Methods and Technologies for Analysing Links Between Musical Sound and Body Motion

Ph.D. thesis

Kristian Nymoen

Abstract

There are strong indications that musical sound and body motion are related. For instance, musical sound is often the result of body motion in the form of sound-producing actions, and muscial sound may lead to body motion such as dance. The research presented in this dissertation is focused on technologies and methods of studying lower-level features of motion, and how people relate motion to sound. Two experiments on so-called *sound-tracing*, meaning representation of perceptual sound features through body motion, have been carried out and analysed quantitatively. The motion of a number of participants has been recorded using state-of-the-art motion capture technologies. In order to determine the quality of the data that has been recorded, these technologies themselves are also a subject of research in this thesis.

A toolbox for storing and streaming music-related data is presented. This toolbox allows synchronised recording of motion capture data from several systems, independently of system-specific characteristics like data types or sampling rates.

The thesis presents evaluations of four motion tracking systems used in research on music-related body motion. They include the Xsens motion capture suit, optical infrared marker-based systems from NaturalPoint and Qualisys, as well as the inertial sensors of an iPod Touch. These systems cover a range of motion tracking technologies, from state-of-the-art to low-cost and ubiquitous mobile devices. Weaknesses and strengths of the various systems are pointed out, with a focus on applications for music performance and analysis of music-related motion.

The process of extracting features from motion data is discussed in the thesis, along with motion features used in analysis of sound-tracing experiments, including time-varying features and global features. Features for realtime use are also discussed related to the development of a new motion-based musical instrument: *The SoundSaber*.

Finally, four papers on sound-tracing experiments present results and methods of analysing people's bodily responses to short sound objects. These papers cover two experiments, presenting various analytical approaches. In the first experiment participants moved a rod in the air to mimic the sound qualities in the motion of the rod. In the second experiment the participants held two handles and a different selection of sound stimuli was used. In both experiments optical infrared marker-based motion capture technology was used to record the motion. The links between sound and motion were analysed using four approaches. (1) A pattern recognition classifier was trained to classify sound-tracings, and the performance of the classifier was analysed to search for similarity in motion patterns exhibited by participants. (2) Spearman's ρ correlation was applied to analyse the correlation between individual sound and motion features. (3) *Canonical correlation analysis* was applied in order to analyse correlations between *combinations* of sound features and motion features in the sound-tracing experiments. (4) Traditional statistical tests were applied to compare sound-tracing strategies between a variety of sounds and participants differing in levels of musical training. Since the individual analysis methods provide different perspectives on the links between sound and motion, the use of several methods of analysis is recommended to obtain a broad understanding of how sound may evoke bodily responses.

Preface

The thesis is written for the Faculty of Mathematics and Natural Sciences at the University of Oslo for the degree of Philosophiae Doctor (Ph.D.). The work has been funded by the Department of Informatics, and included in the research project *Sensing Music-Related Actions* (SMA), which is funded by the Research Council of Norway, with project number 183180. The research has been conducted between 2008 and 2012, under the supervision of Jim Tørresen, and the co-supervision of Rolf Inge Godøy, Alexander Refsum Jensenius, and Mats Høvin. The work has been done within the interdisciplinary research group *fourMs* (Music, Mind, Motion, Machines), involving researchers from the Department of Musicology and the Robotics and Intelligent Systems research group (ROBIN) at the Department of Informatics.

Acknowledgements

I have many people to thank for help and support during the period that I have been working on this thesis. First and foremost, I am grateful to my supervisors who provided invaluable advice and support throughout the whole Ph.D. project. *Jim Tørresen* has pushed me forward and been a supportive sparring partner, and taught me techniques for machine learning and classification. *Rolf Inge Godøy* provided insightful advice on perception and motor theory, and was the one that introduced me to the intriguing field of music cognition. *Alexander Refsum Jensenius* encouraged me towards making this thesis reach its full potential, and taught me about visualisation of motion, and the use of motion tracking technologies in music research.

Next, thank you to my good colleagues in the fourMs and ROBIN research groups in Oslo. Many discussions with fellow Ph.D. student *Ståle A. Skogstad* have kept my motivation high, and Ståle's work on real-time musical applications of motion tracking data has been a great inspiration. Thanks also to *Arjun*, *Arve*, *Gordon*, *Yago*, *Kyrre*, *Anders*, *Even*, *Ripon*, *Yngve*, *Ole Jakob*, *Dirk*, *Mats*, *Kim*, *Alexander*, *Markus*, *Simen*, and *Eivind*. Furthermore, I would like to thank *Tor Halmrast* and *Tellef Kvifte*, whose knowledge about acoustics, sound theory, and organology was a great inspiration during my undergraduate and master studies.

I have been fortunate to have had the opportunity to collaborate with researchers from other institutions than my own. Thanks to *Mariusz Kozak* for great discussions about music perception, synchronisation and visualisations, and for proofreading several of my papers. Thanks to *Baptiste Caramiaux* for introducing me to canonical correlation analysis, and to *Mats Kussner* for guidance regarding correlation of musical time-series. Further, thank you to *Birgitta Burger* for a good collaboration with the motion capture workshop at the NIME conference in Oslo 2011, and to *William Westney* and other participants in the NNIMIPA network, who provided great data recordings for the visualisation section of this thesis.

Moreover, I would like to thank the developers of *Jamoma*, for assisting me in the development of tools for storing and streaming data in the Gesture Description Interchange Format, and *Diemo Schwartz* for the SDIF tools in FTM that made these implementations possible. Also, thank you to all the people who participated in my experiments. It would not have been possible to do this research without you!

My devotion for music has been an important motivation for this research. This interest for music would not have been were it not for a handful of great musicians and close friends with whom I have spent innumerable hours rehearsing and performing: *Thomas Fredriksen*, *Tony André Bogen Heitmann*, *Christoffer Clausen*, *Thomas Wicklund-Larsen*, *Ole Kristian Sakseid*, *Tone Synnøve Alfsen*, *Espen Norbakken*, *Håkon Eivind Larsen*, *Kjetil Hammersmark Olsen*, as well as the great musicians in Kraftverket and Oslo Laptop Orchestra. Thank you!

I am deeply grateful to all my friends and family for understanding my physical and mental absense in the latest years. I'll do my best to make it up to you! *Olav*, *Tine*, *Tonje*, and *Hanne*, thank you for supporting my musical activities for all these years, and for showing interest in my research even when I have been too deep into it to explain properly. Finally, a warm thank you to *Astrid*. Thank you for your patience with me during the months of thesis writing. Also thank you for teaching me Adobe Illustrator, and even making some of the illustrations and a poster for my papers, and for your support and your motivating words in tough times.

Kristian Nymoen
October, 2012

Contents

Chapter 1

Introduction

This chapter introduces the motive for and foundation of this research. Research objectives are stated and the thesis outline presented at the end of the chapter.

1.1 Motive

Have you ever been listening to music and suddenly noticed that your foot is tapping along with the beat? Have you felt the need to put all your energy into that invisible drum kit that surrounds you when nobody is watching? And have you stretched your neck as far as you can to try to sing a pitched tone, or frowned to reach a really low one? I have done all of these and also, as a musician I have experienced how my body moves a lot when I play — much more than what is necessary just to produce the tones I am playing. Moving along with the music and lifting my instrument to emphasise salient phrases adds something to the musical experience that is difficult to explain in words. These things have puzzled me, and made me pursue a path that has emerged in music research in the recent decades, where music-related body motion is studied in order better to understand how music is perceived and processed in mind and body, and why music plays such a large role in the lives of so many people.

My own background is interdisciplinary. I have always been interested in music and started playing and composing music when I was a child. During my years as a Musicology student at the University of Oslo I became increasingly intrigued by research questions of music perception and cognition, and how music-related body motion could provide some of the answers to these. After finishing my Master studies in Musicology, where I focused on music cognition and technology, I started this PhD project in Informatics in 2008. Initially my interest for Informatics was mainly as a means of studying music-related body motion, where quantitative methods like pattern classification and machine learning might be used to learn more about music cognition. However, while my strong interest in music cognition persisted, I also developed a great fascination for research questions in computer science and the quantitative methods themselves. If music-related body motion can help us understand more about music, how can quantitative methods and technologies assist in this research, and when do these techniques fall short?

1.2 Multimodality

An important term in music cognition is *multimodality*. The Oxford Dictionaries defines the term *modality* as 'a particular form of sensory perception', citing visual and auditory modalities as examples [Oxford Dictionaries: "Modality"]. In human-computer interaction, the term has been defined as 'a type of communication channel used to convey or acquire information' [Nigay and Coutaz, 1993]. In the human body, these communication channels are formed by the sense organs, the nerve tracts, the cerebrum, and muscles [Schomaker et al., 1995]. Correspondingly, *multimodality* is the capacity to communicate along different types of communication channels [Nigay and Coutaz, 1993]. A conversation is a typical example of a multimodal phenomenon, where information is communicated through the spoken words as well as bodily gestures. As will be discussed in Section 2.3, there are also certain examples in human perception of interaction between the modalities, a phenomenon known as *cross-modality*.

This thesis takes as given that *music is multimodal*, meaning that music can be communicated through several modalities. Obviously the auditory modality is one of these, but music is more than what we hear. In most cases musical sound is the result of bodily motion in the form of sound-producing actions. Music also often results in body motion such as dance, foot-tapping, head-nodding or the playing of air-instruments. It has been suggested that musical sound *affords* motion, and therefore that by studying the way people move to music, we can gain knowledge about how music is perceived [Clarke, 2005, Godøy, 2010, Leman, 2008]. Other important non-auditory aspects of music are sensations of effort, and visual and tactile cues, as can be seen by the many metaphors that are used to describe musical sound, e.g. 'sharp', 'mellow', 'soft', 'bright', 'dark', 'aggressive', 'smooth' [Godøy, 2003].

A focus on the multimodality of music is one of the characteristics of the field of *Systematic Musicology*. This field is primarily empirical and data-oriented [Parncutt, 2007]. Thus, systematic musicologists conducting research on music and motion often work with quantified representations of both sound and motion data. Quantitative motion data may involve measurement with sensors or cameras, where the positions of limbs are measured at a fixed sampling rate. Furthermore, abstractions of the positional measurements can be gained by calculating the distances between various limbs, or the velocity and acceleration of the limbs. Similar abstractions can be made for audio data, for instance calculating representations of an audio signal that matches some perceptual model of our auditory system. Ways of quantifying and processing sound and motion data will be covered in more detail in the following chapters.

Quantitative data may be of great help to researchers, for instance in capturing nuances of sound and motion that are too subtle for humans to perceive. However, David Huron [1999] recommends that researchers in this be cautious in concluding that a quantitative result is equivalent to the ground truth. He illustrates his point with the rise of musical notation. Musical scores are in principle a quantification of music, and it inspired and facilitated the growth of music theory in the West. However, music notation, says Huron, is not the same as music. It is a simplification, unable to cover the full complexity of music. Consequently, although quantitative methods may facilitate research on music and motion, it is important to not to disregard qualitative analysis in such experiments.

1.3 Interdisciplinarity

As mentioned, my own background is interdisciplinary, and so is the research that is presented in this thesis. In many ways, interdisciplinarity is necessary to capture the full complexity of a multimodal concept such as music: In the Arts and Music Theory we find a tradition of describing for instance the aesthetics of music, and representations of musical pieces from many centuries ago in the form of musical scores. Acoustics and Biology tell us how sound is produced as physical vibrations in a musical instrument, and how these vibrations pass through the air and into our auditory system, where it eventually ends up as nerve impulses that are sent to the brain by tiny hair cells in the cochlea. Neuroscience and Psychology provide means of understanding how sound is processed cognitively, and how it is connected to other modalities. Biomechanics provides means of describing how people move to music, and Mathematics and Information Technology provide tools for capturing and analysing music-related data. This list could be extended further, but my purpose here is merely to show that researchers in Systematic Musicology must wear many hats, being music theorists, psychologists, acousticians, etc., while keeping a focus on both quantitative and qualitative methods.

The readership of this thesis is not assumed to have knowledge of the methods and terminology of the several research disciplines of Systematic Musicology. For this reason the terminology, methods and technologies in Chapters 2, 3, and 4, will be presented in such a way that it is accessible without expert knowledge of quantitative research methods, or with limited knowledge of sound and music.

1.4 Aims and Objectives

The main research objective of this thesis is to:

> *develop methods and technologies for studying links between musical sound and music-related body motion*

This objective may further be divided into three sub-objectives:

Data handling:
to develop solutions for storing and streaming synchronised music-related data

Evaluation of motion tracking technologies:
to evaluate the quality of motion tracking systems used for analysing music-related motion

Sound–action analysis:
to evaluate existing and develop new methods and techniques of analysing bodily responses to short sound excerpts

Studies of music-related motion require tools and methods of analysis that are able to handle the multidimensionality that this area presents. Music-related data may involve audio, musical scores, MIDI-data, annotations, video, motion capture data from various systems, and more. When researchers want to work with some or all of these at the same time, the ability to handle

a large amount of data with different sampling rates and number of dimensions is essential. Preferably, it should be possible to make synchronised recordings of all of these data types, and to play back the data later, just as easily as one would play back a video on a camcorder. Furthermore, evaluation of existing technologies for studying body motion is essential: what degree of precise measurement is possible with high-end motion tracking equipment? And what about low-end tracking technologies like the sensors that are found in ubiquitous mobile technology?

One way of studying music-related body motion is to observe how people move while listening to music. If the body motion is measured quantitatively, for instance by a motion tracking system, effective analysis methods are required. Not all methods of analysing time series can be applied to multidimensional music-related data. Nor is it given that the methods capable of handling multidimensionality can provide as detailed analysis results as those that cannot. Another approach to studying music-related body motion turns the process around. Development of new musical instruments that use body motion to produce sound can teach us how people want to interact with music, and thus also provide knowledge of how properties of bodily motion correspond to sonic properties.

This thesis tries to answer some of the questions posed above, with a main focus on the use of optical marker-based motion tracking technology, and how the data obtained from this can be used to analyse correspondences of sound and motion. I present experiments referred to as *sound-tracing*, where a number of people have moved one or two rigid objects in the air, following the perceptual features of short sound objects. Some of the research could be extended to full-body motion and longer segments of music. However, a focus on short sound objects and simple action responses has enabled the application and evaluation of multiple analysis methods, as well as development of new technologies.

1.5 Thesis Outline

This thesis is a *collection of papers* and thus the eight included research papers constitute the research contribution of the thesis. The first part of the thesis offers an overview of the work that has been carried out and is structured as follows: Chapters 2, 3, and 4 introduce relevant background, including some basic theory on perception and cognition of sound and music in Chapter 2, an overview of technologies of motion tracking in Chapter 3, and a presentation of analytical methods that are applied in the thesis in Chapter 4. Chapter 5 presents an overview of the contents of the research papers, as well as individual summaries and abstracts for each paper. Subsequently, Chapter 6 discusses the findings of the papers and presents conclusions and pointers for future work. Finally, the eight research papers are included at the end of the thesis.

The source code of the software that I have developed as part of this thesis is included digitally, together with the sound files that have been designed for the empirical studies. These will be made available online in the archive for digital publications at the University of Oslo (DUO),[1] and they are also available from my personal web page.[2]

[1] http://www.duo.uio.no/

[2] http://folk.uio.no/krisny/files/knThesisAttachment.zip

Chapter 2

Music Cognition

The term *music cognition* refers to mental processes of perception and processing of music. As argued in the previous chapter, music is multimodal. Consequently, cognitive musical processes must involve not only sound perception, but also a motor modality. Still, sound is obviously an important element of music and, as will be explained in the presentation, there exists evidence for links between sound perception and motor processes in the brain. Discussion of sound constitutes a good point of departure for the presentation of music cognition.

2.1 Sound Descriptors

Sound and music may be described in many different ways. In daily speech, music is commonly described in terms of adjectives, such as 'groovy' or 'smooth', or in terms of an experienced emotional content in the music, e.g. 'sad' or 'passionate'. Other common ways of describing sound is through metaphors, such as 'bright', 'warm', 'big' [Lakoff and Johnson, 1980, Eitan and Timmers, 2010], or through genre labels, such as 'opera', 'hip-hop' or 'jazz'. While all of these sound descriptors connote sound properties, the terms do not give precise information about the sound signal. For this reason, lower-level quantitative features are often used in sound analysis to enable more precise descriptions of nuances in the sound.

When working with quantitative sound features, it is important to be aware of the distinction between *physical* and *perceptual* features. The former describe sound in terms of physical parameters like sound pressure level or the spectral content of the sound wave. Perceptual sound features are designed to describe sound as we hear it, typically by applying a perceptual model that take into account certain limitations of our auditory system.

A sound signal may physically be described as a sum of sinusoidal components with respective frequencies and amplitudes. This is illustrated in Figure 2.1 where the sum of 8 sinusoidal components makes up an audio wave that resembles a sawtooth wave. As such, a sound wave may not only be described as a function of time, but also of frequency. A sound signal in the time domain is usually referred to as a *waveform*, and a signal in the frequency domain is known as a *spectrum*.

The time domain signal is commonly separated into shorter segments, known as *frames*, before features are calculated for each segment. The segments are extracted by multiplying the audio signal with a window function, which smooths the beginning and end of the frame.

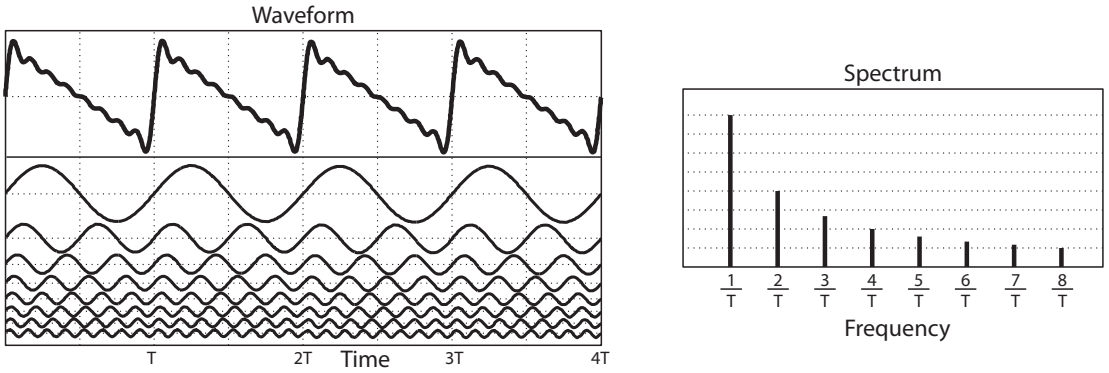


Figure 2.1: The figure shows how the sum of 8 sinusoidal waves resembles a sawtooth wave. The spectrum of the signal is displayed on the right.

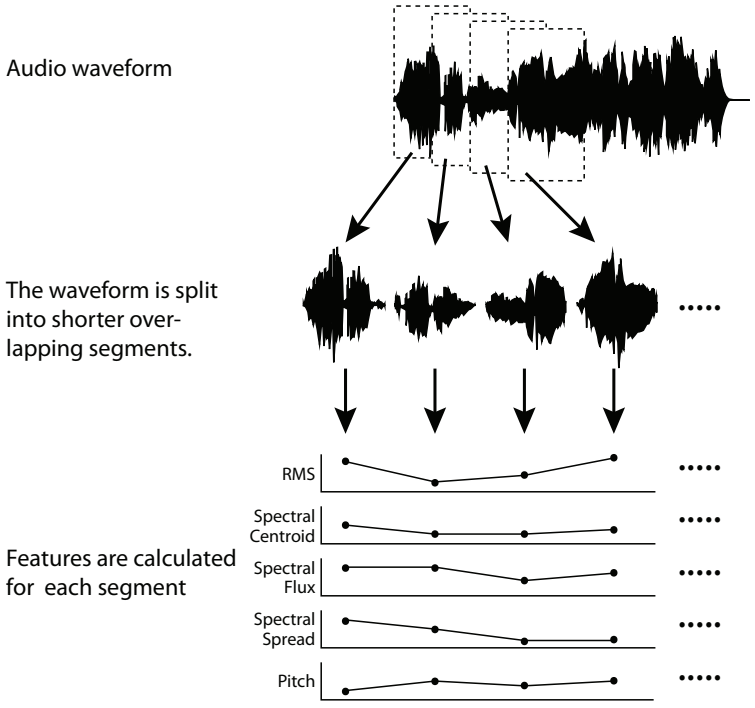


Figure 2.2: Audio features are calculated by segmenting the audio waveform into frames. Each frame is multiplied by a window function to smooth the beginning and end.

Features are calculated upon the waveform and the spectrum within each frame as displayed in Figure 2.2. In this manner time-varying sound descriptors are obtained.

A number of excellent tools for extracting sound features from an audio signal have been developed, and many of them are available free of charge, such as the standalone applications *Praat* [Boersma and Weenink, 2012], *Sonic Visualiser* [Cannam et al., 2010], and *Spear* [Klingbeil, 2005], and the *MIR Toolbox* [Lartillot et al., 2008] and the *Timbre Toolbox* [Peeters et al., 2011] for Matlab.

A detailed listing of specific audio descriptors is beyond the scope of this thesis. Accordingly, the information presented here will concern a limited number of examples. For a comprehensive list of physical and perceptual audio descriptors, please refer to Geoffroy Peeters and others' work on audio features [Peeters, 2004, Peeters et al., 2011]. Adopting terminology put forward by Every [2006], the features I have worked with in the experiments presented in this thesis include *dynamic*, *harmonic*, and *spectral*:

Dynamic features describe the energy of the sound. An example of a physical dynamic feature is the root-mean-square value of the audio signal within a frame. This feature is an important component of the perceptual feature *loudness*. Loudness is not only dependent on energy, but also on the spectral distribution of sound energy [Mathews, 1999a].

Harmonic features concern the periodicity of an audio signal. The frequency whose integer multiples best describe the content of the signal spectrum is known as the *fundamental frequency* of a harmonic signal. This value may be the same as the *zero-crossing frequency* of the waveform. *Pitch* is a perceptual feature which is closely related to the fundamental frequency. However, the perceived pitch can be lower than the actual spectral content of the sound (so-called *missing fundamental*), or in cases with insufficient harmonic content in the audio spectrum, there might not be a perceivable pitch at all [Pierce, 1999].

Spectral features describe the distribution of spectral content in the sound. Within a frame, we can for instance calculate the *spectral centroid*, denoting the barycentre of the spectrum (illustrated in Figure 2.3), or the *spectral flux*, denoting the degree of change in the spectrum between the current and previous timeframe. The perceptual feature *timbre* is to a large extent dependent on spectral content. However, this feature is intrinsically multidimensional, and therefore difficult to quantify. Efforts have been made to develop multidimensional ordering of the timbres of musical instruments [Grey, 1977, McAdams et al., 1995], and methods of analysing timbre through synthesising sound [Risset, 1991]. These methods provide good foundations for reasoning about timbre, but many other aspects of sound will influence this perceptual sound feature, and thus it is not trivial to design a comprehensive model of timbre [Mathews, 1999b].

Low-level sound features as presented above are useful in describing sound signals. Some also describe the perception of sound, by employing a model that simulates limitations in our auditory system. The methods outlined above show how features can be calculated through a *bottom-up approach*, where the time-domain signal is transformed and processed in order to obtain new representations of the signal.

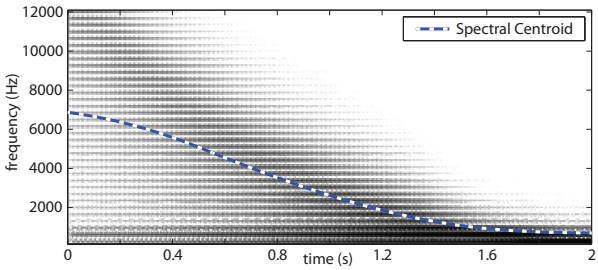


Figure 2.3: Simple illustration of how the *spectral centroid* changes when high-frequency content is filtered out from the signal.

Antonio Camurri has presented a multilayered model for musical signals, outlined in Figure 2.4 [Camurri et al., 2005, Camurri and Moeslund, 2010]. The model illustrates how music can be described at higher levels than the features presented thus far. The basic physical signals (Layer 1) can be represented in terms of low-level continuous features (Layer 2). Furthermore, the signals can usefully be segmented into shorter meaningful units, such as musical phrases or sonic objects (Layer 3). At Layer 4, we find concepts and structures which contain emotive and expressive content of the music. An ideal model of musical signals would we able to translate between all of these layers, so that emotive content could be analysed from physical signals (bottom-up), and physical signals could be synthesised from emotive or metaphorical descriptions (top-down).

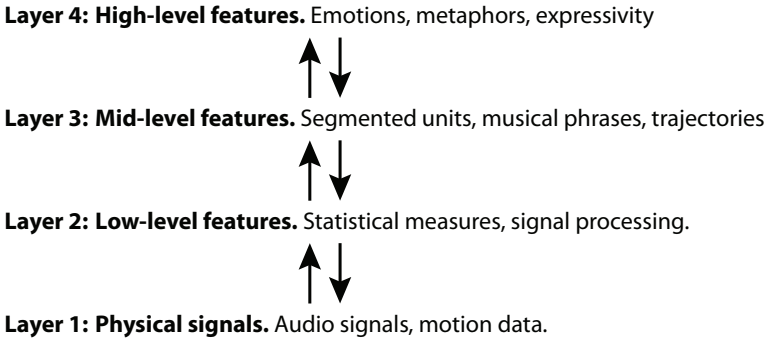


Figure 2.4: My illustration of Camurri's multilayered model for musical signals [Camurri et al., 2005, Camurri and Moeslund, 2010]. With sophisticated signal processing techniques, it should be possible to move between the different layers.

In sound perception, we are not consciously aware of quantitative perceptual features. Rather, our focus is on understanding the continuous auditory input at one of the higher levels, as events and objects, in order to make sense of our surroundings. This will be discussed further in the next section.

2.2 Sound Perception

We all have some understanding of what sound is. We experience it and interact with it every day. We know that clapping our hands together will produce sound, and we can even predict in detail what the hand clap will sound like. We have obtained this so-called *ecological* knowledge by interacting with our bodies in the world [Gibson, 1979, Clarke, 2005]. Our experience with sound enables us to infer the causes of the sounds that we hear, and to determine the direction of, and distance from the sound sources. Albert Bregman [1990] presented a comprehensive

theory on *Auditory Scene Analysis*, describing how we make sense of our surroundings based on auditory input. He investigated a number of principles according to which auditory input is grouped and segregated into so-called *streams*. We can distinguish the sound of a passing vehicle from the voice of a person we are talking to, and even distinguish their voice from the voices of many others in the same room. Furthermore, from auditory input we are able to tell with quite high precision, the gender and age of people around us, as well as the size of passing vehicles and in which direction they are going, and many other details about our surroundings.

The foundations for Bregman's work were laid by a number of German researchers and psychologists in the 19th and early 20th century. Herman von Helmholtz, Wilhelm Wundt and Franz Brentano conducted pioneering work in the fields of Psychoacoustics, Psychology, and Phenomenology, respectively [Leman, 2008]. Later the German *gestalt* psychologists formulated a series of principles for grouping and segmentation of visual stimuli. In the second half of the 20th century, as the technologies for recording, manipulating and playing back sound improved, a large number of researchers contributed to the field of psychoacoustics. In particular, Leon van Noorden's [1975] work on how the perception of tone sequences dependents critically on tempo, and Stephen McAdams' [1984] work on spectral fusion were important contributions to Bregman's work.

2.2.1 Discrete Attention

It is commonly accepted that we do not all the time pay equal attention to all parts of the sound waves that reach us but that we are able to focus our attention on certain parts of the auditory scene. Edmund Husserl's phenomenology described consciousness as a phenomenon consisting of a series of discrete *now-points* in time [Husserl, 1964]. Every conscious "now" is not only an infinitely short time period along a continuous timeline. The "now" also contains awareness of events that occurred just before the present and expectations of what will happen in the near future. This notion in many ways coincides with George A. Miller's concept of *chunks*, which is an explanation of how the continuous input to our sensory system is re-coded and perceived as discrete, holistic units [Miller, 1956].

In his work on the aesthetics for *musique concrete* the French composer and theorist Pierre Schaeffer claimed that music is not perceived as a continuous phenomenon, but rather in terms of *sonic objects*, which are discrete perceptual units, defined by some apparent *cause* [Schaeffer, 1966]. A cause may be an intentional sound-producing action, such as hitting a drum with a mallet, or a naturally occurring event such as the howling sound that is heard when the wind blows past a resonating object. Schaeffer challenged the traditional way of listening to music. He claimed that music should be listened to by disregarding the causes of the sonic objects, and rather focusing on lower-level features of the sound [Schaeffer, 1966].

Schaeffer's approach to sonic objects is in contrast to the bottom-up approach to sound features that was presented in Section 2.1. Schaeffer started with the sonic object and defined types of object according to their onset characteristics and pitch contour. He further inspected low-level features of the different types of object through a number of experiments with sound recordings on magnetic tape, where he manipulated the recordings by cutting them apart and gluing together in new orders, and by increasing and decreasing the speed of the tape [Schaeffer and Reibel, 1967]. If we see this method in light of Camurri's model, we note that Schaeffer's

approach to sonic features, unlike the previously introduced bottom-up approach, is a *top-down* approach. The typology of sonic objects was used as a point of departure, and Schaeffer studied the different types in order to find the features that defined them.

It is well established that sound is processed cognitively as discrete perceptual units and that we are unable simultaneously to be aware of the entire range of physical sound waves that reach us [Alain and Arnott, 2000, Shinn-Cunningham, 2008]. Whether we choose to focus on the perceptual units as short 'sonic objects', or sequences of these that are grouped and segregated into 'auditory streams', they are often identified by stemming from the same cause. Thus several researchers have suggested that we may learn more about the workings of sound perception by looking at more than the auditory modality, and start analysing the causes of sound objects. In music the causes of sound objects are usually human bodily motion in the form of sound-producing actions. This will be covered in the next section.

2.3 Music and Motion

Body motion is an important aspect of music. Several researchers have developed taxonomies to describe various types of music-related motion all of which show how tightly connected music and motion are, e.g. [Cadoz and Wanderley, 2000, Wanderley and Depalle, 2004, Jensenius, 2007a, Jensenius et al., 2010]. These types of motion span from the sound-producing actions of performers on musical instruments to the bodily expression of musical structures through dance, and even our unconscious foot-tapping or head-nodding when listening to music. The combination of music and motion has been the main focus of a large number of recent academic publications, including anthologies edited by Wanderley and Battier [2000], Gritten and King [2006, 2011], Altenmüller et al. [2006], and Godøy and Leman [2010].

Experimental psychologists have shown that our understandings of perceived phenomena might not be obtained through one sensory modality alone, but often through a combination of modalities [Stein and Meredith, 1993, Vroomen and de Gedler, 2000]. This phenomenon is usually referred to as *cross-modality*. A good example is the so-called *McGurk effect*, which explains how we rely on multiple modalities to interpret a spoken syllable. McGurk and MacDonald [1976] showed that when perceivers saw a video of a person saying 'gaga' combined with hearing a person saying 'baba', the spoken word was perceived as 'dada'. Combinatory effects of audition and vision have been the main focus of research on cross-modal perception, but evidence for an interaction between these modalities and a motor modality has also been found. This has contributed to the idea of *embodied music cognition* which has emerged among music researchers in the last decades [Leman, 2008].

An embodied approach to music cognition entails regarding music as not only an auditory phenomenon, but recognising that the human body is an integral part of our experiences of music [Leman, 2008]. This idea builds upon Gibson's ecological approach to visual perception, known as *embodied cognition* [Gibson, 1979], and later the theories of *motor perception* which suggest that perceived phenomena are understood through ecological knowledge of how our own bodies interact with the environment, for instance that we understand speech by projecting the phonemes that we hear onto our own experience with producing phonemes in our vocal system [Liberman and Mattingly, 1985].

2.3.1 Evidence from Neuroscience

In the early 1990's neuroscientists made a discovery that supported the theories of motor perception. A type of neuron, called *mirror neurons* were discovered in the premotor cortex of the brains of macaque monkeys [Pellegrino et al., 1992, Gallese et al., 1996]. These neurons were found to activate not only when the monkeys performed a learned task, but also when observing the experimenter perform the same task. This was later also shown to be true for auditory stimuli, suggesting that the monkey would understand the sound by imagining performing the action that created the sound [Kohler et al., 2002].

In music research motor activity in the brain has been shown in musicians and non-musicians imagining musical performance, and even when just listening to music [Haueisen and Knösche, 2001, Langheim et al., 2002, Meister et al., 2004, Lahav et al., 2007]. And the other way round; activity in the auditory cortex has been found in piano players watching a silent video of piano performance [Haslinger et al., 2005]. These and other findings have inclined researchers to claim that by studying musical activity in more detail, we can learn more about neuroscience in general [Zatorre, 2005, Zatorre and Halpern, 2005].

2.3.2 Sonic Objects are also Action Objects

Interestingly, Schaeffer's theory of sonic objects included a typology based on sound excitation. He described sonic objects as *impulsive*, *sustained*, or *iterative*. Building on the above mentioned work in phenomenology, psychoacoustics, and embodied cognition, Rolf Inge Godøy [2004, 2006] linked these categories to so-called *gestural imagery*, and claimed that visualising or imagining action trajectories is essential to our perception of music. Such trajectories can be seen as a *covert* mirroring of sound-producing actions [Cox, 2006, Godøy et al., 2006a].

Sonic objects and sound-producing actions share the property of being chunked holistic units taken from a continuous phenomenon (sound and motion, respectively). In both modalities, grouping and segmentation of units follow the gestalt principles [Bregman, 1990, Klapp and Jagacinski, 2011]. Limitations of our attention and short-term memory [Pöppel, 1997, Snyder, 2000] and motor abilities [Schleidt and Kien, 1997] constrain these units to about the 0.5 to 5 seconds range. Godøy [2011] suggested a model on which sound and action are analysed at three timescale levels:

- *Sub-chunk level*, meaning continuously varying sound and motion features.
- *Chunk level*, meaning holistically perceived units in the 0.5–5 seconds range.
- *Supra-chunk level*, meaning sequences of concatenated chunks, such as a musical phrase, that consist of several sonic objects.

Sound excitations can coincide with the chunk level, but multiple sound onsets can also be found within a single chunk. This is often best observed by looking closer at the sound-producing actions. For a rapid piano scale, the fast finger actions fuse together into superordinate trajectories in the elbow and shoulder joints [Godøy et al., 2010]. And similarly, for violin bowing actions, increased bowing frequency will cause so-called *phase transitions*, where the principal source for the bowing changes from the elbow joint to the wrist joint [Rasamimanana et al., 2009].

Godøy [2011] argued that chunking of music-related actions happens in terms of *goal-points*. In sound-production, such goal-points can exist at the time of excitation, or in the case of multiple excitations within a chunk, at a salient point in the sequence of excitations. Before and after each goal-point are trajectories leading to it and back again. Musicians prepare the trajectory to the next goal-point before hitting the previous goal-point, and consequently a series of chunks will have overlapping trajectories — an effect known as *coarticulation* [Godøy et al., 2010]. Picture, for instance, how a mallet is struck against a drum. The mallet bounces off the membrane and a drummer will typically initiate a new stroke before the rebound has stopped. If the stroke that follows is on a different drum, the coarticulation between strokes will involve preparatory, so-called *ancillary*, motion in order to move the mallet to the next drum, e.g. by turning the body or by raising the elbow and shoulder in the direction of the other drum.

2.4 Summary

This chapter has shown that music perception is multimodal. Musical sound can be studied by looking at features extracted from the sound signal, either as continuously varying features at the sub-chunk level, or as holistically perceived chunks, or sequences of these. Furthermore, musical sound can be studied by observing the motion people make to music, as an overt expression of the covert mirroring of sound-producing actions that occurs when we listen to music. A natural follow-up question to this is how motion can be studied in a controlled manner. This is addressed in the next chapter.

Chapter 3

Motion Capture

When working with music and body motion it is essential to be able to convey information about how someone or something moves. In daily speech we use words such as 'walking', 'rolling', 'turning', etc., to achieve this. These words, however, do not provide precise descriptions of motion. More detailed representations of motion can be gained through visualisation techniques, such as a video recording, or through a sequence of photographs, drawings or storyboards [Jensenius, 2007a].

Motion capture (mocap) involves the use of a sensing technology to track and store movement. In principle, a pencil drawing on a piece of paper can be called motion capture, since the pencil lead is testimony of the hand motion of the person that made the drawing. However, the most common use of the term refers to tracking and representation of motion in the digital domain.

3.1 Motion Capture Basics

Figure 3.1 shows how motion capture may be divided into three main parts: (1) sensing the motion, (2) processing the sensor data, and (3) storing the processed data. Together, parts 1 and 2 are referred to as *motion tracking*. Rather than being stored, tracking data may be used directly, for instance in realtime interactive applications. Most commercial implementations of tracking technologies include the option of storing data, and so the terms *motion tracking system* and *motion capture system* are often used interchangeably.

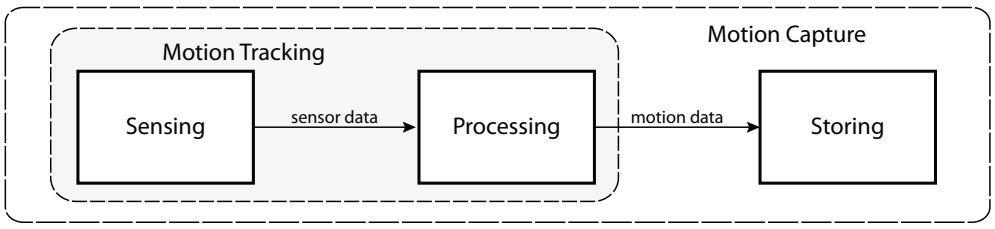


Figure 3.1: Motion tracking involves sensing motion and processing the sensor data. When motion data are stored in order to apply post-processing later, the process is known as *motion capture*.

3.1.1 From Sensor Data to Motion Data

The sensing part of a motion capture system involves measuring some aspect of the motion. This could be done by a large variety of sensors, such as a simple potentiometer or an array of advanced video cameras. In principle, the sensor data can be stored or used directly. However, these data are rarely interesting in themselves, as they typically provide sensor-specific measurements, e.g., resistance in a potentiometer or colour information of camera pixels. Consequently the processing part of a motion capture system translates the raw sensor data into information that describes the motion more significantly, for instance as low-level measures of position or orientation or derivatives of these, such as velocity, acceleration or rotation. Furthermore, certain systems provide motion data specific to the object that is tracked, such as joint angles in a human body.

For positional and orientational measurements the term *degrees of freedom*[1] (DOF) denotes the number of dimensions that are tracked. For instance, 2DOF position would mean the position on a planar surface, and 3DOF position would be the position in three-dimensional space. The description 6DOF is normally used to denote a measurement of an object's three-dimensional position and three-dimensional orientation. 6DOF-tracking is sufficient to represent any position and orientation.

3.1.2 Tracked Objects

Tracking can be applied to point-like objects, such as small spherical *markers*. These are treated as points without volume, and as such only their position (not orientation) can be tracked. A fixed pattern of several markers can be used to identify a *rigid object*. Rigid objects are non-deformable structures whose orientation and position can be tracked. Furthermore, by combining multiple rigid bodies and defining rules for the rotations and translations that can occur between them it is possible to create a *kinematic model*. Such a model may, for instance, represent the human body with the various constraints of the different joints. Such models can even fill in missing data: say, if the data from the lower arm are missing, but the data from the hand and the upper arm are present, the missing data can be estimated by following the kinematic model. Kinematic models might not need position measurements of the different parts: a set of joint angles for the body can be sufficient for a well-defined model. Examples of a marker, a rigid object and a kinematic model are shown in Figure 3.2.

A more formal discussion of how position and orientation can be represented will follow in Section 3.3. First, we shall have a look at the different technologies that are used in motion tracking.

3.2 Motion Tracking Technologies

There is a large variety of motion tracking technologies. The most advanced technologies are capable of tracking motion with very high precision at very high sampling rates. The largest

[1] This should not be confused with the statistical variable *degrees of freedom* (*df*), which is used to denote the size of a tested data set in standardised statistical tests such as *t*-tests and ANOVAs (see Section 4.2). Furthermore, in biomechanics and robotics degrees of freedom (DOF) is usually used to denote the number of rotary and linear joints in kinematic models [Rosenbaum, 2001, Spong et al., 2006].

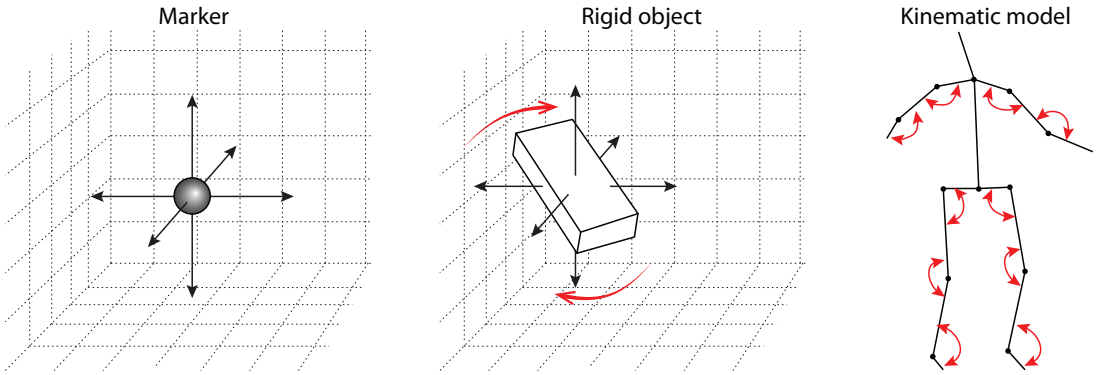


Figure 3.2: The position of a marker can be tracked in three dimensions. A rigid object also allows tracking of orientation. A kinematic model describes the relative position and orientation of connected rigid objects, for instance by joint angles.

appliers of these are the film and gaming industries where they are used for making life-like animations, and researchers who study biomechanics for rehabilitation and sports purposes. At the other end of the scale are ubiquitous low-cost sensor technologies that most people use daily in their mobile phones, laptops, game controllers, and so forth.

This section will give an overview of tracking technologies. The presentation below follows a classification of tracking technologies used by Bishop et al. [2001] where the different systems are sorted according to the physical medium of the technology. The technologies presented in this section include acoustic, mechanical, magnetic, inertial and optical tracking.

Several aspects of each technology will be presented. A description of the sensor technology as well as the algorithms involved in processing the sensor data constitute the technical details of the technology. Furthermore, the technologies differ in use and should be described in terms of the data they provide to the user, as well as their limitations and advantages in various tracking settings. What is more, in the context of this thesis it is interesting to discuss the use of the technologies in musical settings, such as the study of music-related motion or in interactive music systems.

3.2.1 Acoustic Tracking

Acoustic tracking systems calculate position upon the wavelength of an acoustic signal and the speed of sound. Systems based on *time of flight* measure the time between the sending of a signal from a transmitter and its being picked up by a receiver, and systems based on *phase coherence* measure the phase difference between the signal at the transmitter end and the receiver end [Bishop et al., 2001]. The speed of sound in air at 20 °C is about 343 m/s, but it varies with air pressure and temperature. It may therefore be difficult to acquire precise measurements from acoustic tracking systems. A single transmitter combined with a single receiver gives the distance between the two, or in other words the position of the receiver in a sphere around the transmitter. By adding more transmitters the 3D position of the receiver can be found.[2] Figure 3.3 shows how combined distance measurements from two transmitters narrows the possible positions of the receiver down to a circle.

[2]In addition to tracking the receiver position it is also possible to track the position of the transmitter. In this case adding more receivers would enable finding the 3D position.

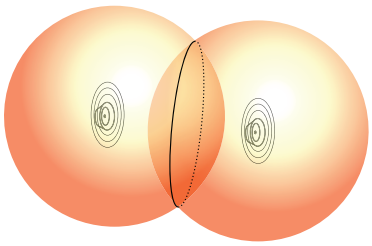


Figure 3.3: Distance measurements from two acoustic transmitters can determine the position of a receiver to be somewhere along a circle.

Acoustic systems usually work in the ultrasonic range and can therefore be used in music-related work without interfering with the musical sound. Still, these systems are not widely used in this area. Among the few examples of those using acoustic tracking are Impett [1994], Vogt et al. [2002] and Ciglar [2010], who included ultrasound sensors in the development of digital musical instruments.

3.2.2 Mechanical Tracking

Mechanical tracking systems are typically based on some mechanical construction which measures angles or lengths between the mechanical parts by using bend sensors or potentiometers. These systems can be worn on the body, for instance by implementing sensors in an exoskeleton or a glove, to obtain a model of the joint angles in the whole body or the hand.

There are other implementations of mechanical tracking systems in which the system is not placed on the body but rather contains a base unit placed at a fixed position in the room. Two examples are input devices such as the 'Phantom Omni' and the 'Gametrak' game controller, sketched in Figure 3.4. The Phantom Omni consists of a movable arm with several joints whose angles are measured by encoders. The Gametrak measures the position of a satellite unit which is attached to the base by a nylon cord. The extension of the nylon cord as well as the angle of the cord are measured, providing positional information for the end of the cord.

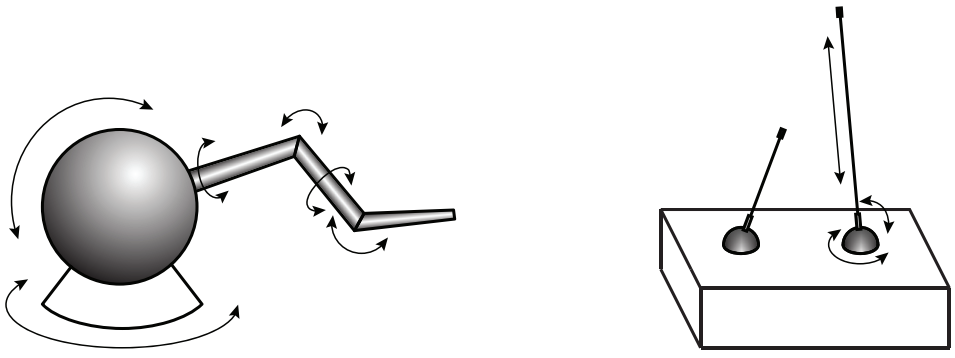


Figure 3.4: Two mechanical motion tracking devices. Left: The Phantom Omni senses the position of the tip of the arm. Right: the Gametrak game controller senses the position of the tip of the nylon cord. The arrows show the measured angles and lengths.

