Neuhoff (Eds.), *The sonification handbook* (p. 273-98). Berlin, Germany: Logos Publishing House.

Hyniewska, S., Niewiadomski, R., Mancini, M., & Pelachaud, C. (2010). Expression of affects in embodied conversational agents. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (p. 213-21). New York, NY: Oxford University Press.

Jee, E.-S., Jeong, Y.-J., Kim, C. H., & Kobayahi, H. (2010). Sound design for emotion and intention expression in socially interactive robots. *Intel Serv Robotics*, *3*, 199-206.

Jee, E.-S., Jeong, Y.-J., Kim, C. H., Kwon, D.-S., & Kobayahi, H. (2009). Sound production for the emotional expression of social interactive robots. In V. A. Kulyukin (Ed.), *Advances in human-robot interaction* (p. 257-72). Vukovar, Croatia: InTech.

Johansson, G. (1975, June). Visual motion perception. *Scientific American*, *232*(6), 76-88.

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology*, *26*(6), 1797-813.

Juslin, P. N. (2001). Communication emotion in music performance: A review and a theoretical framework. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (p. 309-37). New York, NY: Oxford University Press.

Juslin, P. N. (2003). Five facets of musical expression: A psychologist's perspective on music performance. *Psychology of Music*, *31*(3), 273-302.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770-814.

Juslin, P. N., & Sloboda, J. A. (Eds.). (2010). *Handbook of music and emotion: Theory, research, applications.* New York, NY: Oxford University Press.

Juslin, P. N., & Timmers, R. (2010). Expression and communication of emotion in music performance. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (p. 453-89). New York, NY: Oxford University Press.

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*(5), 559-621.

Kapur, A., Tzanetakis, G., Virji-Babul, N., Wang, G., & Cook, P. (2005, September). A framework for sonification of vicon motion capture data. In *Proceedings of the 8th international conference on digial audio effects* (p. 1-6). Madrid, Spain.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., et al. (2010, August). Music emotion recognition: A state of the art review. In *Proceedings of the 11th international society for music information retrieval conference* (p. 255-66). Utrecht, Netherlands.

Kirke, A., & Miranda, E. R. (Eds.). (2013a). *Guide to computing for expressive music performance.* London, UK: Springer-Verlag.

Kirke, A., & Miranda, E. R. (2013b). An overview of computer systems for expressive music performance. In A. Kirke & E. R. Miranda (Eds.), *Guide to computing for expressive music performance* (p. 1-47). London, UK: Springer.

Kramer, G. (Ed.). (1994). *Auditory display: Sonification, audification, and auditory inter-*

faces. Reading, MA: Addison Wesley.

Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J., Miner, N., et al. (1999). *The sonification report: Status of the field and research agenda.* Santa Fe, NM: International Community for Auditory Display (ICAD).

Larkin, O., Koerselman, T., Ong, B., & Ng, K. (2008, June). Sonification of bowing features for string instrument training. In *Proceedings of the 14th international conference on auditory display.* Paris, France.

Larsson, P. (2010). Tools for designing emotional auditory driver-vehicle interfaces. In S. Ystad, M. Aramaki, R. Kronland-Martinet, & K. Jensen (Eds.), *Auditory display: 6th international symposium, CMMR/ICAD 2009, revised papers* (p. 1-11). Berlin, Germany: Springer.

Larsson, P., Västfjäll, D., Olsson, P., & Kleiner, M. (2007, October). When what you hear is what you see: Presence and auditory-visual integration in virtual envrionments. In *Proceedings of the 10th annual international workshop on presence* (p. 11-8). Barcelona, Spain.

Lehman, S. Y., Baker, E., Henry, H. A., Kindschuh, A. J., Markley, L. C., Browning, M. B., et al. (2012). Avalanches on a conical bead pile: Scaling with tuning parameters. *Granular Matter*, *14*(5), 553-61.

Lemaitre, G., Houix, O., Susini, P., Visell, Y., & Franinović, K. (2012). Feelings elicited by auditory feedback from a computationally augmented artifact: The flops. *IEEE Transactions on Affective Computing*, *3*(3), 335-48.

Lombard, M., & Ditton, T. (1997). At the heart of it all: The concept of presence. *Journal of Computer-Mediated Communication*, *3*(2).