Mechanical tracking has been popular in music-related work, particularly for the purpose of developing new musical interfaces. Various exoskeleton implementations have been developed [e.g., de Laubier, 1998, Jordà, 2002, de Laubier and Goudard, 2006] and also a number of glove-instruments [e.g., Fels and Hinton, 1993, Ip et al., 2005, Hayafuchi and Suzuki, 2008,

Fischman, 2011, Mitchell and Heap, 2011]. Furthermore, Zadel et al. [2009] implemented a system for solo laptop musical performance using the Phantom Omni, and Freed et al. [2009] explored a number of musical interaction possibilities for the Gametrak system.

3.2.3 Magnetic Tracking

Magnetic tracking systems use the magnetic field around a sensor. Passive magnetometers can measure the direction and strength of the surrounding magnetic field, the simplest example being a compass which uses the Earth's magnetic field to determine the orientation around the Earth's radial vector. The field varies slightly across the Earth's surface, but this can be compensated for without much effort [Welch and Foxlin, 2002]. Passive magnetometers are widely used in combination with inertial sensors, which will be covered in the next section.

More advanced magnetic systems use an active electromagnetic *source* and a *sensor* with multiple coils. These systems are based on the principle of induction, which explains how an electric current is induced in a coil when it is moved in a magnetic field. To obtain 6DOF tracking a magnetic source with tree coils is used, each perpendicular to the two others [Raab et al., 1979]. Similarly, each sensor consists of three perpendicular coils. The position and orientation of each sensor can be calculated as a function of the strength of the induced signal in each sensor coil [Bishop et al., 2001]. An illustration of the Polhemus Patriot system is shown in Figure 3.5.

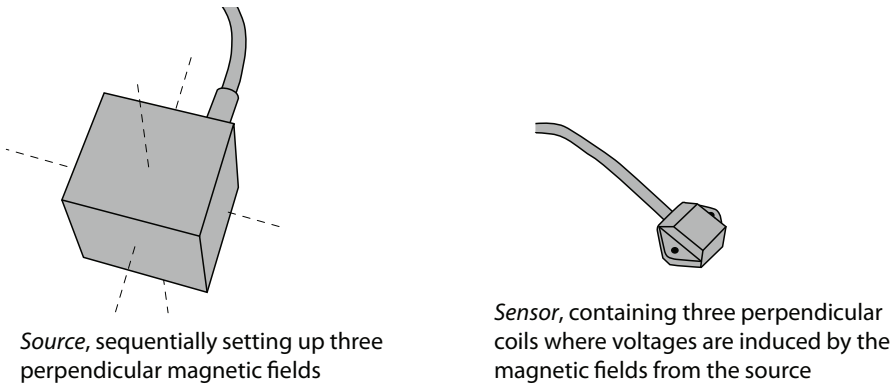


Figure 3.5: The Polhemus Patriot system sets up three perpendicular magnetic fields and tracks the position and orientation of up to two sensors.

Magnetic trackers are able to operate at high sampling rates (more than 200 Hz) with high theoretical accuracy.[3] However, the systems are sensitive to disturbances from ferromagnetic objects in the tracking area. Vigliensoni and Wanderley [2012] showed that the distortion is acceptably low at close distances from the magnetic source. But if a larger area is to be covered, it is necessary to compensate for the distortion of the tracking field [Hagedorn et al., 2007]. This, as concluded by Vigliensoni and Wanderley, may be particularly true for spaces used for musical performance, which often contain ferromagnetic objects. On the positive side, these trackers do not require a clear line-of-sight between the source and the sensor, meaning that the sensors can be hidden under clothes etc.

[3] According to the technical specifications of the Polhemus Liberty system the positional and orientational resolution decrease with increased distance between the source and the sensor. As long as the distance between the sensor and the source is less than 2 m, the system displays submillimeter accuracy [Polhemus Inc.].

Magnetic trackers have been used for analysis of music-related motion by a number of performers and researchers. Trackers from Polhemus have been the most popular, used by e.g. Marrin and Picard [1998], Lin and Wu [2000], Marshall et al. [2002], Ip et al. [2005], Marshall et al. [2006], Maestre et al. [2007] and Jensenius et al. [2008].

3.2.4 Inertial Tracking

Inertial tracking systems include those based on accelerometers and gyroscopes. These sensors are based on the physical principle of *inertia*. Accelerometers measure acceleration based on the displacement of a small "proof-mass" when a force is exerted to the accelerometer. Gravity will contribute to displacement of the proof-mass, and thus the data measured by accelerometers contain the acceleration that is due to gravity (9.8 m/s^2) and any acceleration applied by a user [Bishop et al., 2001]. Gyroscopes apply a similar principle but measure rotational changes. Vibrating parts in the gyroscope resist any torque that is applied to it, and by using vibrating piezoelectric tuning forks in the gyroscopes an electrical signal is emitted when torque is applied [Bishop et al., 2001]. To obtain 6DOF tracking three accelerometers and three gyroscopes are used, with each sensor mounted perpendicularly to the other two.

Inertial tracking systems have certain strong advantages over all the other tracking technologies. Firstly, they are completely self-contained, meaning that they do not rely on external sources such as acoustic ultrasound sensors or cameras which require line-of-sight. Secondly, the sensors rely on physical laws that are not affected by external factors such as ferromagnetic objects or light conditions. Thirdly, the sensors are very small and lightweight, meaning that they are very useful in portable devices; and finally, the systems have low latencies and can be sampled at very high sampling rates [Welch and Foxlin, 2002].

Orientation is gained from inertial tracking systems by integrating the data from the gyroscopes. Any change in orientation also means a change in the direction of the gravity force vector. Position is calculated by first adjusting for any change in the gravity vector, and then integrating the accelerometer data twice [Bishop et al., 2001].

Estimating position from accelerometer data leads us to the downside of inertial sensors; namely *drift*. Even a minor error in data from the gyroscope or the accelerometer will cause a large error in positional estimates. As noted by Welch and Foxlin [2002], a fixed error of 1 milliradian in one of the gyroscopes would cause a gravity compensation error of 0.0098 m/s^2, which after 30 seconds would mean a positional drift of 4.5 metres. For this reason, Welch and Foxlin [2002] conclude, inertial systems work best when combined with other technologies.

Figure 3.6 shows one example of combining inertial sensors with other technologies, namely the Xsens MVN suit [Roetenberg et al., 2009]. The suit uses 17 sensors called *MTx*, fixed at predefined positions on the suit, each containing an accelerometer, a gyroscope and a magnetometer (compass). By combining the sensor signals with a kinematic model, which restricts the positions and orientations of each body segment in relation to the other segments, a full-body model is constructed.

The Xsens MVN suit has been tested and evaluated for use in musical interaction by Skogstad et al. [2011], and actual implementations of the suit in musical interactive systems have been presented by Maes et al. [2010], de Quay et al. [2011] and Skogstad et al. [2012c].

Accelerometers and gyroscopes are now implemented in smart phones and laptops every-

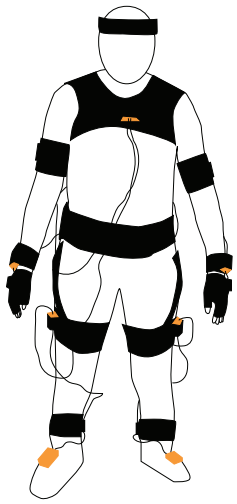


Figure 3.6: The Xsens suit consists of 17 MTx sensors combining inertial sensors and magnetometers. Full body motion capture is obtained through the use of a kinematic model.

where, and the use of inertial sensors in musical performance and research is widespread. This can be seen from the number of laptop orchestras and mobile phone ensembles that have appeared in the recent years [e.g., Trueman et al., 2006, Dannenberg et al., 2007, Wang et al., 2008, Bukvic et al., 2010, Oh et al., 2010].

3.2.5 Optical Tracking

Optical motion tracking systems are based on video cameras and computer vision algorithms. The systems of this type range more widely than do the other types in terms of quality and cost, and various implementations of optical tracking technologies can appear very different to the user.

Optical Sensing

Various types of video camera are used in optical motion tracking. In principle, any digital video camera can be used — in fact, one of the most affordable sensors for conducting motion tracking is a simple web camera. Cameras used in optical motion tracking are either (1) regular video cameras, (2) infrared (IR) video cameras, or (3) depth cameras.

Ordinary video cameras sense light in the visible part of the electromagnetic spectrum. Each pixel in the camera image contains a value corresponding to the amount of light sensed in that particular part of the image. Colour information in each pixel can be represented by using multiple video planes, with the pixel values in each plane representing e.g. the levels of red, green and blue.

Infrared cameras sense light in the infrared part of the electromagnetic spectrum, meaning light with wavelengths above those visible to humans. Some infrared cameras can capture heat radiation, e.g., from humans, but the most common use of infrared cameras in tracking technologies is in a slightly higher frequency range. This is achieved by using some active infrared light source, and either capturing the light from this source directly or as reflections on the tracked objects. Typical implementations consist of a group of infrared light-emitting diodes (LEDs) positioned near the infrared camera and capturing the reflection of this light as it is reflected from small spherical markers.

Depth cameras provide a layer of depth information in addition to the regular two-dimensional image. These cameras use some technology in addition to the regular video camera. One approach is *time-of-flight* cameras, which embed an infrared emitter whose light is reflected off the objects in the field of view. The distance to each pixel is calculated on the speed of light, i.e. the infrared light returns sooner in the case of objects that are closer [Iddan and Yahav, 2001, Ringbeck, 2007]. Another approach, as used in Microsoft's Kinect sensor, is to project a fixed pattern of infrared light and analyse the deformation of this pattern as it is reflected on objects at different distances from the sensor [Freedman et al., 2010].

When not provided by the camera itself depth information can be gained through the use of *stereo cameras*. This involves two cameras mounted next to each other, providing two similar images as shown in Figure 3.7. The figure shows how depth information is found as a correlation function of sideways shifting of the images. The more shift that is required for maximum correlation, the closer to the camera are the pixels in the image. For more details on stereo vision techniques, please refer to [Siegwart and Nourbakhsh, 2004].

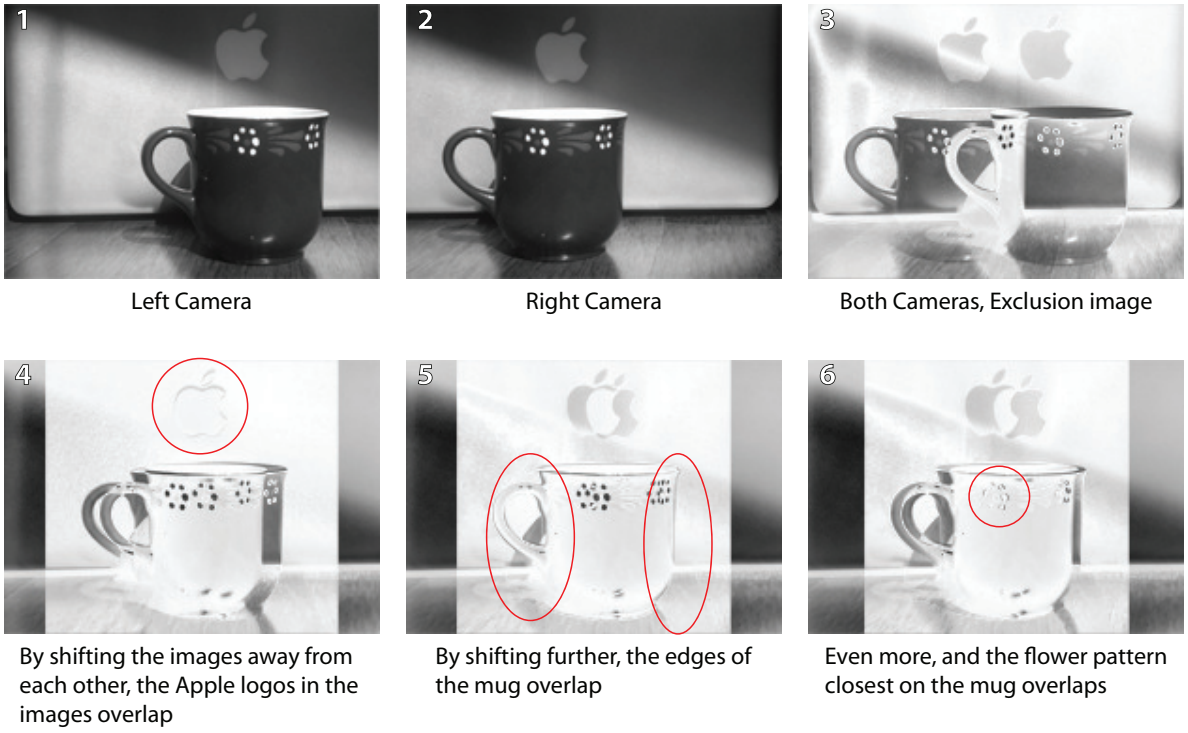


Figure 3.7: Basic illustration of depth extraction from stereo vision

Computer Vision

After obtaining the video data various processing is applied to the video stream. The video processing that is performed in optical tracking systems is primarily dependent on two factors: (1) whether or not the tracking is based on markers and (2) the camera configuration. But in any case the first processing step is to remove unwanted information from the video, i.e. separate the foreground from the background.

When depth information is available the foreground can be isolated by thresholding the depth values, or if we know the colour of the tracked objects, thresholds can be set on the colour

values for each pixel. Other techniques include *background subtraction*, i.e. using a prerecorded background image as reference and detecting any new objects in the image by subtracting the background image from the current image, and *frame difference*, meaning subtracting the previous video frame from the current video frame in order to observe changes in the video image. After the first segmentation step, filtering can be applied and a blob-size[4] threshold can be set in order to remove noise and constrain the tracking to objects of a certain size.

It is useful to distinguish between optical tracking systems that use markers and those that do not. *Markerless* tracking involves tracking whatever is present in the field of view of the camera, e.g. a human body or some object being moved around. The blobs that are detected can be measured in terms of size, centroid, principal axis etc., and these measures can again be matched to some predefined model such as that of a human body, in order to obtain more useful tracking data.

Marker-based tracking technology locates the position of usually spherical or hemispherical markers which can be placed at points of interest. For instance, a human arm can be captured by placing markers on the shoulder, elbow and wrist, or full-body motion tracking can be performed by using larger marker-setups such as Vicon's Plug-in Gait model. Types of marker include *active* light/IR-emitters and *passive* reflective markers which reflect light from an external source. In the case of passive markers the external light sources are typically infrared LEDs mounted around the camera lens.

In marker-based tracking each camera in the system produces a 2D black image with white pixels where markers are observed. This allows efficient separation of the markers from the background by thresholding the pixel values. Furthermore, the markers are treated as points, meaning that only the centroid of each blob is of interest. All in all, this makes the processing of video in marker-based systems quite efficient.

The use of a single camera can provide 2D tracking, or in the case of depth-cameras pseudo-3D tracking — meaning that objects that are hidden behind others in the camera's field of view are not tracked. By using more cameras positioned around the tracked objects full 3D tracking can be obtained. The tracking system is calibrated in order to determine the position and orientation of each camera, usually by moving a calibration wand, meaning a rigid structure with a predefined set of markers attached, around in the tracking area. From the points that are captured simultaneously in multiple cameras the position and orientation of each camera are calculated using so-called *direct linear transformation* [Robertson et al., 2004]. Figure 3.8 shows how the 3D-positions of markers that are seen by multiple cameras can be calculated.

Music-Related Applications

Several systems have been developed for conducting markerless motion capture aimed at music research and musical performance, such as EyesWeb [Camurri et al., 2000], The Musical Gestures Toolbox [Jensenius et al., 2005], and the cv.jit library for Max [Pelletier]. Max objects have also been developed to estimate periodicity in a video image [Guedes, 2006] and create a skeleton model based on video input [Baltazar et al., 2010]. For analysis of marker-based motion capture data Toiviainen's *MoCap Toolbox* is very useful [Toiviainen and Burger, 2011]

[4] A blob is a group of adjacent pixels in an image matching some criterion. In this case the pixels in the blob would match the criterion of having colour values within a certain range.

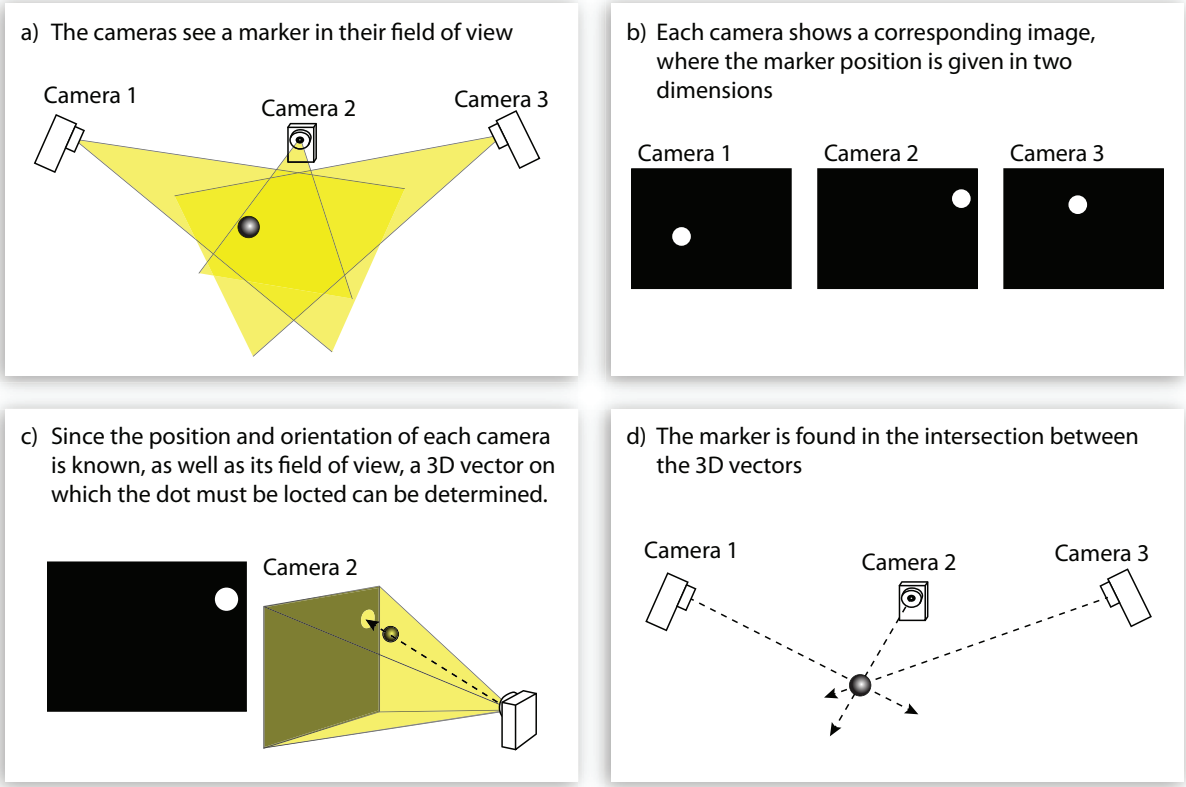


Figure 3.8: Illustration of how 3D marker positions can be calculated by an optical marker-based system.

and includes methods of feature extraction and visualisation which will be further presented in Sections 3.5 and 4.1.

Optical tracking has been popular in analysis of music-related motion. Sofia Dahl [2000, 2004] and later Bouënard et al. [2008] used marker-based motion capture of drummers to observe details of accents in percussion performance. Furthermore, Marcelo M. Wanderley and others studied how musical performance of clarinettists was perceived in different movement conditions [Wanderley, 2002, Wanderley et al., 2005, Nusseck and Wanderley, 2009]. Marker-based motion capture has also been applied in studies of string performance [Ng et al., 2007, Rasamimanana et al., 2009, Schoonderwaldt and Demoucron, 2009] and piano performance [Godøy et al., 2010, Thompson and Luck, 2012]. There are also several examples of the use of optical motion capture to analyse the motion of listeners and dancers [e.g., Camurri et al., 2000, 2003, 2004, Jensenius, 2007a, Leman and Naveda, 2010, Luck et al., 2010a, Toiviainen et al., 2010, Burger et al., 2012, Jensenius and Bjerkestrand, 2012].

The use of optical tracking in musical performance has also been explored. Various frameworks and guidelines for sonification of tracking data have been presented by Bevilacqua et al. [2002], Dobrian and Bevilacqua [2003], Wanderley and Depalle [2004], Kapur et al. [2005], Verfaille et al. [2006], Koerselman et al. [2007], Eckel and Pirro [2009], Grond et al. [2010], Skogstad et al. [2010] and Jensenius [2012c]. Furthermore, several implementations of optical tracking in sound installations or interactive music systems have been presented, e.g., by Leslie et al. [2010], Yoo et al. [2011], Bekkedal [2012], Sentürk et al. [2012] and Trail et al. [2012].

3.3 Tracking Data

Before discussing methods of working with tracking data, I shall briefly present some details of position and orientation representation.

3.3.1 Coordinate Systems

The data obtained from tracking systems constitute either a description of the tracked object in relation to some external reference point or in relation to its own previous state. In many cases the reference used is a *global coordinate system*[5] (GCS) which can sometimes be defined by the user during the calibration of the tracking system, or determined by the position of some hardware, such as a camera or an electromagnetic source [Robertson et al., 2004].

Rigid objects can be assigned a *local coordinate system* (LCS) as shown in Figure 3.9. The LCS is fixed on the object and the axes of the LCS follow the object when it is translated and rotated in space. As will be explained below the orientation of the rigid object can be measured as the orientation of the LCS in relation to the GCS. Similarly, joint angles in a kinematic model are given as the orientation of one rigid object relative to another.

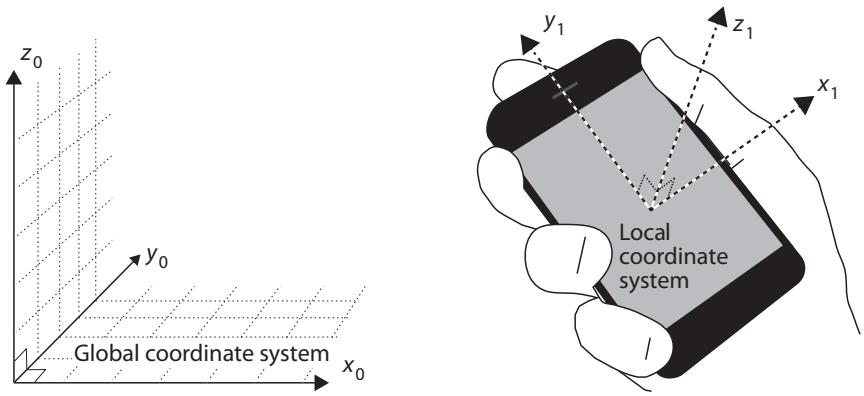


Figure 3.9: A global coordinate system (GCS) is often defined during calibration. Position and orientation measurements are given in relation to the GCS as the position and orientation of a local coordinate system (LCS) with respect to the GCS.

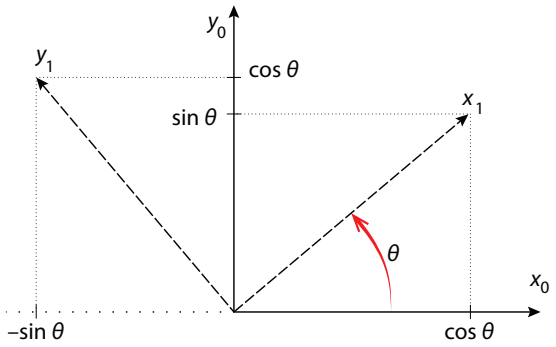
If no global coordinate system is defined, but a local coordinate system exists, the current position and orientation can be reported by reference to the previous position and orientation in a local coordinate system. In principle this also enables definition of a pseudo-global coordinate system at the start of the tracking, and estimation of trajectories in relation to this. However, as mentioned above in the section on inertial sensors, such systems are often sensitive to *drift*, which means that the error in the estimated position and orientation will increase over time.

3.3.2 Representing Orientation

We can find the position of a rigid object by the coordinates of the origin of the LCS in the GCS. Similarly, we can find the orientation of the rigid object by looking at the orientation of the axes of the LCS compared with the axes of the GCS. Figure 3.10 shows how the elements

[5] Also called a *laboratory coordinate system*, *Newtonian frame of reference*, or *absolute reference system*.

of a 2D *rotation matrix*[6] are found by projecting the axes of the LCS (x_1y_1) onto the axes of the GCS (x_0y_0): When the orientation is of the angle θ, the projection of the x-axis of the LCS is at point $(\cos\theta, \sin\theta)$ in the GCS, and the projection of the y-axis is at $(-\sin\theta, \cos\theta)$.



$$R_1^0 = \begin{bmatrix} x_1 \cdot x_0 & y_1 \cdot x_0 \\ x_1 \cdot y_0 & y_1 \cdot y_0 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

Figure 3.10: 2D (planar) rotation. The rotation from coordinate system 0 to coordinate system 1 (written R_1^0) is found by projecting the axes of system 1 onto system 0. The notation on the right shows how this is written as a *rotation matrix*.

In case of a 3D rotation a 3×3 rotation matrix is used. As for the 2D rotation, the rotation matrix is found by projecting the axes of the new coordinate system onto the original system. Figure 3.11 shows how the rotation matrix is found for a rotation of θ around the z_0 axis, followed by a rotation of ψ around the x_1 axis. The rotation matrix for the first rotation (R_1^0), is found by projecting the axes x_1, y_1, z_1 onto x_0, y_0, z_0:

$$R_1^0 = \begin{bmatrix} x_1 \cdot x_0 & y_1 \cdot x_0 & z_1 \cdot x_0 \\ x_1 \cdot y_0 & y_1 \cdot y_0 & z_1 \cdot x_0 \\ x_1 \cdot z_0 & y_1 \cdot z_0 & z_1 \cdot z_0 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and similarly R_2^1, describing the second rotation, is:

$$R_2^1 = \begin{bmatrix} x_2 \cdot x_1 & y_2 \cdot x_1 & z_2 \cdot x_1 \\ x_2 \cdot y_1 & y_2 \cdot y_1 & z_2 \cdot y_1 \\ x_2 \cdot z_1 & y_2 \cdot z_1 & z_2 \cdot z_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{bmatrix}$$

Finally, the rotation matrix R_2^0, denoting a rotation from the initial state to the final state can be found by multiplying the two first rotation matrices:

$$R_2^0 = R_1^0 R_2^1 = \begin{bmatrix} \cos\theta & -\sin\theta\cos\psi & \sin\theta\sin\psi \\ \sin\theta & \cos\theta\cos\psi & -\sin\theta\sin\psi \\ 0 & \sin\psi & \cos\psi \end{bmatrix}$$

Any rotation can be represented by performing three sequential rotations around one axis of the coordinate system in this manner. This is the basis for representing orientation by *Euler angles*, where three angles are used. Euler angles require a specification of axes about which the rotations revolve. For instance, *ZYZ* Euler angles (θ, ψ, ϕ) refer to a rotation of θ around the z-axis, followed by a rotation ψ around the y-axis and a rotation ϕ around the z-axis.

[6] A rotation matrix can also be referred to as Direction Cosine Matrix (DCM) or Orientation Matrix.

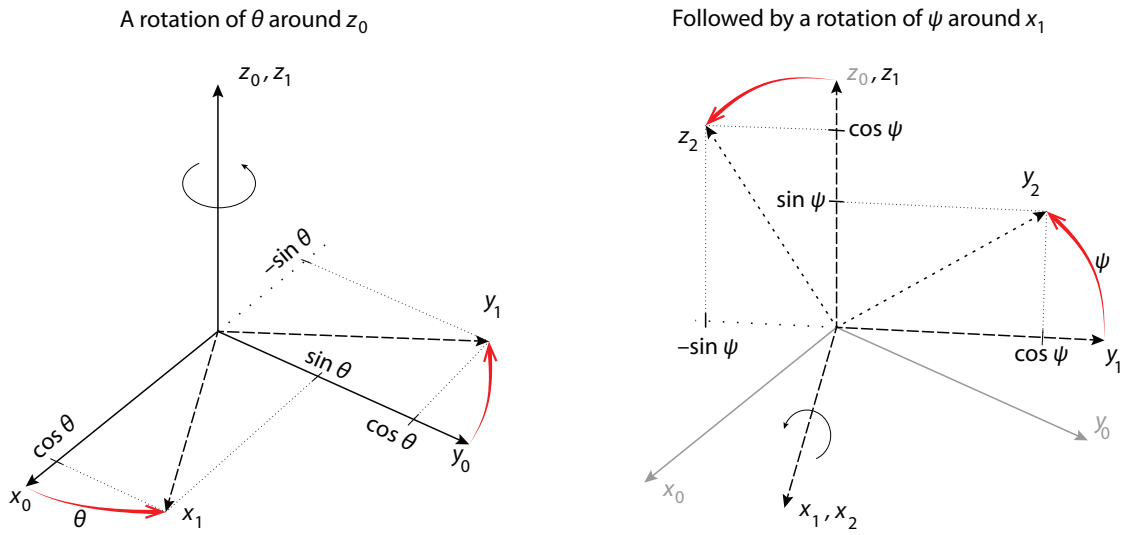


Figure 3.11: 3D rotation made up from two sequential rotations around one axis of the coordinate system. The final rotation matrix R_2^0 is found by multiplying R_1^0 and R_2^1. Any 3D rotation can be represented by three sequential rotations in this manner.

For more details of coordinate systems, representations of orientation, and working with kinematic models, please refer to [Robertson et al., 2004] and [Spong et al., 2006].

3.4 Post-Processing

3.4.1 Tracking Performance

The quality of tracking data provided by the different systems never affords a perfect representation of the real motion. As with all digital data their spatial and temporal resolutions are not infinite and depend on a number of factors related to computational power and limitations in the sensor technology. In addition to the research included in this thesis, Vigliensoni and Wanderley [2012] and Jensenius et al. [2012] have compared motion tracking systems and evaluated their use in musical interaction by measuring accuracy, precision and the temporal stability of the data rate.

The spatial resolution depends on a digitization of a continuous phenomenon. To use a familiar example, a video camera is limited by the number of subdivisions that are measured for the image, i.e. the number of pixels. Furthermore, minor errors in the calibration process can severely affect the spatial resolution [Jensenius et al., 2012]. Also, external factors such as ferromagnetic objects causing disturbance to magnetic trackers can influence the measurements.

The spatial accuracy and precision of tracking systems can be assessed by looking at *noise* and *drift*. Both can be calculated from a static measurement over a period of time. A simple linear regression can be applied to obtain an estimate of a static drift in the system. Or, if the drift is not constant, a better estimate may be obtained by filtering and downsampling the data and observing the extent of change in the data per timeframe.

The level of noise can be measured by the standard deviation (SD) of a static (i.e. without motion) measurement over a time period. If multiple dimensions are tracked, the vector norm of the SDs for each dimension is used. This value is equivalent to the root mean square (RMS) of

the distance from the mean position. One example is given in Figure 3.12 where the calculated noise level is equal to the radius of the circle. Jensenius et al. [2012] also suggested other measures for noise level, including the total spatial range covered and the cumulative distance travelled by a static marker.

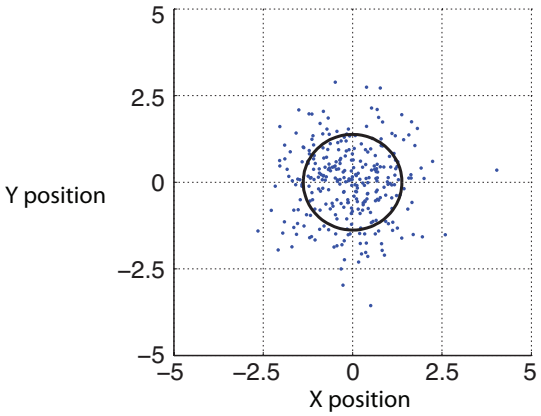


Figure 3.12: Illustration of how noise can be calculated as the standard deviation of a static position recording. The individual dots display 300 position samples (randomly generated for this example), and the circle has a radius equal to the standard deviation of the position samples.

Time is another important performance measure of tracking systems. The systems usually operate at a fixed sampling rate, ranging from a few frames per second up to several thousand frames per second for certain systems [Welch and Foxlin, 2002]. Varying amounts of processing are needed for each timeframe. This processing takes time and thus limits the sampling rate. There may also be time limitations in the sensor technology, such as a regular video camera working in low light conditions, which needs increased shutter time to capture each image.

When tracking data are to be used in real time, temporal stability is important. This is mainly evaluated by *latency* and *jitter*, which in the development of musical interfaces must be kept to a minimum to give the impression of a direct link between the motion and sound [Wessel and Wright, 2002]. The latency of an interactive system is the time delay from when a control action occurs until the system responds with some feedback, for instance the time from when a synthesiser key is pressed until sound is heard. In realtime tracking, latency will increase when processing such as filtering and feature extraction is applied. Any network connection used to stream data between devices will also induce latency. Jitter means any temporal instability in the time interval between data frames. In other words, absence of jitter would mean that the data samples are perfectly periodic.

3.4.2 Gap-Filling

Motion capture recordings may contain gaps, meaning missing frames in the data. This is mostly the case with optical systems, where a marker can be occluded by an arm or moved out of the tracking volume, but can also occur with other systems due, for instance, to packet drops when data are sent over a network.

Gaps in the data can be *gap-filled* by *interpolating* between two points, or by *extrapolating* from a single point if the missing data are at the beginning or end of the recording. Interpolation and extrapolation are achieved by calculating data values at the missing frames from a function where the measured data are used as input. Three interpolation techniques are shown in Figure 3.13. Gap-filling is useful for short gaps, but for longer gaps the trajectory within the gap

may not be possible to estimate mathematically. Such recordings must be treated as incomplete and must sometimes be removed from the dataset.

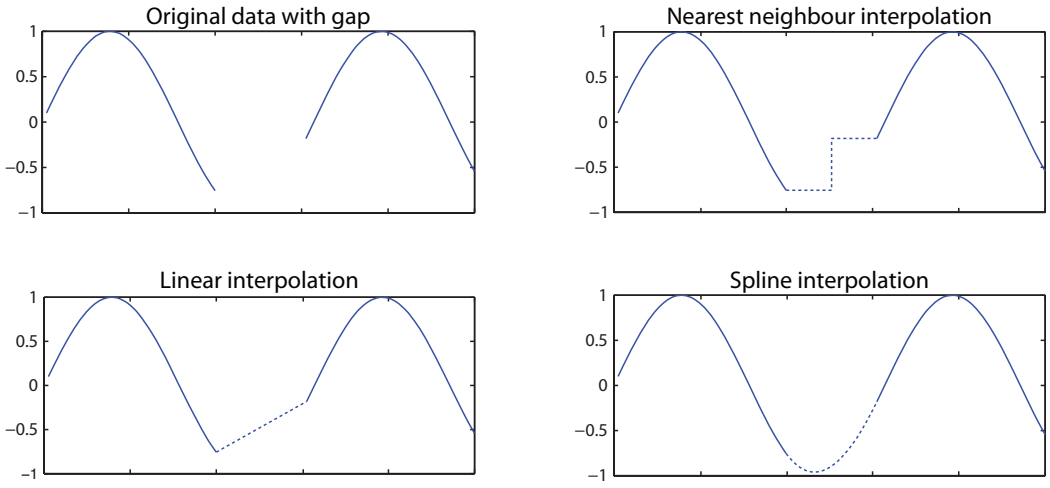


Figure 3.13: Three techniques for gap-filling: nearest neighbour, linear and spline.

3.4.3 Smoothing

Smoothing can be performed by a *moving average* or by more sophisticated *digital filters*. The moving average filter has the advantage of being easy to implement, but it may sometimes attenuate desired signal information and leave unwanted parts of the signal unchanged. The M-point moving average filter is implemented by averaging the past M samples:

$$y_i = \frac{1}{M} \sum_{k=0}^{M-1} x_{i-k}$$

where y_i is the filtered output signal at time i, x is the unfiltered input signal, and M is the number of points for which the moving average is calculated [Smith, 1997].

Better and faster smoothing can be obtained by using more advanced digital filters [Robertson et al., 2004]. Low-pass filters are used to attenuate unwanted noise in the high-frequency range of the spectrum, above the so-called *cut-off frequency*. The frequency band above the cut-off frequency is called *stopband*, and the region below this frequency is called *passband*. The cut-off is never absolute, meaning that there is a *transition band* between the stopband and passband, as shown in Figure 3.14.

Finite impulse-response (FIR) filters implement separate weights (coefficients) for each of the samples in an M-point input signal.

$$y_i = \sum_{k=0}^{M-1} a_k x_{i-k}$$

where a contains the coefficients for weighting the last M samples of x. Moving average filters are a special case of FIR filters, where all coefficients are equal to $1/M$.

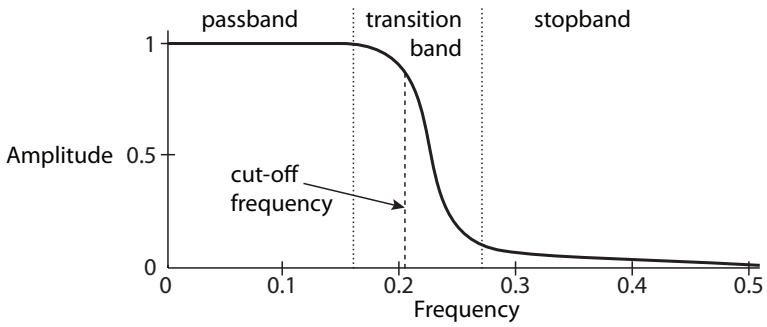


Figure 3.14: The passband, transition band, cut-off frequency and stopband of a digital low-pass filter.

In contrast to FIR filters, *infinite impulse response* (IIR) filters also include weighted versions of the filter output in the calculation. An IIR filter that considers M input samples and N output samples is given by

$$y_i = \sum_{k=0}^{M-1} a_k x_{i-k} + \sum_{k=1}^{N} b_k y_{i-k}$$

where b contains the coefficients for the last N samples of y [Smith, 1997]. IIR filters generally produce narrower transition bands but induce phase distortion, meaning that different parts of the frequency spectrum pass through the filter at different rates. Several standardised filter designs exist, and Matlab-functions for determining the filter coefficients of these are available.[7]

3.5 Feature Extraction

As presented above, there are considerable differences between tracking technologies. Nevertheless, many of the same techniques can be applied to data from different systems. As with the sound features described in Section 2.1, motion features are calculated to obtain more useful information from the raw motion data provided by the tracking system.

The use scenario of the motion data determines the preprocessing and feature extraction that can be applied to motion data. Specifically, when motion tracking is applied to interactive systems where the motion data are used in real time, it is usually important to keep the latency as low as possible. Some processing techniques require a buffering of the signal which induces latency, so trade-offs must often be made between advanced feature extraction algorithms and the amount of latency.

3.5.1 Differentiation

By using basic calculus techniques *velocity* and *acceleration* can be determined from a stream of position data. These are examples of the most basic feature extraction methods for motion data. The simplest way of estimating velocity from position data is to calculate the difference between the current and previous positions (known as the *first finite difference*), multiplied by the sampling rate:

$$v_i = \frac{s_i - s_{i-1}}{\Delta t}$$

[7] e.g. the Matlab functions fir1, fir2, butter, cheby1, cheby2, and ellip

where v_i is the velocity at time i and s is the position in metres. Δt is the time between successive samples (in seconds), and is found by $1/f$ where f is the sampling rate in Hz [Robertson et al., 2004]. More accurate calculations can be obtained by the *central difference* method; however, this induces one more sample delay, which could be undesirable in realtime applications:

$$v_i = \frac{s_{i+1} - s_{i-1}}{2\Delta t}$$

A similar calculation can be made to estimate acceleration from velocity data, and jerk from acceleration data. Such differentiations amplify noise that is present in the signal and therefore data smoothing should be applied before the derivatives are calculated.

3.5.2 Transformations

A stream of position data or its derivatives can be transformed in various ways. By projecting data onto new coordinate systems we can obtain information on relations between tracked objects. The position of a person's hand can, for instance, be projected onto a local coordinate system with the centre in the person's pelvis. This would provide information of the position of the hand relative to the body, independently of whether the person is standing up or lying down.

The dimensionality of the data can, furthermore, be reduced, for instance by calculating the magnitude of a multidimensional vector. The *absolute velocity* of a three-dimensional velocity stream, for instance, is given by the magnitude of the X, Y and Z components of the velocity vector. This value is useful in describing the speed of an object, without paying attention to direction of the velocity vector.

3.5.3 Motion Features

Using basic differentiation and transformation techniques on a raw motion signal is a simple way of calculating salient motion features. This is particularly useful in realtime applications, where low latency is important. Without the need to consider the motion data as representations of human body motion, we can calculate features such as *quantity of motion* by summing the absolute velocities of all the markers, or *contraction index* by calculating the volume spanned by the markers.

A different type of feature can be found by taking into account the labels of the data in the motion capture signal. If two markers represent the two hands of a person, the feature *hand distance* can easily be calculated. Similarly, three markers representing the wrist, elbow and shoulder can be used to calculate the *arm extension*. More sophisticated motion features can be found by taking into account models of the mass of various limbs. One such is the 'Dempster model' [Robertson et al., 2004] which allows calculation of the kinetic or potential energy of the body or a single limb, or estimation of the power in a joint at a certain time.

The features may be purely *spatial*, meaning that they describe positional data without considering how the motion unfolds over time. Examples of this are contraction index and potential energy. Other features are *spatiotemporal*, meaning that they describe how the motion unfolds in space over time. Difference calculations such as the derivative of hand distance are typical examples of this. Finally, a feature such as periodicity is a *temporal* feature, where the spatial

aspect is not described.

Meinard Müller [2007] has proposed a robust set of 7 generic kinematic features for human full body motion. The features are based on relations between joints, which make them work independently of the global position and orientation of the body. Furthermore, the features are boolean[8] which greatly reduces the amount of processing needed to use the features e.g. for search and retrieval in motion capture databases. I present his set of generic features below, with illustrations in Figure 3.15.

F_{plane} defines a plane by the position of three joints and determines whether a fourth joint is in front of or behind this plane. This may be used, for instance, to identify the position of the right ankle in relation to a plane spanned by the centre of the hip, the left hip joint and the left ankle. If a value 1 is assigned when the foot is in front of the plane, and 0 when it is behind, a normal walking sequence would show an alternating 0/1 pattern.

F_{nplane} specifies a vector by the position of two joints and a position along the vector where a plane normal to the vector is defined. For instance, a plane that is perpendicular to the vector between the hip and the neck, located at the head, can be used to determine whether the right hand is raised above the head.

F_{angle} specifies two vectors given by four joints and tests whether the angle between them is within a given range. For instance, the vector between the right ankle and right knee and the vector between the right knee and the right hip could be used to determine the extension of the right knee joint.

F_{fast} specifies a single joint and assumes a value of 1 if the velocity of the joint is above a chosen threshold.

F_{move} defines a vector between two joints and assumes a value of 1 if the velocity component of a third joint is positive in the direction of the defined vector.

F_{nmove} defines a plane between three joints and assumes a value of 1 if the velocity component of a fourth joint is positive in the direction of the vector normal to the plane.

F_{touch} measures the distance between two joints or body segments and assumes a value of 1 if the distance is below a certain threshold.

From the 7 generic features Müller has defined 39 features which contain specific information about the joints and thresholds used. In Müller's research these are used to recognise various full-body actions such as performing a 'cartwheel' or a 'squat'. The 39 boolean features make up a *feature matrix* which describes a single recording. A computer system is used to define so-called *motion templates*, which are real-valued prototypes of the feature matrices that correspond to a certain action. The motion templates are learned by the system by inputting a number of labelled data examples. Motion templates can be used to identify new versions of the same action by using *dynamic time warping* and a distance function which matches the input data to the learned motion templates. Müller also provides a way of visualising the motion templates, which is shown in the next chapter.

[8] Boolean means that the possible values are either 0 or 1.

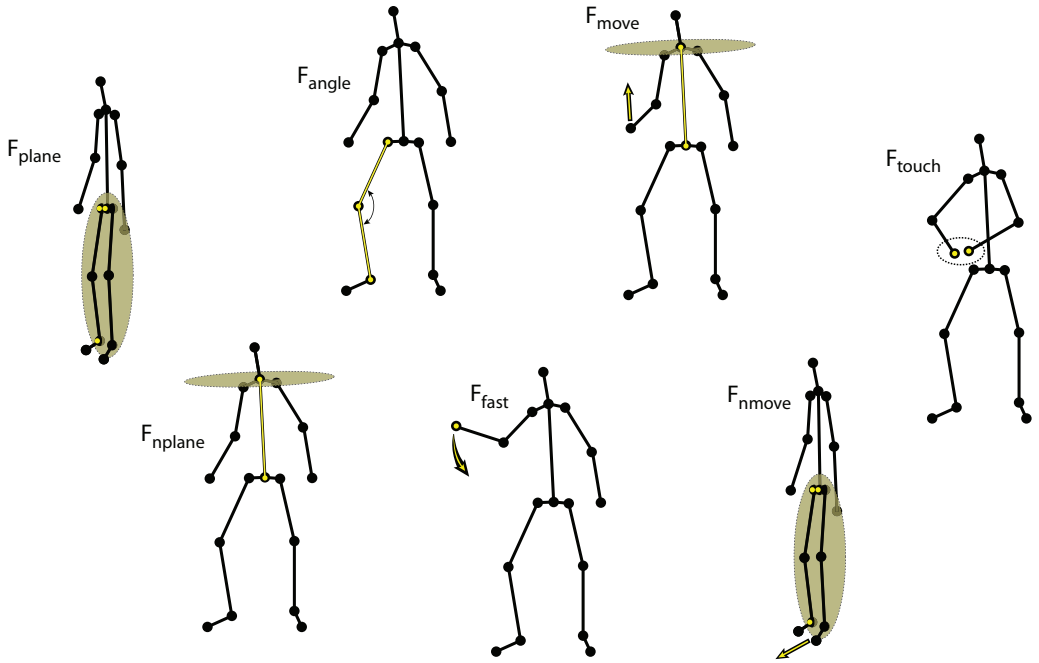


Figure 3.15: Illustrations of Müller's generic kinematic features. The yellow marks denote the joints or vectors that are used in the illustrated implementation of the feature. Refer to the main text for explanation.

3.5.4 Toolboxes

The *MoCap Toolbox* for Matlab, developed at the University of Jyväskylä, includes a variety of feature extraction algorithms for motion capture data [Toiviainen and Burger, 2011]. This includes functions for calculating derivatives, filtering, cumulative distance and periodicity, and models that take the weight of body segments into account, enabling the calculation of potential and kinetic energy. Furthermore, the toolbox has implemented algorithms for calculating the *eigenmovements* of a full body motion capture segment by using *principal component analysis* (PCA) [Duda et al., 2000]. PCA is a method of data reduction applied by projecting the original data onto a set of *principal components*. The first principal component is defined as the vector on which the data in a data set can be projected to explain as much of the variance in the data set as possible. The second principal component is perpendicular to the first and explains as much of the remaining variance as possible. Toiviainen et al. [2010] showed the utility of PCA in motion analysis for a set of motion capture recordings with 3D positions of 20 joints, equivalent to 60 data series. By keeping only the 5 highest ranked principal components, 96.7 % of the variance in the data was explained. The analysis allowed the researchers to distinguish between periodicities in various parts of the body of the subject, and to observe relations between the motion in the different segments.

Other tools have been developed for extracting features from video data and can in principle be used with sensors as simple as an ordinary web camera. Antonio Camurri's EyesWeb software is designed for extracting features from motion data in real time [Camurri et al., 2004]. The software can extract a body silhouette from a video signal, and a number of features can be calculated, most notably the *quantity of motion* and *contraction index*. These have been shown to be pertinent to the experience of emotion in dance [Camurri et al., 2003].

Quantity of Motion is calculated as the number of moving pixels in the silhouette and reflects the overall motion in the image.

Contraction Index denotes the extension of the body and can be estimated by defining a rectangular bounding region around the silhouette (area of motion) and comparing the total number of pixels within this area with the number of pixels covered by the body silhouette.

The *Musical Gestures Toolbox*, developed by Alexander Refsum Jensenius, includes some of the features that are implemented in the EyesWeb software [Jensenius, 2007a]. This software is implemented in Max as modules in the Jamoma framework [Place and Lossius, 2006], and unlike EyesWeb it is open source. The toolbox includes modules for preprocessing video, calculating features such as the quantity of motion, area of motion, the barycentre of the motion in the image, and also smoothing and scaling of the data. The toolbox also contains numerous modules for visualising motion, which will be covered in Section 4.1.

3.6 Storing and Streaming Music-Related Data

We can distinguish between two main use scenarios for tracking data. Firstly, as explained in Section 3.1, *motion capture* involves storing the tracking data in order later to apply analysis or import the data in animation software. Secondly, *realtime tracking* involves using the tracking data directly, within a very short time period after the motion occurs. Realtime tracking is used, for instance, in interactive systems such as motion-based computer games like Microsoft Kinect. When a user performs an action it is reflected in the movement of an avatar some milliseconds later, after the necessary processing has been completed.

In music-related contexts tracking data are often just one part of the total amount of data involved. In addition to motion data, music-related data include video, audio and symbolic representations of musical sound such as MIDI-data or sensor data from electronic musical instruments. Furthermore, music researchers and performers use features that are extracted from the tracking data. These may be simple, time-varying transformations, such as relations between body limbs, or distinct events such as sound-producing actions or musical phrases and also higher-level features such as descriptions of the emotive content of the music. The diversity of these data is challenging: sampling rates range typically from 44.1 kHz for audio and down to less than one event per second for event-based data such as MIDI, and dimensionality varies from a single number per sample for audio data to more than one million pixel values for one frame of video data. An overview with some typical examples of music-related data, adopted from [Jensenius et al., 2008], is presented in Table 3.1. Thus for storing data we need a format that can to handle the different data types, and for streaming we need a protocol that enables simple routing of the different types of data in realtime applications.

3.6.1 The Gesture Description Interchange Format

Most commercial mocap systems provide proprietary file formats for storing tracking data, with the option to export the data to a more open format. These solutions are sufficient in most

Table 3.1: The data types used in the experiment presented by Jensenius et al. [2008]. The different numbers of sensors, sampling rates, bit resolutions and channels per device are challenging to handle with standard protocols for storing and streaming tracking data.

Input	Sampling rate	Sensors	Channels	Bit resolution
Accelerometer	60 Hz	9	3 DOF	32
Polhemus tracking	60 Hz	2	6 DOF	32
Bioflex EMG	100 Hz	2	1 DOF	7
High-speed video	86 Hz	1	320 × 240	8
Audio	44100 Hz	1	2 (Stereo)	16
MIDI	Event-based	1	3	7

motion capture settings. However, in research on music and motion the standard formats often fall short since they are not able to handle the wide variety of data at hand [Jensenius, 2007a].

Jensenius et al. [2006b] proposed the *Gesture Description Interchange Format* (GDIF) as a multi-layered approach to structuring music-related data. The various layers in GDIF contain different representations of the data, with the most basic *acquisition layers* containing raw sensor data, and sensor data where some simple processing (e.g. filtering) has been applied. Next, the *descriptive layers* describe the motion in relation to the body, in relation to a musical instrument or in relation to the environment. Then the *functional* and *meta layers* contain descriptions of the functions of the various actions in a recording (sound-producing, communicative, etc.), and abstract representations, higher-level features and metaphors.

GDIF was mainly proposed as a concept and idea for structuring music-related data, and not as a file format *per se*. In a panel session at the International Computer Music Conference in 2007 the *Sound Description Interchange Format* (SDIF) was suggested as a possible format for the implementation of GDIF [Jensenius et al., 2007]. As shown in Figure 3.16, SDIF tackles the challenge of synchronising data with different sampling rates by organising the data into time-tagged frames in individual streams [Wright et al., 1998]. SDIF also allows data with different dimensionality in the individual streams. The use of SDIF as a storage format for music-related data has been explored by several researchers [e.g., Jensenius et al., 2008, Peters et al., 2009, Bresson and Schumacher, 2011] and is currently the most used format in GDIF development.

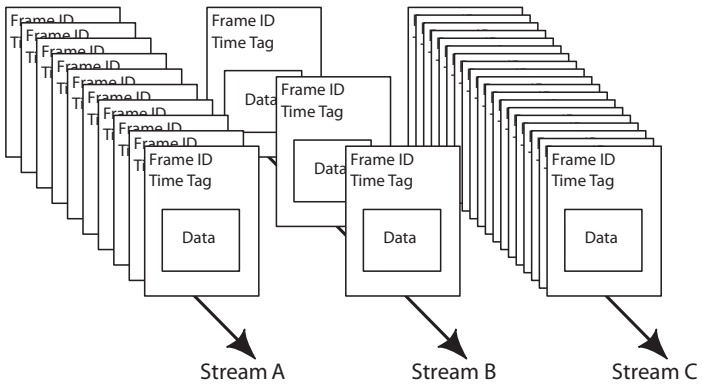


Figure 3.16: The Sound Description Interchange Format arranges data into individual streams containing time-tagged frames.

More recently researchers in the SIEMPRE EU FP7 ICT project have developed a system that allows synchronised recordings of data from several devices using SMPTE time-coding [Gillian et al., 2011]. An XML-based file format and synchronisation protocol has been developed for storing synchronised recordings of audio, video and text-based sensor and mocap data.

The system also includes a solution for uploading recordings to a server, and visualisation tools for video, motion capture, sensor data and audio using EyesWeb. A similar database solution for classifying and performing search and retrieval of music-related actions has been proposed by Godøy et al. [2012], and is currently under development at the University of Oslo.

3.6.2 Open Sound Control

One application of realtime tracking in music is in interactive systems such as digital musical instruments or other interactive sound installations. This may entail streaming the data from a tracking system and *mapping* features extracted from the data to a synthesiser. Adopting terms from Miranda and Wanderley [2006], the motion data and extracted features are referred to as *gestural variables* and the parameters available for controlling the sound output of the synthesiser are called *synthesis parameters*. With the large amount of data that is communicated, and also with different representations of the data, it is important to have a structure for communicating between gestural variables and synthesis parameters.

The Open Sound Control (OSC) protocol, introduced by Wright and Freed [1997], has become the leading protocol for communicating music-related data in research on novel musical instruments. A main idea in OSC is to structure music-related data hierarchically, for instance to facilitate mapping between gesture variables and synthesis parameters in digital musical instruments. The hierarchical structure is reflected in the so-called *OSC-address* which is sent together with the data. Each level is separated in the OSC-address by a slash " / ". One example could be the following OSC-namespace for synthesis parameters in a musical instrument:

- `/synthesiser/1/oscillator/1/frequency`
- `/synthesiser/1/oscillator/1/amplitude`
- `/synthesiser/1/oscillator/2/frequency`
- `/synthesiser/1/oscillator/2/amplitude`

Here, '`/synthesiser`' is at the top level, and the '`/1`' indicates that we are referring to the first of possibly several synthesisers. The '`/frequency`' and '`/amplitude`' of two oscillators can be controlled. Thus to set the frequency of the first oscillator to 220 Hz, we would use the control message '`/synthesiser/1/oscillator/1/frequency 220`'.

Synthesis parameters are only one aspect of OSC messages. OSC is also a good way of structuring gesture variables. The Qualisys motion tracking system[9] has native support for OSC, and researchers have developed applications for interfacing with several other tracking systems via OSC, e.g. Vicon,[10] Nintendo Wii,[11] and Xsens MVN [Skogstad et al., 2011]. For full body motion capture data examples of OSC addresses might include:

- `/hand/left/velocity`
- `/head/position`

[9] http://www.qualisys.com
[10] http://sonenvir.at/downloads/qvicon2osc/
[11] http://www.osculator.net/

Various tools have been developed for using OSC-formatted data in the development of musical instruments, for instance the Open Sound Control objects for Max provided by CNMAT.[12] The *Digital Orchestra Toolbox*, developed by Joseph Malloch et al. [2007], also includes a mapping tool that simplifies mapping between OSC-formatted gesture variables and synthesis parameters. Malloch's mapping tool was later included in Jamoma which also includes several other tools for mapping between control data and sound [Place et al., 2008].

3.7 Summary

This chapter has introduced a variety of motion tracking technologies with a main focus on optical infrared marker-based motion tracking. Some general concepts in motion tracking have been introduced. Tracked objects include markers, rigid objects or kinematic models, and the type of object defines the type of tracking data provided. Positions and orientations can be described in relation to a global or local coordinate system defined by the tracked object itself or by another object.

The chapter also introduced basic processing techniques for motion data, including gap-filling and smoothing. Some feature extraction techniques were introduced, with basic differentiation and transformation, and Müller's motion features as examples of how boolean features can be extracted from relation of body limbs. Further, some features available in toolboxes for working with music-related motion were introduced. Finally, I presented some of the challenges of storing and synchronising music-related data, and basic theory on how motion tracking can be used in real time for musical applications.

[12] http://cnmat.berkeley.edu/downloads

Chapter 4

Methods of Analysis

Chapter 3 having explained how motion can be captured by various tracking technologies, this chapter will introduce the methods that have been applied in the thesis to analyse correspondences between sound and body motion. Several of the methods presented here are well-known, and more comprehensive details of these methods can be found in most textbooks on statistics. In my own analysis I have used existing software to run statistical tests and for classification, and therefore only a basic introduction to the methods is offered here, as a background to the analysis results and assessments that are made in the papers included in this thesis.

Stanley S. Stevens [1966] introduced the term *cross-modality matching*, denoting the process of matching some sensory input in two modalities. Steven's use of the technique involved an experiment in which participants were asked to adjust the sound level of a tone to match the strength of a vibration applied to their finger, and the other way around — adjusting the strength of the vibration according to the apparent loudness of the tone. The analyses presented in several of the papers included in this thesis are based on a variant of the cross-modality matching approach, in studies referred to as *sound-tracing*. Experiment participants were asked to match their body motion to some auditory input (i.e. to 'trace the sound'). Analysis of the data involves comparing features of the sound objects used as stimuli with features of the recorded motion.

Most of the sound stimuli used in the experiments have durations of less than 5 seconds and each constitutes a single sound object. The relations between sound and motion are analysed on a chunk timescale level and a sub-chunk timescale level (ref. the discussion in Section 2.3.2), but not as multiple concatenated chunks. Analysis at the sub-chunk timescale level is concerned with comparing features that contain numerical values in each timeframe. Borrowing terminology from Peeters et al. [2011], I refer to them as *time-varying features*. Other features describe an entire object; for instance, the mean acceleration of an action or the categorical labelling of a sound object as 'pitched'. These features consist of a single value or a single description for an entire object and are referred to as *global features*. Figure 4.1 displays various examples of the two main feature types involved.

4.1 Visualisation of Motion Data

A requirement of analysis of music-related data is to have good visualisation techniques. In addition to providing qualitative assessments from tracking data, good visualisation techniques

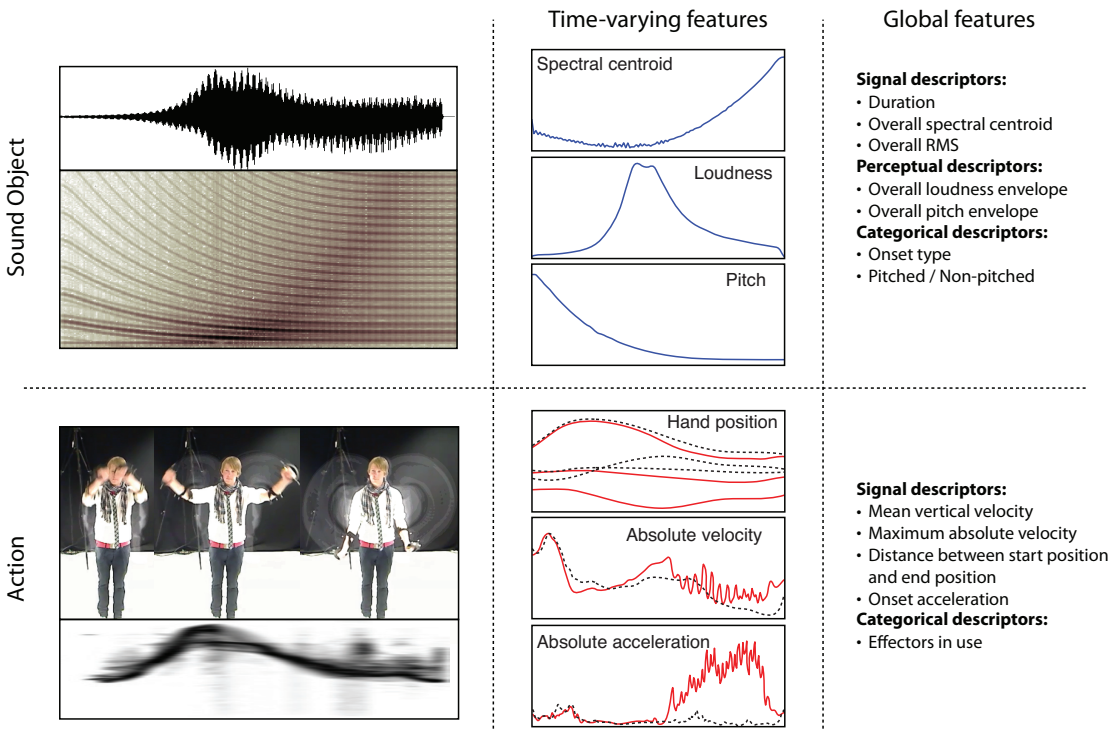


Figure 4.1: A sound object with a corresponding action (sound-tracing) and feature examples. *Time-varying features* contain separate values for each frame, and *global features* are either overall numerical calculations based on the time-varying features or non-numerical classifications of the objects.

facilitate conveying analysis results to other researchers, project funders and the general public. Further, visualisations are an essential aid in developing hypotheses that can be tested quantitatively [Moore and McCabe, 2006]. Displaying motion data over time is not trivial, particularly because of the large number of dimensions that a motion capture recording typically contains. In some cases a simple plot of absolute velocity over time is sufficient, but if 3D marker positions, velocities and accelerations are to be displayed for multiple markers, a timeline plot soon becomes unreadable. This section will cover the background of the visualisations I have used in my own work, including two techniques that I have developed.

4.1.1 The Challenge of Motion Data Visualisation

Motion data span both *time* and *space*, and it is important to have visualisation techniques that cover both of these domains. Time is one-dimensional, and spatial position three-dimensional, and in the end we want techniques that display all these dimensions on a two-dimensional medium, namely paper.

A straight forward and quite common way of plotting motion data is with time on the horizontal axis and position the vertical axis. In Figure 4.2 this is shown for a single marker on the right wrist of a pianist. The plot provides precise temporal information and when zooming in it is also easy to read the precise position of the wrist at a certain time.

Although Figure 4.2 gives precise information about the motion of the hand marker, dividing the original single trajectory into three lines seem to run counter to intuition. Furthermore, motion data usually consists of more than a single marker, and attempting to plot all the markers in a mocap recording on a timeline is in most cases too cumbersome. Figure 4.3 shows this, by

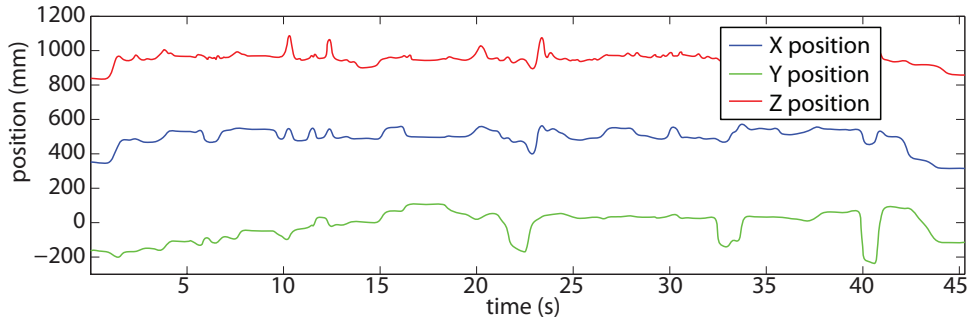


Figure 4.2: A common way of plotting three-dimensional marker data in time and space.

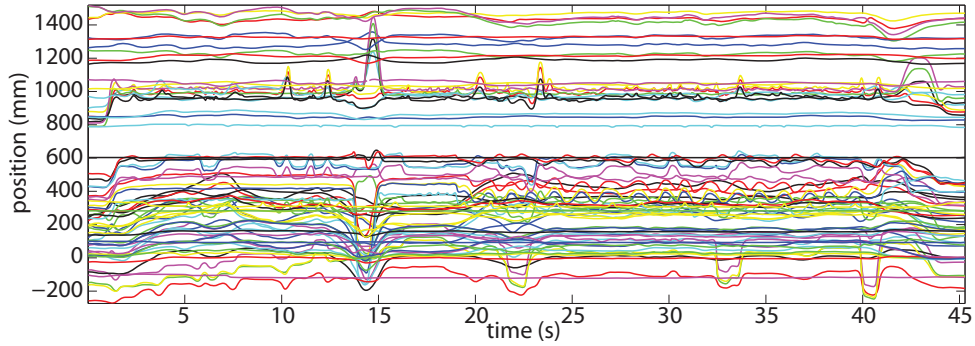


Figure 4.3: Plots of X, Y and Z positions of 24 markers from motion capture of a short piano performance. Although the plot provides some impression of salient moments (e.g. between 14 and 15 seconds), it is too complex to provide any detailed information.

plotting the X, Y and Z positions of all the 24 markers of the same piano performance.

Several visualisation techniques are able to present marker data in a more intuitive manner than Figure 4.2, and there are also techniques for displaying full-body motion without the need of plots as in Figure 4.3. There is often a trade-off between intuition and precise representations of time and space in these techniques. It takes time to become familiar with some of the methods, while others can be understood without any explanation.

Returning to the terms introduced in Section 2.3.2, we can relate the visualisation techniques to three timescale levels: *sub-chunk*, *chunk* and *supra-chunk*. Visualisations at the sub-chunk level display motion in an instant, or over a very short period of time. Such visualisations typically show a static pose and therefore the spatial aspect is important. At the supra-chunk level visualisations of long time periods may often be at the expense of spatial information. In some visualisations at the chunk level the time-span is reduced enough to be able to combine good representations of both time and space.

The relation between visualisations and the three timescale levels is particularly evident in the visualisation techniques implemented in the Musical Gestures Toolbox [Jensenius, 2007a] which was introduced in Section 3.5.3. I shall illustrate these techniques before continuing with visualisation techniques for three-dimensional motion data.

4.1.2 Motion in Video Files

Jensenius' tools for analysing motion in video contain several techniques for visualising motion [Jensenius, 2012a]. The toolbox is based on differentiating and filtering video frames, and

algorithms for visualising the video as it unfolds over time. Three of the methods are listed below and are illustrated in Figure 4.4.

Motion images display the changes in the current from the previous video frame. Various filtering and thresholding techniques can be applied to remove unwanted noise from the motion image.

Motion history images display a combination of several motion images extracted from a sequence of video frames, for instance by averaging the pixel value across all of the motion images. Jensenius implemented various ways of calculating motion history images, which all show different qualities of the analysed video.

Motiongrams are displayed by collapsing each motion image frame down to one-dimensional images, either horizontal or vertical. The collapsing is done by averaging the pixel values across one of the dimensions. The one-dimensional image that is produced is plotted on a timeline and provides a visual impression of the evolution of motion in the video.

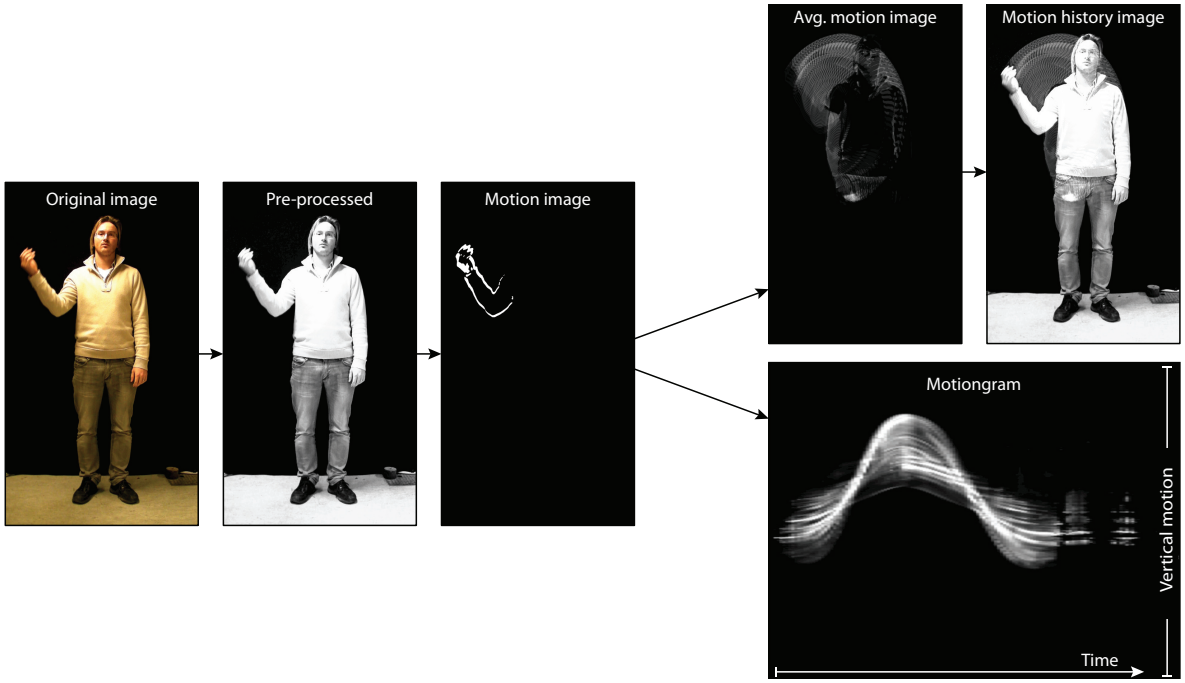
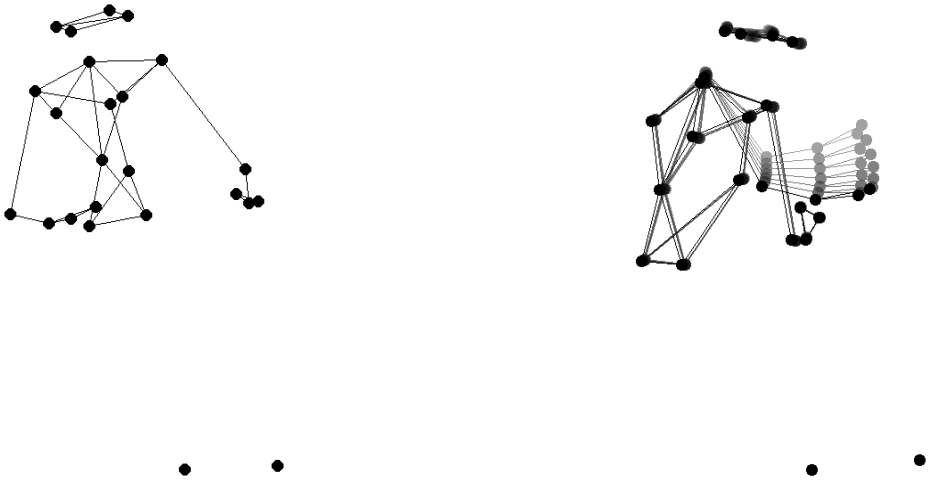


Figure 4.4: Jensenius' techniques for visualising motion. (Adapted from [Jensenius, 2012b])

In the images shown in Figure 4.4 the motion image shows which part of the body is moving in this instant. The image shows precisely which pixels are different between two successive frames of video data, and is a sub-chunk visualisation of motion. The motion history image shows a slightly longer timespan, providing a quite intuitive description of the spatial trajectory of the hand. However, the image does not show precisely which pixels have changed in each timeframe. Finally, a motiongram can be made for longer segments of movement, and motion can be displayed with as high temporal precision as the framerate of video file. However, the spatial information has been reduced, since the motiongram can only display one dimension at a time. Furthermore, the motiongram is less intuitive than the motion history image, because most people are not used to looking at one-dimensional images unfolding over time.



(a) (b)

Figure 4.5: The figure shows the 24 markers in the piano recording plotted in Figure 4.2 and Figure 4.3, displaying the head, torso and arms with interconnected lines as well as the feet. (a) illustrates a pose without time-information and can be seen as a cross-section of Figure 4.3 at time = 4 seconds. (b) shows how multiple sequential poses can be superimposed to display trajectories over a time-period.

4.1.3 3D Motion Data

In the case of 3D motion capture data the various suppliers of motion tracking equipment provide proprietary environments for visualising their data.[1] This may involve a 3D view of markers with interconnected lines. It is also normal to be able to show marker trajectories for the past and coming frames in the 3D view. Furthermore, the programs typically contain timeline views of the individual markers with position, velocity and acceleration. These visualisations are useful in getting an initial overview of the motion data; however, the solutions are inadequate if we want to apply various processing techniques to the data that are not implemented in the proprietary motion capture software.

Toiviainen's MoCap Toolbox provides a variety of scripts for plotting motion data [Toiviainen and Burger, 2011]. Individual marker positions and processed data can be plotted on timelines, and marker positions in any timeframe can be plotted in *point-light displays*, as shown in Figure 4.5(a). Such point-light displays have been shown to retain salient perceptual information about the motion, allowing people to recognise the gender of a person, or the affect of bodily gestures [Kozlowski and Cutting, 1977, Pollick et al., 2001]. The toolbox also includes a feature for collecting a sequence of such poses in a video file. By using image processing software the point-light displays can be put together into an intuitive visualisation of motion trajectories at the chunk-level. Figure 4.5(b) shows an example of this where multiple sequential poses have been superimposed.

The supra-chunk level can be illustrated through basic Matlab functions by plotting the position in each timeframe in a scatterplot. However, the plot quickly becomes too complex when more than a single marker is included. Figure 4.6 shows how the position of the same

[1] e.g. Naturalpoint Arena for OptiTrack, Xsens MVN Studio, and Qualisys Track Manager

marker as in Figure 4.2 can be plotted in a more intuitive manner than with the time-series plot. Again, there is a trade-off between precise data and intuition — position and temporal information are present in the plot but cannot be read as precisely as in the time-series plot of Figure 4.2. Supra-chunk trajectory plots are useful for observing how a single marker moves over a longer time period and I have made these for one of our lab publications so far [Jensenius et al., 2012].

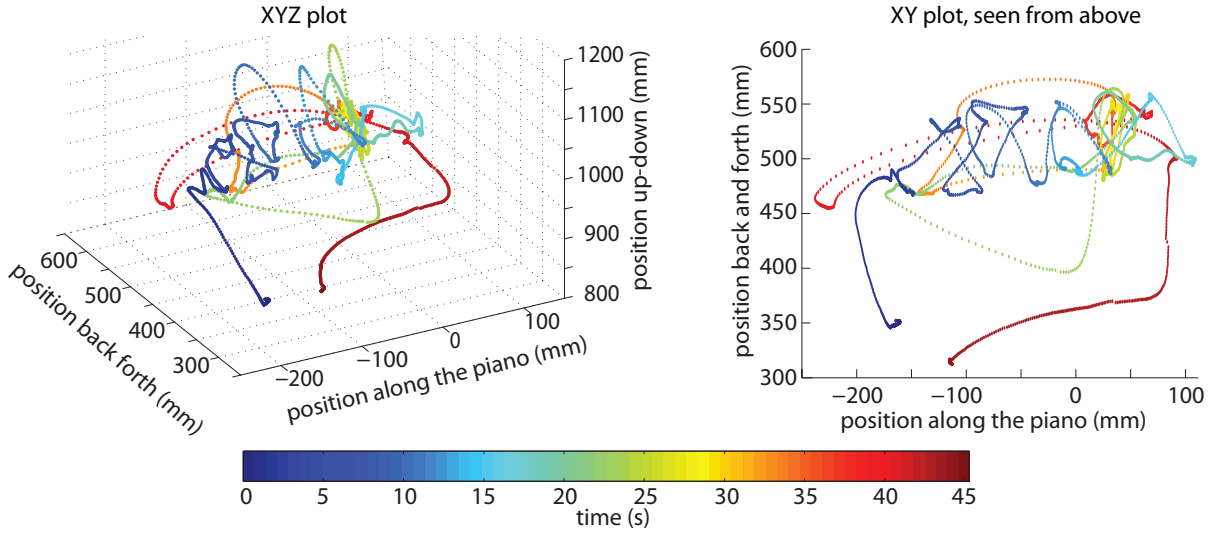


Figure 4.6: The trajectory of a single marker can be shown in 2D or 3D plots. Time can be shown by colour-coding the trajectory. The marker shown in the plots is the same right wrist marker as in Figure 4.2.

4.1.4 High-Dimensional Feature Vectors and Multiple Data Series

When it is desirable to visualise an entire full-body mocap recording or a set of time-varying features describing the data, colour information can be used to indicate the position of each marker, or the magnitude of the features.

Meinard Müller [2007] used colour information to visualise 39 features in his *motion templates*. In this technique each feature is assigned a separate row in a matrix and the time-frames are shown in the columns. This allows studying a high number of dimensions on a timeline, and provides an overview of patterns in the mocap data. An example is shown in Figure 4.7.

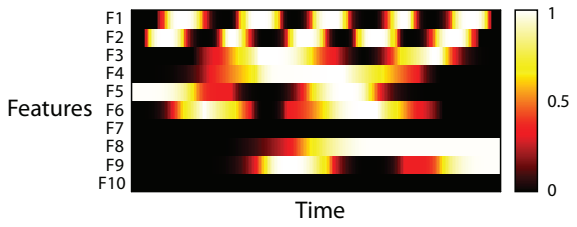


Figure 4.7: Example of Müller's visualisation technique for motion templates, showing the ten first features (not based on actual data). The top two rows show values that alternate between 0 and 1, something that could represent some feature of the left and right foot respectively in a walking pattern.

A similar technique can also be used to show positions of a larger number of markers by assigning the markers to individual rows and projecting the spatial coordinates onto a colourspace [Jensenius et al., 2009]. Figure 4.8 shows a so-called *mocapgram* of the 24 markers in the same piano performance as used in the plots above. Marker names following Vicon's plugin gait[2] convention are shown on the left. The XYZ coordinates have been projected onto red, green and blue, respectively and the values in each row are normalised. Although we can not tell the precise position of the markers from the plot, certain clear patterns can be seen — for instance the large trajectories in the right arm (RELB,RWEI,RHAO,RHAI) at 22, 33 and 40 seconds. Note also the almost binary pattern in the right toe (RTOE) when the sustain pedal is pressed.

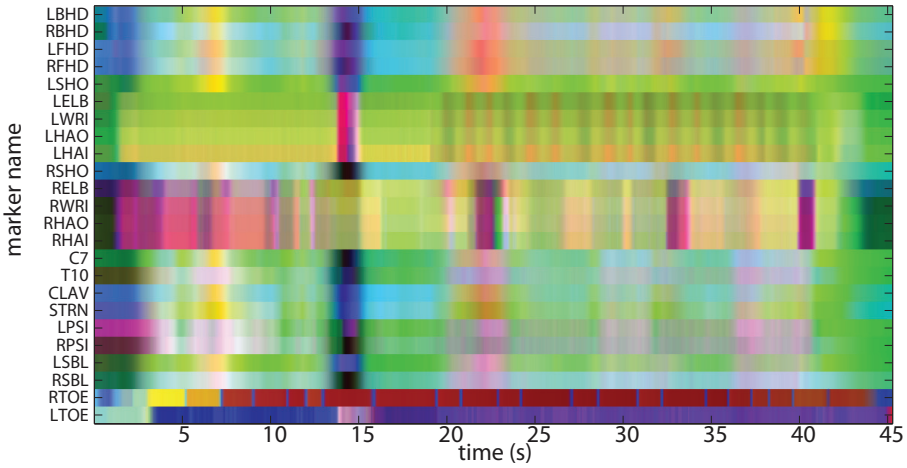


Figure 4.8: Mocapgram showing 3D position coordinates mapped onto a colourspace.

In my own research I needed to display the results of a large number of motion capture sequences in order to show general tendencies in the data. I developed mocapgrams further, in a script in Matlab for visualising data [Kozak et al., 2012, Nymoen et al., 2012]. Figure 4.9 is adopted from Paper VII, and shows how multiple motion capture recordings can be compared (in this case only 5 recordings). The use of these plots in the paper involved comparing motion capture data with a sound stimulus. The stimulus started 0.5 seconds after the start of the motion capture recording and ended 0.5 seconds before the recording ended. As shown in the figure the value of each data series is given as a shade of grey, here normalised between 0 and 1. The mean value of the 5 data series at each time-frame is shown as a dashed line, with two dotted lines showing the standard deviation. The units for the mean value plot are on the left axis.

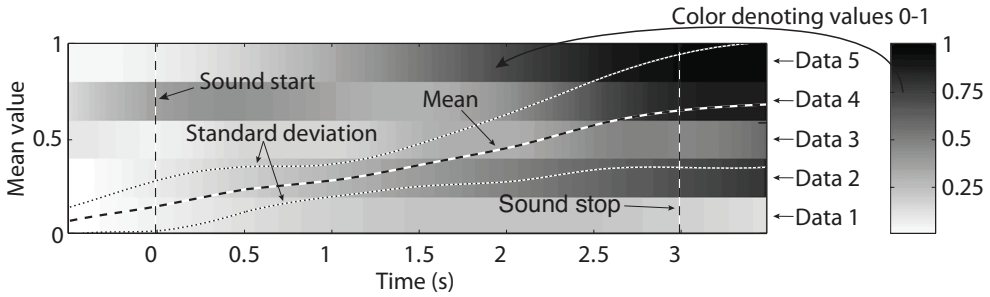


Figure 4.9: Mocapgram example, adopted from [Nymoen et al., 2012].

[2] http://fourms.wiki.ifi.uio.no/MoCap_marker_names

The mocapgrams do not give precise information on the value in each time-series since the different shades of grey may be difficult to distinguish. However, the temporal information is as precise as in any time-series plot, and the plots facilitate illustration of the distribution of a large number of time-series. Figure 4.10 shows an example of how this technique can display a larger number of mocap recordings. The figure shows the absolute acceleration of a rigid object in 122 recordings, all of which are sound-tracings of sound objects with impulsive onsets.

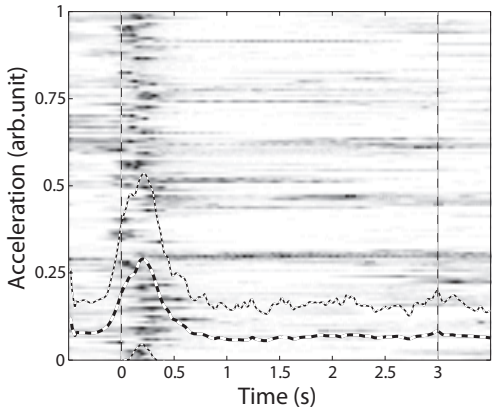


Figure 4.10: Mocapgram showing 122 data series of acceleration data. The data stem from sound-tracings of sounds with impulsive onsets.

4.1.5 Realtime Visualisation

The MoCap Toolbox is an excellent tool for working with recorded motion capture data. However, I missed an interactive 3D visualisation functionality, which could allow playing back motion capture data at different tempi, synchronised with sound files, with support for scrubbing back and forth in the recording and looping short segments of the motion capture data. I therefore implemented an addon to Toivianen's MoCap toolbox, which allows 3D display of motion capture data with scrubbing, looping, zooming, rotating and tempo adjustments, synchronised with audio. The implementation with an example recording is available for download at the fourMs website, along with a video that demonstrates the functionality.[3] Figure 4.11 shows a screenshot of this tool in action, and more details on the tool are provided in Section 5.3.

While visualisations of sound and motion features are useful, they are rarely a sufficient means of analysis. The sections below cover various quantitative methods that can be applied in experiments on correspondences between sound and motion features.

4.2 Statistical Tests

We can use visualisation techniques or simple statistical measures such as mean and standard deviation to get an indication of differences between various groups of data. However, the indications obtained from inspecting visualisations alone should preferably be tested quantitatively. Take as an example a comparison of the body mass of male Danish and Swedish citizens. Just by walking around the streets of Denmark and Sweden we could get a visual impression of the difference (or similarity) between the two populations, but to check the accuracy of our impression we would need to measure the body mass of the people. Since we cannot possibly measure

[3] http://fourms.uio.no/downloads/software/mcrtanimate

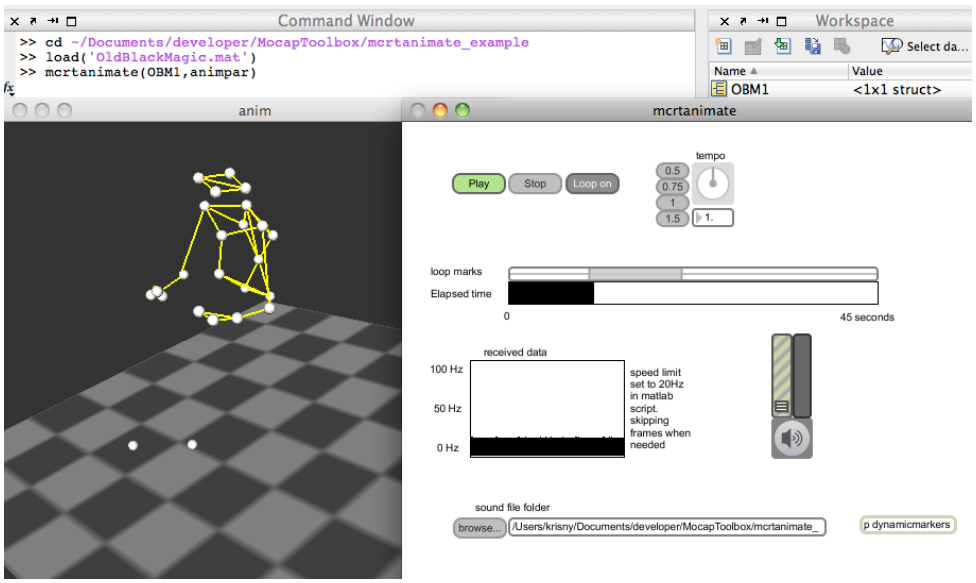


Figure 4.11: My implementation of interactive 3D animation for Toiviainen's MoCap Toolbox.

this for every male person in these countries, we select a subset from each country, called a *sample*. If the samples consist of a hundred Danes and a hundred Swedes chosen at random, the mean mass of the Danes and Swedes will probably be different by a small amount, and there will be some variation within the groups of Swedes and Danes. If the difference between the means is large and the variation within each group is small, we can be quite certain that there is a difference between the populations. However, if the difference is small and the variation within each group is large, we cannot generalise the result to count for the entire Danish and Swedish populations.

Similar problems are commonly faced in many research areas. Various statistical tests can be used to assess the *statistical significance* of the difference between two samples. In other words these tests estimate the probability that there is a difference between two populations based on a sample drawn from the populations. In some of my papers results from *t-test*[4] and *analysis of variance* (ANOVA) are reported. The tests have been applied to compare global motion features for various groups of motion capture recordings; for instance, to assess the statistical significance of the difference between *onset acceleration* for sound-tracings related to sounds with a soft onset and sounds with an impulsive onset.

The statistical tests discussed here assume that the data samples in each set are normally distributed. The results from the tests are thus exactly correct only for normal populations, something which is never the case in real life [Moore and McCabe, 2006]. If we use a larger sample size, the standard deviations of the set will approach the true standard deviation of the population. Thus the robustness of statistical tests increases with the sizes of the samples that are tested. Moore and McCabe [2006] state that even clearly skewed (i.e. not normally distributed) populations can be tested with *t*-tests when the sample size is larger than 40.

[4] Also called Student's *t*-test, after the inventor W. Gosset who was prevented by his employer from publishing under his own name. He published this technique under the pseudonym "Student" [Moore and McCabe, 2006].

4.2.1 *t*-test

A *t*-test can be used to test the statistical significance of the difference between random samples from two populations. The process involves defining a *null hypothesis*, stating that the means of the populations are equal, and this null-hypothesis is verified or falsified upon the *t*-test. For the sake of comparing results between experiments three measures are provided when reporting the results of *t*-tests: (1) The *degrees of freedom* (*df*)[5] is calculated from the sample size, and describes the number of values that are free to vary. (2) The *t-statistic* is calculated from the sample sizes as well as the standard deviations and mean values of the samples. (3) The *p-value* is the probability that the null-hypothesis is true, and is derived from the *t*-statistic.

The *p*-value denotes the probability that the two samples stem from populations with equal mean values. The sizes, means and standard deviations of the samples are used to estimate this probability. The *p*-value is used to infer whether the difference between the two distributions is statistically significant. A *significance level* (α) is defined and if *p* is less than this value, the result is said to be statistically significant at level α. Typical levels for α are between 0.001 and 0.05 [Moore and McCabe, 2006].

4.2.2 Analysis of Variance

In many cases Analysis of Variance (ANOVA) rather than the *t*-test is applicable. Like the *t*-test ANOVA tests for statistically significant differences between groups, but can take multiple groups into account. In other words while a *t*-test can be used to assess the statistical significance of the difference in *two* sample means, an ANOVA can be applied to test whether the observed difference in mean values of *several* groups is statistically significant.

Furthermore, ANOVA allows measurement of the significance of several factors, or features, at once. For instance, in the example with Danes and Swedes presented above, the age of those measured could be added in the analysis. This would allow us to infer whether there is a difference in body mass between Danes and Sweden and, further, whether age is related to mass.

ANOVAs do not use the *t*-statistic but rather an *F-statistic*. This statistic is based on the variations *within* each group and *between* the groups [Moore and McCabe, 2006]. In addition to the *F*-statistic the degrees of freedom and the *p*-value are specified when reporting ANOVA results.

4.3 Correlation

In real life we daily encounter variables that are related. The value of the volume knob on a hi-fi system is related to the sound level of the output, and the number of floors in a building is related to the number of steps in the stairs. If we want to determine whether or not there is a relation between two variables in a data set and how strong the relation is, we need a measure to describe how the variables *correlate*.

Correlation is a measure of the direction and strength of the relationship between two quantitative variables [Moore and McCabe, 2006]. The value of a correlation is between -1 and 1,

[5] Not to be confused with the term *degrees of freedom* (DOF) in motion tracking.

where a correlation of 1 denotes a full dependence between the two variables, and -1 denotes a full negative dependence between the variables.

Several methods are available for determining correlation coefficients. Firstly, the *Pearson correlation coefficient* measures the linear dependence between variables. When the two variables are plotted on separate axes in a scatterplot a Pearson correlation coefficient of 1 means that all the samples in the two variables follow a straight ascending line, and similarly a correlation coefficient of -1 shows as a straight descending line, as shown on the left in Figure 4.12 [Zou et al., 2003]. Non-linear correlations may also exist, for instance if one of the input variables stems from a skewed distribution. This is particularly true in music-related research, where several sound features scale logarithmically (e.g. loudness and pitch). For non-linear relations, the *Spearman* ρ measure is more applicable than the Pearson correlation. Non-linearity is achieved by ranking (ordering) the input variables and calculating the Pearson correlation from the rank, rather than the variable value [Spearman, 1904]. The result is that a continuously rising or falling tendency in a scatter plot will have correlation coefficients of 1 and -1 respectively, as shown on the right in Figure 4.12.

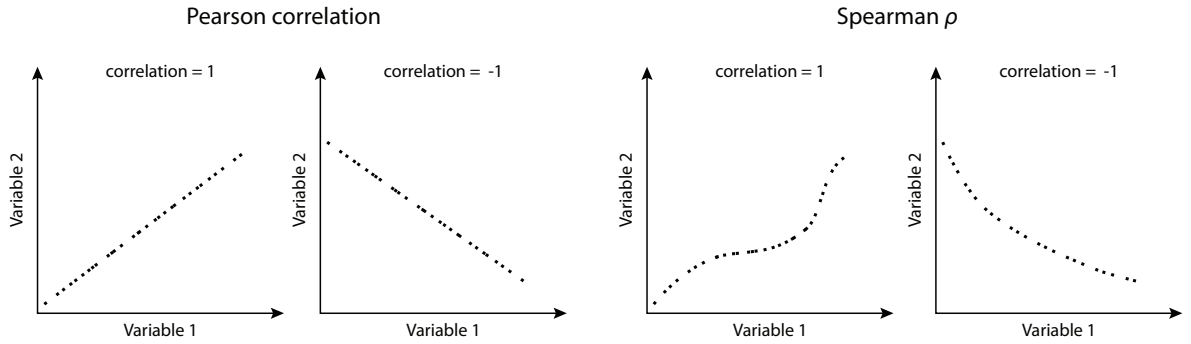


Figure 4.12: The difference between Pearson correlation and Spearman ρ. Pearson correlation measures the linear relation between the variables, and Spearman ρ uses a ranking of the variables to measure the monotonic relation between them.