Matsumoto, D. (2009). Display rules. In D. Sander & K. R. Scherer (Eds.), *Oxford companion to emotion and the affective sciences* (p. 124). New York, NY: Oxford University Press.

McCartney, J. (1996, August). Supercollider: A new real time synthesis language. In *Proceedings of the international computer music conference* (p. 257-8). Hong Kong, China.

McGookin, D., & Brewster, S. (2011). Earcons. In T. Hermann, A. Hunt, & J. G. Neuhoff (Eds.), *The sonification handbook* (p. 339-61). Berlin, Germany: Logos Verlag.

Miranda, E. R., & Wanderley, M. M. (2006). *New digital music instruments: Control and interaction beyond the keyboard.* Middleton, WI: A-R Editions, Inc.

Nakra, T. M. (2000). Searching for meaning in gestural data. In M. M. Wanderley & M. Battier (Eds.), *Trends in gestural control of music* (p. 269-99). Paris, France: IRCAM.

Naveda, L., & Leman, M. (2010). The spatiotemporal representation of dance and music gestures using topological gesture analysis. *Music Perception: An Interdisciplinary Journal*, *28*(1), 93-111.

Norman, D. A. (2004). *Emotional design: Why we love (or hate) everyday things.* New York, NY: Basic Books.

Nusseck, M., & Wanderley, M. M. (2009). Music and motion—how music-related ancillary body movements contribute to the experience of music. *Music Perception: An Interdisciplinary Journal*, *26*(4), 335-53.

Ogihara, M., & Kim, Y. (2012). Mood and emotion classification. In T. Li, M. Ogihara, & G. Tzanetakis (Eds.), *Music data mining* (p. 135-67). Boca Raton, FL: CRC Press.

Pauletto, S., & Hunt, A. (2004, January). Interactive sonification in two domains: Helicopter flight analysis and physiotherapy movement analysis. In *Proceedings of the 1st international workshop on interactive sonification.* Bielefeld, Germany.

Picard, R. (1997). *Affective computing.* Cambridge, MA: The MIT Press.

Picard, R. (2009). Affective computing. In D. Sander & K. R. Scherer (Eds.), *The oxford companion to emotion and the affective sciences* (p. 11-5). New York, NY: Oxford University Press.

Picard, R., & Daily, S. B. (2005, April). Evaluating affective interactions: Alternatives to asking what users feel. In *CHI worskhop on evaluating affective interfaces.* Portland, OR.

Preti, C., & Schubert, E. (2011, June). Sonification of emotions II: Live music in a pediatric hospital. In *Proceedings of the 17th international conference on auditory display.* Budapest, Hungary.

Ramsay, J., & Silverman, B. (2005). *Functional data analysis* (2nd ed.). New York, NY: Springer.

Rasamimanana, N., Bernardin, D., Wanderley, M. M., & Bevilacqua, F. (2009). String bowing gestures at varying bow stroke frequencies: A case study. In M. Sales Dias, S. Gibet, M. Wanderley, & R. Bastos (Eds.), *Gesture-based human-computer interaction and simulation* (p. 216-26). Berlin, Germany: Springer-Verlag.

Risset, J.-C. (1969). Pitch control and pitch paradoxes demonstrated with computer synthesized sounds. *Journal of the Acoustical Society of America*, *36*(1A), 88.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161-78.

Savard, A. (2009). *When gestures are perceived through sounds: A framework for sonification of musicians' ancillary gestures.* Unpublished master's thesis, McGill University, Montréal, Canada.

Scherer, K. R. (2004). Which emotions can be induced by music? what are the underlying mechanisms? and how can we measure them? *Journal of New Music Research*, *33*(3), 239-51.

Scherer, K. R., & Coutinho, E. (2013). How music creates emotion: A multifactorial process approach. In T. Cochrane, B. Fantini, & K. R. Scherer (Eds.), *The emotional power of music* (p. 121-45). New York, NY: Oxford University Press.

Schmidt, E. M., & Kim, Y. E. (2011, October). Modeling musical emotion dynamics with conditional random fields. In *Proceedings of the 12th international society for music information retrieval conference* (p. 777-82). Miami, FL.

Schmidt, E. M., Scott, J., & Kim, Y. E. (2012, October). Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion. In *Proceedings of the 13th international society for music information retrieval conference* (p. 325-30). Porto, Portugal.