4.3.1 Correlation and Music-Related Time-Series

Emery Schubert [2002] and later also several other researchers [e.g., Vines et al., 2006, Upham, 2012, Kussner, 2012] have presented critical views on the common practice in music cognition research of uncritically applying the Pearson correlation measure to time-series of music-related data without taking into account the serial nature of the data. Specifically, the correlation coefficients cannot be tested for statistical significance because the value of each sample is not drawn randomly from a normal distribution. This is because the value at each time step will be dependent on the value in the immediately preceding time steps. Take as an example a 200 Hz motion capture recording — it is impossible to have ones arms fully stretched in one time step and then fully contracted in the next time step (5 milliseconds later). Consequently the sample value in each time-frame is likely to be close to the previous sample value, and unlikely to be far away from that value. This effect is known as *serial correlation*.

Some approaches have been suggested to make correlation measures more applicable when analysing time-series in music research. For instance, the serial correlation may be lowered by

downsampling the data series, or by applying the correlation analysis to the first-order difference (derivative) of the data series [Schubert, 2002]. Furthermore, Spearman ρ has been suggested as a more appropriate measure than Pearson correlation, since the ranking of sample values in Spearman ρ prevents the inflation of the correlation coefficient that occurs with Pearson correlation [Schubert, 2002].

Upham [2012] argues that the correlation coefficients themselves can be useful measures, but that one cannot uncritically report on the statistical significance of correlations between data-series, for instance by running statistical tests on the correlation coefficients. Schubert [2002] also argues that inspecting the correlation coefficients can be useful as an assessment of the distribution of correlations within a single data set. However, because of the problems with serial correlation the coefficients should not be used for comparison of data sets that have been gathered in different circumstances.

4.3.2 Cross-Correlation

The correlation between two variables is a measure of the relation between them. We may not be interested in this relation *per se*, but rather how it is affected by some other factor. For instance, we can examine how the correlation coefficient between two time-series changes if we shift one of the time-series back or forth in time. In this manner the correlation between the time-series becomes a function of a time-shift (lag) applied to one of them. This process, called *cross-correlation*, is shown in Figure 4.13.

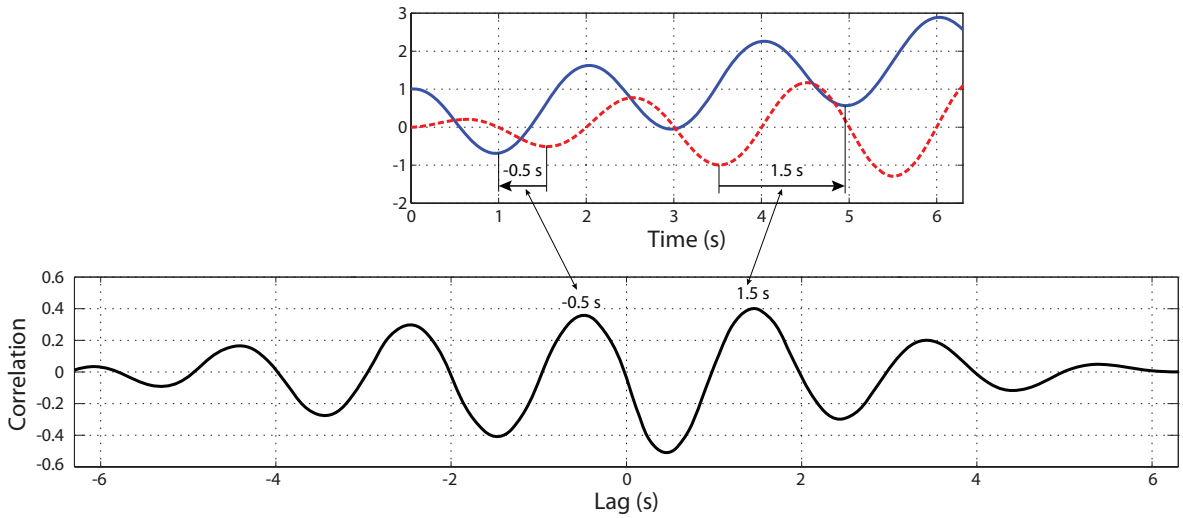


Figure 4.13: Illustration of cross-correlation. Both of the functions in the top plot have a periodic tendency at 0.5 Hz, with a phase difference of the quarter of a wavelength (0.5 s). The correlation is highest when the red dashed line is shifted back 0.5 s or forward 1.5 s.

Cross-correlation applied to two related time-series can give an indication of any time lag between them. In my research I have applied this technique to the orientation data[6] from two tracking systems running in parallel in order to analyse the latency of one system as compared with the other. Cross-correlation can also be applied to find periodicities within a single time-series. In other words we can find repeating patterns in the time-series by calculating its correla-

[6] Actually the first order difference of orientation data. This is presented in more detail in Paper III.

tion with itself as a function of a time lag, a process known as *autocorrelation*. If the time-series is periodic, the resulting cross-correlation function will have peaks at every wavelength.

4.3.3 Canonical Correlation

The correlation approaches discussed above measure the relation between two variables. Canonical correlation analysis (CCA) is slightly different in that it measures the relation between two *sets* of variables [Hotelling, 1936]. As shown by Caramiaux et al. [2010] CCA can be applied to a set of sound features and a set of motion features to analyse how several sound features and several motion features relate to each other. In this case CCA finds two sets of basis vectors, one for the sound features and the other for the motion features, such that the correlations between the projections of the features onto these basis vectors are mutually maximized [Borga, 2001].[7]

CCA is illustrated in Figure 4.14. The first projection of sound and motion features onto their respective basis vectors is that in which the correlation between the projected features is maximised. These projections are known as the *first canonical variates*.[8] The *second canonical variates* follow by projecting the features onto basis vectors that are orthogonal to the first basis vectors, i.e. the second canonical variates are uncorrelated to the first variates. This is repeated until all the dimensions in the sound features or motion features are covered (e.g. if there are 4 sound features and 3 motion features, 3 sets of canonical variates are calculated).

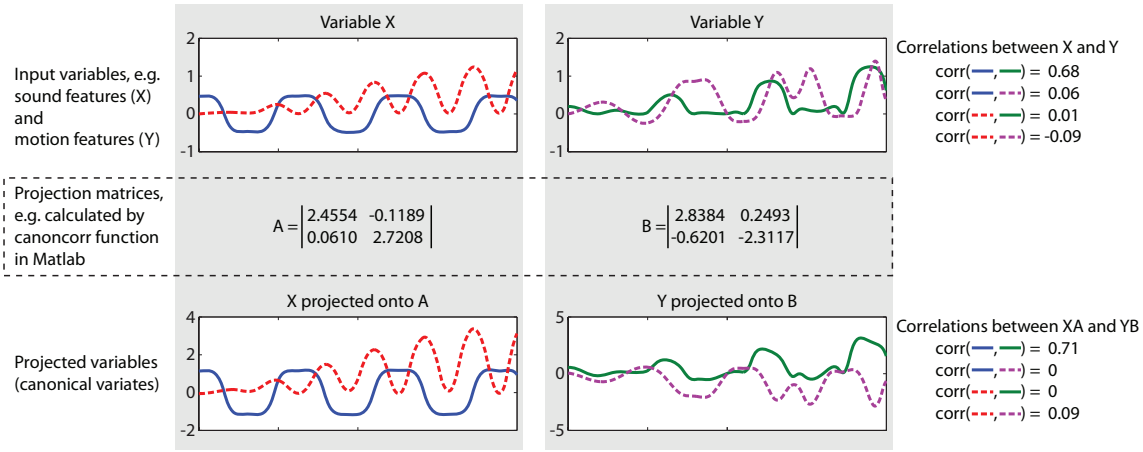


Figure 4.14: Illustration of canonical correlation. The correlations between the variables at the top are between -0.09 and 0.68. By projecting the variables onto new spaces two projected variables are found. The maximum correlation between the two sets is explained between the first canonical variates (0.71), and the correlation between the first and second variate is 0. A similar example applied to sound and motion features is shown in Paper VIII.

In my papers I have followed the approach of Caramiaux et al. [2010] and inspected the *canonical loadings* when interpreting the results of a canonical correlation analysis. This in-

[7] To readers familiar with *Principal Component Analysis* (PCA), CCA may be understood as a similar phenomenon. PCA operates on a set of variables within a single data set, explaining as much as possible of the variance in the first principal component. Then second principal component then explains as much of the remaining variance as possible, and so forth. Rather than explaining variance within a single set of variables, CCA tries to explain the maximum correlation *between two sets* of variables in the first canonical variates, and then as much as possible of the "remaining" correlation in the second canonical variates.

[8] In my papers I have referred to these as *canonical components*, but the term *canonical variates* seems to be more commonly used.

volves calculating the correlation between the input features and their corresponding canonical variate. A high canonical loading between an input variable and a canonical variate indicates that the input variable is pertinent to the correlation described by the particular canonical variate.

One weakness of canonical correlation analysis, especially if a large number of features are used, is the possibility of "overfitting" the CCA to the data. This means that the CCA might give very good solutions that are not due to actual correlations, but rather to small levels of noise in the data that are exploited by the CCA [Melzer et al., 2003]. For this reason limited numbers of sound and motion features have been used in my analyses.

4.4 Pattern Recognition-Based Classification

An analysis approach distinctly different from the correlation methods presented above considers an entire data set and implements a classifier algorithm to search for patterns within the set. Fortunately, a wide variety of ready-made implementations of computer classifiers is available, so these methods can be applied without detailed knowledge of the algorithms involved. In my work I have analysed the motion recordings with a *Support Vector Machine* (SVM) classifier. This technique was chosen because it typically matches or outperforms other classification techniques in terms of error rate [Burges, 1998]. I have used the software *Rapidminer* to implement the classifiers in my research [Mierswa et al., 2006]. This software includes a wide range of classifiers and a user interface which greatly facilitates the classification task. SVM is implemented in Rapidminer by the *LIBSVM* library [Chang and Lin, 2011], which also contains useful scripts for optimising certain parameters of the classifier. Basic concepts of computer classifiers and support vector machines are outlined below, as well as details of how classification results can be analysed.

In computer-based classification each instance in a data set is usually represented by a *class ID* and a *feature vector*. The class ID is equal among all instances in a class, and the feature vector is specific to each instance. If we want to classify fruit, and look at the class 'apple', all apples will have 'apple' as their class ID, but features such as 'colour', 'size' and 'shape' will vary. The data set is typically split into two subsets: a *training set* and a *validation set*. The classifier uses the data in the training set to develop rules for what is common between the instances in a class, and what distinguishes these instances from other classes. Continuing the fruit example above, a classifier may primarily use 'shape' to distinguish bananas from apples, but other features like 'size' or 'color' may be necessary to differentiate apples from peaches or oranges.

4.4.1 Support Vector Machines

A Support Vector Machine (SVM) classifier is trained to find a hyperplane in the feature space between the classes of training data [Duda et al., 2000]. Figure 4.15 shows the location of the optimal hyperplane between two classes, where three instances make up the so-called *support vectors*, which are equally close to the hyperplane.

It is often the case that the training data are not linearly separable. When this is so, the support vector machine increases the dimensionality of the feature space by a *kernel function*. This process is illustrated in Figure 4.16.

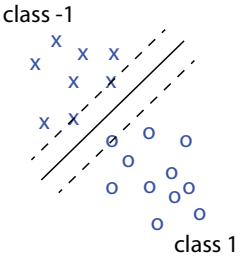


Figure 4.15: The optimal hyperplane (which in this 2-dimensional case means a line) is located between the *support vectors*. The classes are named -1 and 1, corresponding to the way in which this hyperplane is derived, where the two margins (dashed lines) are found -1 and 1 times a certain vector from the hyperplane [Duda et al., 2000].

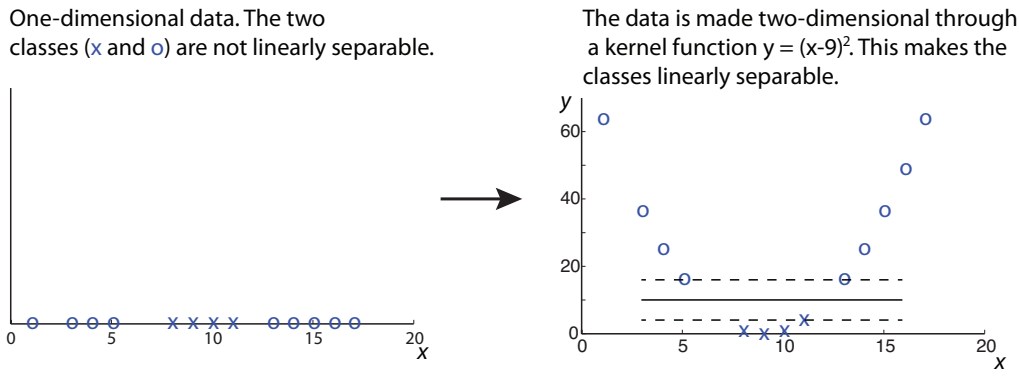


Figure 4.16: The two classes in the one-dimensional data in the left plot are not linearly separable. By adding another dimension $y = (x - 9)^2$ it is possible to identify support vectors.

4.4.2 Validating the Classifier

After the training process the performance of the classifier is evaluated by classifying the instances in the validation set. The evaluation can be measured using terms from the field of document retrieval, namely *precision* and *recall* [Salton and Lesk, 1968]. Continuing with the fruit classification example above, let us say that we want to retrieve all the apples from a fruit basket. We pick fruit from the basket; mostly apples but also a few oranges. We fail to notice some of the apples in the basket. *Precision* then denotes the ratio between the number of apples picked and the total number of fruits we picked (including oranges). *Recall* denotes the ratio between the number of apples picked and the total number of apples that were present in the basket in the first place.

Applied to computer classification, this measure shows correctly classified instances rather than correctly retrieved documents (or fruit), and we get precision and recall measures for each class. We define *class precision* (CP) and *class recall* (CR) for class i as:

$$\mathrm{CP}_i = \frac{||R_i \cap A_i||}{||A_i||} \qquad \text{and} \qquad \mathrm{CR}_i = \frac{||R_i \cap A_i||}{||R_i||},$$

where $||A_i||$ denotes the number of examples classified as i, and $||R_i||$ denotes the total numbers of examples in class i. In other words CP denotes the ratio between correctly classified examples and all the examples the classifier *predicted* to be in the specific class. CR denotes the ratio between correctly classified examples and the *true* number of examples in class i. Figure 4.17 shows how both measures are necessary to get a good assessment of the performance of the classifier: 100 % class precision could mean that the class has been drawn too narrowly, and a 100 % class recall could mean that the class has been defined too broadly.

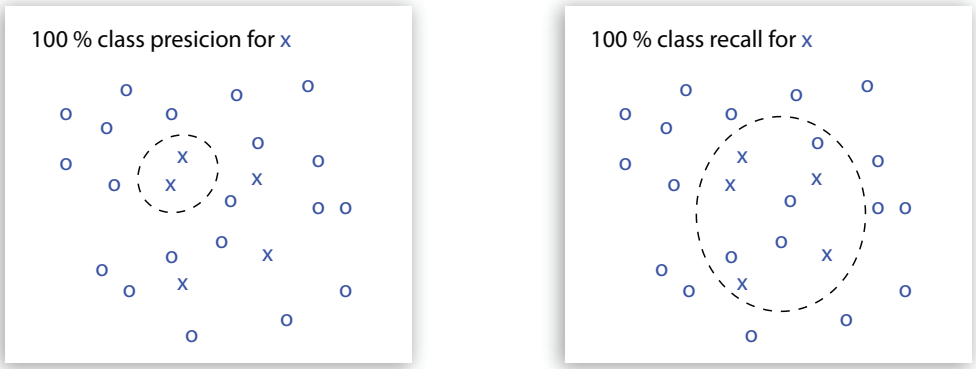


Figure 4.17: In the figure to the left 100 % class precision is obtained. However, several examples that should have been included are left out. To the right all the examples have been included in the classification. However, a number of incorrect examples are also included.

When the data set is of limited size a technique called *cross-validation* can be used to obtain a larger number of examples in the validation set [Duda et al., 2000]. That is, multiple classifications and validations are performed and the examples present in the validation set are different each time. In my experiments I applied the *leave-one-out* principle which entails using the entire data set but one example for training the classifier, and subsequently performing validation with the remaining example. The process is repeated as many times as there are examples in the data set, such that each example is used once for validation.

More detailed results than the precision and recall are obtained by inspecting the classifier results in a *confusion matrix*. This matrix shows the distribution of the examples in the validation set and how they were classified. An example of what the confusion matrix looks like is given in Table 4.1. Systematic classification errors may be revealed by the confusion matrix and such errors may suggest that there are similarities between classes. Examples of how this can be applied to sound and motion analysis will be given in the included Papers V and VIII.

Table 4.1: Confusion matrix showing a classification result. Each row shows the classifications (predictions) made by the classifier and each column shows the actual classes of the examples. The correctly classified examples are found along the diagonal marked in grey. This particular table suggests that classes 1 and 3 have some similarities.

	True 1	True 2	True 3	Class Precision
Predicted 1	6	0	5	55 %
Predicted 2	1	10	1	83 %
Predicted 3	3	0	4	57 %
Class Recall	60 %	100 %	40 %	

4.5 Summary

This chapter has introduced various methods of analysing correspondences between sound and motion. Sound and action objects can be described with time-varying features, meaning features that describe how the sound or motion evolves at regular time-intervals. They can also be

described by global features, meaning a single value or typological description that describes the entire object.

The chapter presented techniques for visualising motion data and how the visualisations can be applied to obtain an overview of general tendencies within a single motion recording or a set of recordings. The visualisations may be useful in combination with statistical tests, such as *t*-tests and ANOVAs, which can be applied to test the significance of tendencies in the data. Furthermore, the chapter examined how various correlation measures can be applied to evaluate the correlation between sound and motion features. While correlation coefficients can usually be tested for statistical significance, this is not recommended for continuous sound and motion features given the serial nature of the data. Finally, the use of a computer-based classifier was introduced, with an example of how a confusion matrix can be analysed to get an indication of similar classes.

Chapter 5

Research Summary

This chapter provides summaries of the papers included in the thesis. The overall structure of the papers is outlined in Section 5.1, together with a presentation of how the papers relate to the research objectives. Following in Section 5.2 are shorter and more specific descriptions of each paper, including abstracts. Finally, in Section 5.3, I present the software that I have developed and made available to other researchers while working on this thesis.

5.1 Overview

The eight papers included in this thesis cover tools and methods both for capturing and analysing music-related motion. The work that has been carried out involves prototyping and development of software, evaluation of technologies, collection and analysis of data and an evaluation of existing methods of comparing lower-level features of sound and body motion.

As stated in Section 1.4 the main objective in this thesis is to develop methods and technologies for studying music-related motion, with the purpose of analysing cross-modal correspondences between sound and body motion. Below I use the three sub-objectives to discuss the requirements of fulfilling this objective.

5.1.1 Sub-objective 1: Data Handling

As discussed in Section 3.6 researchers on music and motion work with a large variety of data. Music-related data can be numerical, e.g. motion tracking data, audio and video, with differences in dimensionality and sampling rate. Music-related data can also take the form of nominal descriptions such as genre labels or verbal descriptions, or metadata describing an experimental setup.

The variety of data involved in music research differs from experiment to experiment. One experiment might involve full body motion capture of a cellist performing in a string quartet, while another the tracking of the head marker of each individual in the ensemble. What is more, data from listeners can be recorded, e.g. in sound-tracing experiments using markerless motion capture and electromyography sensors. These examples are but a few scenarios of research on music-related motion, and the possibilities for different setups are endless.

This shows that there is a need for solutions for storing music-related data. The implementation must be able to keep track of different simultaneous data streams, and flexible enough to

be used with different experimental setups. Next it must be straight forward to use, affording a possibility to make multiple recordings in quick succession such that an experiment can run smoothly without unnecessary waiting time for the participants.

Contribution

The first sub-objective in this thesis is primarily met in Paper I. This paper presents a toolbox for capturing and playing back synchronised music-related data, based on work related to the Gesture Description Interchange Format. The released version of the toolbox follows data types according to the current GDIF specification.[1] It is, however, flexible, and new data types can be added by modifying a text file. What is more, the software is open source, implemented in Max as modules for the Jamoma[2] framework, so users who need increased flexibility may make the necessary adjustments to have the software meet their needs. I have developed scripts for importing and parsing GDIF recordings into Matlab data structures, for analysis e.g. with the MoCap Toolbox. These scripts are supplements to the toolbox presented in Paper I, and are presented in Section 5.3.2.

In addition to the solutions for recording and playing back music related data, Paper IV presents a solution for realtime handling of music-related data. The software calculates features from a stream of position data and streams the features to a sound synthesiser using Open Sound Control (OSC).

5.1.2 Sub-objective 2: Evaluation of Motion Tracking Technologies

As explained in Chapter 3 there is available a wide range of technologies for tracking and capturing motion. Earlier discussion in this thesis has illuminated various strengths and weaknesses of the technologies. Still, there are important differences in the various implementations of the same tracking technologies, of which researchers should be aware.

In musical realtime applications of tracking data the stability of the tracking systems is more important than when tracking data is used in non-realtime. In non-realtime the data can be processed after recording and it is not critical if time is spent on computationally intensive filtering techniques or gap-filling. For designers of interactive musical applications, however, processing steps that induce more than a few milliseconds delay is typically unacceptable.

There is a need to identify strengths and weaknesses of using various motion capture systems in musical applications. In any case the degree of noise and drift should be identified and taken into account when motion data are analysed or used in musical interaction. Furthermore, in realtime applications it is essential to be aware of timing performance, both in terms of latency and jitter, and the general data quality, such as susceptibility to frame drops or marker swaps.

Contribution

In this thesis quantitative evaluations of various motion tracking systems are provided in Papers II and III. These papers are mainly technology-orientated, focusing on aspects such as noise, drift, latency and jitter. To a certain extent Paper IV, also contributes to this sub-objective,

[1] http://xdif.wiki.ifi.uio.no/Data_types
[2] http://www.jamoma.org

although from an instrument design perspective. This paper presents a musical application of motion tracking data in realtime and shows how state-of-the-art motion tracking technology provides an interesting paradigm for sound-interaction, more so than an affordable Wiimote based on inertial sensors.

5.1.3 Sub-objective 3: Sound–Action Analysis

As argued in Chapter 2 there is strong evidence that the experience of music is a multimodal phenomenon, incorporating both auditory and motor sensations. Consequently visualisation of action trajectories, and particularly those related to sound-producing actions, has been advocated as an important part of music cognition. However, the existing body of music research, analysing lower-level features of people's spontaneous motion response to short sound objects, is quite limited. Most of the research conducted in this area has been concerned with longer periods of musical sound, for instance analysing periodicities or emotional responses to musical sound [e.g., van Noorden and Moelants, 1999, Toiviainen et al., 2010]. Furthermore, most of the research on links between short sound objects and body motion has gathered responses in terms of metaphors or visualisation of motion rather than actual motion data [e.g., Eitan and Granot, 2006, Merer et al., 2008, Kohn and Eitan, 2012].

It is difficult to identify any single reason why the research done on this particular topic has been quite limited. Most likely, there are several reasons. Firstly, accurate motion data may be difficult to obtain as they require advanced technologies which are not available to all researchers. It may be hoped, though, that this situation is about to change, as better low-cost tracking technologies are developed. Secondly, the relation between two multidimensional phenomena, such as motion and sound, is difficult to analyse. This interdisciplinary research requires knowledge both of quantitative data processing and of music theory, for the application of appropriate methods of analysis.

There is a need to extend research on how lower-level features of bodily motion are related to musical sound. This will contribute to the further development of theories of sound-action relations. There is still a need to develop robust methods for analysing relationships between sound and body motion. As explained in Section 4.3 current methods of correlation of time-varying musical features can not be tested for statistical significance and thus it is difficult for researchers to evaluate the significance of their findings. This calls for an increased number of empirical studies which can be used to test the approaches that have been suggested (e.g. by Schubert [2002]) for improving current analysis methods.

Contribution

Papers V, VI, VII and VIII, are all concerned with analysis of data collected in two sound-tracing experiments, using short sound objects as stimuli, and free-air motion as response. The papers include empirical results, but are to a large extent methodologically orientated. The papers discuss design of sound-tracing experiments, and introduce a variety of features calculated from motion data. A variety of analysis methods is presented and evaluated, both for time-varying features, and for global features of actions and sound objects. The software introduced in Sections 5.3.4 and 5.3.3 also provides tools for visualising data with a high number

of dimensions. This will be essential to expand future sound-tracings studies to incorporate full-body motion data.

Lastly, Paper IV shows how cross-modal relations can be studied through the development of new interfaces for musical expression. The paper also serves as a possible application scenario for the other work reported in this thesis.

5.2 Papers

This section presents details of the motive and contributions of each paper together with abstracts.

5.2.1 Paper I

A Toolbox for Storing and Streaming Music-Related Data
K. Nymoen and A.R. Jensenius.
In *Proceedings of SMC 2011 8th Sound and Music Computing Conference "Creativity rethinks science"*, pages 427–430, Padova University Press 2011.

Abstract: Simultaneous handling and synchronisation of data related to music, such as score annotations, MIDI, video, motion descriptors, sensor data, etc. require special tools due to the diversity of the data. We present a toolbox for recording and playback of complex music-related data. Using the Sound Description Interchange Format as a storage format and the Open Sound Control protocol as a streaming protocol simplifies exchange of data between composers and researchers.

The work presented in this paper started several years ago with Jensenius' proposal of the Gesture Description Interchange Format (GDIF) [Jensenius et al., 2006b, Jensenius, 2007a,b]. More than a file format, the GDIF proposal advanced an idea of how music-related data could be structured into several layers, spanning from raw sensor and audio data to higher-level features, closely corresponding to Camurri's [2005] model for music-related data (ref. Section 2.1). At a panel discussion at the International Computer Music Conference in 2007 the Sound Description Interchange Format (SDIF) was suggested as one of several possible file formats for storing GDIF data [Jensenius et al., 2007].

I became involved in GDIF development in late 2007 and Jensenius, Godøy, and I published a paper in 2008 reporting on the use of SDIF in the study of the hand and upper-body motion of pianists [Jensenius et al., 2008]. At the time this approach required substantial programming to set up an environment for storing and streaming the data. Accordingly, I started the implementation of a toolbox within the Jamoma framework, with the purpose of streamlining recording and playback of music-related data [Nymoen, 2008b]. This implementation was developed further by Jensenius [2009] and me, and Paper I presents the current state of this toolbox.

5.2.2 Paper II

Comparing Inertial and Optical MoCap Technologies for Synthesis Control
S.A. Skogstad, K. Nymoen, and M.E. Høvin.
In *Proceedings of SMC 2011 8th Sound and Music Computing Conference "Creativity rethinks science"*, pages 421–426, Padova University Press 2011.

Abstract: This paper compares the use of two different technologies for controlling sound synthesis in real time: the infrared marker-based motion capture system OptiTrack and Xsens MVN, an inertial sensor-based motion capture suit. We present various quantitative comparisons between the data from the two systems and results from an experiment where a musician performed simple musical tasks with the two systems. Both systems are found to have their strengths and weaknesses, which we will present and discuss.

The work carried out in this paper originated in an interest in developing a musical interface based on motion tracking data. Since various advanced tracking technologies are available in the labs at the University of Oslo, Ståle A. Skogstad and I started systematic testing of two of these technologies.

We made multiple recordings of a person wearing the Xsens motion capture suit and a full-body set of markers for the NaturalPoint OptiTrack infrared mocap system. The recordings involved various basic tasks such as sitting still, walking and jumping. A musician was recruited to a small experiment in which both of the tracking technologies were applied to control a software sound synthesiser. The musician was given simple music-related tasks such as following a simple melody and triggering tones and we compared the data from each recording as well as the verbal feedback of the musician.

The experiment revealed strengths and weaknesses of both motion tracking systems. The Xsens suit can track full-body motion in a large space; it is portable and quite easy to set up. The OptiTrack system is limited to the visual range of the cameras and needs camera stands and lots of cables which makes it less portable and more cumbersome to set up. The Xsens system suffers from a potentially substantial amount of drift, which makes Xsens position data less accurate than OptiTrack data. Furthermore, Xsens is limited to certain tracking configurations (e.g. full-body or upper-body), while OptiTrack is more flexible, allowing tracking of user-defined rigid bodies.

With OptiTrack we found deviations in position of a steady marker when the marker was occluded in one of the cameras (but still tracked by the system). This gave static displacements of up to 0.5 mm and spikes of more than 1 mm. Generally, the level of noise in the OptiTrack position data was not very high, but if the position data are differentiated to calculate velocity and acceleration, the noise increases drastically. This has led Skogstad to continue working on optimal low-latency filters and differentiators for realtime use [Skogstad et al., 2012a,b]. The work done in this paper also inspired two more articles comparing tracking technologies, the first (Paper III in this thesis) presenting a study of data quality in a mobile device, and the second investigating data quality of the Qualisys and OptiTrack mocap systems [Jensenius et al., 2012].

5.2.3 Paper III

Comparing Motion Data from an iPod Touch to a
High-End Optical Infrared Marker-Based Motion Capture System
K. Nymoen, A. Voldsund, S.A. Skogstad, A.R. Jensenius, and J. Torresen.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 88–91, University of Michigan 2012.

Abstract: The paper presents an analysis of the quality of motion data from an iPod Touch (4th gen.). Acceleration and orientation data derived from internal sensors of an iPod is compared to data from a high end optical infrared marker-based motion capture system (Qualisys) in terms of latency, jitter, accuracy and precision. We identify some rotational drift in the iPod, and some time lag between the two systems. Still, the iPod motion data is quite reliable, especially for describing relative motion over a short period of time.

My research group in Oslo is involved in the European research project *EPiCS*[3] one of whose goals involves developing a mobile device for active music. The active music system involves several mobile devices which contribute to the musical output, either controlled by autonomous agents or by human users [Chandra et al., 2012]. Body motion has been chosen as the main input paradigm from human users to the system as this is thought to represent more spontaneous movements than the touch screen interface. An *iPod Touch* was considered as development platform for the active music device. Thus in order to determine whether the iPod met the requirements for such a device it was essential to evaluate the quality of the motion data obtained from the iPod.

The iPod Touch was equipped with four markers and tracked as a rigid object with a Qualisys optical tracking system. Recordings of the iPod sensor data and motion tracking data were made using the toolbox presented in Paper I, and the data from the two systems were analysed in Matlab.

The analysis showed that the accuracy of the iPod orientation data was quite high, albeit with some drift around the gravity vector. The acceleration data were also quite accurate; however, the error was too high for estimation of position from the acceleration data. The data samples from the iPod arrived less regularly than the Qualisys data, and on average 43 ms later, indicating that there is substantially more latency and jitter in the iPod data than in the Qualisys system.

5.2.4 Paper IV

SoundSaber — A Motion Capture Instrument
K. Nymoen, S.A. Skogstad and A.R. Jensenius.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 312–315, University of Oslo 2011.

---

[3] European Union FP7, project 257906, Engineering Proprioception in Computer Systems.

Abstract: The paper presents the SoundSaber — a musical instrument based on motion capture technology. We present technical details of the instrument and discuss the design development process. The SoundSaber may be used as an example of how high-fidelity motion capture equipment can be used for prototyping musical instruments, and we illustrate this with an example of a low-cost implementation of our motion capture instrument.

The SoundSaber is a musical instrument based on optical infrared marker-based motion tracking technology. The development of the SoundSaber started in 2009, after the fourMs group at the University of Oslo acquired a NaturalPoint OptiTrack motion capture system. During my undergraduate and graduate studies I experimented with various ways of interacting with sound through sensors and custom-built musical instruments [e.g., Nymoen, 2008a] and sound interaction with motion-based interfaces was widely explored in our research group [Jensenius et al., 2006a, Jensenius, 2007a, Jensenius and Voldsund, 2012]. The newly acquired motion capture system motivated a search for new ways of interacting with sound.

The initial development phase consisted of trying out different types of object equipped with reflective markers that could be used to control a sound synthesiser developed in Max5. Synthesis techniques were tested along with ways in which the tracking data could be mapped to synthesis parameters.

Development of musical interfaces is interesting not only for the novelty of the interface itself but also because well-designed systems of interaction between motion and sound can tell us something about how sound and motion are related. I believe that motion-based interfaces for sound interaction become engaging to interact with when their interaction model fits our mental model of how sound corresponds to motion. Development of new musical interfaces can be seen as an *analysis-by-synthesis*[4] in the sense that the mappings in an interesting musical interface reflect what we perceive as "natural" couplings between actions and sound.

By using a rigid object as controller the SoundSaber can be used without the need for a tedious setup each time a new user would like to try the instrument. This has made it possible to demonstrate the SoundSaber at various conferences, research fairs and exhibitions, for instance as shown in Figure 5.1 where the Norwegian Minister of Education and Research interacts with the SoundSaber. A more detailed lists of exhibitions and public demos of the SoundSaber can be found on the fourMs website.[5]

A low-cost version of the SoundSaber was also implemented, using a *Wiimote*[6] instead of the optical infrared marker-based tracking technology. I found that people tended to tilt the Wiimote while moving it up and down. Because of this correlation between tilt angle and vertical position, the orientation data could be used to some extent to simulate the positional control data provided by the optical system.

[4] Analysis-by-synthesis means analysis of a phenomenon by synthesising the same phenomenon from basic building blocks. The technique has been used widely in linguistics [e.g., Fant, 1961, Halle and Stevens, 1962] and also in research on musical timbres [Risset, 1991].

[5] http://fourms.uio.no/projects/sma/subprojects/Soundsaber

[6] A Wiimote is a game controller for Nintendo Wii

Figure 5.1: Norwegian Minister of Education and Research, Kristin Halvorsen, moving a SoundSaber on a visit to the University of Oslo in April 2012. *Photo: Yngve Hafting*

5.2.5 Paper V

Searching for Cross-Individual Relationships between
Sound and Movement Features Using an SVM Classifier
K. Nymoen, K. Glette, S.A. Skogstad, J. Torresen, and A.R. Jensenius.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 259–262, Sydney University of Technology 2010.

Abstract: In this paper we present a method for studying relationships between features of sound and features of movement. The method has been tested by carrying out an experiment with people moving an object in space along with short sounds. 3D position data of the object was recorded and several features were calculated from each of the recordings. These features were provided as input to a classifier which was able to classify the recorded actions satisfactorily, particularly when taking into account that the only link between the actions performed by the different subjects was the sound they heard while making the action.

This is my first paper on "sound-tracing" which presents a novel approach to analysing correspondences between sound and body motion by employing a pattern recognition classifier to search for similarities between people's motion responses to various sound objects. Inspired and influenced by a previous sound-tracing experiment conducted by my advisors who had studied people's rendering of short sound objects on a digital tablet [Godøy et al., 2006b] I designed an experiment which extended the response data to three-dimensional motion.

A data set was collected, in which 15 participants listened to 10 sound files and moved a rod (the SoundSaber controller) in free air while pretending that their motion would create the sound they heard. A number of features were calculated from the mocap data, and we applied a pattern classifier to the features using the sound file that inspired the motion as class identifier.

Thus by analysing the classification results presented in a confusion matrix (ref. Section 4.4) we were able to infer similarities between motion recordings from different subjects.

The classifier performed with an overall classification accuracy of $78.6\% \pm 7.3\%$, indicating that the extracted features contained salient information about the sound-tracings. We performed the same classification task using only the features that described vertical motion. This greatly reduced the overall classification accuracy. The precision and recall of the sounds with changing pitch, however, remained quite high. This indicated that the participants linked vertical motion with pitch trajectories.

5.2.6 Paper VI

Analyzing sound tracings: a multimodal approach to music information retrieval
K. Nymoen, B. Caramiaux, M. Kozak, and J. Torresen.
In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 39–44, ACM 2011.

Abstract: This paper investigates differences in the gestures people relate to pitched and non-pitched sounds, respectively. An experiment has been carried out where participants were asked to move a rod in the air, pretending that moving it would create the sound they heard. By applying and interpreting the results from Canonical Correlation Analysis we are able to determine both simple and more complex correspondences between features of motion and features of sound in our data set. Particularly, the presence of a distinct pitch seems to influence how people relate gesture to sound. This identification of salient relationships between sounds and gestures contributes as a multi-modal approach to music information retrieval.

One year after the publication of the previous paper on sound-tracing I met Baptiste Caramiaux, who also did research on music-related body motion. Caramiaux et al. [2010] had previously applied Canonical Correlation Analysis (CCA) to a set of sound and motion recordings and wanted to explore the technique further.

As explained in Section 4.3 canonical correlation can be used to find the linear combination of a set of sound features and a set of motion features which explains the maximum correlation of the two. This means that while the method applied in Paper V mainly provided an indication of similarity between sound-tracings at a high level, CCA could be applied to analyse correspondences between the lower-level features of the sound objects and the sound-tracings. We applied CCA to the same data set as in Paper V, to analyse the relations between sound and motion features in more detail than the previous publication.

The CCA verified the finding from Paper V, where vertical position and pitch were found to be closely related. For non-pitched sound objects we did not see similarly strong canonical loadings as for pitch, and therefore inspected the strongest canonical components individually. This analysis showed that brightness corresponded to vertical position, and loudness to horizontal position and velocity.

5.2.7 Paper VII

A Statistical Approach to Analyzing Sound Tracings
K. Nymoen, J. Torresen, R.I. Godøy, and A.R. Jensenius.
In S. Ystad, M. Aramaki, R. Kronland-Martinet, K. Jensen, and S. Mohanty (eds.)
Speech, Sound and Music Processing: Embracing Research in India, Lecture Notes in Computer Science vol. 7172, pages 120–145. Springer, Berlin Heidelberg 2012.

Abstract: This paper presents an experiment on sound-tracing, meaning an experiment on how people relate motion to sound. 38 participants were presented with 18 short sounds, and instructed to move their hands in the air while acting as though the sound was created by their hand motion. The hand motion of the participants was recorded, and has been analyzed using statistical tests, comparing results between different sounds, between different subjects, and between different sound classes. We have identified several relationships between sound and motion which are present for the majority of the subjects. A clear distinction was found in onset acceleration for motion to sounds with an impulsive dynamic envelope compared to non-impulsive sounds. Furthermore, vertical movement has been shown to be related to sound frequency, both in terms of spectral centroid and pitch. Moreover, a significantly higher amount of overall acceleration was observed for non-pitched sounds as compared to pitched sounds.

The data set used in Papers V and VI has some limitations. Firstly, the sound stimuli used did not allow for controlling opposite conditions. For example, if a link was found between sounds with rising pitch and a certain motion feature, the experiment did not include a sound object with inverted pitch envelope to determine whether the same relation held in the opposite case. Secondly, the motion response was collected using the SoundSaber interface. The manipulation of an interface is conceptually quite different from making empty hand gestures, and arguably the use of the SoundSaber rather than free movement of the hands induced a certain disembodiment in the response.

Following advice from a reviewer of one of the previous papers I conducted a second sound-tracing experiment in which the sounds were designed by a parametric tree of features (shown in Figure 5.2). The sound stimuli thus included sound objects with rising, falling and steady pitch as well as no pitch. Versions of each pitch category with rising, falling and steady brightness envelopes were made. Sound objects with different onset characteristics were also made. The motion responses of the participants were gathered by using two handles equipped with reflective markers. The handles allowed more varied responses than the SoundSaber provided and also decreased the disembodiment of the response caused by the SoundSaber. Arguably the disembodiment could have been reduced even further by putting markers directly on the body. However, this would have increased setup time per participant and also increased the probability of frame drops.

The paper discussed experimental design in sound-tracing experiments, including issues that emerged as a result of increasing the dimensionality of the response data. For instance, participants were provided with two handles but sometimes used only one. Conceptually, there is not much difference between performing a specific action with the left or right hand, a phenomenon

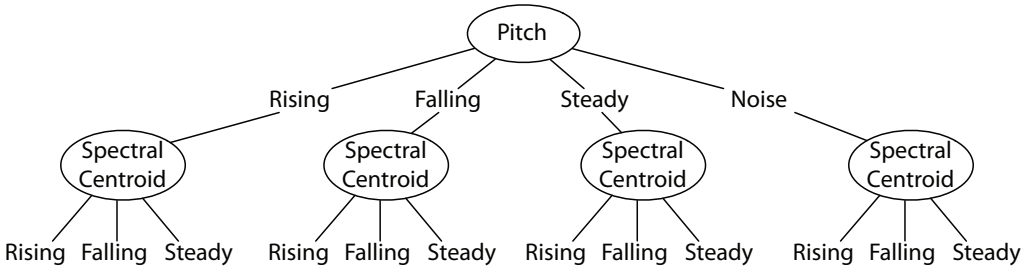


Figure 5.2: 12 of the sound objects in the second sound-tracing experiment were designed using a parametric tree of features. This made it possible to compare sound objects with certain feature envelopes to the inverse case. The figure is adopted from Paper VIII.

called *motor equivalence* in motor theory [Rosenbaum, 2001]. However, the numerical differences in motion data between actions performed with either of the two hands are potentially very large.

The paper introduced mocapgrams as a technique for visualising motion features from a large data set. The mocapgrams provided a qualitative overview of the sound-tracings, which again facilitated formulation of null-hypotheses. The hypotheses were tested for statistical significance with *t*-tests and ANOVAs and several significant results were shown in the paper. However, the *p*-values provided by the tests were not corrected for repeated measurements. When several tests are performed on the same data set the probability of incorrectly discarding the null-hypothesis (Type I error) increases. In other words there is a probability that some of the findings that are presented in this paper are due to chance, and the *p*-values should not be taken "literally". I was made aware of the problem in the first review of Paper VIII (presented below), and the necessary corrections have been included in this paper.

After correcting for repeated measures three main relations between sound and motion were shown to be statistically significant at $\alpha = 0.05$: (1) people's average vertical velocity to sounds with rising pitch or rising spectral centroid was significantly higher than for the sounds with falling pitch or spectral centroid. (2) The onset acceleration for sounds with an impulsive onset was significantly higher than for non-impulsive sounds. (3) The mean acceleration of sound-tracings of pitched sounds was significantly lower than sound-tracings of non-pitched sounds. The results stated in the paper were discussed in relation to findings of other researchers, and in the light of metaphors commonly used in describing sound.

5.2.8 Paper VIII

Analysing Correspondence Between Sound Objects and Body Motion
K. Nymoen, R.I. Godøy, A.R. Jensenius, and J. Torresen.
To appear in *ACM Transactions on Applied Perception*

Abstract: Links between music and body motion can be studied through experiments of so-called *sound-tracing*. One of the main challenges in such research is to develop robust analysis techniques that are able to deal with the multidimensional data that musical sound and body motion present. The paper evaluates four different analysis methods applied to an experiment in which participants moved their hands

following perceptual features of short sound objects. Motion capture data has been analysed and correlated with a set of quantitative sound features using four different methods: (a) a pattern recognition classifier, (b) *t*-tests, (c) Spearman's ρ correlation, and (d) canonical correlation. The paper shows how the analysis methods complement each other, and that applying several analysis techniques to the same data set can broaden the knowledge gained from the experiment.

Through the work described in the previous publications on sound-tracing, I have gained an overview of various methods of analysing this type of data. The analysis carried out in Paper VII only took global features of sonic objects into account, and new perspectives could be obtained by also using time-varying features in the analysis. Paper VIII uses the data set from Paper VII, and applies the pattern recognition and CCA methods from Papers V and VI. Spearman ρ is also included as a measure of one-to-one relationship between individual sound and motion features.

The paper discusses the choice of analysis method and argues that even though various methods are available for this research, no single method is able to capture the full multidimensionality of body motion and sound. The pattern classification method provided quite low classification accuracy, on average 22.5 %. However, the misclassifications were systematic and sound-tracings related to sounds with similar sound features were often confused. When making combined classes, including sound-tracings of sounds with similar sound features, the average classification accuracy was 69 %. This indicated that sound-tracings based on similar sounds were similar.

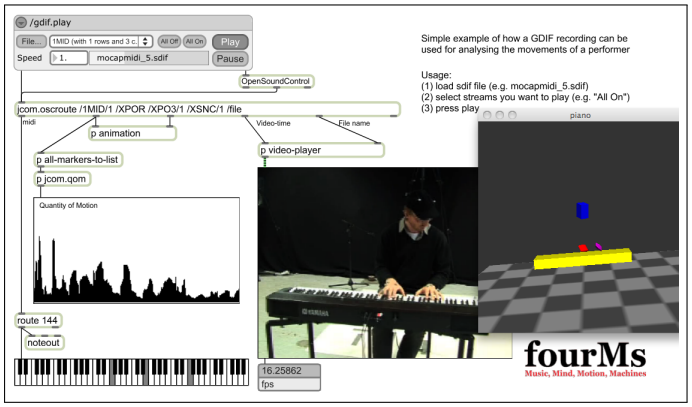
As discussed in Section 5.2.7, the statistical results from Paper VII were adjusted for repeated measurements in this paper. Furthermore, a canonical correlation analysis suggested that the motion features hand distance and vertical position were pertinent in the experiment. Moreover, a Spearman ρ analysis verified the close correspondence between vertical position and pitch as was found in the previous paper, and also indicated a link between spectral centroid and vertical position.

5.3 Developed Software

Part of my research has been in the development of software and tools for storing, streaming, analysis, sonification and visualisation of motion capture data. Some of the tools have been released for public use and will be introduced in this section. The software is open source and can be downloaded from the fourMs website at the University of Oslo.[7]

[7] http://fourms.uio.no/downloads/software/

5.3.1 JamomaGDIF



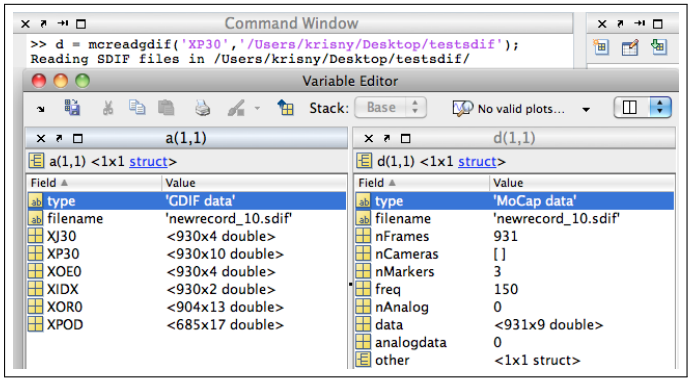
Built with	Max5 Jamoma 0.5.2 FTM 2.5.0
Authors	Kristian Nymoen Alexander Refsum Jensenius
Developed	2008–2012

JamomaGDIF is a toolbox for storing and streaming music-related data. The tools are developed as part of the development of a Gesture Description Interchange Format and uses the Sound Description Interchange Format for storing data (ref. presentation in Section 3.6). The software is developed in Max using Jamoma and FTM.[8] Since details on the implementation of and motivation for the development are covered in Paper I, I shall only provide a brief description of the main functions of the toolbox here.

gdif.record captures data from different data streams which may have different sampling rates and dimensionality. For instance, motion tracking data, MIDI, video and audio can be stored. The data are stored using the **ftm.sdif.write** object for Max, developed by Diemo Schwarz, which structures the data into time-tagged frames [Schnell et al., 2005].

gdif.play streams data from prerecorded files as Open Sound Control formatted text strings. The module creates an **ftm.track** object for each data stream in the SDIF file and each of these can be enabled or disabled individually from a drop-down menu.

5.3.2 GDIF in Matlab



Built with	Matlab 2011b Easdif 3.10.5 MoCap Toolbox 1.3
Author	Kristian Nymoen
Developed	2011-2012

I have developed scripts for importing the files recorded with JamomaGDIF into Matlab for further processing and analysis. In principle the scripts import any file in SDIF format but they are mainly developed for the purpose of working with GDIF data and further development of the scripts will also be aimed at data structured according to the current GDIF specification.[9]

[8] http://ftm.ircam.fr
[9] http://xdif.wiki.ifi.uio.no/Data_types

The scripts are based on importing SDIF-files using Easdif.[10] The data obtained when using Easdif to import SDIF files is difficult to work with directly, especially so for motion data. The GDIF scripts parse the data into a structure which is more intuitive and efficient to work with.

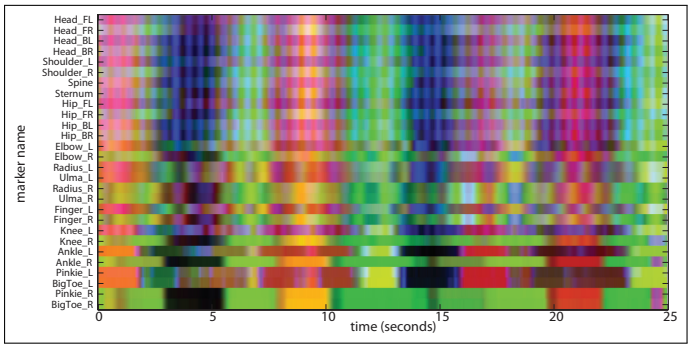
At the time of writing, two scripts are provided:

readgdif.m reads all SDIF files in a single directory, and puts each file into an individual entry of a *struct array*. The individual data streams in each file are assigned an individual field in the struct.

mcreadgdif.m reads a single stream of an SDIF file and parses the data into a data structure compatible with the Jyväskylä MoCap Toolbox.

In future development more Matlab functions for processing and analysing motion data read by the **readgdif.m** function will be provided. Until then the MoCap Toolbox provides good functions for conducting analysis on a single stream of data.

5.3.3 Mocapgrams



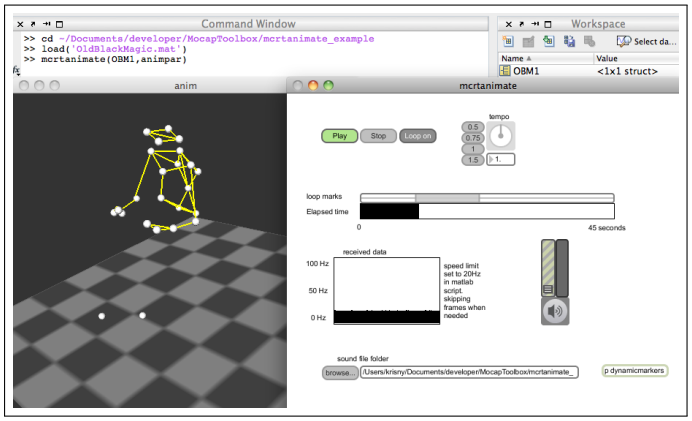
Built with	Matlab 2011b MoCap Toolbox 1.3
Authors	Kristian Nymoen, Alexander Refsum Jensenius Ståle A. Skogstad
Developed	2012

Visualising a large amount of time-varying multidimensional data is difficult — especially if the visualisation is limited to a two-dimensional medium, which is the case in most scientific publications. Section 4.1 introduced *motiongrams* as one possible solution for plotting a large set of motion capture data. I have prepared a script for plotting mocapgrams of motion data in MoCap Toolbox data structures.

The Matlab function **mcmocapgram.m** takes a mocap data structure as input and plots a mocapgram of all the markers in the data structure. An optional argument specifying whether the time axis should display seconds (default) or frames can be given. If marker names are present in the mocap data structure, these are displayed on the Y axis.

[10] http://www.ircam.fr/sdif/download/Easdif

5.3.4 Interactive Animation in the MoCap Toolbox



Built with	Max6 Matlab 2011b MoCap Toolbox 1.3
Author	Kristian Nymoen
Developed	2012

One of the challenging aspects of working with post-processing and analysis of motion capture data is to get an intuitive understanding of the data. In the MoCap Toolbox motion data can be plotted on a time-line or viewed as point-light displays in each frame. Furthermore, multiple point-light displays may be concatenated into a video file to obtain a more intuitive impression of how the motion unfolds over time. If there are audio data related to the mocap recording, this may be added using video-editing software.

While the tools for visualisation, animation and analysis that are implemented in the MoCap Toolbox are good, I missed support for fast and interactive visualisation of mocap data synchronised with related audio files. If able to play back the mocap data with interactive control possibilities such as setting loop marks, scrubbing back and forth, adjusting tempo, zoomimg, rotating and panning, researchers would not have to rely on two-dimensional plots of time-series, or rendering of a video file to see the motion unfold. The software **mcrtanimate** is developed to meet these needs.

A patch for visualising mocap data with an arbitrary number of markers has been implemented in Max, and a function for streaming mocap data from the MoCap Toolbox format has been developed in Matlab. A local UDP[11] network link between Matlab and the Max patch is initiated by calling the **mcrtanimate.m** function in Matlab. One or two arguments are given — first, the reference to the variable that contains the mocap data, and second, an optional animation parameter (animpar) structure. Both are standard data types in the MoCap Toolbox, where the latter contains information on the connections between markers. This causes lines to be drawn in the visualisation patch between the markers specified in the animpar structure, as shown in Figure 5.3.

Through a graphical user interface (GUI) in Max, the user may start and stop playback, set loop marks, adjust the playback tempo and scrub back and forth in the recording. Additionally, the GUI allows zooming, translating and rotating. All in all this provides a simple way to get an intuitive impression of the mocap data while working with it. It also makes it possible rapidly to see the effects of various forms of processing done in the MoCap Toolbox — for instance, to obtain a qualitative impression of how filtering affects the data.

Support for synchronised playback of sound files has also been implemented. For now, the sound file and the motion capture recordings must be of the same length, but in future work I plan to implement more sophisticated methods of playing back synchronised audio that

[11] User Datagram Protocol (UDP) is a standard protocol for network communication.

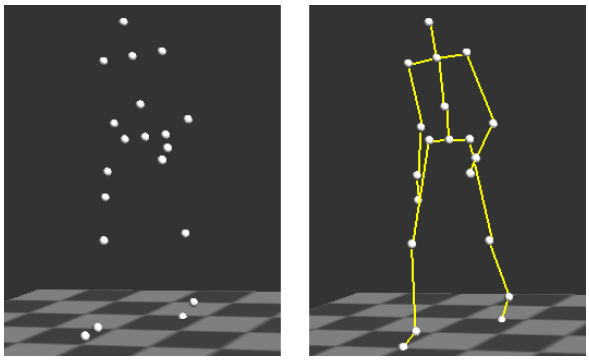
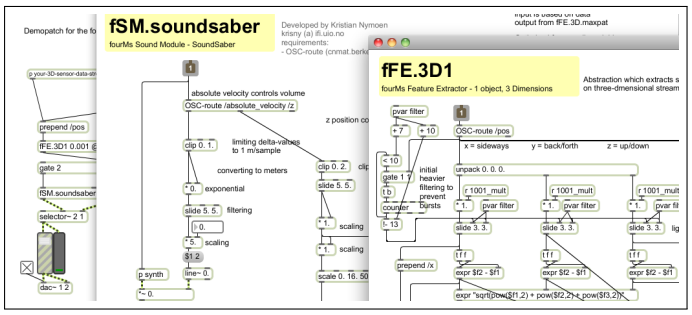


Figure 5.3: The figure shows the difference between an animation with and without the optional animpar arument.

is of a length different from that of the mocap data. Ideally, this would be implemented in sequencer-like software, where audio, video, motion data and other music-related data can be easily manipulated, and also shared among researchers.

The current implementation of **mcrtanimate** relies on the *Instrument Control Toolbox* for Matlab to communicate between Max and Matlab. A future implementation will attempt to avoid the use of the Instrument Control Toolbox as it is expensive and not all universities have access to it. In addition to the Max patch a compiled version for Mac OS is available for users without a valid Max licence.

5.3.5 Motion Tracking Synthesiser



Built with	Max5
Author	Kristian Nymoen
Developed	2008-2012

The development of the SoundSaber involved designing a Max patch to calculate motion features in real time from a stream of position data and another patch for sound synthesis. Together with one more patch for sound sysnthesis, based on looping of audio files, these patches have been released as a generic toolbox for developing musical instruments based on motion tracking data. A schematic layout of how the patches are used in the SoundSaber is offered in Figure 5.4.

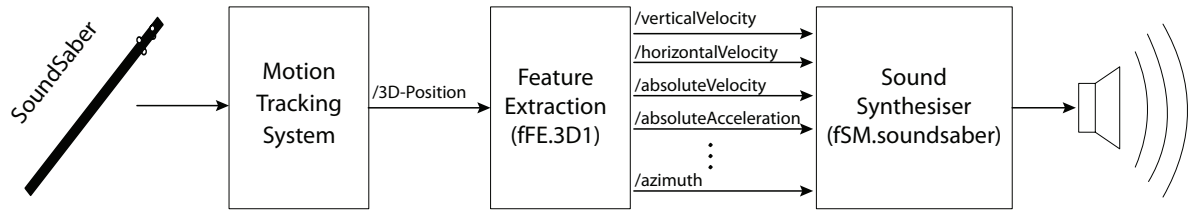


Figure 5.4: The figure shows the data flow from the SoundSaber controller, tracked by the motion tracking system and sent into the **fFE.3D1** patch for feature extraction. The output from the feature extractor is sent to the **fSM.soundsaber** patch.

The Motion Tracking Synthesiser package is implemented as three Max-patches which communicate through OSC messages:

fFE.3D1 calculates various features from a single stream of three-dimensional position data. The features calculated from the position stream include vertical velocity, horizontal velocity, absolute acceleration, distance from origin, angle from origin and more.

fSM.soundsaber and **fSM.loop** are sound modules that work together with the **fFE.3D1** patch. Both take the output of **fFE.3D1** as input, and output an MSP signal which can be processed further or sent directly to the audio interface in Max. **fSM.soundsaber** is based on a pulse train and various delay lines with feedback and is described more in detail in Paper IV. **fSM.loop** loads an audio sample into a buffer and the absolute velocity calculated by **fF3.3D1** is used to control the playback speed of the loop.

5.4 List of Publications

Papers Included in the Thesis

I A Toolbox for Storing and Streaming Music-Related Data.
K. Nymoen and A.R. Jensenius.
In *Proceedings of SMC 2011 8th Sound and Music Computing Conference "Creativity rethinks science"*, pages 427–430, Padova University Press 2011.

II Comparing Inertial and Optical MoCap Technologies for Synthesis Control.
S.A. Skogstad, K. Nymoen, and M.E. Høvin.
In *Proceedings of SMC 2011 8th Sound and Music Computing Conference "Creativity rethinks science"*, pages 421–426, Padova University Press 2011.

III Comparing Motion Data from an iPod Touch to a High-End Optical Infrared Marker-Based Motion Capture System.
K. Nymoen, A. Voldsund, S.A. Skogstad, A.R. Jensenius, and J. Torresen.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 88–91, University of Michigan 2012.

IV SoundSaber — A Motion Capture Instrument.
K. Nymoen, S.A. Skogstad and A.R. Jensenius.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 312–315, University of Oslo 2011.

V Searching for Cross-Individual Relationships between Sound and Movement Features Using an SVM Classifier.
K. Nymoen, K. Glette, S.A. Skogstad, J. Torresen, and A.R. Jensenius.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 259–262, Sydney University of Technology 2010.

VI Analyzing Sound Tracings: A Multimodal Approach to Music Information Retrieval.
K. Nymoen, B. Caramiaux, M. Kozak, and J. Torresen.
In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 39–44, ACM 2011.

VII A Statistical Approach to Analyzing Sound Tracings.
K. Nymoen, J. Torresen, R.I. Godøy, and A.R. Jensenius.
In S. Ystad, M. Aramaki, R. Kronland-Martinet, K. Jensen, and S. Mohanty (eds.) *Speech, Sound and Music Processing: Embracing Research in India*, volume 7172 of Lecture Notes in Computer Science, pages 120–145. Springer, Berlin Heidelberg 2012.

VIII Analysing Correspondence Between Sound Objects and Body Motion.
K. Nymoen, R.I. Godøy, A.R. Jensenius, and J. Torresen.
To appear in *ACM Transactions on Applied Perception*.

Other Papers

- Godøy, R. I., Jensenius, A. R., Voldsund, A., Glette, K., Høvin, M. E., Nymoen, K., Skogstad, S. A., and Torresen, J. (2012). Classifying music-related actions. In *Proceedings of 12th International Conference on Music Perception and Cognition*, pp. 352–357. Thessaloniki, Greece.

- Chandra A, Nymoen, K., Voldsund, A., Jensenius, A.R., Glette, K. and Torresen, J. (2012), Enabling Participants to Play Rhythmic Solos Within a Group via Auctions. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval*, pp. 674–689. Queen Mary University, London.

- Kozak, M., Nymoen, K. and Godøy, R.I. (to appear, 2012). The Effects of Spectral Features of Sound on Gesture Type and Timing. In E. Efthimiou, G. Kouroupetroglou, and S.-E. Fotinea, editors, *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, volume 7206 of Lecture Notes in Computer Science/LNAI. Springer, Berlin Heidelberg.

- Jensenius, A.R., Nymoen, K., Skogstad, S.A. and Voldsund, A. (2012), A Study of the Noise-Level in Two Infrared Marker-Based Motion Capture Systems in *Proceedings of the 9th Sound and Music Computing Conference*, pp. 258–263. Aalborg University, Copenhagen.

- Skogstad, S.A., Nymoen, K., de Quay, Y., and Jensenius, A.R. (2012) Developing the Dance Jockey System for Musical Interaction with the Xsens MVN Suit. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 226–229. University of Michigan, MI.

- Skogstad S.A., Nymoen, K., de Quay, Y. and Jensenius, A.R. (2011). OSC Implementation and Evaluation of the Xsens MVN Suit. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 300–303. University of Oslo.

- Nymoen, K. (2011). Comparing Sound Tracings Performed to Sounds with Different Loudness Envelopes. In *Proceedings of the Joint International Symposium on Computer Music Modeling and Retrieval and Frontiers of Research on Speech and Music*, pp. 225–229. Utkal University, Bhubaneswar, India.

- Godøy, R. I., Jensenius, A. R. and Nymoen, K. (2010). Chunking in Music by Coarticulation. In Acta Acustica united with Acustica. Vol 96(4), pp. 690–700

- Skogstad, S.A., Jensenius, A.R. and Nymoen, K. (2010). Using IR Optical Marker Based Motion Capture for Exploring Musical Interaction. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 407–410. University of Technology, Sydney.

- Nymoen, K. and Jensenius, A.R. (2009). A Discussion of Multidimensional Mapping in Nymophone2. In *Proceedings of the 2009 International Conference on New Interfaces for Musical Expression*, pp. 94–97. Carnegie Mellon University, Pittsburg, PA.

- Jensenius, A.R., K. Nymoen and R.I. Godøy (2008): A Multilayered GDIF-Based Setup for Studying Coarticulation in the Movements of Musicians. In *Proceedings of the International Computer Music Conference*, pp. 743–746. SARC, Belfast.
- Nymoen, K. (2008). A Setup for Synchronizing GDIF Data Using SDIF-files and FTM for Max. COST — Short-Term Scientific Mission, report. McGill University, Montreal.

Chapter 6

Discussion

This chapter discusses and contextualises the work that has been presented in the thesis. The results concerned with evaluation and development of technologies are discussed in Section 6.1. Subsequently, *sound-tracing* as a research method is discussed in Section 6.2, including feature extraction, methods of analysis, techniques of visualisation and the results of my experiments. Furthermore, conclusions and pointers for future extensions of this research are offered.

6.1 Technologies

A little more than one century ago music could only be experienced by way of a live musical performance. Experiencing musical sound and music-related motion implied a direct transfer of auditory and visual sensations from a musician to the perceiver. *Mediation technologies* [Leman, 2008] such as those of broadcasting, recording and reproduction of sound and video, have drastically changed music consumption over the last hundred years, allowing people to enjoy music without the presence of a musical performer. These technologies influence how the music is presented, e.g. the sound quality depends on the sound reproduction system and consequently, experience of music depends on the mediation technologies involved.

Mediation also occurs when technologies are used to quantify sound and body motion in music research. Such quantification enables researchers to reproduce the music-related body motion in analysis software, which facilitates precise comparison of several musical performances. However, like the mediation technologies of music consumption, technologies for tracking, storing and visualising motion are not transparent. In other words the technologies used in research on music-related body motion influence how the motion is presented to the researcher. This does not mean that technologies should be avoided, but researchers should be aware of the distinction between actual body motion and the mediated representation provided by technology.

6.1.1 Evaluation of Motion Tracking Technologies

My evaluation of tracking technologies has shown that they possess distinct strengths and weaknesses, which may also mean that they are suited to different uses. On the one hand, tracking technologies provide a reduced digital representation of the real body motion. For instance,

an ordinary video recording may reduce three-dimensional full-body motion into coloured pixels on a two-dimensional video plane. Even current state-of-the-art technologies for full-body motion tracking are incapable of tracking every moving part of the human body. On the other hand, these technologies provide more detailed results than we would be able to see with the naked eye. Some technologies allow researchers to inspect motion nuances at a sub-millimetre level and, what is more, phenomena that are intrinsically time-dependent can be observed independently of time. Such considerations must be taken into account in order to deduce which aspects of motion the technologies are able to represent and which aspects they leave out.

The papers included in this thesis mainly discuss three tracking technologies. (1) optical infrared marker-based motion capture, through studies of tracking systems from NaturalPoint and Qualisys, (2) the Xsens MVN system, as an example of combining inertial and magnetic tracking technologies with a kinematic model, and (3) an iPod touch, as an example of mobile tracking technology with inertial sensors. Other affordable systems of motion tracking such as ordinary web-cameras, the Wiimote (Nintendo) and the Kinect (Microsoft) have not been studied in the same systematical manner in my research, and are therefore not discussed in the same level of detail. However, having positive, although more exploratory experiences with these technologies I want to pursue analysis of these in the future.

The key findings of the evaluation of tracking technologies include:

Reduction The tracking technologies reduce the actual full body motion into a simpler quantitative representation. Through the use of a kinematic model the motion can be displayed in an avatar. This can be done with optical infrared marker-based technology and with full-body suits like Xsens MVN. The positions of limbs or joints can also be displayed on point light displays (ref. Section 4.1), for instance by using the *mcrtanimate* software presented in Section 5.3.4.

Flexibility Optical infrared marker-based technology allows placement of an arbitrary number of markers, meaning that it provides more flexibility in terms of the tracking configuration than the Xsens. For instance, optical systems can capture finger motion with markers on each finger joint, while the Xsens is limited to certain models like full-body or upper-body with a single sensor on each hand. In contrast, the iPod Touch only reports data from one location. Still, the iPod is flexible in the sense that it may for instance be held to report hand motion or strapped on other parts of the body. All three technologies provide three-dimensional numerical data, albeit with different precision and accuracy, as will be discussed next.

Drift My analyses showed that the data from tracking technologies based on inertial sensors were characterised by a substantial degree of positional drift over a given period. This means that the Xsens system should not be used for analyses where positional precision over time is important. However, the drift is small enough to represent relative positional displacements within a time period of a few seconds. With the iPod the drift of the estimated position data was much larger, and position data estimated from this device should be avoided altogether.

Noise The noise levels of the Xsens position data and the iPod orientation data were higher than their equivalents provided by the optical marker-based tracking systems. However,

the acceleration data calculated by double differentiating the position data from Qualisys and OptiTrack were much noisier than acceleration data from the Xsens and the iPod, and needed filtering to display similar results. The reasons for this are probably two-fold. Firstly, both the iPod and Xsens use accelerometer data which do not require differentiation to estimate acceleration. Secondly, there might be stronger filtering of data inside the Xsens software and the iPod sensor fusion algorithm than what is used in Qualisys and OptiTrack.

Stability Optical tracking technologies require line-of-sight, meaning that data are lost when a marker is occluded or moved out of the predefined tracking space. Furthermore, spikes in the data will occur when the set of cameras that see a marker changes. Optical infrared marker-based technologies are also sensitive to "light-pollution", meaning that foreign light sources or reflective material in the tracking space might cause tracking error. Xsens and the iPod have no such restrictions. According to our results the range limit of the Bluetooth connection between the Xsens suit and the host computer is more than 50 metres.

Intrusiveness The different systems may be more or less "intrusive", both on the body and on a stage if used in performance. The least intrusive in both cases is the iPod, as most people carry such mobile devices anyway. Also, the iPod is a fully integrated system with no need for cables or external hardware. The Xsens suit contains two big transmitters with batteries, and lots of cables and sensors that could feel intrusive to the person wearing it, especially so for a musician where the cables and sensors can interfere with the playing technique. Moreover, the suit is tiresome to wear for several hours. In a stage setup, however, the suit can be worn under clothes in order not to disturb the visual appearance of performers [Skogstad et al., 2012c]. The reflective markers used by the optical systems are less intrusive to the person wearing them as they can be quite small and attached directly to clothes or skin. However, in a performance setup the cameras must be positioned in a manner that gives optimal tracking performance, meaning that they will usually dominate the performance area, which might be disturbing to the audience.

Availability Finally, the cost and availability of the systems differ greatly. Although more affordable solutions exist, the technologies used in state-of-the-art motion tracking systems are more expensive than most research groups can afford without dedicated grants. The iPod Touch and other devices such as mobile phones or game controllers (e.g. Nintendo Wii or Microsoft Kinect) are readily available. Since many people own such devices these days, it is now possible to develop applications that would analyse the music-related motion of users in their homes and elsewhere. This could enable research on music and motion involving a very large number of people.

6.1.2 Software for Storing and Streaming Music-Related Motion Data

The software developed and presented in this thesis is provided to assist research on music-related motion. Consequently this software too is a mediator between the actual motion and the data used by the researcher. As argued above, the data representations of the motion are simplifications of the actual motion, yet these representations offer more detail than the researcher

would be able to infer by just observing the motion as it occurred. As with tracking technologies, users of software for storing and streaming music-related motion data should be aware of how mediation influences data. I offer three observations on the developed software:

Data synchronisation retains multimodality. The software that I have developed for working with music-related motion depends on the data that is provided by the tracking systems. This means that data stored using JamomaGDIF will have the same strengths and weaknesses as were found with the tracking technologies above. However, by combining video, audio, MIDI, motion tracking and other sensors in synchronised recordings the multimodal nature of the musical performance is somewhat retained. Thus the JamomaGDIF software may provide researchers with a collection of data that is closer to the actual musical performance than what a motion tracking system can provide by itself.

Offline streaming disregards interactivity. One of the suggested use scenarios for JamomaGDIF is prototyping of new musical interfaces. Since both motion and sound are phenomena unfolding in time, it is usually necessary to have the stream of motion data available when programming to test the interface in real time. For instance, the data from a new interface containing a motion tracking system and a force-sensing resistor can be recorded in JamomaGDIF, and subsequently the two data streams can be streamed from the recorded file while the developer experiments with the mappings in the instrument prototype. However, if JamomaGDIF is used in this way, one important aspect of the musical interface is easily neglected, namely *interactivity*. In most musical interfaces there is a feedback loop between motion and sound and users will adjust their motion according to the sonic feedback. Thus if pre-recorded data are used in the development of interfaces, it is important to test the interface regularly with realtime interaction as part of the development process.

Feature extraction may increase intuition. JamomaGDIF does not perform any processing on the recorded data, meaning that data that are streamed from the software are the same as was provided by the tracking systems. The developed *Motion Tracking Synthesiser*, described in Section 5.3.5, provides features that may be closer to an intuitive interpretation of motion than the raw data from the tracking system. For instance, the absolute velocity feature is closely related to the kinetic energy, which again is a result of mechanical work carried out by the person tracked.

6.2 The Sound-Tracing Approach

A large part of the research presented in this thesis consists of studies of bodily responses to musical sound. The sound-tracing research approach lies between the extremes of being *controllable* and *ecologically valid* as will be discussed below.

In research experiments in auditory psychophysics, such as presented by Bregman and Ahad [1996], the sound stimuli are typically strictly controlled. By using, for instance, sine tones and noise bursts researchers can easily control sound features like frequency and duration with high precision. Controlled stimuli are desired in order to draw inferences from experimental results with a high degree of confidence. However, the use of such stimuli in music research has been

criticised for lacking *ecological validity* [Rumsey, 2002, De Gelder and Bertelson, 2003, Jones, 2010]. The other extreme for sound stimuli in music research is the use of sound from entire musical works. In such studies some musical features, like *tempo*, can usually be controlled quite well. However, a precise description of the entire soundscape is difficult to make. A large number of time-varying signal features can be calculated, in addition to cultural and social aspects that may influence experience of musical sound. This multidimensionality reduces the controllability of the sound stimuli.

An equally wide range of possibilities exist for bodily responses. If full body motion is tracked, the angles of each joint, with corresponding limb positions, velocities and acceleration values, present a large number of dimensions and thus a response scenario with low degree of control. The other, highly controlled, extreme would be to restrict the bodily response to moving a slider or pressing a button. Again, the notion of ecological validity should be considered: are presses of a button suitable objects of investigation if the objective is to study links between musical sound and body motion?

I believe that different degrees of controllability and ecological validity should be allowed in experiments on music and motion. My aim is not to disapprove of any type of stimulus or response, but to emphasise that it is important to be aware of this trade-off. The sound-tracing approach focusing on short sound objects and simple free-air motion, lies between the extremes of controllability and ecological validity. For instance, the sound stimuli are controllable because of their short durations and controlled sound parameters, and they are ecologically valid in music research, since they correspond to sonic objects which are potential building blocks of a piece of music (ref. Section 2.2.1). Musical sound may be seen as a sequence of sound objects, and the body motion of a performing musician as a sequence of *sound-producing*, *communicative*, and *sound-facilitating* actions [Jensenius et al., 2010]. My experiments have used the manipulation of two handles or a rod as motion response, which is similar to *sound-producing actions* that typically involve manipulating a musical instrument with the hands. Because a sonic object often is the result of a sound-producing action, and because both of these correspond to the chunk timescale level (ref. Section 2.3.2), this particular trade-off between controllability and ecological validity is an interesting object of study.

6.2.1 Motion Features

Discussion of controllability and ecological validity in sound-tracing experiments is also relevant to the process of extracting and selecting motion features. As humans we use other descriptions of motion than the raw data provided by a motion capture system. Metaphors and verbal descriptions such as ‘falling’, ‘expanding’, ‘impulsive’, ‘smooth’ or ‘shaking’ are more common in daily language than the use of precise positional measures. Thus one of the important steps in the analysis of motion is converting the raw position data into features that describe the body motion in a way that is closer to human descriptions. Still, the use of numerical features rather than verbal metaphors retains some degree of *controllability* in sound-tracing experiments.

A number of features have been used in previous research on music-related motion, including first and second order derivatives [Schoonderwaldt et al., 2006, Luck et al., 2010b], energy measurements of different parts of the body [Toiviainen et al., 2010], quantity of motion and

contraction index [Camurri et al., 2003], timing perspectives like inter-onset-intervals or bodily resonances [Dahl, 2000, van Noorden, 2010], and relations between limbs [Müller and Röder, 2006, Klapp and Jagacinski, 2011]. More features than those used in this thesis could have been extracted; however, the use of more features also makes it more difficult to control the results. Furthermore, as discussed in Section 4.3, the use of many features may result in overfitting of the canonical correlation analysis. Positional features and their derivatives were shown in a sound-tracing experiment by Caramiaux et al. [2010] to be the most pertinent motion features. Similar features have been used in all the analyses presented in this thesis.

6.2.2 Methods of Analysis

As mentioned in Section 5.1.3, limited research on sound-tracing has previously been conducted. Consequently, methods of analysis for such research are sparse and no standardised implementations of analysis techniques are available in current software to analyse music-related motion (e.g. the MoCap Toolbox or the Musical Gestures Toolbox). For this reason I have implemented my own set of tools in Matlab, including scripts for feature extraction, normalisation, visualisation and the various analysis methods presented in this thesis. The presented analysis techniques show strengths and weaknesses and should preferably be used in conjunction. The main results of my methodical research include:

Visualising data. Visualisation of data from sound-tracing experiments is challenging due to the potentially large number of recordings in the experiment and the large number of feature dimensions for each sound-tracing. The *mcmocapgram* and *mcrtanimate* software tools introduced in Sections 5.3.3 and 5.3.4 can assist research in this area by creating generic visualisations.

Recognition of sounds from motion data can be automated. My experiments have shown that pattern classification, through Support Vector Machines, can be applied to a set of sound-tracings to reveal which sound stimuli induced similar motion responses. Such an overview may provide some basic indications of which sound and motion features were pertinent to a particular classification. However, a detailed knowledge of the relation between sound and motion features is not provided by pattern recognition analysis. The satisfying performance of the classifier should motivate further research on solutions for *query-by-gesture* in music databases [Leman, 2008, Godøy and Jensenius, 2009, Caramiaux et al., 2011].

Correlation of time-varying features. The Spearman ρ correlation coefficient has been suggested as a better method than the more commonly applied Pearson correlation for estimating the relation between two music-related time-series [Schubert, 2002]. However, due to the problems with serial correlation, the results found by this method cannot be tested with traditional statistical tests. In my experiments the results from correlations between sound and motion features were useful when applied in conjunction with other methods of analysis of sound-tracings.

Complex relations between sound and motion. Spearman ρ can provide a measure of the correlation between one sound feature and one motion feature. However, more complex

correlations may exist as is frequently seen in musical instruments. The connections between sound-producing actions and instrumental sound are rarely one-to-one mappings [Kvifte, 1989]. Rather, one-to-many, many-to-one, or many-to-many mappings exist, and the multidimensional links between control action and sound features can make a musical instrument more intriguing to play [Hunt et al., 2002]. Canonical correlation may be applied to assess how a *set* of motion features correlates with a *set* of sound features. However, as with other coefficients from correlation of time-series, the results cannot be tested for statistical significance. Furthermore, the multidimensional correlations shown by CCA may be difficult to interpret due to their complexity.

Global versus time-varying features. Global features of sound-tracings may be applicable for hypothesis testing and assessment of statistical significance. This means that a value such as the average velocity for one group of sound-tracings may be compared with another group of sound-tracings. By using short sound objects and action responses that are in themselves chunks, such global features can be extracted quite easily. However, if longer segments of musical sound and bodily responses are used, it is more difficult to extract global features that characterise the overall feature trajectories. Hence there is a need to develop robust methods of analysis that are able to take time-varying features into account.

6.2.3 Empirical Results

The research presented in this thesis to a large extent focuses on technologies for and methods of studying music-related body motion. However, some cognitive results have also been obtained through the two sound-tracing experiments. Some of the results were found in only a few of the analyses, while one result was particularly evident in all of my papers on sound-tracing, namely the correlation between pitch and verticality.

Pitch and verticality The correlation between pitch and verticality has already been shown by several researchers. This relation is common both in everyday speech, as can be seen by the expressions *high* and *low* pitch, and in Western musical notation where the vertical position of a note on the score indicates the pitch of a tone. Huron et al. [2009] showed that the pitch of sung tones correlates with the vertical position of the eyebrows, and several other experiments have indicated that people link pitch to verticality [Eitan and Granot, 2004, 2006, Eitan and Timmers, 2010, Kohn and Eitan, 2012].

Arnie Cox [1999] explained the correspondence of pitch and verticality through a metaphoric understanding of *up* as *more*. The 'more', in this context, is of the increased effort that is necessary to produce higher pitched vocal sounds. However, this metaphor has been criticised for only being applicable in certain circumstances. Eitan and Timmers [2010] e.g. have shown that if pitch is mapped to 'mass', 'size' or 'quantity', it may be more appropriate to understand *up* as *less*. The latter claim contradicts a finding of Walker [2000], where experiment participants linked an increase in pitch to increase in all of the observed dimensions, including size, temperature, pressure and velocity.

The use of free air body motion may have made the link between spatial verticality and pitch apparent to participants in my experiments. Metaphors such as those describing

pitch in terms of temperature, mass or quantity, were not possible. Still, my research has not focused on metaphors other than those of body motion, and my results show that for music-related body motion there seems to be a strong link between pitch and spatial verticality.

Spectral centroid and verticality. The spectral centroid envelopes of the sound stimuli were also found to be related to vertical position. This is not very surprising, as both the pitch and the spectral centroid of musical sound correspond to frequencies within the audible frequency range. Also, in most musical instruments, pitch and spectral centroid are not independent sound features [Kvifte, 1989, Chowning, 1999]. In my sound stimuli, however, the applied bandpass filter allowed controlling spectral centroid quite independently of pitch.

For sound stimuli without a distinct pitch the correlation between vertical position and spectral centroid was stronger than for stimuli with a stable pitch. Furthermore, when pitch and spectral centroid moved in opposite directions, most participants moved their hands up for rising, and down for falling pitch. This indicates that the presence of a perceivable pitch influences, or even overrides the link between spectral centroid and vertical position.

Dynamic envelope and acceleration. For sound objects classified as *impulsive*, meaning sound objects with a quickly increasing and slowly decreasing dynamic envelope, most participants responded with correspondingly impulsive actions, seen by acceleration peaks at the time of the sound onsets. Similar peaks were not found for *non-impulsive* sound objects. This corresponds well with Godøy's [2003] suggestion that musical sound is perceived and understood as mental visualisations of sound-producing actions.

Noise and acceleration. Paper VII showed a significantly higher mean acceleration for motion performed to sound stimuli without a perceivable pitch than with sound stimuli with a perceivable pitch. It seems that the sounds with a perceivable pitch provided the participants with something to 'hold on to', meaning that the motion trajectories of these sound-tracings were smooth. Sounds based on noise, however, did not provide the same reference and caused irregular and jerky motion.

Cultural background. It should be mentioned that none of the experiments took cultural background of the participants into account. The majority of the participants in all experiments were of Western background. Arguably, the results of the experiment might be tightly linked with Western culture, and if the same experiments were conducted in other cultures, different results might have been obtained, as suggested by Ashley [2004].

Adjustment of statistical results. As presented in section 5.2.7 the results from Paper VII were adjusted in Paper VIII to correct for repeated measurements of the data. Consequently, several of the results were redefined as non-significant. Arguably, this adjustment may have resulted in type II errors, meaning that significant results may have been defined as non-significant. Several methods may be applied to correct for this *experiment-wise error rate*, and different methods should be applied depending on whether the tests are

independent of, or dependent on each other. In my experiment certain tests were independent; for instance testing for differences in onset acceleration for impulsive versus non-impulsive sounds is independent of testing for differences in vertical motion for different pitch envelopes. Other tests were dependent on each other, for instance testing differences in vertical position for all subjects and then testing the same for only expert subjects. Nevertheless, in this case possible type II errors are more acceptable than incorrectly rejecting the null-hypotheses. As argued in Paper VIII, new tests should be applied in future experiments to learn whether or not these results were significant.

6.3 Conclusion

This PhD project has been concerned with development and evaluation of methods of and technologies for studying links between musical sound and music-related body motion. This includes research on technologies for tracking body motion and development of a new solution for storing multidimensional music-related data. Participants' motion responses to short sound objects have been studied in two experiments on sound-tracing and the usefulness of four methods of analysis of such experiments has been evaluated. The following points conclude the research:

- Various motion tracking technologies provide data with different limitations. The properties of the technologies determine their proper use in music research and performance. While the quality of the tracking data to some extent correlates with the cost of systems, more affordable solutions may be adequate as long as the researcher pays attention to the strengths and weaknesses of the technology. This was shown by the accurate orientation data of the iPod Touch, and the SoundSaber Wiimote implementation presented in this thesis.

- The findings of the sound-tracing studies are in accordance with previous results of research into use of metaphors to describe sound. The close association of pitch and verticality coincides with a ubiquitous understanding of pitch in Western culture. An association was also shown to exist between impulsive sounds and high onset acceleration in sound-tracings, and between noisiness and overall acceleration.

- My evaluation of methods of analysis for sound-tracing studies has shown that a pattern recognition classifier may be trained to recognise sound objects based on the actions that people associate with the sounds. Although perfect classification results were not obtained, the classifier grouped together actions that were based on similar sounds. This shows that similar sounds entailed similar motion responses in my sound-tracing experiments.

- To analyse how the time-varying features of sound relate to time-varying features of motion responses correlation analysis may be applied. Spearman ρ compares one feature with another, and canonical correlation compares a set of features with those of another set. However, these correlation analyses are not suited for statistical testing and consequently, there is still a need to develop new methods of analysis that are able to handle time-varying multidimensional data from sound-tracing experiments.

- The JamomaGDIF software presented in this thesis is a step towards flexible handling of multidimensional music-related data, and may assist researchers in their work on music-related body motion. The Sound Description Interchange Format and Open Sound Control are promising solutions, and future development of the Gesture Description Interchange Format should explore these protocols further.

6.4 Future Work

Taking on such a project has been extensive in both content and form. While I have been able to answer some questions, many still remain. As mentioned previously, the evaluated tracking technologies do not cover all current available technologies, and further research should also evaluate the potential of affordable tracking solutions like the Kinect, Wiimotes, and webcameras.

In my experiments, a main focus was been kept on short sound objects as stimuli and action responses of equivalent length, reduced to free-air manipulation of one or two rigid objects. The research should eventually be continued by expanding sound-tracing experiments to incorporate full-body motion as response data, and using longer excerpts of musical sound as stimuli. Müller's features for full body motion data (ref. Section 3.5.3) should be explored for this purpose since the features are robust and computationally efficient. There is a need to develop robust methods of segmentation of continuous motion data, such that longer recordings of continuous motion can be analysed as a sequence of concatenated motion chunks. More methods of analysing time-varying motion features and sound features should be explored. Machine learning techniques such as Dynamic Time Warping, Hidden Markov Models or Hierarchical Temporal Memory, which are capable of learning time-varying patterns could provide solutions to this problem. The results of the sound-tracing experiments should be verified by an 'opposing' experiment, having participants watch animations of the recorded motion, and match the sound file to motion.

The tools for working with music-related data should be developed further. A wide range of existing audio sequencer software (Cubase, Logic, ProTools) and video editing software (Final Cut, iMovie, Windows Movie Maker), enables interactive editing of sound and video. A similar system should be developed for music-related data including audio, sensor data, MIDI, video, annotations, scores and more. JamomaGDIF, with corresponding Matlab scripts and the developed interactive animation tool for the MoCap Toolbox are small steps along the way to making a multimodal sequencer software for music research. Such a sequencer should ideally allow editing synchronised audio, motion and video data, with possibilities for e.g. gap-filling, filtering and scaling of motion data.

Bibliography

C. Alain and S. Arnott. Selectively attending to auditory objects. *Frontiers in Bioscience*, 5: 202–212, 2000.

E. Altenmüller, M. Wiesendanger, and J. Kesselring, editors. *Music, Motor Control and the Brain*. Oxford University Press, Oxford, New York, 2006.

R. Ashley. Musical pitch space across modalities: Spatial and other mappings through language and culture. In *Proceedings of the 8th International Conference on Music Perception and Cognition*, pages 64–71, Evenston, IL., USA, 2004.

A. Baltazar, C. Guedes, B. Pennycook, and F. Gouyon. A real-time human body skeletonization algorithm for max/msp/jitter. In *Proceedings of the International Computer Music Conference*, pages 96–99, New York, 2010.

E. Bekkedal. Music Kinection: Sound and Motion in Interactive Systems. Master's thesis, University of Oslo (to appear), 2012.

F. Bevilacqua, J. Ridenour, and D. J. Cuccia. 3D motion capture data: motion analysis and mapping to music. In *Proceedings of the Workshop/Symposium on Sensing and Input for Media-centric Systems*, University of California, Santa Barbara, 2002.

G. Bishop, G. Welch, and B. Allen. Tracking: Beyond 15 minutes of thought. *SIGGRAPH 2001 Courses*, 2001. URL http://www.cs.unc.edu/~tracker/ref/s2001/tracker/ (Accessed June 24, 2012).

P. Boersma and D. Weenink. Praat: doing phonetics by computer (software). version 5.3.19, 2012. URL http://www.praat.org/ (Accessed June 29, 2012).

M. Borga. Canonical correlation: a tutorial. Technical report, Linköping University, 2001.

A. Bouënard, M. M. Wanderley, and S. Gibet. Analysis of Percussion Grip for Physically Based Character Animation. In *Proceedings of the International Conference on Enactive Interfaces*, pages 22–27, Pisa, 2008.

A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA, 1990.

A. S. Bregman and P. Ahad. Demonstrations of auditory scene analysis: The perceptual organization of sound. Audio CD and booklet, Distributed by MIT Press, 1996.

J. Bresson and M. Schumacher. Representation and interchange of sound spatialization data for compositional applications. In *Proceedings of the International Computer Music Conference*, pages 83–87, Huddersfield, 2011.

I. Bukvic, T. Martin, E. Standley, and M. Matthews. Introducing L2Ork: Linux Laptop Orchestra. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 170–173, Sydney, 2010.

B. Burger, M. R. Thompson, G. Luck, S. Saarikallio, and P. Toiviainen. Music moves us: Beat-related musical features influence regularity of music-induced movement. In *Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) and 8th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM)*, pages 183–187, Thessaloniki, 2012.

C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

C. Cadoz and M. M. Wanderley. Gesture — Music. In M. M. Wanderley and M. Battier, editors, *Trends in Gestural Control of Music*, pages 71–94. Ircam—Centre Pompidou, Paris, France, 2000.

A. Camurri and T. B. Moeslund. Visual gesture recognition. from motion tracking to expressive gesture. In R. I. Godøy and M. Leman, editors, *Musical Gestures: Sound, Movement, and Meaning*, pages 238–263. Routledge, 2010.

A. Camurri, S. Hashimoto, M. Ricchetti, A. Ricci, K. Suzuki, R. Trocca, and G. Volpe. EyesWeb: Toward gesture and affect recognition in interactive dance and music systems. *Computer Music Journal*, 24(1):57–69, 2000.

A. Camurri, I. Lagerlöf, and G. Volpe. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1–2):213–225, 2003.

A. Camurri, B. Mazzarino, and G. Volpe. Analysis of expressive gesture: The EyesWeb expressive gesture processing library. In A. Camurri and G. Volpe, editors, *Gesture-based Communication in Human-Computer Interaction*, volume 2915 of *LNAI*, pages 460–467. Springer, Berlin Heidelberg, 2004.

A. Camurri, G. Volpe, G. De Poli, and M. Leman. Communicating expressiveness and affect in multimodal interactive systems. *Multimedia, IEEE*, 12(1):43 – 53, 2005.

C. Cannam, C. Landone, and M. Sandler. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of ACM Multimedia*, pages 1467–1468, Firenze, Italy, October 2010.

B. Caramiaux, F. Bevilacqua, and N. Schnell. Towards a gesture-sound cross-modal analysis. In S. Kopp and I. Wachsmuth, editors, *Gesture in Embodied Communication and Human-Computer Interaction*, volume 5934 of *Lecture Notes in Computer Science*, pages 158–170. Springer, Berlin Heidelberg, 2010.

B. Caramiaux, F. Bevilacqua, and N. Schnell. Sound selection by gestures. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 329–330, Oslo, 2011.

A. Chandra, K. Nymoen, A. Voldsund, A. R. Jensenius, K. Glette, and J. Torresen. Enabling participants to play rhythmic solos within a group via auctions. In *Proceedings of the 9th International Symposium on Computer Music Modelling and Retrieval*, pages 674–689, London, 2012.

C. Chang and C. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

J. Chowning. Perceptual fusion and auditory perspective. In P. R. Cook, editor, *Music, Cognition, and Computerized Sound*, pages 261–275. MIT Press, Cambridge, MA, USA, 1999.

M. Ciglar. An Ultrasound Based Instrument Generating Audible and Tactile Sound. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 19–22, Sydney, 2010.

E. F. Clarke. *Ways of listening: An ecological approach to the perception of musical meaning*. Oxford University Press, New York, 2005.

A. W. Cox. *The Metaphoric Logic of Musical Motion and Space*. PhD thesis, University of Oregon, 1999.

A. W. Cox. Hearing, feeling, grasping gestures. In A. Gritten and E. King, editors, *Music and gesture*, pages 45–60. Ashgate, Aldershot, UK, 2006.

S. Dahl. The playing of an accent – preliminary observations from temporal and kinematic analysis of percussionists. *Journal of New Music Research*, 29(3):225–233, 2000.

S. Dahl. Playing the accent-comparing striking velocity and timing in an ostinato rhythm performed by four drummers. *Acta Acustica united with Acustica*, 90(4):762–776, 2004.

R. B. Dannenberg, S. Cavaco, E. Ang, I. Avramovic, B. Aygun, J. Baek, E. Barndollar, D. Duterte, J. Grafton, R. Hunter, C. Jackson, U. Kurokawa, D. Makuck, T. Mierzejewski, M. Rivera, D. Torres, and A. Y. and. The carnegie mellon laptop orchestra. In *Proceedings of the International Computer Music Conference*, pages 340–343, Copenhagen, 2007.

B. De Gelder and P. Bertelson. Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences*, 7(10):460–467, 2003.

S. de Laubier. The meta-instrument. *Computer Music Journal*, 22(1):25–29, 1998.

S. de Laubier and V. Goudard. Meta-instrument 3: a look over 17 years of practice. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 288–291, Paris, France, 2006.

Y. de Quay, S. Skogstad, and A. Jensenius. Dance jockey: Performing electronic music by dancing. *Leonardo Music Journal*, pages 11–12, 2011.

C. Dobrian and F. Bevilacqua. Gestural control of music: using the vicon 8 motion capture system. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 161–163, Montreal, 2003.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2000.