Schröder, M., Burkhardt, F., & Krstulović, S. (2010). Synthesis of emotional speech. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (p. 222-31). New York, NY: Oxford University Press.

Schubert, E. (2010). Continuous self-report methods. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (p. 223-53). New York, NY: Oxford University Press.

Schubert, E., Ferguson, S., Farrar, N., & McPherson, G. E. (2011, June). Sonification of emotion I: Film music. In *Proceedings of the 17th international conference on auditory display.* Budapest, Hungary.

Schutz, M., & Lipscomb, S. (2007). Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception*, *36*(6), 888-97.

Serafin, S., Franinović, K., Hermann, T., Lemaitre, G., Rinott, M., & Rocchesso, D. (2011). Sonic interaction design. In T. Hermann, A. Hunt, & J. G. Neuhoff (Eds.), *The sonification handbook* (p. 87-110). Berlin, Germany: Springer-Verlag.

Spence, C., & Soto-Faraco, S. (2010). Auditory perception: Interactions with vision. In C. Plack (Ed.), *The oxford handbook of auditory science: Hearing* (Vol. 3, p. 271-96). New York, NY: Oxford University Press.

Supper, A. (2012). The search for the “killer application”: Drawing the boundaries around the sonification of scientific data. In T. Pinch & K. Bijsterveld (Eds.), *The oxford handbook of sound studies* (p. 249-70). New York, NY: Oxford University Press.

Tajadura-Jiménez, A. (2008). *Embodied psychoacoustics: Spatial and multisensory determinants of auditory-induced emotion.* Unpublished doctoral dissertation, Chalmers University of Technology, Göteborg, Sweden.

Tajadura-Jiménez, A., Larsson, P., Väljamäe, A., Västfjäll, D., & Kleiner, M. (2010). When room size matters: Acoustic influences on emotional responses to sounds. *Emotion*, *10*(3), 416-22.

Tajadura-Jiménez, A., Väljamäe, A., Asutay, E., & Västfjäll, D. (2010). Embodied auditory perception: The emotional impact of approaching and receding sound sources. *Emotion*, *10*(2), 216-29.

Tajadura-Jiménez, A., & Västfjäll, D. (2008). Auditory-induced emotion: A neglected channel for communication in human-computer interaction. In C. Peter & R. Beale (Eds.), *Affect and emotion in HCI* (p. 63-74). Berlin, Germany: Springer-Verlag.

Toiviainen, P., Luck, G., & Thompson, M. R. (2010). Embodied meter: Hierarchical eigenmodes in music-induced movement. *Music Perception: An Interdisciplinary Journal*, *28*(1), 59-70.

Västfjäll, D. (2003). The subjective sense of presence, emotion recognition, and experienced emotions in auditory virtual environments. *CyberPsychology & Behavior*, *6*(2), 181-8.

Västfjäll, D. (2012). Emotional reactions to sounds without meaning. *Psychology*, *3*(8), 606-9.

Västfjäll, D., Kleiner, M., & Gärling, T. (2003). Affective reactions to interior aircraft sounds. *Acta Acustica United with Acustica*, *89*, 693-701.

Västfjäll, D., Larsson, P., & Kleiner, M. (2002). Emotion and auditory virtual environments: Affect-based judgements of music reproduced with virtual reverberation times. *CyberPsychology & Behavior*, *5*(1), 19-32.

Verfaille, V. (2003). *Effects audionumériques adaptatifs: Théorie, mise en oeuvre et usage en création musicale numérique.* Unpublished doctoral dissertation, Université Aix-

Marseille II, Marseille, France.

Verfaille, V., Quek, O., & Wanderley, M. (2006, June). Sonification of musician's ancillary gestures. In *Proceedings of the 12th international conference on auditory display* (p. 194-7). London, UK.

Verfaille, V., Wanderley, M. M., & Depalle, P. (2006). Mapping strategies for gestural and adaptive control of digital audio effects. *Journal of New Music Research*, *35*(1), 71-93.

Vickers, P. (2011). Sonification for process monitoring. In T. Hermann, A. Hunt, & J. G. Neuhoff (Eds.), *The sonification handbook* (p. 455-91). Berlin, Germany: Logos Verlag.

Vickers, P., & Hogg, B. (2006, June). Sonification abstraite/sonification concrète: An 'aesthetic perspective space' for classifying auditory diplays in the ars musica domain. In *Proceedings of the 12th international conference on auditory display* (p. 210-6). London, UK.