G. Eckel and D. Pirro. On artistic research in the context of the project embodied generative music. In *Proceedings of the International Computer Music Conference*, pages 541–544, Montreal, 2009.

Z. Eitan and R. Granot. Musical parameters and images of motion. In *Proceedings of the Conference on Interdisciplinary Musicology*, pages 15–18, Graz, 2004.

Z. Eitan and R. Y. Granot. How music moves: Musical parameters and listeners' images of motion. *Music Perception*, 23(3):pp. 221–248, 2006.

Z. Eitan and R. Timmers. Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition*, 114(3):405 – 422, 2010.

M. R. Every. *Separation of musical sources and structure from single-channel polyphonic recordings*. PhD thesis, University of York, 2006.

G. Fant. Speech analysis and synthesis. Technical report, Royal Institute of Technology, Stockholm, 1961.

S. Fels and G. Hinton. Glove-talk: A neural network interface between a data-glove and a speech synthesizer. *IEEE Trans. Neural Networks*, 4(1):2–8, 1993.

R. Fischman. Back to the parlour. *Sonic Ideas*, 3(2):53–66, 2011.

A. Freed, D. McCutchen, A. Schmeder, A. Skriver, D. Hansen, W. Burleson, C. Nørgaard, and A. Mesker. Musical applications and design techniques for the gametrak tethered spatial position controller. In *Proceedings of the 6th Sound and Music Computing Conference*, pages 189–194, Porto, 2009.

B. Freedman, A. Shpunt, M. Machline, and Y. Arieli. Depth mapping using projected patterns. Patent Application, 2010. US 12522171.

V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti. Action recognition in the premotor cortex. *Brain*, 119(2):593–609, 1996.

J. J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1979.

N. Gillian, P. Coletta, B. Mazzarino, and M. Ortiz. Techniques for data acquisition and multimodal analysis of emap signals. EU FP7 ICT FET SIEMPRE, Project No. 250026, Deliverable report 3.1, May 2011.

R. I. Godøy. Motor-mimetic music cognition. *Leonardo Music Journal*, 36(4):317–319, 2003.

R. I. Godøy. Gestural imagery in the service og musical imagery. In A. Camurri and G. Volpe, editors, *Gesture-Based Communication in Human-Computer Interaction: 5th International Gesture Workshop, GW 2003, Genova, Italy, April 15-17, 2003, Selected Revised Papers, LNAI 2915*, pages 55–62. Springer, Berlin Heidelberg, 2004.

R. I. Godøy. Gestural-sonorous objects: embodied extensions of Schaeffer's conceptual apparatus. *Organised Sound*, 11(02):149–157, 2006.

R. I. Godøy. Gestural affordances of musical sound. In R. I. Godøy and M. Leman, editors, *Musical Gestures: Sound, Movement, and Meaning*, chapter 5, pages 103–125. Routledge, New York, 2010.

R. I. Godøy. Sonic feature timescales and music-related actions. In *Proceedings of Forum Acusticum*, pages 609–613, Aalborg, 2011. European Acoustics Association.

R. I. Godøy and A. R. Jensenius. Body movement in music information retrieval. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 45–50, Kobe, Japan, 2009.

R. I. Godøy and M. Leman, editors. *Musical Gestures: Sound, Movement, and Meaning*. Routledge, New York, 2010.

R. I. Godøy, E. Haga, and A. R. Jensenius. Playing "air instruments": Mimicry of sound-producing gestures by novices and experts. In S. Gibet, N. Courty, and J.-F. Kamp, editors, *International Gesture Workshop. Revised Selected Papers*, volume 3881/2006, pages 256–267. Springer, Berlin Heidelberg, 2006a.

R. I. Godøy, E. Haga, and A. R. Jensenius. Exploring music-related gestures by sound-tracing. a preliminary study. In *2nd ConGAS International Symposium on Gesture Interfaces for Multimedia Systems*, Leeds, UK, 2006b.

R. I. Godøy, A. R. Jensenius, and K. Nymoen. Chunking in music by coarticulation. *Acta Acustica united with Acustica*, 96(4):690–700, 2010.

R. I. Godøy, A. R. Jensenius, A. Voldsund, K. Glette, M. E. Høvin, K. Nymoen, S. A. Skogstad, and J. Torresen. Classifying music-related actions. In *Proceedings of 12th International Conference on Music Perception and Cognition*, pages 352–357, Thessaloniki, 2012.

J. M. Grey. Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5):1270–1277, 1977.

A. Gritten and E. King, editors. *Music and gesture*. Ashgate, Aldershot, UK, 2006.

A. Gritten and E. King, editors. *New perspectives on music and gesture*. Ashgate, Aldershot, UK, 2011.

F. Grond, T. Hermann, V. Verfaille, and M. Wanderley. Methods for effective sonification of clarinetists' ancillary gestures. In S. Kopp and I. Wachsmuth, editors, *Gesture in Embodied Communication and Human-Computer Interaction*, volume 5934 of *Lecture Notes in Computer Science*, pages 171–181. Springer, Berlin Heidelberg, 2010.

C. Guedes. Extracting musically-relevant rhythmic information from dance movement by applying pitch-tracking techniques to a video signal. In *Proceedings of the Sound and Music Computing Conference SMC06*, pages 25–33, Marseille, 2006.

J. Hagedorn, S. Satterfield, J. Kelso, W. Austin, J. Terrill, and A. Peskin. Correction of location and orientation errors in electromagnetic motion tracking. *Presence: Teleoperators and Virtual Environments*, 16(4):352–366, 2007.

M. Halle and K. Stevens. Speech recognition: A model and a program for research. *IRE Transactions on Information Theory*, 8(2):155 –159, 1962.

B. Haslinger, P. Erhard, E. Altenmüller, U. Schroeder, H. Boecker, and A. Ceballos-Baumann. Transmodal sensorimotor networks during action observation in professional pianists. *Journal of cognitive neuroscience*, 17(2):282–293, 2005.

J. Haueisen and T. R. Knösche. Involuntary motor activity in pianists evoked by music perception. *Journal of cognitive neuroscience*, 13(6):786–792, 2001.

K. Hayafuchi and K. Suzuki. Musicglove: A wearable musical controller for massive media library. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 241–244, Genova, 2008.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

A. Hunt, M. M. Wanderley, and M. Paradis. The importance of parameter mapping in electronic instrument design. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 1–6, Singapore, 2002.

D. Huron. The new empiricism: Systematic musicology in a postmodern age. *The 1999 Ernest Bloch Lectures*, 1999. URL http://www.musicog.ohio-state.edu/Music220/Bloch.lectures/3.Methodology.html (Accessed June 6, 2012).

D. Huron, S. Dahl, and R. Johnson. Facial expression and vocal pitch height: Evidence of an intermodal association. *Empirical Musicology Review*, 4(3):93–100, 2009.

E. Husserl. *The Phenomenology of Internal Time Consciousness*. (trans. J.S. Churchill.) Indiana University Press, Bloomington, 1964.

G. Iddan and G. Yahav. 3d imaging in the studio (and elsewhere). In B. D. Corner, J. H. Nurre, and R. P. Pargas, editors, *Three-Dimensional Image Capture and Applications IV*, volume 4298 of *Proceedings of SPIE*, pages 48–55, 2001.

J. Impett. A meta-trumpet(er). In *Proceedings of the International Computer Music Conference*, pages 147–150, Aarhus, 1994.

H. Ip, K. Law, and B. Kwong. Cyber composer: Hand gesture-driven intelligent music composition and generation. In *Proceedings of the IEEE 11th International Multimedia Modelling Conference*, pages 46–52. Melbourne, 2005.

A. R. Jensenius. *Action–Sound : Developing Methods and Tools for Studying Music-Related Bodily Movement*. PhD thesis, University of Oslo, 2007a.

A. R. Jensenius. GDIF Development at McGill. Short Term Scientific Mission Report, COST Action 287 ConGAS. McGill University, Montreal 2007b. URL http://urn.nb.no/URN:NBN:no-21768 (Accessed October 10, 2012).

A. R. Jensenius. Motion capture studies of action-sound couplings in sonic interaction. Short Term Scientific Mission Report, COST Action IC0601 SID. Royal Institute of Technology, Stockholm, 2009. URL http://urn.nb.no/URN:NBN:no-26163 (Accessed October 10, 2012).

A. R. Jensenius. Some video abstraction techniques for displaying body movement in analysis and performance. *Leonardo Music Journal* (to appear), 2012a.

A. R. Jensenius. Evaluating how different video features influence the quality of resultant motiongrams. In *Proceedings of the Sound and Music Computing Conference*, pages 467–472, Copenhagen, 2012b.

A. R. Jensenius. Motion-sound interaction using sonification based on motiongrams. In *Proceedings of The Fifth International Conference on Advances in Computer-Human Interactions*, pages 170–175, Valencia, 2012c.

A. R. Jensenius and K. A. V. Bjerkestrand. Exploring micromovements with motion capture and sonification. In A. L. Brooks et al., editors, *Arts and Technology*, volume 101 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 100–107. Springer, Berlin Heidelberg, 2012.

A. R. Jensenius and A. Voldsund. The music ball project: Concept, design, development, performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 300–303, Ann Arbor, 2012.

A. R. Jensenius, R. I. Godøy, and M. M. Wanderley. Developing tools for studying musical gestures within the max/msp/jitter environment. In *Proceedings of the International Computer Music Conference*, pages 282–285. Barcelona, 2005.

A. R. Jensenius, R. Koehly, and M. Wanderley. Building low-cost music controllers. In R. Kronland-Martinet, T. Voinier, and S. Ystad, editors, *Computer Music Modeling and Retrieval*, volume 3902 of *Lecture Notes in Computer Science*, pages 123–129. Springer, Berlin Heidelberg, 2006a.

A. R. Jensenius, T. Kvifte, and R. I. Godøy. Towards a gesture description interchange format. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 176–179, Paris, 2006b.

A. R. Jensenius, A. Camurri, N. Castagne, E. Maestre, J. Malloch, D. McGilvray, D. Schwarz, and M. Wright. Panel: the need of formats for streaming and storing music-related movement and gesture data. In *Proceedings of the International Computer Music Conference*, pages 13–16, Copenhagen, 2007.

A. R. Jensenius, K. Nymoen, and R. I. Godøy. A multilayered GDIF-based setup for studying coarticulation in the movements of musicians. In *Proceedings of the International Computer Music Conference*, pages 743–746, Belfast, 2008.

A. R. Jensenius, S. A. Skogstad, K. Nymoen, J. Torresen, and M. E. Høvin. Reduced displays of multidimensional motion capture data sets of musical performance. In *Proceedings of ESCOM 2009: 7th Triennial Conference of the European Society for the Cognitive Sciences of Music*, Jyväskylä, Finland, 2009.

A. R. Jensenius, M. M. Wanderley, R. I. Godøy, and M. Leman. Musical gestures: Concepts and methods in research. In R. I. Godøy and M. Leman, editors, *Musical Gestures: Sound, Movement, and Meaning*. Routledge, New York, 2010.

A. R. Jensenius, K. Nymoen, S. A. Skogstad, and A. Voldsund. How still is still? a study of the noise-level in two infrared marker-based motion capture systems. In *Proceedings of the Sound and Music Computing Conference*, pages 258–263, Copenhagen, 2012.

M. R. Jones. Music perception: Current research and future directions. In M. Riess Jones, R. R. Fay, and A. N. Popper, editors, *Music Perception*, volume 36 of *Springer Handbook of Auditory Research*, pages 1–12. Springer, New York, 2010.

S. Jordà. Afasia: the Ultimate Homeric One-man-multimedia-band. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 102–107, Dublin, 2002.

A. Kapur, G. Tzanetakis, N. Virji-Babul, G. Wang, and P. R. Cook. A framework for sonification of vicon motion capture data. In *Proceedings of the International Conference on Digital Audio Effects*, pages 47–52, Madrid, 2005.

S. T. Klapp and R. J. Jagacinski. Gestalt principles in the control of motor action. *Psychological Bulletin*, 137(3):443–462, 2011.

M. Klingbeil. Software for spectral analysis, editing, and synthesis. In *Proceedings of the International Computer Music Conference*, pages 107–110, Barcelona, 2005.

T. Koerselman, O. Larkin, and K. Ng. The mav framework: Working with 3d motion data in max msp / jitter. In *Proceedings of the 3rd International Conference on Automated Production of Cross Media Content for Multi-channel Distribution (AXMEDIS 2007). Volume for Workshops, Tutorials, Applications and Industrial, i-Maestro 3rd Workshop*, Barcelona, 2007.

E. Kohler, C. Keysers, M. A. Umiltà, L. Fogassi, V. Gallese, and G. Rizzolatti. Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, 297(5582):846–848, 2002.

D. Kohn and Z. Eitan. Seeing sound moving: Congruence of pitch and loudness with human movement and visual shape. In *Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) and 8th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM)*, pages 541–546, Thessaloniki, 2012.

M. Kozak, K. Nymoen, and R. I. Godøy. The effects of spectral features of sound on gesture type and timing. In E. Efthimiou, G. Kouroupetroglou, and S.-E. Fotinea, editors, *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication. 9th International Gesture Workshop - GW 2011, May 2011, Athens, Greece. Revised selected papers.*, volume 7206 of *Lecture Notes in Computer Science/LNAI*. Springer (to appear), Berlin Heidelberg, 2012.

L. Kozlowski and J. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Attention, Perception, & Psychophysics*, 21(6):575–580, 1977.

M. Kussner. Creating shapes: musicians' and non-musicians' visual representations of sound. In U. Seifert and J. Wewers, editors, *Proceedings of SysMus11: Fourth International Conference of students of Systematic Musicology*, Osnabrück, epOs-Music (to appear), 2012.

T. Kvifte. *Instruments and the Electronic Age*. Solum, Oslo, 1989.

A. Lahav, E. Saltzman, and G. Schlaug. Action representation of sound: audiomotor recognition network while listening to newly acquired actions. *The journal of neuroscience*, 27(2):308–314, 2007.

G. Lakoff and M. Johnson. *Metaphors We Live By*. University of Chicago Press, Chicago, IL., 1980.

F. Langheim, J. Callicott, V. Mattay, J. Duyn, and D. Weinberger. Cortical systems associated with covert music rehearsal. *NeuroImage*, 16(4):901–908, 2002.

O. Lartillot, P. Toiviainen, and T. Eerola. A matlab toolbox for music information retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, editors, *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–268. Springer, Berlin Heidelberg, 2008.

M. Leman. *Embodied Music Cognition and Mediation Technology*. The MIT Press, 2008.

M. Leman and L. A. Naveda. Basic gestures as spatiotemporal reference frames for repetitive dance/music patterns in samba and charleston. *Music Perception*, 28(1):71–91, 2010.

G. Leslie, D. Schwarz, O. Warusfel, F. Bevilacqua, B. Zamborlin, P. Jodlowski, and N. Schnell. Grainstick: A collaborative, interactive sound installation. In *Proceedings of the International Computer Music Conference*, pages 123–126, New York, 2010.

A. M. Liberman and I. G. Mattingly. The motor theory of speech perception revised. *Cognition*, 21(1):1 – 36, 1985.

E. Lin and P. Wu. Jam master, a music composing interface. In *Proceedings of Human Interface Technologies*, pages 21–28, Vancouver, BC, 2000.

G. Luck, S. Saarikallio, B. Burger, M. Thompson, and P. Toiviainen. Effects of the big five and musical genre on music-induced movement. *Journal of Research in Personality*, 44(6):714 – 720, 2010a.

G. Luck, P. Toiviainen, and M. R. Thompson. Perception of expression in conductors' gestures: A continuous response study. *Music Perception*, 28(1):47–57, 2010b.

P.-J. Maes, M. Leman, M. Lesaffre, M. Demey, and D. Moelants. From expressive gesture to sound. *Journal on Multimodal User Interfaces*, 3:67–78, 2010.

E. Maestre, J. Janer, M. Blaauw, A. Pérez, and E. Guaus. Acquisition of violin instrumental gestures using a commercial EMF tracking device. In *Proceedings of the International Computer Music Conference*, pages 386–393, Copenhagen, 2007.

J. Malloch, S. Sinclair, and M. M. Wanderley. From controller to sound: Tools for collaborative development of digital musical instruments. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 66–69, New York, 2007.

T. Marrin and R. Picard. The "conductor's jacket": A device for reocording expressive musical gestures. In *Proceedings of the International Computer Music Conference*, pages 215–219, Ann Arbor, 1998.

M. Marshall, M. Rath, and B. Moynihan. The virtual bodhran: the vodhran. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 1–2, Dublin, 2002.

M. Marshall, N. Peters, A. R. Jensenius, J. Boissinot, M. M. Wanderley, and J. Braasch. On the development of a system for gesture control of spatialization. In *Proceedings of the International Computer Music Conference*, pages 360–366, New Orleans, 2006.

M. Mathews. What is loudness? In P. R. Cook, editor, *Music, Cognition, and Computerized Sound*, pages 71–78. MIT Press, Cambridge, MA, USA, 1999a.

M. Mathews. Introdction to timbre. In P. R. Cook, editor, *Music, Cognition, and Computerized Sound*, pages 79–87. MIT Press, Cambridge, MA, USA, 1999b.

S. McAdams. *Spectral Fusion, Spectral Parsing, and the Formation of Auditory Images*. PhD thesis, Stanford University, 1984.

S. McAdams, S. Winsberg, S. Donnadieu, G. Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58:177–192, 1995.

H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 12 1976.

I. Meister, T. Krings, H. Foltys, B. Boroojerdi, M. Müller, R. Töpper, and A. Thron. Playing piano in the mind—an fmri study on music imagery and performance in pianists. *Cognitive Brain Research*, 19(3):219 – 228, 2004.

T. Melzer, M. Reiter, and H. Bischof. Appearance models based on kernel canonical correlation analysis. *Pattern Recognition*, 36(9):1961–1971, 2003.

A. Merer, S. Ystad, R. Kronland-Martinet, and M. Aramaki. Semiotics of sounds evoking motions: Categorization and acoustic features. In R. Kronland-Martinet, S. Ystad, and K. Jensen, editors, *Computer Music Modeling and Retrieval. Sense of Sounds*, number 4969 in Lecture Notes in Computer Science, pages 139–158. Springer, Berlin Heidelberg, 2008.

I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. Yale: Rapid prototyping for complex data mining tasks. In L. Ungar, M. Craven, D. Gunopulos, and T. Eliassi-Rad, editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, 2006.

G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63:81–97, 1956.

E. R. Miranda and M. Wanderley. *New Digital Musical Instruments: Control And Interaction Beyond the Keyboard*. A-R Editions, Inc., Middleton, WI, 2006.

T. Mitchell and I. Heap. Soundgrasp: A gestural interface for the performance of live music. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 465–468, Oslo, 2011.

D. S. Moore and G. P. McCabe. *Introduction to the practice of statistics*. W.H. Freeman and Company, New York, 5th edition, 2006.

M. Müller. *Information retrieval for music and motion*. Springer, Berlin Heidelberg, 2007.

M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *SCA '06: Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 137–146, Aire-la-Ville, Switzerland, 2006.

K. Ng, O. Larkin, T. Koerselman, B. Ong, D. Schwarz, and F. Bevilacqua. The 3D augmented mirror: motion analysis for string practice training. In *Proceedings of the International Computer Music Conference*, pages 53–56, Copenhagen, 2007.

L. Nigay and J. Coutaz. A design space for multimodal systems: concurrent processing and data fusion. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 172–178, Amsterdam, 1993.

M. Nusseck and M. Wanderley. Music and motion-how music-related ancillary body movements contribute to the experience of music. *Music Perception*, 26(4):335–353, 2009.

K. Nymoen. The Nymophone2 – a study of a new multidimensionally controllable musical instrument. Master's thesis, University of Oslo, 2008a.

K. Nymoen. A setup for synchronizing GDIF data using SDIF-files and FTM for Max. Short Term Scientific Mission Report, COST Action IC0601 SID. McGill University, Montreal, 2008b. URL http://urn.nb.no/URN:NBN:no-20580 (Accessed October 10, 2012).

K. Nymoen, J. Torresen, R. Godøy, and A. R. Jensenius. A statistical approach to analyzing sound tracings. In S. Ystad, M. Aramaki, R. Kronland-Martinet, K. Jensen, and S. Mohanty, editors, *Speech, Sound and Music Processing: Embracing Research in India*, volume 7172 of *Lecture Notes in Computer Science*, pages 120–145. Springer, Berlin Heidelberg, 2012.

J. Oh, J. Herrera, N. J. Bryan, L. Dahl, and G. Wang. Evolving the mobile phone orchestra. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 82–87, Sydney, 2010.

Oxford Dictionaries: "Modality". URL http://oxforddictionaries.com/definition/english/modality (Accessed September 21, 2012).

R. Parncutt. Systematic musicology and the history and future of western musical scholaship. *Journal of Interdisciplinary Music Studies*, 1(1):1–32, 2007.

G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Tech. rep., IRCAM, Paris, 2004.

G. Peeters, B. Giordano, P. Susini, N. Misdariis, and S. McAdams. The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916, 2011.

G. Pellegrino, L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti. Understanding motor events: a neurophysiological study. *Experimental Brain Research*, 91:176–180, 1992.

J.-M. Pelletier. cv.jit — Computer Vision for Jitter (software). URL http://jmpelletier.com/cvjit/ (Accessed June 27, 2012).

N. Peters, T. Lossius, J. Schacher, P. Baltazar, C. Bascou, and T. Place. A stratified approach for sound spatialization. In *Proceedings of 6th Sound and Music Computing Conference*, pages 219–224, Porto, 2009.

J. Pierce. Introduction to pitch perception. In P. R. Cook, editor, *Music, Cognition, and Computerized Sound*, chapter 5. MIT Press, Cambridge, MA, USA, 1999.

T. Place and T. Lossius. Jamoma: A Modular Standard for Structuring Patches in Max. In *Proceedings of the International Computer Music Conference*, pages 143–146, New Orleans, 2006.

T. Place, T. Lossius, A. R. Jensenius, and N. Peters. Flexible control of composite parameters in max/msp. In *Proceedings of the International Computer Music Conference*, pages 233–236, Belfast, 2008.

Polhemus Inc. Liberty brochure. URL http://www.polhemus.com/polhemus_editor/assets/LIBERTY.pdf (Accessed June 25, 2012).

F. Pollick, H. Paterson, A. Bruderlin, and A. Sanford. Perceiving affect from arm movement. *Cognition*, 82(2):B51–B61, 2001.

E. Pöppel. A hierarchical model of temporal perception. *Trends in cognitive sciences*, 1(2): 56–61, 1997.

F. Raab, E. Blood, T. Steiner, and H. Jones. Magnetic position and orientation tracking system. *IEEE Transactions on Aerospace and Electronic Systems*, 15(5):709–718, 1979.

N. Rasamimanana, D. Bernardin, M. Wanderley, and F. Bevilacqua. String bowing gestures at varying bow stroke frequencies: A case study. In M. Sales Dias, S. Gibet, M. Wanderley, and R. Bastos, editors, *Gesture-Based Human-Computer Interaction and Simulation*, volume 5085 of *Lecture Notes in Computer Science*, pages 216–226. Springer, Berlin Heidelberg, 2009.

T. Ringbeck. A 3d time of flight camera for object detection. In *Proceedings of the 8th Conference on Optical 3D Measurement Techniques*, pages 1–10, ETH Zürich, 2007.

J.-C. Risset. Timbre analysis by synthesis: representations, imitations, and variants for musical composition. In G. De Poli, A. Piccialli, and C. Roads, editors, *Representations of musical signals*, pages 7–43. MIT Press, Cambridge, MA, 1991.

D. G. E. Robertson, G. E. Caldwell, J. Hamill, G. Kamen, and S. N. Whittlesey. *Research Methods in Biomechanics*. Human Kinetics, 2004.

D. Roetenberg, H. Luinge, and P. Slycke. Xsens MVN: Full 6DOF Human Motion Tracking Using Miniature Inertial Sensors. Technical report, Xsens Technologies B.V., 2009. URL http://www.xsens.com/images/stories/PDF/MVN_white_paper.pdf (Accessed March 27, 2011).

D. Rosenbaum. *Human motor control*. Academic Press, San Diego, 2001.

F. Rumsey. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *J. Audio Eng. Soc*, 50(9):651–666, 2002.

G. Salton and M. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36, 1968.

P. Schaeffer. *Traité des objets musicaux*. Éditions du Seuil, 1966.

P. Schaeffer and G. Reibel. *Solfège de l'objet sonore*. ORTF, Paris, France, INA-GRM 1998 edition, 1967.

M. Schleidt and J. Kien. Segmentation in behavior and what it can tell us about brain function. *Human nature*, 8(1):77–111, 1997.

N. Schnell, R. Borghesi, D. Schwarz, F. Bevilacqua, and R. Müller. FTM – complex data structures for Max. In *Proceedings of the International Computer Music Conference*, pages 9–12, Barcelona, 2005.

L. Schomaker, J. Nijtmans, A. Camurri, F. Lavagetto, P. Morasso, C. Benoit, T., J. Robert-Ribes, A. Adjoudani, I. Defée, S. Münch, K. Hartung, and J. Blauert. A taxonomy of multimodal interaction in the human information processing system. Report of the esprit project 8579 MIAMI, Nijmegen University, The Netherlands, 1995.

E. Schoonderwaldt and M. Demoucron. Extraction of bowing parameters from violin performance combining motion capture and sensors. *The Journal of the Acoustical Society of America*, 126(5):2695–2708, 2009.

E. Schoonderwaldt, N. Rasamimanana, and F. Bevilacqua. Combining accelerometer and video camera: reconstruction of bow velocity profiles. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 200–203, Paris, 2006.

E. Schubert. Correlation analysis of continuous emotional response to music. *Musicae Scientiae*, Special issue 2001–2002:213–236, 2002.

S. Sentürk, S. W. Lee, A. Sastry, A. Daruwalla, and G. Weinberg. Crossole: A gestural interface for composition, improvisation and performance using kinect. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Ann Arbor, 2012.

B. G. Shinn-Cunningham. Object-based auditory and visual attention. *Trends in Cognitive Sciences*, 12(5):182 – 186, 2008.

R. Siegwart and I. Nourbakhsh. *Introduction to autonomous mobile robots*. MIT Press, 2004.

S. A. Skogstad, A. R. Jensenius, and K. Nymoen. Using IR optical marker based motion capture for exploring musical interaction. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 407–410, Sydney, 2010.

S. A. Skogstad, K. Nymoen, Y. de Quay, and A. R. Jensenius. OSC implementation and evaluation of the Xsens MVN suit. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 300–303, Oslo, 2011.

S. A. Skogstad, S. Holm, and M. Høvin. Designing Digital IIR Low-Pass Differentiators With Multi-Objective Optimization. In *Proceedings of IEEE International Conference on Signal Processing*, Beijing Jiaotong University (to appear), 2012a.

S. A. Skogstad, S. Holm, and M. Høvin. Designing Low Group Delay IIR Filters for Real-Time Applications. In *Proceedings of the International Conference on Engineering and Technology*, Cairo (to appear), 2012b.

S. A. Skogstad, K. Nymoen, Y. de Quay, and A. R. Jensenius. Developing the dance jockey system for musical interaction with the Xsens MVN suit. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 226–229, Ann Arbor, 2012c.

S. W. Smith. *The scientist and engineer's guide to digital signal processing*. California Technical Publishing, San Diego, 1997.

B. Snyder. *Music and Memory. An Introduction*. MIT Press, Cambridge, MA, 2000.

C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

M. Spong, S. Hutchinson, and M. Vidyasagar. *Robot modeling and control*. John Wiley & Sons, New York, 2006.

B. E. Stein and M. A. Meredith. *The merging of the senses*. The MIT Press, Cambridge, MA, 1993.

S. S. Stevens. A metric for the social consensus. *Science*, 151(3710):530–541, 1966.

M. Thompson and G. Luck. Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music. *Musicae Scientiae*, 16(1):19–40, 2012.

P. Toiviainen and B. Burger. *MoCap toolbox manual*. University of Jyväskylä, 2011. URL https://www.jyu.fi/music/coe/materials/mocaptoolbox/MCTmanual (Accessed June 29, 2012).

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

S. Trail, M. Dean, G. Odowichuk, T. F. Tavares, P. Driessen, W. A. Schloss, and G. Tzanetakis. Non-invasive sensing and gesture control for pitched percussion hyper-instruments using the kinect. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Ann Arbor, 2012.

D. Trueman, P. Cook, S. Smallwood, and G. Wang. Plork: Princeton laptop orchestra, year 1. In *Proceedings of the International Computer Music Conference*, pages 443–450, New Orleans, 2006.

F. Upham. Limits on the application of statistical correlations to continuous response data. In *Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) and 8th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM)*, pages 1037–1041, Thessaloniki, 2012.

L. van Noorden. *Temporal Coherence in the Perception of Tone Sequences*. PhD thesis, Technical University Eindhoven, 1975.

L. van Noorden. The functional role and bio-kinetics of basic and expressive gestures in activation and sonification. In R. I. Godøy and M. Leman, editors, *Musical Gestures: Sound, Movement, and Meaning*, pages 154–179. Routledge, New York, 2010.

L. van Noorden and D. Moelants. Resonance in the perception of musical pulse. *Journal of New Music Research*, 28(1):43–66, 1999.

V. Verfaille, O. Quek, and M. Wanderley. Sonification of musicians' ancillary gestures. In *Proceedings of the International Conference on Auditory Display*, pages 194–197, London, 2006.

G. Vigliensoni and M. M. Wanderley. A quantitative comparison of position trackers for the development of a touch-less musical interface. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 103–108, Ann Arbor, 2012.

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

F. Vogt, G. Mccaig, M. A. Ali, and S. S. Fels. Tongue 'n' Groove: An Ultrasound based Music Controller. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 181–185, Dublin, 2002.

J. Vroomen and B. de Gedler. Sound enhances visual perception: Cross-modal effects on auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance*, 26(5):1583–1590, 2000.

B. Walker. *Magnitude estimation of conceptual data dimensions for use in sonification*. PhD thesis, Rice University, Houston, TX, 2000.

M. M. Wanderley. Quantitative analysis of non-obvious performer gestures. In I. Wachsmuth and T. Sowa, editors, *Gesture and Sign Language in Human-Computer Interaction*, volume 2298 of *Lecture Notes in Computer Science*, pages 241–253. Springer, Berlin Heidelberg, 2002.

M. M. Wanderley and M. Battier, editors. *Trends in Gestural Control of Music*. IRCAM — Centre Pompidou, Paris, 2000.

M. M. Wanderley and P. Depalle. Gestural control of sound synthesis. In *Proceedings of the IEEE*, volume 92, pages 632–644, 2004.

M. M. Wanderley, B. W. Vines, N. Middleton, C. McKay, and W. Hatch. The musical significance of clarinetists' ancillary gestures: An exploration of the field. *Journal of New Music Research*, 43(1):97–113, 2005.

G. Wang, G. Essl, and H. Penttinen. Do mobile phones dream of electric orchestras. In *Proceedings of the International Computer Music Conference*, Belfast, 2008.

G. Welch and E. Foxlin. Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Computer Graphics and Applications*, 22(6):24–38, 2002.

D. Wessel and M. Wright. Problems and prospects for intimate musical control of computers. *Computer Music Journal*, 26:11–22, 2002.

M. Wright and A. Freed. Open sound control: A new protocol for communicating with sound synthesizers. In *Proceedings of the International Computer Music Conference*, pages 101–104, Thessaloniki, 1997.

M. Wright, A. Chaudhary, A. Freed, D. Wessel, X. Rodet, D. Virolle, R. Woehrmann, and X. Serra. New applications of the sound description interchange format. In *Proceedings of the International Computer Music Conference*, pages 276–279, Ann Arbor, 1998.

M. Yoo, J. Beak, and I. Lee. Creating musical expression using kinect. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 324–325, Oslo, 2011.

M. Zadel, S. Sinclair, and M. Wanderley. Haptic feedback for different strokes using dimple. In *Proceedings of the International Computer Music Conference*, pages 291–294, Montreal, 2009.

R. Zatorre. Music, the food of neuroscience? *Nature*, 434:312–315, 2005.

R. Zatorre and A. Halpern. Mental concerts: musical imagery and auditory cortex. *Neuron*, 47 (1):9–12, 2005.

K. Zou, K. Tuncali, and S. Silverman. Correlation and simple linear regression. *Radiology*, 227 (3):617–628, 2003.

Glossary

2D Two dimensions.

3D Three dimensions.

Canonical Loading The correlation between a canonical component and an input feature in canonical correlation analysis.

Canonical Variates Also known as canonical components. The resulting vectors from projecting the input features onto projection matrices in canonical correlation analysis.

CCA Canonical Correlation Analysis.

Class Precision (CP) The number of correctly classified instances of a class divided by the total number of instances classified as the same class.

Class Recall (CR) The number of correctly classified instances of a class divided by the true number of instances in the class.

DCM Direction Cosine Matrix, also called Rotation Matrix or Orientation Matrix.

Degrees of Freedom (DOF) A term used in *kinematics:* The number of dimensions in which an object can be tracked. For instance 3DOF position, or 6DOF position and orientation.

Degrees of Freedom (*df*) A term used in *statistics:* A statistical variable related to statistical tests, such as the *t*-test. The variable describes the number of values that are free to vary in the calculation of a statistic.

FTM A set of objects for Max. Developed at IRCAM, Paris.

GCS Global Coordinate System.

GDIF Gesture Description Interchange Format.

Jamoma A set of modular patches and objects for Max.

Kinect An optical markerless motion tracking device developed by PrimeSense for Microsoft. Intended for use with the Xbox 360 console, but can also interface with other computers.

Kinematic model A set rigid objects, with corresponding rules for how the objects relate to each other, e.g. joint angles.

LCS Local Coordinate System.

Max A graphical programming environment, also known as Max/MSP/Jitter, Max5 or Max6.

mocap Motion capture. Sometimes also MoCap, for instance in the name *MoCap Toolbox*.

MoCap Toolbox Matlab Toolbox for processing, visualising and analysing motion data.

Open Sound Control (OSC) A protocol for streaming (primarily music-related) data.

OptiTrack An optical infrared marker-based motion tracking system. Produced by NaturalPoint Inc., Oregon, USA.

Qualisys An optical infrared marker-based motion tracking system. Produced by Qualisys AB, Sweden.

Rigid object An object for which both position and orientation can be tracked.

RMS Root-mean-square. A measure of the magnitude of a varying signal. May be used to describe the dynamic envelope of a sound signal.

SDIF Sound Description Interchange Format.

Spectral Centroid The barycentre (centre of gravity) of a spectrum. In sound perception, the spectral centroid is one way of describing brightness.

SVM Support Vector Machine.

Training set Data set used to train a classifier.

Validation set Data set used to validate a classifier.

Xsens MVN A full body motion capture suit based on inertial sensors and magnetometers. Produced by Xsens Technologies, Belgium.

Wiimote Game controller for the Nintendo Wii. Used in an alternative low-cost implementation of the SoundSaber.

Papers

I A Toolbox for Storing and Streaming Music-Related Data.
K. Nymoen and A.R. Jensenius.
In *Proceedings of SMC 2011 8th Sound and Music Computing Conference "Creativity rethinks science"*, pages 427–430, Padova University Press 2011.

II Comparing Inertial and Optical MoCap Technologies for Synthesis Control.
S.A. Skogstad, K. Nymoen, and M.E. Høvin.
In *Proceedings of SMC 2011 8th Sound and Music Computing Conference "Creativity rethinks science"*, pages 421–426, Padova University Press 2011.

III Comparing Motion Data from an iPod Touch to a High-End Optical Infrared Marker-Based Motion Capture System.
K. Nymoen, A. Voldsund, S.A. Skogstad, A.R. Jensenius, and J. Torresen.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 88–91, University of Michigan 2012.

IV SoundSaber — A Motion Capture Instrument.
K. Nymoen, S.A. Skogstad and A.R. Jensenius.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 312–315, University of Oslo 2011.

V Searching for Cross-Individual Relationships between Sound and Movement Features Using an SVM Classifier.
K. Nymoen, K. Glette, S.A. Skogstad, J. Torresen, and A.R. Jensenius.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 259–262, Sydney University of Technology 2010.

VI Analyzing Sound Tracings: A Multimodal Approach to Music Information Retrieval.
K. Nymoen, B. Caramiaux, M. Kozak, and J. Torresen.
In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 39–44, ACM 2011.

VII A Statistical Approach to Analyzing Sound Tracings.
K. Nymoen, J. Torresen, R.I. Godøy, and A.R. Jensenius.
In S. Ystad, M. Aramaki, R. Kronland-Martinet, K. Jensen, and S. Mohanty (eds.) *Speech, Sound and Music Processing: Embracing Research in India*, volume 7172 of Lecture Notes in Computer Science, pages 120–145. Springer, Berlin Heidelberg 2012.

VIII Analysing Correspondence Between Sound Objects and Body Motion.
K. Nymoen, R.I. Godøy, A.R. Jensenius, and J. Torresen.
To appear in *ACM Transactions on Applied Perception*.

Paper I

A Toolbox for Storing and Streaming Music-related Data.
K. Nymoen and A.R. Jensenius.
In *Proceedings of SMC 2011 8th Sound and Music Computing Conference "Creativity rethinks science"*, pages 427–430, Padova University Press 2011.

A TOOLBOX FOR STORING AND STREAMING MUSIC-RELATED DATA

Kristian Nymoen
fourMs - Music, Mind, Motion, Machines
Department of Informatics
University of Oslo
krisny@ifi.uio.no

Alexander Refsum Jensenius
fourMs - Music, Mind, Motion, Machines
Department of Musicology
University of Oslo
a.r.jensenius@imv.uio.no

ABSTRACT

Simultaneous handling and synchronisation of data related to music, such as score annotations, MIDI, video, motion descriptors, sensor data, etc. requires special tools due to the diversity of the data. We present a toolbox for recording and playback of complex music-related data. Using the Sound Description Interchange Format as a storage format and the Open Sound Control protocol as a streaming protocol simplifies exchange of data between composers and researchers.

1. INTRODUCTION

In this paper we introduce a set of tools that have been developed for working with music-related data. Our goal with this software is primarily to provide a set of tools for researchers working with music-related body motion, but we also see the potential for using the tools in other research areas. We started working on the tools in 2008, and the development has continued over the last years together with our research on music and movement [1, 2, 3]. The need for a common method of storing and sharing data related to musical movement was discussed at a panel session at the International Computer Music Conference 2007 [4], and further emphasised at a seminar in May 2010 at IRCAM, Paris, where several researchers from around the world working with music and motion, and sound spatialisation were present. A common denominator for this seminar was to come closer to a scheme for describing spatio-temporal aspects of music. The tools we are presenting were revised after this seminar with the intention of making them easy to use for the research community.

Section 2 introduces previous research and gives an overview of why these tools are needed, and what has already been done in the field. In section 3, the different types of data we are working with are discussed. Section 4 introduces the tools. Finally, in section 5, we conclude and point out the future directions of the development.

2. BACKGROUND AND MOTIVATION

In our research on music-related body motion, we are often faced with situations where we want to study data from several devices at the same time. We will start this section by looking at two use cases that summarise some of the challenges in the field and the tools needed.

2.1 Use cases

a. The music researcher
A researcher is interested in studying the movement of a pianist by using an optical infrared motion capture system and record MIDI events from the piano. By themselves, the MIDI and motion capture data is trivial to record. However, synchronising the two, and being able to play them back later, or even scrubbing through the recording, keeping the MIDI-data and the motion capture data aligned, is not as trivial. Motion capture data is typically recorded at a sampling rate of 100–500 Hz, while the MIDI data stream is event driven and only needs to be stored each time a MIDI event takes place. Thus, using a common sampling rate for MIDI data and motion capture data would mean recording a lot of redundant data. The setup becomes even more complex when the researcher wants to record data from other sensors and audio/video as well.

b. The composer
A composer wants to develop a system for modifying sound in real-time. Let us say that the composer has hired a violin player who is wearing a position sensor and using a bow equipped with an accelerometer. She wants to develop a system that modifies the violin sound in real-time, based on output from the position sensor and the bow accelerometer data. Having the musician available at all times to perform can be expensive, as the musician would typically have to spend quite a lot of time waiting for the composer to make adjustments in the mapping between motion and sound. The composer would benefit from being able to record both the sound and the sensor data, and to play them back as a single synchronised performance.

Both of these examples show us that there is a need for a flexible system that is able to record different types of data from an arbitrary number of devices simultaneously. Further complexity is added when multiple representations of the same data is required. For instance, the researcher could be interested in the coordinates of the hands of a piano player in relation to a global coordinate system, but

also in relation to a coordinate frame defined by the position of the piano, or the center of mass in the pianist's body. The natural complexity of music introduces needs for various simultaneous representations of the same data.

Existing formats for working with these types of data have advantages and disadvantages, and there is no agreement between researchers on how to share music-related motion data. For motion capture data, the most widespread format is C3D. [1] Unfortunately, C3D does not allow for storing or synchronising music-related data and media. The Gesture Motion Signal [2] format has been developed to handle low level data in a musical context, but does not handle higher level data. The latter is handled well with the Performance Markup Language, [3] but this format does not meet our requirements when it comes to audio and video synchronisation.

An approach similar to or own has been implemented in OpenMusic [5]. Bresson et al. have implemented a solution for storing and streaming sound spatialisation data in the Sound Description Interchange Format (SDIF). This seems to be a promising solution, and we hope to keep collaborating on SDIF descriptors for spatio-temporal data.

2.2 GDIF

The Gesture Description Interchange Format (GDIF) has been proposed for handling the diversity of data related to music and motion [6]. The name GDIF might be somewhat misleading, as this is neither a format per se, nor is it limited to only gesture-related data. Rather, it is a concept and an idea for how data, and particularly data related to musical movement, can be described and shared among different researchers.

This concept includes a hierarchical structure, where the raw data (i.e. the data that one receives directly from the sensor or interface) is stored at the bottom layer. Above this layer is a so-called *cooked layer*, where certain processing has taken place. This can be anything from simple filtering or transformations, to more advanced analysis. Other layers may include segmentations or chunks [7] and even higher-level descriptors such as expressivity, affect and mood.

So far, GDIF development has been concerned with conceptual issues, and it has been up to the user to define how to implement storage and streaming. Some guidelines have been suggested, one of them being the approach implemented in the system we are presenting in this paper. We are using the Sound Description Interchange Format for storing and the Open Sound Control protocol for streaming GDIF data [4]. These formats will be presented in sections 2.3 and 2.4.

2.3 SDIF

The Sound Description Interchange Format (SDIF) was proposed by researchers at IRCAM and CNMAT and has been suggested as a format for storing GDIF data [4, 8]. This file format describes a sequence of time-tagged *frames*. Each frame consists of an identifier indicating what type of frame it is, the frame size, the actual data and zero-padding to make the frame size a multiple of eight bytes [9]. The frames are further structured into *streams*. These streams are series of frames, and all streams share a common timeline. Inside each frame, the actual data is stored as strings, bytes, integers or floating point values in one or more 2D matrices.

2.4 Open Sound Control

Open Sound Control (OSC) is a protocol for real-time audio control messages [10]. Conceptually, OSC shares many similarities with the SDIF format, as it describes a way of streaming time-tagged bundles of data. Each bundle contains one or more *OSC messages*, each message containing an *OSC address* and the actual data in a list format. The OSC address contains a hierarchical structure of human readable words, separated by slashes, making it simple to work with and share data between researchers and musicians (e.g. `/mySynth/pitch 120`).

3. DATA TYPES

We are working with many different sorts of data. Part of GDIF development is to define data types that are as generic and at the same time as well defined as possible. In other words, data types in GDIF recordings must be defined in such a way that they are open enough for different use, and at the same time detailed enough to leave little or no doubt about what sort of data that is contained in a GDIF stream.

Frames and matrices in SDIF streams are identified by a four letter type tag. This introduces some challenges when it comes to describing data. By convention, the first letter should be X for non-standard SDIF streams, leaving us with three letters to define the frame type and matrix type we are working with. Although it makes sense to distinguish between the two, our current implementation makes no distinction between the frame type and the matrix type. This means that the current system only allows a single data matrix inside each frame, and the frame automatically adapts the type tag from the matrix it contains. This has been sufficient in our use so far, but it would make more sense to let the frame type identify the stream (e.g. according to input device) and the matrix types define the data within each matrix (e.g. position, orientation, etc.).

For our matrix type tags, we have chosen to let the second letter determine the main data category, e.g. "P" for position data. The third letter denotes the dimensionality of the data, e.g. "2" if we are only tracking horizontal position. The fourth letter lets us know if the stream contains delta values of the original data. This number denotes derivative level, for instance "1" if the stream is the first derivative of the original data. This means that an XP32 matrix would contain 3-dimensional data, of the second derivative from the original position stream (i.e. acceleration).

We are sometimes interested in the absolute value of a vector, i.e. the length of the vector independent of the direction. This type of matrix is denoted by replacing the

[1] `http://www.c3d.org/`
[2] `http://acroe.imag.fr/gms/`
[3] `http://www.n-ism.org/Projects/pml.php`

third letter in the type tag with an "A". To formalise, this gives us the general case:

$$\mathrm{XPjd}[n] = \mathrm{XPj(d-1)}[n] - \mathrm{XPj(d-1)}[n-1]$$

$$\mathrm{XPAd}[n] = \sqrt{\sum_{i=1}^{j} \mathrm{XPjd}[n][i]^2}$$

and as an example, the specific case:

$$\mathrm{XP31}[n] = \mathrm{XP30}[n] - \mathrm{XP30}[n-1]$$

$$\mathrm{XPA1}[n] = \sqrt{\sum_{i=1}^{3} \mathrm{XP31}[n][i]^2}$$

where d denotes the derivative level, *n* denotes the frame index in a sequence of frames, *i* denotes the dimension index at frame *n*, and *j* denotes dimensionality of the stream.

In addition to streams describing position, velocity, etc., GDIF data types include everything from raw data from sensors to higher level descriptors. Table 1 displays a selection of the GDIF data types we are currently working with. A more complete list of data types can be found at the wiki that has been set up for GDIF and SpatDIF development. [4] It should be noted that these are our suggestions, and we welcome a discussion on these data types.

Table 1. A selection of GDIF data types.

Tag	Description
XIDX	Referring to a certain event in a series of events, e.g. triggering a sound sample from a sample bank.
XP30	3-dimensional position stream.
XP31	3-dimensional position stream. 1st derivative. (i.e. velocity calculated from position data)
XPA1	x-dimensional position stream. Absolute value of 1st derivative.
XOQ0	Orientation stream, four quaternion values.
XA30	3D acceleration stream. Used when working with systems that provide acceleration data as raw data.
1MID	MIDI stream, already defined in the SDIF standard
XEMG	Electromyography sensor input.
XMQ0	Quantity of motion stream.
XMA1	Area of motion stream. First derivative.

The system accepts all measurement units. However, we recommend using the International System of Units (SI) whenever this is possible. This will make it easier for researchers to share GDIF recordings.

4. IMPLEMENTATION

The tools presented in this paper are based on the SDIF tools in the FTM library, [5] mainly `ftm.sdif.write` for recording and `ftm.track` for playback [11]. They are implemented in Max as modules in the Jamoma [6] framework. These frameworks provide solutions for OSC and SDIF. The two main modules in the toolbox are the recording module and the playback module.

[4] `http://xdif.wiki.ifi.uio.no/Data_types`
[5] `http://ftm.ircam.fr`
[6] `http://www.jamoma.org`

The recording module, based on `ftm.sdif.write`, is designed for writing matrix-formatted data into separate streams in an SDIF file (Figure 1).

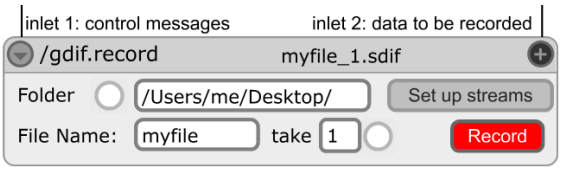


Figure 1. The record module

Different streams are separated by different OSC namespaces (e.g. `\stream\0`, `\stream\1`). The internal components of the recording module are created dynamically based on the user's selection of streams from a drop-down menu in the GUI. The user may customise the stream types that are available in the drop-down menu by editing a simple text file. Using a script language that is specific to the Max environment, stream definition commands and data descriptions are generated dynamically and sent to the `ftm.sdif.write` object whenever the user inputs a command or selects streams. The internally used OSC-routing objects as well as the `ftm.sdif.write` object are also created dynamically whenever the user chooses a different selection of data types. Figure 2 displays a simplified flowchart of how the record module works internally.

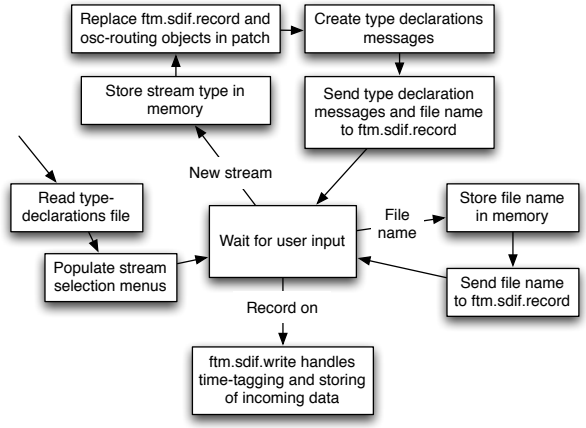


Figure 2. Simplified flowchart of the scripting system in the record module

The playback module displayed in Figure 3 is based on the `ftm.track` object. When an SDIF file is loaded into the playback module, an `ftm.track` object is created for each stream in the file. The data that is streamed from each track object is converted from the FTM float matrix format to Open Sound Control bundles using the OSC tools developed at CNMAT [10]. OSC does not support streaming matrices, hence each matrix row is separated as an instance number with its own OSC sub-address, e.g. first row gets the address `/XPOS/1`, second row `/XPOS/2`, etc. The user may set a custom buffer size for the OSC time tag to compensate for network latency and jitter. This buffer is set to a default value of 10 milliseconds.

The modules provide the user with a simple user inter-

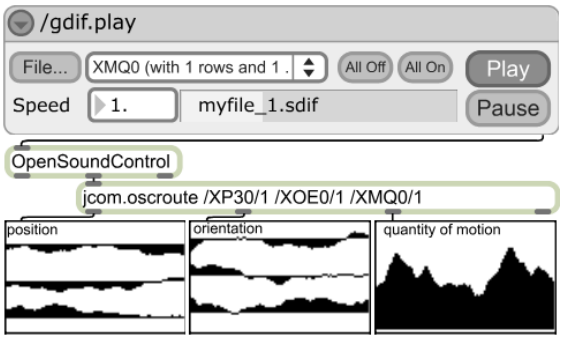


Figure 3. The playback module streaming a *3D position stream*, an *euler orientation* stream and a *quantity of motion* stream

face. Rather than having to record data into separate unsynchronised buffers, the user can now record data into a single container without worrying about synchronisation. The presented tools are open source, and can be downloaded by checking out the Jamoma repository from github, [7] or directly from the project website. [8] Since the data is recorded as SDIF files, users may benefit from tools like EasDIF [9] for analysis and post processing.

5. CONCLUSIONS AND FUTURE WORK

This paper has presented challenges we are facing when studying music-related body motion, and our solution to some of these problems in the form of a software toolbox. This toolbox includes a flexible module for making synchronized recordings of music-related data, and a module for playing back the data in real-time. The implementation makes the GDIF recording setup fast and easy, and makes this type of technology available to less experienced Max users.

Future development includes:

- Separating frame types as independent definitions. This will allow describing the stream type according to the device (e.g. a motion capture stream), and each frame can contain different data matrices (e.g. a position matrix and an orientation matrix).
- Human readable OSC namespace for data from the playback module (currently using the SDIF type tag).
- Integration of the Jamoma dataspaceLib for conversion between different data representations [12].
- Implementing simple data processing like automatic filtering and calculating absolute values.
- Develop a sequencer-like visual display, allowing zooming, editing, etc.
- Database for storing large collections of GDIF data.

6. ACKNOWLEDGEMENTS

Thanks to the developers of Max, Jamoma, FTM and OSC for providing a good frameworks for implementing these tools. Thanks also to the reviewers for valuable feedback.

[7] http://github.com/jamoma/Jamoma
[8] http://www.fourms.uio.no/software/jamomagdif/
[9] http://sourceforge.net/projects/sdif/

7. REFERENCES

[1] A. R. Jensenius, "GDIF development at McGill," McGill University, Montreal, Canada, COST ConGAS – STSM report, 2007.

[2] K. Nymoen, "A setup for synchronizing GDIF data using SDIF-files and FTM for Max," McGill University, Montreal, Canada, COST SID – STSM report, 2008.

[3] A. R. Jensenius, "Motion capture studies of action-sound couplings in sonic interaction," KTH, Stockholm, Sweden, COST SID – STSM report, 2009.

[4] A. R. Jensenius, A. Camurri, N. Castagne, E. Maestre, J. Malloch, D. McGilvray, D. Schwarz, and M. Wright, "Panel: the need of formats for streaming and storing music-related movement and gesture data," in *Proceedings of the 2007 International Computer Music Conference*, Copenhagen, 2007.

[5] J. Bresson, C. Agon, and M. Schumacher, "Représentation des données de contrôle pour la spatialisation dans openmusic," in *Actes de Journées d'Informatique Musicale (JIM'10)*, 2010.

[6] A. R. Jensenius, T. Kvifte, and R. I. Godøy, "Towards a gesture description interchange format," in *Proceedings of the 2006 International Conference on New Interfaces for Musical Expression*. Paris, France: Paris: IRCAM – Centre Pompidou, 2006, pp. 176–179.

[7] R. I. Godøy, "Reflections on chunking in music," in *Systematic and Comparative Musicology: Concepts, Methods, Findings. Hamburger Jarbuch für Musikwissenschaft*. P. Lang, 2008, vol. 24, pp. 117–132.

[8] M. Wright, A. Chaudhary, A. Freed, S. Khoury, and D. L. Wessel, "Audio applications of the sound description interchange format standard," in *AES 107th Convention*, 1999.

[9] M. Wright, A. Chaudhary, A. Freed, D. Wessel, X. Rodet, D. Virolle, R. Woehrmann, and X. Serra, "New applications of the sound description interchange format," in *Proceedings of the 1998 International Computer Music Conference*, Ann Arbor, 1998, pp. 276–279.

[10] M. Wright, A. Freed, and A. Momeni, "OpenSound Control: state of the art 2003," in *Proceedings of the 2003 conference on New Interfaces for Musical Expression*, Montreal, Canada, 2003, pp. 153–160.

[11] N. Schnell, R. Borghesi, D. Schwarz, F. Bevilacqua, and R. Müller, "FTM – complex data structures for Max," in *Proceedings of the 2005 International Computer Music Conference*, Barcelona, 2005, pp. 9–12.

[12] T. Place, T. Lossius, A. R. Jensenius, N. Peters, and P. Baltazar, "Addressing classes by differentiating values and properties in OSC," in *Proceeding of the 8th International Conference on New Instruments for Musical Expression*, 2008.

Paper II

Comparing Inertial and Optical MoCap Technologies for Synthesis Control.
S.A. Skogstad, K. Nymoen, and M.E. Høvin.
In *Proceedings of SMC 2011 8th Sound and Music Computing Conference "Creativity rethinks science"*, pages 421–426, Padova University Press 2011.

COMPARING INERTIAL AND OPTICAL MOCAP TECHNOLOGIES FOR SYNTHESIS CONTROL

Ståle A. Skogstad and Kristian Nymoen
fourMs - Music, Mind, Motion, Machines
Department of Informatics
University of Oslo
{savskogs,krisny}@ifi.uio.no

Mats Høvin
Robotics and Intelligent Systems group
Department of Informatics
University of Oslo
matsh@ifi.uio.no

ABSTRACT

This paper compares the use of two different technologies for controlling sound synthesis in real time: the infrared marker-based motion capture system *OptiTrack* and *Xsens MVN*, an inertial sensor-based motion capture suit. We present various quantitative comparisons between the data from the two systems and results from an experiment where a musician performed simple musical tasks with the two systems. Both systems are found to have their strengths and weaknesses, which we will present and discuss.

1. INTRODUCTION

Motion capture (MoCap) has become increasingly popular among music researchers, composers and performers [1]. There is a wide range of different MoCap technologies and manufacturers, and yet few comparative studies between the technologies have been published. Where one motion capture technology may outperform another in a sterilized laboratory setup, this may not be the case if the technologies are used in a different environment. Optical motion capture systems can suffer from optical occlusion, electromagnetic systems can suffer from magnetic disturbance, and so forth. Similarly, even though one motion capture system may be better than another at making accurate MoCap recordings and preparing the motion capture for offline analysis, the system may not be as good if the task is to do accurate motion capture in real time, to be used for example in controlling a sound synthesizer.

In this paper we compare the *real-time* performance of two motion capture systems (Figure 1) based on different technologies: Xsens MVN which is based on inertial sensors, and OptiTrack which is an infrared marker-based motion capture system (IrMoCap). Some of our remarks are also relevant to other motion capture systems than the ones discussed here, though the results and discussions are directed only toward OptiTrack and Xsens.

We will return to a description of these technologies in section 3. In the next section we will give a brief overview of related work. Section 4 will present results from comparisons between the two motion capture systems, which are then discussed in section 5.

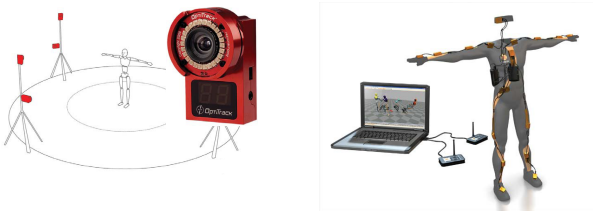


Figure 1. The NaturalPoint OptiTrack system (left) and the Xsens MVN system (right).

2. RELATED WORK AND BACKGROUND

Motion capture technologies have been used in musical contexts for a long time, and during the 00's we saw several examples of using various motion capture technologies for real-time control of sound. This includes electromagnetic motion capture [2], video-based motion capture [3], optical marker-based motion capture [4] and inertial motion capture [5], to mention a few.

Several researchers have reported on differences between motion capture technologies. Most of these reports, however, have been related to offline analysis for medical or animation purposes. Cloete et al. [6] have compared the kinematic reliability of the Xsens MVN suit with an IrMoCap system during routine gait studies. They conclude that the Xsens MVN system is comparable to IrMoCap systems but with shortcomings in some angle measurements. They also point out several practical advantages with the Xsens suit, like its wireless capabilities and quick set-up time. Another experiment by Thies et al. [7] found comparable acceleration values from two Xsens sensors and an IrMoCap system, and showed that calculating acceleration from the IrMoCap position data introduced noise. One of the conclusions from this experiment was that filtering methods need to be investigated further.

Miranda and Wanderley have pointed out some strengths and weaknesses with electromagnetic and optical motion capture systems [1]: Electromagnetic systems are able to track objects, even if it is not within the direct line of sight of external cameras. On the other hand, these systems need cables which may be obtrusive. Optical systems are superior to many other systems in terms of sampling rate, since they may track markers at sampling rates of more than 1000 Hz, and systems using passive markers have no need for obtrusive cables. Still, these systems need a direct line of sight between markers and cameras, and a passive

marker system may not be able to uniquely identify each marker.

Possibilities, strengths and weaknesses for real-time motion capture in musical contexts are discussed individually for IrMoCap and full-body inertial sensor systems in [8] and [9]. In this paper we will compare the real-time abilities of the two technologies.

2.1 Initial remarks on requirements when using MoCap for real-time control of music

A musical instrument is normally controlled with excitation and modification actions [10]. We can further distinguish between two types of excitations: discrete (i.e. trigger), or continuous (like bowing a string instrument). Dobrian [11] identifies two types of control data: triggers and streams of discrete data representing a sampling of a continuous phenomenon. Following these remarks, we are looking for a system able to robustly trigger sound events with good temporal accuracy, and to continuously control a system with good spatial accuracy and little noise. Consequently, we have chosen to emphasize three properties: spatial accuracy, temporal accuracy and system robustness. We will come back to measurements and discussion of these properties in sections 4 and 5.

3. TECHNOLOGIES

3.1 NaturalPoint OptiTrack

NaturalPoint OptiTrack is an optical infrared marker-based motion capture system (IrMoCap). This technology uses several cameras, equipped with infrared light-emitting diodes. The infrared light from the cameras is reflected by reflective markers and captured by each camera as 2D point-display images. By combining several of these 2D images the system calculates the 3D position of all the markers within the capture space. A calibration process is needed beforehand to determine the position of the cameras in relationship to each other, and in relationship to a global coordinate system defined by the user.

By using a combination of several markers in a specific pattern, the software can identify rigid bodies or skeletons. A *rigid body* refers to an object that will not deform. By putting at least 3 markers on the rigid body in a unique and non-symmetric pattern, the motion capture system is able to recognize the object and determine its position and orientation. A *skeleton* is a combination of rigid bodies and/or markers, and rules for how they relate to each other. In a human skeleton model, such a rule may be that the bottom of the right thigh is connected to the top of the right calf, and that they can only rotate around a single axis. In the NaturalPoint motion capture software (Arena), there exist 2 predefined skeleton models for the human body. It is not possible to set up user-defined skeletons.

3.2 The Xsens MVN

The Xsens MVN technology can be divided into two parts: (1) the sensor and communication hardware that are responsible for collecting and transmitting the raw sensor data, and (2) the Xsens MVN software engine, which interprets and reconstructs the data to full body motion while trying to minimize positional drift.

The Xsens MVN suit [12] consists of 17 inertial MTx sensors, which are attached to key areas of the human body. Each sensor consists of 3D gyroscopes, accelerometers and magnetometers. The raw signals from the sensors are connected to a pair of Bluetooth 2.0-based wireless transmitters, which again transmit the raw motion capture data to a pair of wireless receivers.

The data from the Xsens MVN suit is fed to the MVN software engine that uses sensor fusion algorithms to produce absolute orientation values, which are used to transform the 3D linear accelerations to global coordinates. These in turn are translated to a human body model which implements joint constraints to minimize integration drift. The Xsens MVN system outputs information about body motion by expressing body postures sampled at a rate up to 120Hz. The postures are modeled by 23 body segments interconnected with 22 joints.

4. MEASUREMENTS

We carried out two recording sessions to compare the OptiTrack and Xsens systems. In the first session, a series of simple measurements were performed recording the data with both Xsens and OptiTrack simultaneously. These recordings were made to get an indication of the differences between the data from the systems. In the second session (Section 4.5), a musician was given some simple musical tasks, using the two MoCap systems separately to control a sound synthesizer.

4.1 Data comparison

Our focus is on comparing real-time data. Therefore, rather than using the built-in offline recording functionality in the two systems, data was streamed in real-time to a separate computer where it was time-stamped and recorded. This allows us to compare the quality of the data as it would appear to a synthesizer on a separate computer. Two terminal applications for translating the native motion capture data to Open Sound Control and sending it to the remote computer via UDP were used.

We have chosen to base our plots on the unfiltered data received from the motion capture systems. This might differ from how a MoCap system would be used in a real world application, where filtering would also be applied. Using unfiltered data rather than filtered data gives an indication of how much pre-processing is necessary before the data can be used for a musical application.

The Xsens suit was put on in full-body configuration. For OptiTrack, a 34-marker skeleton was used. This skeleton model is one of the predefined ones in the Arena software. Markers were placed outside the Xsens suit, which made it necessary to adjust the position of some of the markers slightly, but this did not alter the stability of the OptiTrack system.

Both systems were carefully calibrated, but it was difficult to align their global coordinate systems perfectly. This

is because OptiTrack uses a so-called L-frame on the floor to determine the global coordinate system, whereas Xsens uses the position of the person wearing the suit during the calibration to determine the origin of the global coordinate system. For this reason, we get a bias in the data from one system compared to the other. To compensate for this, the data has been adjusted so that the mean value of the data from the two systems more or less coincide. This allows us to observe general tendencies in the data.

4.2 Positional accuracy and drift

When comparing the Xsens and the OptiTrack systems there is one immediately evident difference. OptiTrack measures absolute position, while the sensors in the Xsens MVN suit can only observe relative motion. With Xsens, we are bound to experience some positional drift even though the system has several methods to keep it to a minimum [9].

4.2.1 Positional accuracy - still study

Figure 2 shows the position of the left foot of a person sitting in a chair without moving for 80 seconds. The upper plot shows the horizontal (XY) position and the lower plot shows vertical position (Z) over time. In the plot it is evident that Xsens suffers from positional drift, even though the person is sitting with the feet stationary on the floor. Xsens reports a continuous change of data, with a total drift of more than 0.2 m during the 80 seconds capture session. Equivalent plots of other limbs show similar drift, hence there is little relative drift between body limbs.

This measurement shows that OptiTrack is better at providing accurate and precise position data in this type of clinical setup. However, for the vertical axis, we do not observe any major drift, but the Xsens data is still noisier than the OptiTrack data.

4.2.2 Positional accuracy - walking path

The left plot in Figure 3 displays the horizontal (XY) position of the head of a person walking along a rectangular path in a large motion capture area recorded with Xsens. The plot shows a horizontal positional drift of about 2 meters during the 90 seconds capture session. Xsens shows

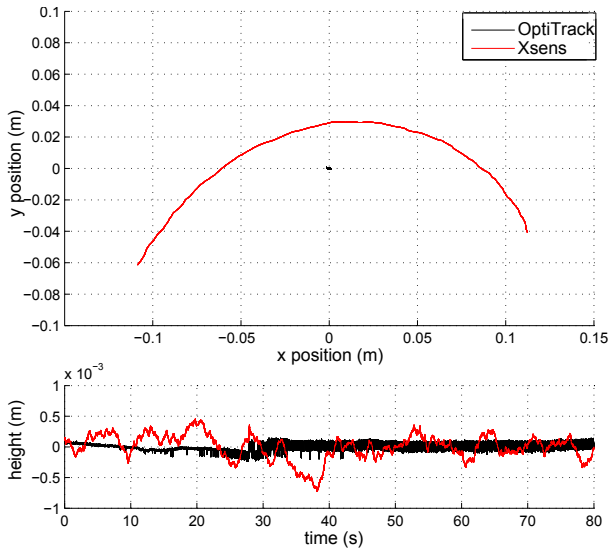


Figure 2. Horizontal and vertical plots of a stationary foot.

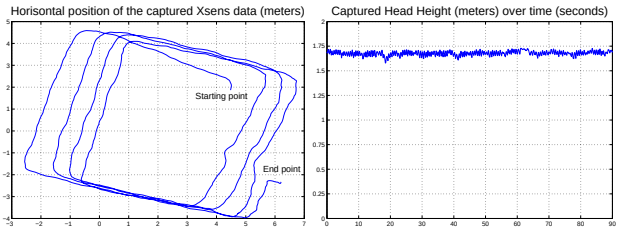


Figure 3. Recording of the horizontal (left) and vertical (right) position of the head.

no drift in the vertical direction (Z), as can be seen in the right plot. This is expected since the MVN engine maps the data to a human body model and assumes a fixed floor level. Because of the major horizontal drift we can conclude that Xsens MVN is not an ideal MoCap system if absolute horizontal position is needed.

4.2.3 Camera occlusion noise

The spatial resolution of an IrMoCap system mainly relies on the quality of the cameras and the calibration. The cameras have a certain resolution and field of view, which means that the spatial resolution of a marker is higher close to the camera than far away from the camera. The calibration quality determines how well the motion capture system copes with the transitions that happen when a marker becomes visible to a different combination of cameras. With a "perfect" calibration, there might not be a visible effect, but in a real situation we experience a clearly visible change in the data whenever one or more cameras fail to see the marker, as shown in Figure 4. When a marker is occluded from a camera, the 3D calculation will be based on a different set of 2D images.

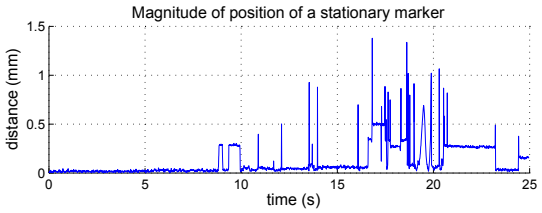


Figure 4. OptiTrack: Magnitude of the distance from the mean position of a stationary marker. The disturbances in the last part of the measurement is caused when a person moves around the marker, and thus blocks the marker in one or more cameras at a time. FrameRate 100 Hz

4.2.4 Xsens floor level change

If the motion capture area consists of different floor levels, like small elevated areas, the Xsens MVN engine will match the sensed raw data from the suit against the floor height where the suit was calibrated. This can be adjusted in post-processing, but real-time data will suffer from artifacts during floor level changes, as shown in Figure 5.

4.3 Acceleration and velocity data

In our experience, velocity and acceleration are highly usable motion features for controlling sound. High peaks in absolute acceleration can be used for triggering events, while velocity can be used for continuous excitation.

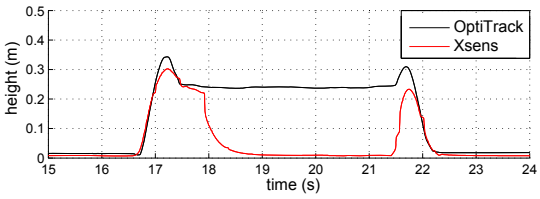


Figure 5. Recording of the vertical position of the left foot of a person, stepping onto an elevated area (around 0.25 m high). When the user plants his left foot on the object, the Xsens MVN engine will eventually map the stationary foot to floor level (18 to 19 s).

A difference between the two MoCap systems is that the Xsens system can offer velocity and acceleration data directly from the MVN engine [9]. When using the OptiTrack system we need to differentiate position data to estimate velocity and acceleration. If the positional data is noisy, the noise will be increased by differentiation (act as an high-pass filter), as we can see from Figure 6. The noise resulting from optical occlusion (see Section 4.2.3) is probably the cause for some of OptiTrack's positional noise.

Even though the Xsens position data is less accurate, it does offer smoother velocity and, in particular, acceleration data directly. We can use filters to smooth the data from the OptiTrack system; however, this will introduce a system delay, and hence increased latency.

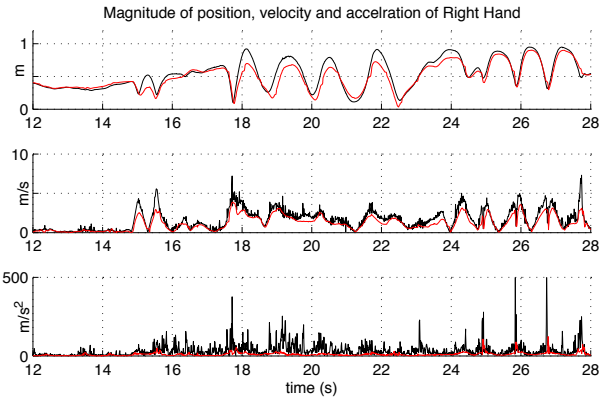


Figure 6. Velocity and acceleration data quality comparison (OptiTrack in black and Xsens in red).

4.4 Action-to-sound: latency and jitter

Low and stable latency is an important concern for *real-time* musical control [13], particularly if we want to use the system for triggering temporally accurate musical events. By *action-to-sound latency* we mean the time between the sound-producing action and the sonic reaction from the synthesizer.

To be able to measure the typical expected latency in a setup like that in Figure 7 we performed a simple experiment with an audio recorder. One computer was running one of the MoCap systems and sent OSC messages containing the MoCap information about the user's hands. A patch in Max/MSP was made that registered hand claps

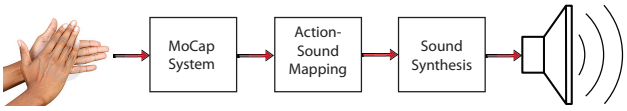


Figure 7. The acoustic hand clap and the triggered sound were recorded to measure latency of the systems.

based on MoCap data and triggered a *click* sound for each clap. The time difference between the acoustic hand clap and the triggered sound should indicate the typical expected latency for the setup.

Both MoCap systems were run on the same PC [1]. The sound-producing Max/MSP patch was run on a separate Mac laptop [2] and received OSC messages from the MoCap systems through a direct Gbit Ethernet link. All experiments used the same firewire connected sound card, *Edirol FA-101*, as output source. The hand claps and the click output from the Max patch was recorded with a microphone. Statistical results from the time delays between hand claps and corresponding click sound in the recorded audio files are given in Table 1. The values are based on 30 claps each. In this experiment, OptiTrack had a faster sound output response and a lower standard deviation than Xsens. The standard deviation is included as an indication of the jitter performance of the MoCap systems, since lower standard deviation indicates higher temporal precision.

Higher Xsens latency and jitter values are probably partly due to its use of Bluetooth wireless links. The Xsens MVN system also offers a direct USB connection option. We performed the same latency test with this option; and the results indicate that the connection is around 10-15 milliseconds faster, and has a lower jitter performance, than the Bluetooth link.

The upper bounds for "intimate control" have been suggested to be 10ms for latency and 1ms for its variations (jitter) [13]. If we compare the bounds with our results, we see that both systems have relatively large latencies. However, in our experience, a latency of 50ms is still usable in many cases. The high jitter properties of the Xsens system are probably the most problematic, especially when one wants high temporal accuracy.

	min	mean	max	std. dev.
OptiTrack	34	42.5	56	5.0
Xsens Bluetooth	41	52.2	83	8.4
Xsens USB	28	37.2	56	6.9

Table 1. Statistical results of the measured action-to-sound latency, in milliseconds.

4.5 Synthesizer control

In a second experiment, a musician was asked to perform simple music-related tasks with the two motion capture

[1] Intel 2.93 GHz i7 with 8GB RAM running Win 7
[2] MacBook Pro 10.6.6, 2.66 GHz Duo with 8GB RAM

systems. Three different control mappings to a sound synthesizer were prepared:

- Controlling pitch with the distance between the hands
- Triggering an impulsive sound based on high acceleration values
- Exciting a sustained sound based on the velocity of the hand

For the pitch mapping, the task was to match the pitch of one synthesizer to the pitch of another synthesizer moving in the simple melodic pattern displayed in Figure 8, which was repeated several times. This task was used to evaluate the use of position data from the two systems as the control data.

For the triggering mapping, the task was to follow a pulse by clapping the hands together. This task was given to evaluate acceleration data from the two systems as the control data, and to see if the action-to-sound latency and jitter would make it difficult to trigger events on time.

The excitation mapping was used to follow the loudness of a synthesizer, which alternated between ”on” and ”off” with a period of 1 second. This task was used to evaluate velocity data as control data.

The *reference sound* (the sound that the musician was supposed to follow) and the *controlled sound* (the sound that was controlled by the musician) were played through two different loudspeakers. The two sounds were also made with different timbral qualities so that it would be easy to distinguish them from each other. The musician was given some time to practice before each session. To get the best possible accuracy, both systems were used at their highest sampling rates for this experiment: Xsens at 120 Hz, and OptiTrack at 100 Hz.

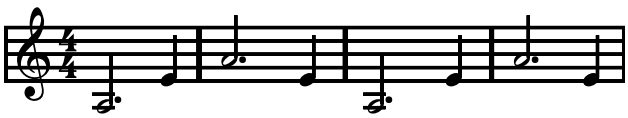


Figure 8. The simple melody in the pitch-following task. This was repeated for several iterations.

4.5.1 Pitch-following results

We found no significant difference between the performances with the two systems in the pitch-following task. Figure 9 displays an excerpt of the experiment, which shows how the participant performed with both Xsens and OptiTrack. The participant found this task to be difficult, but not more difficult for one system than the other. Also, the data shows no significant difference in the performances with the two systems. This indicates that the quality of relative position values (between markers/limbs) is equally good in the two systems for this kind of task.

4.5.2 Triggering results

Table 2 shows the results of the latency between the reference sound and the controlled sound for the triggering test. They are based on 40 hand claps for each of the two

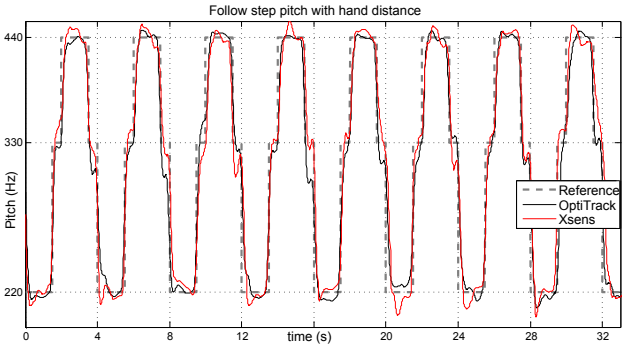


Figure 9. There was no significant difference between the two systems for the pitch-following task.

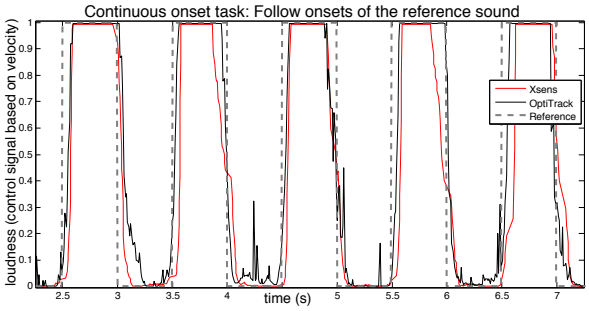


Figure 10. The major difference between the two systems in the continuous onset task was the noisy data from the OptiTrack system, which made it difficult to be quiet between the onsets. Apart from this, there was no big difference between the two systems.

MoCap systems. As we can see, the *mean* latency value is almost equal for Xsens and OptiTrack. Xsens has a higher standard deviation, which may indicate that the Xsens jitter shown in Table 1 makes it difficult for the user to make a steady trigger pulse.

	min	mean	max	std. dev.
OptiTrack	18.5	45.2	77.1	13.8
Xsens	2.6	44,7	96.3	28.3

Table 2. Statistical results, in milliseconds, of the measured time differences between reference signal and control signal.

4.5.3 Continuous onset results

For the continuous onset task, where the loudness of the sound was controlled by the absolute velocity of the right hand, we also observed a time delay between the onset of the reference tone and the onset of the sound played by our performer. This delay was present for both systems. In this task, the OptiTrack system suffered from noise, which was introduced when calculating the absolute velocity of the unfiltered OptiTrack data, as described in Section 4.3 (see Figure 10). The musician said that this made it more difficult to be quiet between the reference tones, and that this task was easier to perform with the Xsens system.

5. DISCUSSION

We have seen several positive and negative aspects with the quantitative measurements of the two technologies. In this section we will summarize our experiences of working with the two systems in a music-related context.

The main assets of the Xsens suit is its portability and wireless capabilities. The total weight of the suit is approximately 1.9 kg and the whole system comes in a suitcase with the total weight of 11 kg. Comparably, one could argue that a 8-camera OptiTrack setup could be portable, but this system requires tripods, which makes it more troublesome to transport and set up. OptiTrack is also wireless, in the sense that the user only wears reflective markers with no cables, but the capture area is restricted to the volume that is covered by the cameras, whereas Xsens can easily cover an area with a radius of more than 50 meters. When designing a system for real-time musical interaction based on OptiTrack, possible marker dropouts due to optical occlusion or a marker being moved out of the capture area must be taken into account. For Xsens, we have not experienced complete dropouts like this, but the Bluetooth link is vulnerable in areas with heavy wireless radio traffic, which may lead to data loss. Nevertheless, we consider Xsens to be the more robust system for on-stage performances.

OptiTrack has the benefit of costing less than most other motion capture technologies with equivalent resolution in time and space. The full Xsens suit is not comfortable to wear for a longer time period, whereas OptiTrack markers impose no or little discomfort. On the other hand, OptiTrack markers can fall off when tape is used to attach them. Also, OptiTrack's own solution for hand markers, where a plastic structure is attached to the wrist with Velcro, tends to wobble a lot, causing very noisy data for high acceleration movement, something we experienced when we set up the hand clapping tests. Xsens has a similar problem with the foot attachments of its sensors, which seems to cause positional artifacts.

Sections 4.2 to 4.5 show a number of differences between Xsens and OptiTrack. In summary, OptiTrack offers a higher positional precision than Xsens without significant drift, and seemingly also lower latency and jitter. Xsens delivers smoother data, particularly for acceleration and velocity. Our musician subject performed equally well in most of the musical tasks. However, the noisy OptiTrack data introduced some difficulties in the continuous onset task, and also made it challenging to develop a robust algorithm for the triggering task. Furthermore, Xsens jitter made the triggering task more difficult for the musician.

6. CONCLUSIONS

Both OptiTrack and Xsens offer useful MoCap data for musical interaction. They have some shared and some individual weaknesses, and in the end it is not the clinical data that matters, but the intended usage. If high positional precision is required, OptiTrack is preferable over Xsens, but if acceleration values are more important, Xsens provide less noisy data without occlusion problems. Overall, we find Xsens to be the most robust and stage-friendly MoCap system for real-time synthesis control.

7. REFERENCES

[1] E. R. Miranda and M. Wanderley, *New Digital Musical Instruments: Control And Interaction Beyond the Keyboard*. A-R Editions, Inc., 2006.

[2] J. Michel Couturier and D. Arfib, "Pointing fingers: Using multiple direct interactions with visual objects to perform music," in *Proc . NIME*, 2003, pp. 184–188.

[3] G. Castellano, R. Bresin, A. Camurri, and G. Volpe, "Expressive control of music and visual media by full-body movement," in *Proc. NIME*. New York, USA: ACM, 2007, pp. 390–391.

[4] F. Bevilacqua, J. Ridenour, and D. J. Cuccia, "3d motion capture data: motion analysis and mapping to music," in *Proc. Workshop/Symposium SIMS*, California, Santa Barbara, 2002.

[5] P.-J. Maes, M. Leman, M. Lesaffre, M. Demey, and D. Moelants, "From expressive gesture to sound," *Journal on Multimodal User Interfaces*, vol. 3, pp. 67–78, 2010.

[6] T. Cloete and C. Scheffer, "Benchmarking of a full-body inertial motion capture system for clinical gait analysis," in *EMBS*, 2008, pp. 4579 –4582.

[7] S. Thies, P. Tresadern, L. Kenney, D. Howard, J. Goulermas, C. Smith, and J. Rigby, "Comparison of linear accelerations from three measurement systems during reach & grasp," *Medical Engineering & Physics*, vol. 29, no. 9, pp. 967 – 972, 2007.

[8] S. A. Skogstad, A. R. Jensenius, and K. Nymoen, "Using IR optical marker based motion capture for exploring musical interaction," in *Proc. NIME*, Sydney, Australia, 2010, pp. 407–410.

[9] S. A. Skogstad, K. Nymoen, Y. de Quay, and A. R. Jensenius, "Osc implementation and evaluation of the xsens mvn suit," in *Proc of NIME*, Oslo, Norway, 2011.

[10] A. R. Jensenius, M. M. Wanderley, R. I. Godøy, and M. Leman, "Musical gestures: concepts and methods in research," in *Musical Gestures: Sound, Movement, and Meaning*, R. I. Godøy and M. Leman, Eds. New York: Routledge, 2010, pp. 12–35.

[11] C. Dobrian, "Aesthetic considerations in the use of 'virtual' music instruments," in *Proc. Workshop on Current Research Directions in Computer Music*, 2001.

[12] D. Rosenberg, H. Luinge, and P. Slycke, "Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors," *Xsens Technologies*, 2009.

[13] D. Wessel and M. Wright, "Problems and prospects for intimate musical control of computers," in *Proc. NIME*, Seattle, USA, 2001.

Paper III

Comparing Motion Data from an iPod Touch to a High-End Optical Infrared Marker-Based Motion Capture System.
K. Nymoen, A. Voldsund, S.A. Skogstad, A.R. Jensenius, and J. Torresen.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 88–91, University of Michigan 2012.

Comparing Motion Data from an iPod Touch to an Optical Infrared Marker-Based Motion Capture System

Kristian Nymoen,[1] Arve Voldsund,[1,2] Ståle A. Skogstad,[1]
Alexander Refsum Jensenius,[2] and Jim Torresen[1]
[1]Department of Informatics, University of Oslo, Norway
[2]Department of Musicology, University of Oslo, Norway
{krisny,savskogs,jimtoer}@ifi.uio.no,{arve.voldsund,a.r.jensenius}@imv.uio.no

ABSTRACT

The paper presents an analysis of the quality of motion data from an iPod Touch (4th gen.). Acceleration and orientation data derived from internal sensors of an iPod is compared to data from a high end optical infrared marker-based motion capture system (Qualisys) in terms of latency, jitter, accuracy and precision. We identify some rotational drift in the iPod, and some time lag between the two systems. Still, the iPod motion data is quite reliable, especially for describing relative motion over a short period of time.

1. INTRODUCTION

With advances in mobile technology during the last years, mobile devices have become increasingly popular for musical interaction. In this paper we will focus on Apple's iOS devices, which come with a variety of sensors, depending on the type and model: touch screen, accelerometer, gyroscope, GPS, and magnetometer. Additionally, pre-processed data extracted from the raw sensor data, e.g. orientation and acceleration, is made available through the iOS SDK.

The motivation for the present study is to learn more about the quality of the motion data from an iPod Touch. Several researchers have reported on strengths and weaknesses of iOS devices, e.g. [9, 11], but, these are rarely quantified. In order to know how precisely a motion feature can be reproduced, how fast an action can be recognized, and so forth, we need quantitative evaluations of the data.

Some musical parameters may require high precision and accuracy, while other parameters do not, and with the proper knowledge about the quality of the iPod data, we can make more qualified decisions when mapping motion parameters to musical parameters. This paper evaluates data from an iPod Touch by comparing it to data from a state-of-the-art optical marker-based motion capture (mocap) system from Qualisys, through analyses of timing (i.e. latency and jitter), as well as accuracy and precision, hereunder drift and noise of orientation and acceleration data.

2. BACKGROUND

In the last decade or so, we have seen an increased interest of mobile phones for musical applications in the NIME community and elsewhere. PDAs [18] and Nokia phones [7] have been used, in addition to the increasing number of applications developed for iOS devices in the last years, e.g. [4, 12, 19]. Recently, mobile devices have also become the main computing platform of certain formalised ensembles, e.g. [14].

Several general purpose environments for working with music on mobile phones have been developed, including versions of PureData for PDAs [10], mobile phones [15], and the *libpd* port of PureData to iOS and Android [2]. Moreover, the Synthesis ToolKit (STK) has been ported to Symbian [6], and iOS [3], and the urMus environment has been designed for rapid design of mobile musical instruments [5].

In [8], Essl and Rohs present a design space based on using sensors on mobile devices for developing musical instruments. They emphasise the importance of considering the properties of the sensors at hand. Specifically for gyroscopes and accelerometers, which are used in the iPod Touch discussed in the present paper, they mention that these sensors are good for measuring relative motion, but that the lack of an absolute frame of reference makes absolute measurements difficult. Through the experiments presented in the next chapter, we have aimed to quantify such measures.

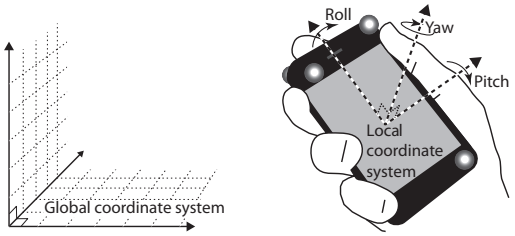


Figure 1: The iPod, defined as a rigid body, enables Qualisys tracking of orientation and position.

3. EXPERIMENT

We have used data from a Qualisys optical infrared marker-based mocap system as a reference when evaluating the iPod data. Our setup consisted of 9 Oqus 300 cameras, operating at a sampling rate of 150 Hz. The system is reported to have a high spatial resolution. However, this resolution depends on the distance between the object that is being captured and the mocap cameras, in addition to the calibration quality [17].

The iPod (Figure 1) was equipped with four reflective markers ($\varnothing$ = 12 mm). The configuration of the markers was used to define the iPod as a rigid object, with centre position at the geometric centre of the markers. In this manner, we used the optical motion capture system to record the position and the orientation of the iPod.

3.1 iPod

We used an iPod Touch, 4th generation, running iOS version 4.3.5, for the experiment. The device contains a three-

NIME'12, May 21 – 23, 2012, University of Michigan, Ann Arbor.

axis accelerometer and gyroscope, which is used to calculate certain motion features on the iPod. We have not used the raw data values from the accelerometer and gyroscope, but rather utilised the motion features that are available through the `CMDeviceMotion` class in the iOS Developer library [1]: *attitude*, *rotationRate*, *gravity*, and *userAcceleration*. The reason for using these features is that they are intended to be conceptually similar to the data provided by the Qualisys system, as opposed to what raw sensor data (e.g. from an accelerometer) would be.

We have developed an application for accessing these data from the iPod. The motion features were sampled at 60 Hz, and packed into OpenSound Control (OSC) bundles. These were sent via UDP over a wifi network set up by the recording computer. The 60 Hz sampling rate was set in the iPod application at the development stage, and other sampling rates have not been tried in this paper.

3.2 Recordings

The data from the Qualisys system was sent as OSC bundles via UDP over Ethernet to the recording computer. The iPod data and Qualisys data were recorded in Max5, as separate streams in an SDIF file, to obtain synchronized recordings of the motion data [13]. The recorded data types are presented in Table 1, these were provided natively from the devices. In this table, *global* means that the data stream is given in relation to some global, fixed, coordinate system, and *local* means that the data stream is measured in relation to the local coordinate system of the iPod (Figure 1).

The iPod was held in one hand, and a total of 22 recordings were made. These included short recordings of tilting the iPod around each of the rotational axes individually, as well as longer, spontaneous rotational and shaking gestures (durations ≈ 4–23 seconds). Furthermore, a ten minute recording was made where the iPod was lying still. Orientation was recorded both as Euler angles and 3×3 Direction Cosine Matrix (DCM).[1] Since the coordinate systems from the iPod and Qualisys were not aligned, the iPod orientation data was adjusted by hand to match the Qualisys data during postprocessing.

Table 1: Recorded motion data

Qualisys		iPod	
Orientation	Global	Orientation	Global
Position	Global	User Acceleration	Local
Marker pos.	Global	Gravity	Local
		Rotation rate	Local

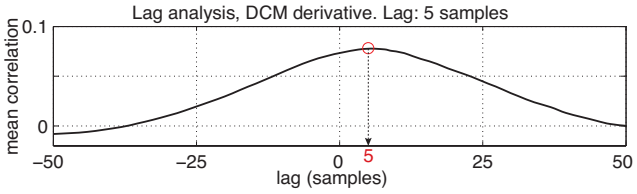
4. ANALYSIS

We start the data analysis by looking at issues related to timing, including latency and jitter. Subsequently, we move on to accuracy and precision of rotational and positional data. For the analysis, we selected a subset of the recordings where there were no gaps in the motion capture data (i.e. the rigid body was tracked at every frame). The results presented in this section are discussed in Section 5.

4.1 Timing

4.1.1 Lag

We observed a time lag between the data from Qualisys and the iPod. To analyse this, we performed cross-correlation on the derivatives of the DCM elements, for eight recordings. Cross-correlation measures the similarity between the two data streams as a function of a time lag applied to one of the streams. Using the derivatives removes inaccurately high correlation scores of stationary extreme-value elements. To achieve an equal number of samples in the data streams, the iPod data was up-sampled to 150 Hz using cubic spline interpolation before the derivative was calculated. By averaging the cross correlations, we achieved an estimate of the time lag between Qualisys and the iPod for each recording, as shown for one recording in Figure 2, the figure also shows that for the eight recordings, the mean lag between Qualisys and iPod data was 43 ms, standard deviation (SD) 8 ms.



	mean	SD	min	max
Lag (ms)	43	8.0	33	53

Figure 2: The plot shows the averaged cross-correlation between the DCM elements of the iPod versus Qualisys for one recording. In this recording, the lag is 5 samples (∼33 ms). The table below shows lag statistics for 8 recordings. Qualisys and iPod correlation is highest when shifted by 43 ms.

4.1.2 Jitter

For applications where the iPod sensor data is sent to an external device, it can be crucial that the timing of received data packets is stable. To evaluate the temporal stability of the system, we measure jitter, as the variation in the time interval between received OSC bundles, in a sequence of 1000 samples. Figure 3 shows a histogram and statistics of the time intervals between successive samples. The standard deviations give indications of the amount of jitter in the data streams. This measure is high for both systems, suggesting that variations in the network connections between the sensing devices and the receiving computer might be partly responsible for this. Still, the standard deviation is notably higher for the iPod than for the Qualisys system, suggesting that the iPod is less reliable when it comes to delivering data packets at regular time intervals.

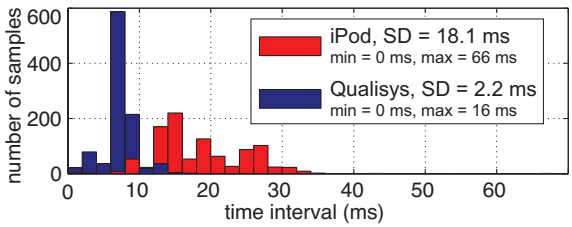


Figure 3: Histogram of the time interval between 1000 successive received samples.