Vinciarelli, A., Pantic, M., Heylen, F., Pelachaud, C., Poggi, I., D'Errico, F., et al. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, *3*(1), 69-87.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition*, *101*(1), 80-113.

Walker, B. N., & Nees, M. A. (2011). Theory of sonification. In T. Hermann, A. Hunt, & J. G. Neuhoff (Eds.), *The sonification handbook* (p. 9-39). Berlin, Germany: Logos Verlag.

Wallis, I., Ingalls, T., & Campana, E. (2008, September). Computer-generating emotional music: The design of an affective music algorithm. In *Proceedings of the 11th international conference on digital audio effects* (p. 1-6). Espoo, Finland.

Wanderley, M. M. (1999). Non-obvious performer gestures in instrumental music. In *Gesture based communication in human-computer interaction* (p. 37-48). Berlin, Germany: Springer-Verlag.

Wanderley, M. M. (2002). Quantitative analysis of non-obvious performer gestures. In I. Wachsmuth & T. Sowa (Eds.), *Gesture and sign language in human-computer interaction* (p. 241-53). Berlin, Germany: Springer-Verlag.

Wanderley, M. M., Vines, B. W., Middleton, N., McKay, C., & Hatch, W. (2005). The musical significance of clarinetists ancillary gestures: An exploration of the field.

Journal of New Music Research, *34*(1), 97-113.

Wilson, S., Cottle, D., & Collins, N. (Eds.). (2011). *The supercollider book.* Cambridge, MA: MIT Press.

Winters, R. M. (2009, May). *The musical mapping of chaotic attractors* (Tech. Rep.). Physics Department: The College of Wooster.

Winters, R. M. (2010, June). *The two higgs doublet model and sonification: Using sound to understand the origin of mass* (Tech. Rep.). Physics Department: The College of Wooster.

Winters, R. M. (2011a, June). 1/f noise and auditory aesthetics: Sonification of a driven bead pile. In *Proceedings of the 17th international conference on auditory display.* Budapest, Hungary.

Winters, R. M. (2011b, December). *Documentation of the sonification desktop* (Tech. Rep.). Input Devices and Music Interaction Lab: McGill University.

Winters, R. M. (2011c, August). *Literature review and new directions for sonification of musicians' ancillary gestures* (Tech. Rep.). Input Devices and Music Interaction Lab: McGill University.

Winters, R. M., Blaikie, A., & O'Neil, D. (2011, June). Simulating the electroweak phase transition: Sonification of bubble nucleation. In *Proceedings of the 17th international conference on auditory display.* Budapest, Hungary.

Winters, R. M., Hattwick, I., & Wanderley, M. M. (2013, June). Integrating emotional data into music performance: Two audio environments for the emotional imaging composer. In *Proceedings of the 3rd international conference on music and emotion.* Jyväskylä, Finland.

Winters, R. M., Savard, A., Verfaille, V., & Wanderley, M. M. (2012). A sonification tool for the analysis of large databases of expressive gesture. *International Journal of Multimedia and Its Applications*, *4*(6), 13-26.

Winters, R. M., & Wanderley, M. M. (2012a, April). *High speed sonification of pitch in large corpora of music.* Input Devices and Music Interaction Lab. (In Preparation)

Winters, R. M., & Wanderley, M. M. (2012b, June). New directions for sonification of expressive movement in music. In *Proceedings of the 18th international conference on auditory display.* Atlanta, Georgia.

Winters, R. M., & Wanderley, M. M. (2013, June). Sonification of emotion: Strategies for continuous auditory display of arousal and valence. In *Proceedings of the 3rd* *international conference on music and emotion.* Jyväskylä, Finland.

Winters, R. M., & Wanderley, M. M. (2014). Sonification of emotion: Strategies and results from the intersection with music. *Organised Sound*, *19*(1), Accepted.

Yang, Y.-H., & Chen, H. H. (2011). *Music emotion recognition.* Boca Raton, FL: CRC Press.

Zwaag, M. D. van der, Fairclough, S., Spiridon, E., & Westerink, J. H. (2011). The impact of music on affect during anger inducing drives. In S. D'Mello (Ed.), *Affective computing and intelligent interaction* (p. 407-16). Berlin, Germany: Springer-Verlag.