4.2 Accuracy and Precision

4.2.1 Orientation Data

It has been shown that Spearman's rank correlation is suitable for comparing data with serial correlation [16]. We applied this to the 9 DCM elements of the iPod and Qualisys to analyse accuracy of the orientation data. Again, a cubic spline was used to upsample the iPod data, and the data was time-shifted and trimmed according to the iPod lag, as described in Section 4.1.1.

Figure 4 shows a histogram of the correlation coefficients for the 9 DCM elements for 8 recordings. 2/3 of the correlation coefficients are above 0.96, which indicates that in

[1] Rotation Matrix / Orientation Matrix

general, the iPod reproduces the "true" orientation of the device satisfactorily. A few of the elements in the histogram have low correlation coefficients. This may be explained by a low variance in the particular DCM element, which again causes a poor signal-to-noise ratio. The 8 recordings involved simple rotations around single axes, as well as composite rotations, with durations between 4 and 10 seconds.

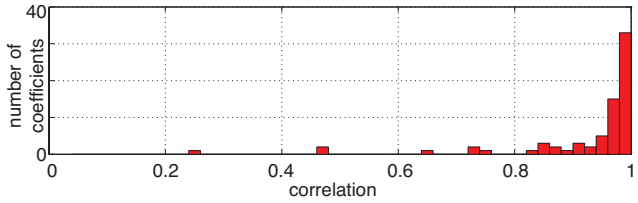


Figure 4: Histogram of correlation coefficients when correlating the orientation coordinates (DCM) of the iPod to Qualisys.

To analyse rotational drift, a gradient for each of the Euler angle coordinates was extracted by linear regression of the 10 minute recording of the iPod lying still in the motion capture space. A small amount of drift was observed in the orientation data. This, together with analysis of the rotational noise is shown in Table 2. The noise measurements are the RMS level of the Euler coordinates (in degrees), after subtracting the average drift, and centering around the mean value (removing offset). Note that compared to the Qualisys system, the iPod performs quite well when it comes to *roll* and *pitch*, with superior drift performance, and equal noise level, but the *yaw* measurements from the iPod are less accurate and less precise. The average yaw drift of a still recording is 70.6×10^{-5} deg/s which is equivalent to a drift of 2.5 deg/h. An additional effort to force the device to give inaccurate yaw data by shaking the device violently for 23 s, resulted in a yaw drift of 11.5 deg.

Table 2: Rotational drift and noise (Euler, degrees)

	Drift (10^{-5}deg/s)			Noise, RMS (= SD)		
	Roll	*Pitch*	*Yaw*	*Roll*	*Pitch*	*Yaw*
iPod	-0.61	1.05	70.6	0.028	0.018	0.153
Qualisys	-17.2	7.24	8.95	0.029	0.018	0.010

4.2.2 Acceleration

Acceleration data from the iPod is limited by the range of the accelerometer that provides the data. Apple has not officially released these specifications for the iPod, but it can quite easily be measured. Since the raw accelerometer data (which includes the gravity vector) and the user acceleration values are not identical, the range of acceleration values depends on the current orientation of the device. A recording of heavy shaking of the iPod provided maximum and minimum values of acceleration in the range -29 m/s^2 to $+29$ m/s^2, which is equivalent to ±3 g.

Table 3 shows acceleration data statistics for the 10 minute recording of the iPod lying still. The table shows high standard deviations and max/min values for unfiltered Qualisys data. This is because even small noise in the position data will become large when the derivative is calculated [17]. However, a filtered version, using a simple 5 sample averaging filter on each derivative level significantly improves this. As shown, the iPod has a certain offset in this data, even though internal processing on the device is supposed to remove the gravity component in the acceleration data. The standard deviations from the iPod are slightly higher than the filtered Qualisys data.

Figure 5 shows that the acceleration data from the two systems match well (Qualisys is filtered as mentioned above). This will be discussed more in the next section.

Table 3: iPod acceleration noise, unit: 10^{-3} m/s^2

	mean	SD	min	max
iPod X	5.3	18.5	-71.1	84.8
Y	0.7	15.8	-67.9	61.8
Z	160.7	22.6	33.9	303.8
Qualisys X	0.005	261.5	-1613	1637
unfiltered Y	0.001	272.4	-2431	2236
Z	0.003	358.1	-2502	2745
Qualisys X	0.000	10.4	-49.0	71.3
filtered Y	0.000	12.3	-73.3	61.3
Z	0.000	16.7	-77.6	87.3

4.2.3 Position, Velocity, and Acceleration

By integrating the acceleration data from the iPod, and differentiating the position data from Qualisys, we have estimated the accelerations, velocities and the positions measured by the two systems. Acceleration values from the iPod are given in local coordinates (cf. Section 3.2), while the second derivative of Qualisys position data provides acceleration in a global coordinate system. Hence, the iPod acceleration vector was transformed to a global coordinate system. This means that any orientational drift also influenced calculations of position.

Figure 5 shows an example of a short recording containing a simple vertical translation followed by a pitch rotation combined with vertical translation. The figure shows some drift in velocity, and a lot of drift in position. The figure also shows an attempt to correct for the positional drift through filtering, but long filters can induce unacceptable amounts of delay. In the figure, a 100 samples FIR filter is used, which corrects for some of the drift, but in most real-time settings a filter of this length would cause too much latency.

Figure 6 shows similar plots of the 10 minute still recording. There was a small offset of 0.16 m/s^2 in the iPod acceleration data, which was removed before estimating velocity and position. Even after removing the offset, the drift is significant. After one minute, the error of the position estimate is more than 1 m, and after 10 minutes, it is more than 60 m.

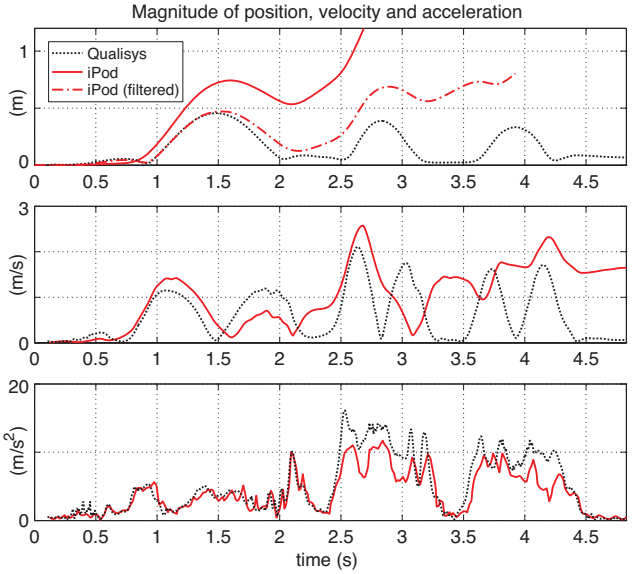


Figure 5: Plots of a short motion sequence, with magnitude of position, velocity and acceleration for iPod and Qualisys data. The filtered version of iPod position data has been time-shifted forward by 51 samples to compensate for filter latency.

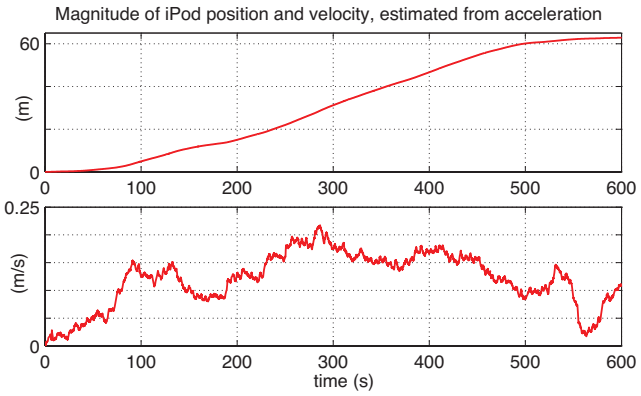


Figure 6: Plots of the magnitude of estimated position and velocity of the iPod lying still for 10 min.

5. DISCUSSION

We have presented how the motion data from an iPod compares to data from a high-end motion capture system. These results influence how we use the iPod motion data in music applications. Our analysis of lag in orientation data showed that there is an average of 43 ms between the time when an orientation is measured in the Qualisys system and when it is measured on the iPod. There may be several reasons for this; the different sampling rates of the two systems might have played some role, but we find it reasonable to assume that the processing done on the iPod to extract the orientation from the sensor raw data is the main cause. This means that it might be unsuitable to use orientation to control musical features that demand high temporal precision.

In addition to the lag, orientation data was evaluated in terms of accuracy and precision. For *roll* and *pitch* coordinates, the accuracy and precision are high, and sufficient for continuous control of sound. *Yaw*, on the other hand, does not show equally good results, and should be used with caution. The drift is still low enough to assume that it is suitable for measuring relative rotations over short time periods. In future work, it would be interesting to compare this with newer iPhone models which contain magnetometers.

The data jitter from the iPod is significantly higher than for Qualisys, despite the fact that the iPod sent less data at a lower sampling rate. This might be important to consider if the iPod is used for direct control of sound on a separate computer. The jitter could be compensated for by buffering, but this again would cause increased latency.

As expected, our attempt to estimate the position of the iPod from the acceleration data resulted in large errors, since the noise propagates a lot when the signal is integrated. Still, we notice that some positional features can be inferred from the iPod acceleration data. Especially for shorter segments, it is possible to tell when the iPod is moved in one plane, but the estimates are too imprecise to estimate when the device reaches back to the starting position. As seen in the lower plot in Figure 5, the acceleration data from the iPod is quite responsive, and is well suited for controlling musical parameters that require high temporal precision.

6. ACKNOWLEDGMENTS

This research has received funding from EU FP7 under grant agreement no. 257906 (EPiCS), and the Norwegian Research Council, project no. 183180 (SMA).

7. REFERENCES

[1] Apple Inc. iOs Developer Library, CMDeviceMotion. `http://developer.apple.com/library/IOs/#documentation/CoreMotion/Reference/CMDeviceMotion_Class/Reference/Reference.html`.

[2] P. Brinkmann, P. Kirn, R. Lawler, C. McCormick, M. Roth, and H.-C. Steiner. Embedding pure data with libpd. In *Proceedings of the Pure Data Convention*, Weimar, Germany, 2011.

[3] N. J. Bryan, J. Herrera, J. Oh, and G. Wang. MoMu: A mobile music toolkit. In *Proc. of Int. Conf. New Interfaces for Musical Expression*, pages 174–177, Sydney, 2010.

[4] N. J. Bryan and G. Wang. Two turntables and a mobile phone. In *Proc. of Int. Conf. New Interfaces for Musical Expression*, pages 179–184, Oslo, 2011.

[5] G. Essl. UrMus – An environment for mobile instrument design and performance. In *Proc. of the International Computer Music Conference*, pages 270–277, New York, 2010.

[6] G. Essl and M. Rohs. Mobile STK for Symbian OS. In *Proc. of the International Computer Music Conference*, pages 278–281, New Orleans, 2006.

[7] G. Essl and M. Rohs. ShaMus – A sensor-based integrated mobile phone instrument. In *Proc. of International Computer Music Conference*, pages 27–31, Copenhagen, 2007.

[8] G. Essl and M. Rohs. Interactivity for mobile music-making. *Org. Sound*, 14:197–207, 2009.

[9] G. Essl, G. Wang, and M. Rohs. Developments and challenges turning mobile phones into generic music performance platforms. In *Proc. of Mobile Music Workshop*, pages 11–14, Vienna, 2008.

[10] G. Geiger. PDa: Real time signal processing and sound generation on handheld devices. In *Proc. of Int. Computer Music Conference*, Singapore, 2003.

[11] A. R. Jensenius. Some challenges related to music and movement in mobile music technology. In *Proc. of Mobile Music Workshop*, pages 19–22, Vienna, 2008.

[12] N. Kruge and G. Wang. MadPad: A crowdsourcing system for audiovisual sampling. In *Proc. of Int. Conf. New Interfaces for Musical Expression*, pages 185–190, Oslo, 2011.

[13] K. Nymoen and A. R. Jensenius. A toolbox for storing and streaming music-related data. In *Proc. of Int. Sound and Music Computing Conference*, pages 427–430, Padova, 2011.

[14] J. Oh, J. Herrera, N. J. Bryan, L. Dahl, and G. Wang. Evolving the mobile phone orchestra. In *Proc. of Int. Conf. New Interfaces for Musical Expression*, pages 82–87, Sydney, 2010.

[15] G. Schiemer and M. Havryliv. Pocket Gamelan: a PureData interface for mobile phones. In *Proc. of Int. Conf. New Interfaces for Musical Expression*, pages 156–159, Vancouver, 2005.

[16] E. Schubert. Correlation analysis of continuous emotional response to music. *Musicae Scientiae*, Special issue 2001–2002:213–236, 2002.

[17] S. A. Skogstad, K. Nymoen, and M. E. Høvin. Comparing inertial and optical mocap technologies for synthesis control. In *Proc. of Int. Sound and Music Computing Conference*, pages 421–426, Padova, 2011.

[18] A. Tanaka. Mobile music making. In *Proc. of Int. Conf. New Interfaces for Musical Expression*, pages 154–156, Singapore, 2004.

[19] G. Wang. Designing Smule's Ocarina: The iPhone's magic flute. In *Proc. of Int. Conf. New Interfaces for Musical Expression*, pages 303–307, Pittsburgh, 2009.

Paper IV

SoundSaber — A Motion Capture Instrument.
K. Nymoen, S.A. Skogstad and A.R. Jensenius.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 312–315, University of Oslo 2011.

SoundSaber - A Motion Capture Instrument

Kristian Nymoen, Ståle A. Skogstad
fourMs group - Music, Mind, Motion, Machines
Department of Informatics
University of Oslo, Norway
{krisny, savskogs}@ifi.uio.no

Alexander Refsum Jensenius
fourMs group - Music, Mind, Motion, Machines
Department of Musicology
University of Oslo, Norway
a.r.jensenius@imv.uio.no

ABSTRACT

The paper presents the SoundSaber - a musical instrument based on motion capture technology. We present technical details of the instrument and discuss the design development process. The SoundSaber may be used as an example of how high-fidelity motion capture equipment can be used for prototyping musical instruments, and we illustrate this with an example of a low-cost implementation of our motion capture instrument.

1. INTRODUCTION

We introduce the SoundSaber, a musical instrument based on optical infrared marker-based motion capture technology. *Motion capture* (mocap) involves recording motion, and translating it to the digital domain [10]. *Optical* motion capture means that the system is based on video cameras, and we distinguish between *marker-based* and *markerless* systems which work without markers. We will refer to musical instruments based on optical motion capture as *mocap instruments*.

Optical infrared marker-based mocap technology is superior to most other methods of motion capture with respect to temporal and spatial resolution. Some systems can track markers at a rate of more than 1000 frames per second, and in most cases they provide a spatial resolution in the sub-millimeter range. On the other hand, this technology is expensive, and better suited for laboratory use than for stage performances. A wide range of other less expensive and portable mocap technologies exists, like accelerometer-based sensor systems and computer vision. These provide different types of data, usually with lower frame rate and spatial resolution than optical infrared mocap.

A large amount of the research that is done in our lab involves the exploration of motion capture systems for musical interaction, ranging from high-end technologies to solutions like web-cameras and accelerometers. This involves studies of the different technologies separately, and also experiments on how the experience from interactive systems based on high-end mocap technology can be transferred to low-cost mocap technologies.

We present the SoundSaber as an example of how a seemingly simple sound synthesiser may become interesting through the use of high quality motion capture technology and an intuitive action-sound model. With a system that is able to register very subtle motion at a high sampling rate, it is possible to create an instrument that comes close to the control intimacy of acoustic instruments [11]. These ideas are presented through reflections that have been made while developing the instrument. Included in the presentation are some thoughts and experiences from how optical motion capture technology can be used to prototype new interfaces for musical expression.

In Section 2 we lay out a general theory on digital musical instruments and use of mocap for sound generation. Section 3 presents the SoundSaber, including considerations and evaluations that have been made in the process of development. In Section 4 we illustrate how the instrument was "ported" to another technology and compare the results to the original SoundSaber. Section 5 provides conclusions and directions for future work.

NIME'11, 30 May–1 June 2011, Oslo, Norway.

2. MOCAP INSTRUMENT CONTROLLERS

Most digital musical instruments consist of a controller with sensors, a sound synthesiser, and a defined *mapping* between the control data from the sensors and the input parameters of the synthesiser [5]. Mocap instruments are slightly different in that the controller is separate from the sensor technology. This distinction between the sensors and the controller present an interesting opportunity because almost any object can be used to communicate with the mocap system: a rod, a hand, an acoustic instrument, etc.

This makes it possible to try out objects with different physical properties and shapes, hence also different *affordances*. In design literature, the affordance of an object is a term used to describe the perceived properties of how this object could possibly be used [6]. For an object used in a mocap instrument, the affordance may refer to a "pool" of different control actions that could be associated with it, e.g. whether it should be held with one or both hands. Following this, physical properties of the object, such as size, inertia, etc., will also influence how it can be handled. The possibility of quickly swapping objects may be a useful tool for prototyping new digital musical instruments.

The data from the motion capture system can be processed in several ways, see [1] and [10] for discussion on how motion capture data can be mapped to musical parameters. The GrainStick installation at IRCAM used mocap technology to generate sound in yet another way, using the metaphor of a virtual rainstick being held between two objects [4]. Our choices for data processing in the SoundSaber will be presented in Sections 3.2 to 3.5.

3. THE SOUNDSABER

The different components of the SoundSaber are illustrated in Figure 1. The position of the controller is captured by the motion capture system, which sends position data to a

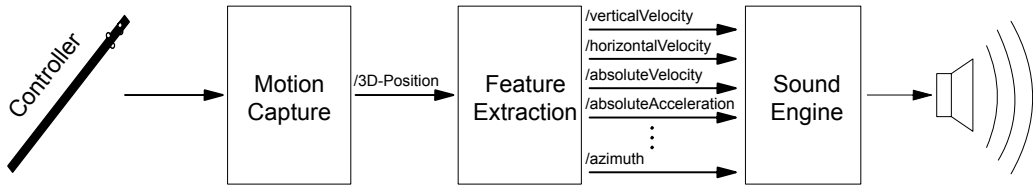


Figure 1: The figure shows the different parts of the SoundSaber instrument from controller, via motion capture technology and feature extraction to a sound synthesiser.

Max/MSP patch that calculates various features from the data. The features calculated by this patch are mapped to various control parameters in a synthesiser.

One of the advantages of digital musical instruments is that it is simple to try out different technologies for each part of the instrument. In our own work, we have experimented with different motion capture systems and controllers. Currently, we have two versions of the SoundSaber: The original one, based on optical motion capture, and another wii-controller (wiimote) implementation.

We will start the presentation of the SoundSaber by describing the controller, followed by a presentation of the motion capture technology, feature extraction and the synthesiser. Even though the different parts of the instrument are presented separately, they have been developed together, both simultaneously and iteratively.[1]

3.1 The controller

The SoundSaber controller that we are currently using is a rod, roughly 120 cm in length with a diameter of 4 cm, and is shown in Figure 2. Four markers are placed in one end of the rod, and the motion capture system recognizes these as a single *rigid object*, tracking position and orientation of the tip of the rod. The rod is heavy enough to give it a reasonable amount of inertia, and at the same time light enough so that it does not feel too heavy, at least not when it is held with both hands. The shape and mass of the rod also make it natural to perform large and smooth actions. We have observed that the majority of people who have tried the instrument performed gestures that imitate fencing. The reason for this may be their association of these gestures with the name of the instrument in combination with the physical properties and affordance of the controller.

Figure 2: The SoundSaber controller

3.2 Motion capture

We have been using different motion capture systems for the SoundSaber. Initially we used an 8-camera OptiTrack system from NaturalPoint, which can stream real-time data at a rate of 100 Hz. The OptiTrack software uses the proprietary NatNet protocol for data streaming. We used a client developed by Nuno Diniz at IPEM in Ghent for translating NatNet data to Open Sound Control (OSC) over UDP. OSC simplifies the communication between the motion capture system and the layers for feature extraction, mapping and sound synthesis.

More recently, we have been using a high-end motion capture system from Qualisys. This system has a higher spatial resolution than OptiTrack, and it is able to stream data at higher sampling rates. The Qualisys system also has native support for Open Sound Control.

[1]For video examples of the SoundSaber, please visit `http://www.youtube.com/fourmslab`

3.3 Feature extraction

We have implemented a tool in Max/MSP for real-time feature extraction from position data. Our approach is similar to the Motion Capture Music toolbox, developed by Dobrian et al. [1], with some differences. Our tool is structured as one single module, and outputs data as OSC messages. OSC formatting of these features simplifies the mapping between the motion features and the control features in the synthesiser.

Thus far, difference calculations, dimensionality reduction and transformations between different coordinate systems have been implemented. Based on a three-dimensional position stream the patch calculates:

- Velocity in a single direction, e.g. vertical velocity
- Velocity in a two-dimensional subspace, e.g. horizontal velocity
- Absolute velocity, as the vector magnitude of the three velocity components
- Change in absolute velocity
- Acceleration in a single direction
- Absolute acceleration
- Polar equivalent of the cartesian input coordinates, providing horizontal angle, elevation, and distance from the origin

3.4 Sound synthesis

As the name *SoundSaber* suggests, we initially had an idea of imitating the sound of the lightsaber from the Star Wars movies. The development of the synthesiser was more or less a process of trial and error to find a sound that would have some of the perceptual qualities that are found in the lightsaber sound.

The SoundSaber synthesiser is implemented in Max/MSP. Figure 3 shows a schematic illustration of the synthesiser, where a pulse train (a sequence of impulses or clicks) with a frequency of 1000 Hz is sent through two delay lines with feedback loops. The delay times for the delay lines can be adjusted by the user, resulting in variations in harmonic content. Furthermore, the output from the delay lines is sent to a ring modulator where it is modulated by a sinusoidal oscillator. The user can control the frequency of this oscillator in the range between 40 and 100 Hz. The ring modulated signal and the output from the delay lines are added together and sent through an amplitude control, then another feedback delay line and finally through a bandpass filter where the user controls bandwidth and frequency.

3.5 Mapping

Several considerations have been made regarding the action-sound relationship in the SoundSaber. Naturally, we have

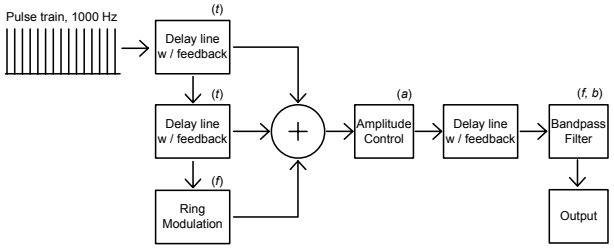


Figure 3: The SoundSaber synthesiser. Letters in parentheses denote user-controllable parameters: (t) = time, (f) = frequency, (a) = amplitude and (b) = bandwidth.

not been limited to mimicking action-sound relationships from traditional musical instruments, but at the same time we do appreciate some of the constraints that acoustic instruments provide. For instance, the sound of an instrument is almost always the result of an energy transfer from a sound-producing action to mechanical vibrations.

Since our approach to the physical design of this instrument has been simple, using only the position of a single point on the controller as the basis for feature extraction, we have chosen a simple approach when mapping motion features to control parameters. This is what Hunt and Wanderley call *explicit mapping*, meaning direct couplings between motion features and control parameters in the sound synthesiser [2].

When designing mapping for a mocap instrument, it is important to understand what the motion features actually describe. Motion features calculated from a stream of data describing position of a controller in a room can be one (or a combination) of the following:

Relative to the room meaning that the axes of the room influence the motion feature. An example of this is the vertical velocity component of the motion.

Relative to the controller itself typically referring to difference calculations, e.g. the absolute velocity.

Relative to another controller describing how the controller relates to other controllers in the room. For instance the distance to another SoundSaber.

In the current SoundSaber implementation, we have only used data that describes the controller in relation to the room or to itself. But we believe that the perspective of how the instrument relates to other controllers presents interesting possibilities in making collaborative musical instruments.

One of the considerations we have made is regarding motion in the horizontal plane. Should it make a difference whether the instrument is being moved along the X-axis or the Y-axis? In our opinion, the SoundSaber should respond equally whether the musician is on the left side or the right side of a stage, and also behave in the same manner no matter which direction the performer is facing, as is the case for any hand-held acoustic instrument. Therefore we reduced the two dimensions of horizontal velocity to a single absolute horizontal velocity, and let this mapping govern one of the timbral control parameters in the synthesiser (the delay time of the first delay line).

Vertical motion, on the other hand, is different. Our previous experiments have shown that people tend to relate vertical motion to changes in frequency, such as changes in pitch and spectral centroid [7, 8]. No matter which direction the performer is facing, gravity will act as a natural reference. In light of this, we have chosen to let the vertical position control the frequency of the ring modulation and the bandpass filter, and the vertical velocity control the delay time of the second delay line.

Another action-sound relationship which has been confirmed in our previous experiments, is the correspondence between velocity and loudness [7]. Hunt and Wanderley noted that increased input energy is required to increase sound energy in acoustic instruments, and received better results for a digital musical instrument where users had to feed the system with energy to generate sound, rather than just positioning a slider to adjust sound level [2]. With this in mind, we wanted an increase in kinetic energy to result in an increase in sound energy. Therefore, we let the absolute velocity control the amplitude of the synthesiser.

We have implemented a simple mapping for sound spatialisation. Spatial sound is obviously related to the room, so we used motion features related to the room in this mapping. More specifically, we sent the polar position coordinates of the rod to a VBAP control system [9], so the musician can control sound spatialisation by pointing the SoundSaber towards different loudspeakers.

3.6 SoundSaber evaluation

Neither the feature extraction, the explicit mapping strategy, nor the synthesiser of the SoundSaber are particularly sophisticated or novel by themselves. At the same time, after observing how people interact with the instrument, we feel confident to say that such interaction is engaging for the user. We believe that the most important reason for this are the considerations that were made to obtain a solid coupling between control actions and sound.

In addition to the rod, we tried using three other objects for controlling the SoundSaber synthesiser. For two of these, we simply changed the rod with another object and used the same motion capture technology, meaning that the only difference was the object itself. First, we tried a small rod, which was best suited for single-hand use, and also had less inertia and thus higher mobility. Second, we tried using a small handle with markers. This handle reduced the distinction between the controller and the performer, because the motion of the controller was basically equal to the hand motion of the performer. Both of these solutions were less satisfying than the large rod because the loudness control in the synthesiser had a fairly long response time, making it more suitable for controllers with more inertia. Also, the deep and full sound of the SoundSaber works better with a larger object. Third, as mentioned above, we made an implementation of the SoundSaber using a Nintendo Wii controller which will be discussed in more detail below.

Furthermore, we believe that the considerations of how motion features related to sound were important. The use of vertical position (which is only relative to the room) to adjust spectral centroid via a bandpass filter, and of absolute velocity (which is only relative to the object itself) to control loudness appeared to work well.

Nevertheless, the complexity of the control input, and the motion capture system's ability to capture motion nuances are perhaps the most important reasons why it is engaging to interact with the SoundSaber. Even though separate motion features were selected and mapped to different control parameters, the motion features themselves are related to each other. As an example, consider what happens when the performer makes a change in vertical position of the rod to adjust the spectral centroid. This action will also imply a change in the motion features "vertical velocity" and "absolute velocity".

When the spatial and temporal resolution of the motion

capture system is high, the instrument responds to even the smallest details of the performer's motion. For a reasonably sized object like the SoundSaber rod, we are satisfied with a spatial resolution of 1 mm and a frame rate of 100 Hz, but for smaller and more responsive objects we might require even higher resolution to capture the nuances of the actions these objects afford.

4. TOWARDS PORTABILITY

Because of the expensive hardware, the implementation of the SoundSaber based on optical motion capture is not available to everyone. One motivation for this research is to make instruments that are based on high-end technology available to a broader audience. Thus, we need less expensive and preferably also more portable solutions.

Of the many affordable sensor solutions, we chose to use a Nintendo wii-controller (wiimote) for our low-cost implementation. The wiimote provides a different set of control possibilities than optical motion capture, and the major challenges with porting the SoundSaber to the wiimote are related to processing the data from the controller and mapping strategies. A survey by Kiefer et al. ([3]) showed that the wiimote could be well suited for continuous control, which makes it an interesting test case for the SoundSaber.

4.1 Wiimote implementation

We used OSCulator[2] for communication between the wiimote and the computer. OSCulator provides estimates of orientation and absolute acceleration of the wiimote.

Orientation data can be seen as similar to the position data from the motion capture system, in the sense that it describes a state of the device within a single time-frame. Because of this similarity, change in orientation was mapped to the amplitude control. Although ideally the orientation data from the wiimote should not change unless there was an actual change in the orientation of the wiimote, the fact is that these values changed quite a lot even for non-rotational motion. Because of a significant amount of noise in the data, we used one of the push-buttons on the wiimote as an on/off button, to prevent the instrument from producing sound when the controller was lying still.

The angle between the floor and an imagined line along the length axis of the wiimote is called *pitch*. We let this value and its derivative control the synthesis parameters that originally were controlled by vertical position and vertical velocity, meaning the first delay line, frequency of the bandpass filter and the frequency of the ring modulator. Finally, we let the estimate of the dynamic acceleration control the second delay line in the synthesis patch.

4.2 Evaluation of the wiimote implementation

The wiimote implementation of the SoundSaber was, as expected, not as satisfying as the version based on optical motion capture. In our experience the orientation values needed some time to "settle". By this we mean that sudden actions affected these parameters quite a lot, and they did not settle at stable values until after the wiimote stopped moving. As a result, an action that was meant to cause a sudden increase in frequency would cause a sudden increase in loudness when the action started, and then a sudden increase in frequency when the wiimote was being held steady pointing up.

Using the tilt parameter *pitch* with the wiimote is conceptually quite different from the original mapping, where vertical position was used. However, we were surprised by how well this worked for slower motion. During a demonstration, one subject was moving the wiimote up and down with his arm fully stretched out, not realising that by doing this, he also pointed the wiimote up and down. The subject was puzzled by this and asked how we were able to extract vertical position values from the accelerometer in the wiimote.

In our opinion, the most important differences between the high-end implementation and the wiimote version are the size of the controller and the accuracy of the data. The wiimote data is too noisy for accurate control, and the size and shape of the wiimote afford one-handed, rapid impulsive actions, in contrast to the rod which is more suited for larger and slower actions. The wiimote implementation would probably benefit from using another synthesis module that is better suited for its affordances.

5. CONCLUSIONS AND FUTURE WORK

In this paper we presented the SoundSaber and our thoughts on how optical motion capture technology can be used for prototyping musical instruments. Our experience shows us that even a quite simple synthesiser and simple control signal are sufficient to create an interesting musical instrument, as long as the action-sound coupling is perceptually robust.

We will continue our work on the SoundSaber and other mocap instruments. It would be interesting to investigate whether the instrument would benefit from attaching an FSR. Furthermore, we see intriguing challenges and research questions related to developing the SoundSaber into a collaborative instrument, as well as an adaptive instrument that will adjust to different performers and situations.

6. REFERENCES

[1] C. Dobrian and F. Bevilacqua. Gestural control of music: using the vicon 8 motion capture system. In *Proceedings of NIME 2003*, pages 161–163, Montreal, Canada, 2003.

[2] A. Hunt and M. M. Wanderley. Mapping performer parameters to synthesis engines. *Organised Sound*, 7(2):97–108, 2002.

[3] C. Kiefer, N. Collins, and G. Fitzpatrick. Evaluating the wiimote as a musical controller. In *Proceedings of ICMC 2008*, Belfast, Northern Ireland, 2008.

[4] G. Leslie et al. Grainstick: A collaborative, interactive sound installation. In *Proceedings of ICMC 2010*, New York, USA, 2010.

[5] E. R. Miranda and M. Wanderley. *New Digital Musical Instruments: Control And Interaction Beyond the Keyboard*. A-R Editions, Inc., 2006.

[6] D. A. Norman. *The Design of Everyday Things*. Basic Books, New York, 1988.

[7] K. Nymoen. Comparing sound tracings performed to sounds with different sound envelopes. In *Proceedings of FRSM - CMMR 2011*, pages 225 – 229, Bhubaneswar, India, 2011.

[8] K. Nymoen, K. Glette, S. A. Skogstad, J. Torresen, and A. R. Jensenius. Searching for cross-individual relationships between sound and movement features using an SVM classifier. In *Proceedings of NIME 2010*, pages 259 – 262, Sydney, Australia, 2010.

[9] V. Pulkki. Generic panning tools for max/msp. In *Proceedings of ICMC 2000*, pages 304–307, 2000.

[10] S. A. Skogstad, A. R. Jensenius, and K. Nymoen. Using IR optical marker based motion capture for exploring musical interaction. In *Proceedings of NIME 2010*, pages 407–410, Sydney, Australia, 2010.

[11] D. Wessel and M. Wright. Problems and prospects for intimate musical control of computers. *Computer Music Journal*, 26:11–22, September 2002.

[2] `http://www.osculator.net/`

Paper V

Searching for Cross-Individual Relationships between Sound and Movement Features using an SVM Classifier.
K. Nymoen, K. Glette, S.A. Skogstad, J. Torresen, and A.R. Jensenius.
In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 259–262, Sydney University of Technology 2010.

Searching for Cross-Individual Relationships between Sound and Movement Features using an SVM Classifier

Kristian Nymoen, Kyrre Glette, Ståle Skogstad, Jim Torresen, Alexander R. Jensenius†

University of Oslo
Department of Informatics,
Pb 1080 Blindern, 0316 Oslo, Norway
{krisny, kyrrehg, savskogs, jimtoer}@ifi.uio.no

†University of Oslo
Department of Musicology
Pb 1017, Blindern, 0315 Oslo, Norway
a.r.jensenius@imv.uio.no

ABSTRACT

In this paper we present a method for studying relationships between features of sound and features of movement. The method has been tested by carrying out an experiment with people moving an object in space along with short sounds. 3D position data of the object was recorded and several features were calculated from each of the recordings. These features were provided as input to a classifier which was able to classify the recorded actions satisfactorily, particularly when taking into account that the only link between the actions performed by the different subjects was the sound they heard while making the action.

1. INTRODUCTION

What are the underlying links between movement and sound? We believe that the way we perceive sounds and their sound-producing actions are related, and that this relationship may be explored by observing human movement to sound. Auditory sensations are often perceived as mental images of what caused the sound. This idea of a *gestural-sonic object* is built upon motor theory in linguistics and neuroscience [8]. This belief has motivated an experiment to explore how sound and body movement are related: Is it possible to discover cross-individual relationships between how we perceive features of sound and features of movement by studying how people choose to move to sounds? The term *cross-individual* here denotes relationships that are found in the majority of the subjects in this experiment. Further, we use *movement* to denote continuous motion, and *action* to denote a segment of motion data.

Several papers have focused on training a machine learning system to recognize a specific action. This paper, however, presents a technique for discovering correlations between sound features and movement features. We investigate the use of a machine learning system to classify the actions that subjects link to certain sounds, here denoted as *sound-tracings* [9]. The features used for classification are evaluated, and the results of presenting various subsets of those features to the classifier are explored. This makes it possible to discover how a classification of sound-tracings based on certain *action* features is able to distinguish between *sounds* with certain characteristics. At the same time the classifier may be unable to distinguish between sounds with other characteristics. For instance, one of our hypotheses has been that features related to velocity would distinguish well between sounds with different loudness envelopes. Another hypothesis is that the features related to vertical displacement would distinguish between sounds with different pitch envelopes. An analysis of the classifier's performance can provide information on natural relationships between sounds and actions. This is valuable information in our research on new musical instruments.

Section 2 gives a brief overview of related research, including some notes on previous use of machine learning to classify music-related movement. Section 3 gives an overview of the method used. Section 4 presents the classification of the data, including feature extraction from the movement data and some results on reducing the number of inputs to the classifier. Finally, in section 5 we discuss the method used in the light of the results presented in section 4, and provide some conclusions and plans for future work on this material.

2. RELATED WORK

Machine learning and pattern recognition of motion data have been applied in musical contexts in various ways. Early works on applying neural networks to recognize actions to be mapped to sound synthesis parameters were presented in the early 1990s [5, 10]. In the last decade, various other machine learning implementations of mapping motion capture data to sound synthesis have been presented. This includes toolkits for machine learning in PureData [3] and Max/MSP [1], and a tool for on-the-fly learning where the system is able to learn new mappings, for instance during a musical performance [6].

Although mapping applications seem to have been the most used implementation of machine learning on motion data in musical contexts, some analytical applications exist as well. In *EyesWeb*, Camurri et al. have implemented recognition of expressivity in what they call 'musical gestures' [2]. Machine learning has also been applied to instrumental actions, like extraction of bowing features and classification of different bow strokes in violin performance [12, 13].

A significant amount of work has been done on information retrieval of motion capture data within research fields related to computer animation [11]. Much of the work in this field has been on classification of different actions in a motion database (e.g. distinguishing a kicking action from a jumping action). For this sort of classification Müller and Röder have introduced *motion templates* [11]. This method is based on spatio-temporal relationships between various parts of the body. They present a sophisticated method for recognizing specific actions, a method which is independent

NIME2010, June 15-18, 2010, Sydney, Australia

from numerical differences in the raw data.

The research presented in this paper distinguishes itself from the previously mentioned ones in that it aims to recognize certain unknown features of the actions rather than the actions themselves. The approach is analytical, with a goal of discovering cross-individual relationships between features of sound and features of movement.

A similar experiment to the one presented in this paper was carried out in 2006, where subjects were presented with short sounds and instructed to sketch sound-tracings on a Wacom tablet [9]. This data was initially studied qualitatively, and has recently also been processed quantitatively in an unpublished manuscript which inspired this paper [7].

3. METHOD

3.1 Setup

In our experiment we used a 7 camera *Optitrack* infrared motion capture system for gathering position data of reflective markers on a rod. A sampling rate of 100 Hz was used, and data was sent in real-time to Max/MSP for recording.

3.2 Observation Experiment

Fifteen subjects, with musical experience ranging from no performance experience to professional musicians, were recruited. These were 4 females and 11 males, selected among university students and staff. The subjects were presented with ten sounds and asked to move a rod in space along with each sound, as if they themselves were creating the sound. The rod was roughly 120 cm long with a diameter of 4 cm (Figure 1). Before recording the movement data, the subjects listened to the sound twice (or more if they requested it), to allow them to make up their mind on what they thought would be a natural connection between the sound and the movement. A metronome was used so that the subjects could know at what time the sound started. The motion capture recording started 500 ms before the sound, and was stopped at the end of the sound file. Thus, all the motion capture recordings related to a single sound were of equal length which made it easier to compare the results from different subjects. We made three recordings of each action from each subject. Some of the recordings were discarded, due to the subject moving the rod out of the capture volume, which caused gaps in the data. Hence, there are between 42 and 45 data recordings of actions performed to each sound.

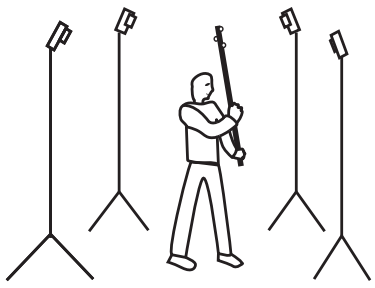


Figure 1: The figure shows a subject holding the rod with reflective markers in one end. Motion capture cameras are surrounding the subject.

The recorded data was the 3D position at the end of the rod, in addition to video. The subjects also filled out a small questionnaire where they were asked whether they considered themselves to be novices, intermediates or music experts, and whether they found anything in the experiment to be particularly difficult.

3.3 Sounds

The sounds used in the experiment all had one or more distinct features (e.g. rising pitch or varying sound intensity), which we believed would make the users move differently to the different sounds. A brief overview of the sounds is presented in Table 1, and the sounds are available online.[1] Some of the sounds were quite similar to each other, e.g. with only subtle differences in the timing of loudness peaks. As we shall see, actions performed to these similar sounds were often mistaken for each other by the classifier. Sounds 1 and 2 are similar, where the loudness and the center frequency of a bandpass filter sweeps up and down three times. The difference between the sounds is the timing of the peaks, which gives a slightly different listening experience. Sounds 9 and 10 are also quite similar to each other, with the same rhythmic pattern. The difference between the two is that Sound 9 has a steady envelope, while Sound 10 has impulsive attacks with a decaying loudness envelope after each attack.

Table 1: Simple description of the sounds used in the experiment

Sound	Pitch	Spectral Centroid	Loudness	Onsets
1	Noise	3 sweeps	3 sweeps	3
2	Noise	3 sweeps	3 sweeps	3
3	Falling	Rising	Steady	1
4	Rising	Falling	Steady	1
5	Noise	Rising	Steady	1
6	Noise	Rising / Complex	Steady	1
7	Noise	Rising, then falling	Steady	1
8	Rising	Complex	Steady	1
9	Noise	Steady	Rhythm:♩♩♫♩ Static (on/off)	5
10	Noise	Complex	Like 9, with decaying slopes	5

3.4 Software

For classification we used *RapidMiner*,[2] a user-friendly toolbox for data mining, classification and machine learning. A brief test of the various classification algorithms in RapidMiner indicated that Support Vector Machines (SVM) would provide the highest classification accuracies, so this was chosen for the experiments. RapidMiner uses the LIBSVM[3] library for SVMs. The python-script *grid.py* is provided with LIBSVM and was used for finding the best parameters for the algorithm. This script performs a simple grid search to determine the best parameters.

When training and validating the system, *cross-validation* was used due to the limited number of data examples. This means that two complementary subsets are randomly generated from the full data set. One subset of the data examples is used for training the classifier, and the other is used as a validation set to measure the performance of the classifier [4]. This process was repeated ten times with different subsets. Finally, the performance results ere averaged across all performance evaluations. Matlab was used for preprocessing and feature extraction.

4. ANALYSIS

The analysis process consists of two main parts: the feature extraction and the classification. In our opinion, the former is the most interesting in the context of this paper, where the goal is to evaluate a method for comparing movement features to sound features. The features selected are features that we believed would distinguish between the sounds.

[1] http://folk.uio.no/krisny/nime2010/
[2] http://rapid-i.com/
[3] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

4.1 Feature Extraction

When extracting the movement features, it is important to note that two actions that seem similar to the human eye do not need to be similar numerically. This implies that the features should be based on relative, rather than absolute data. In our setup, we have a recording of only a single point in space, and thus we cannot calculate spatial relations as suggested by Müller et al.[11], but we can look at temporal relations. Hence, we have chosen to base the features presented here on time-series based on the derivative of the position data. Since we have no reference data on the position of the subject (only the rod), we cannot tell whether horizontal movement is forwards or sideways. Thus horizontal movement along either of the two horizontal axes should be considered as equivalent. However, the vertical component of the movement can be distinguished from any horizontal movement, since gravity is a natural reference.

The 3D position data was used to calculate the following features from the recorded data:

- *VelocityMean* and *VelocityStD* are the mean and standard deviation of the vector length of the first derivatives of the 3D position data.
- *AccelerationMean* is the mean value of the vector length of the second derivative of the 3D position data.
- *TurnMean* and *TurnStD* are the mean value and the standard deviation of change in direction between the samples, i.e. the angle between the vector from sample n to n+1, and the vector from n+1 to n+2.
- *PreMove* is the cumulative distance before the sound starts. This is a period of 50 samples in all recordings.
- *vVelocityMean* is the mean value of the derivatives of the vertical axis. As opposed to *VelocityMean*, this feature can have both positive (upwards) and negative (downwards) values.
- *vEnergy* is an exponentially scaled version of *vVelocityMean*, meaning that fast movement counts more than slow movement. For example, fast movement downwards followed by slow movement upwards would generate a negative value, even if the total distance traveled upwards and downwards is the same.

Finally, each recording was divided into four equally sized segments, e.g. to be able to see how the first part of the action differed from the last part. The variables *segmentVelMean* — the mean velocity of each segment, and *segmentShake* — a measure based on autocorrelation to discover shaking, were calculated.

In the next section we will present the classification results, and investigate if classifications based on different subsets of features will reveal relationships between sound features and action features.

4.2 Results

When all the movement features were fed to the classifier, a classification accuracy of 78.6% ± 7.3% was obtained. This should be interpreted as the precision of recognizing the *sound* that inspired a certain action, based on features extracted from the *movement* data. Sound 7 was the one with the best accuracy, where the algorithm classified the 95.2% of the actions correctly, as shown in Table 2. The classifier misinterpreted some of the actions made to similar sounds, but still the lowest individual classification accuracy was as high as 68.9%. The table columns show the true actions, and the rows show the predictions of the classifier. The diagonal from top left to lower right indicates the correctly classified instances (marked in grey). We define *class recall (CR)* and *class precision (CP)* of class i as:

$$CR_i = \frac{||R_i \cap A_i||}{||R_i||} * 100\% \qquad CP_i = \frac{||R_i \cap A_i||}{||A_i||} * 100\%$$

$||A_i||$ denotes the number of instances classified as i, and $||R_i||$ denotes the total numbers of instances in class i. Then CP is the probability that a certain prediction made by the classifier is correct, and CR is the probability that the classifier will provide the correct result, given a certain class.

When reducing the features fed to the classifier to only include the two features related to vertical displacement, i.e. *vVelocityMean* and *vEnergy*, the total classification accuracy was reduced to 36%. However, the sounds with a distinct rising or falling pitch had significantly less change in classification accuracy than other sounds. For Sounds 3 and 4, we obtained a class recall of 79.1% and 51.2%, respectively. In addition to this we obtained a class recall of

Table 2: Classification accuracies for the individual sounds, when using all sound features. CP and CR denote class precision and class recall in percent, respectively. t1–t10 are the true classes, p1–p10 are the predictions made by the classifier.

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	CP
p1	34	6	1	1	1	0	0	0	0	4	72.3
p2	9	36	0	0	0	1	0	2	0	0	75.0
p3	0	0	36	2	0	2	0	0	0	0	90.0
p4	0	0	2	32	1	0	1	3	0	0	82.1
p5	0	0	1	2	31	6	1	2	1	0	70.5
p6	1	0	3	0	6	32	0	1	2	0	71.1
p7	0	0	0	0	1	0	40	3	0	0	90.9
p8	1	0	0	6	3	1	0	34	0	0	75.6
p9	0	1	0	0	2	2	0	0	36	6	76.6
p10	0	0	0	0	0	0	0	0	6	34	85.0
CR	75.6	83.7	83.7	74.4	68.9	72.7	95.2	75.6	80.0	77.3	

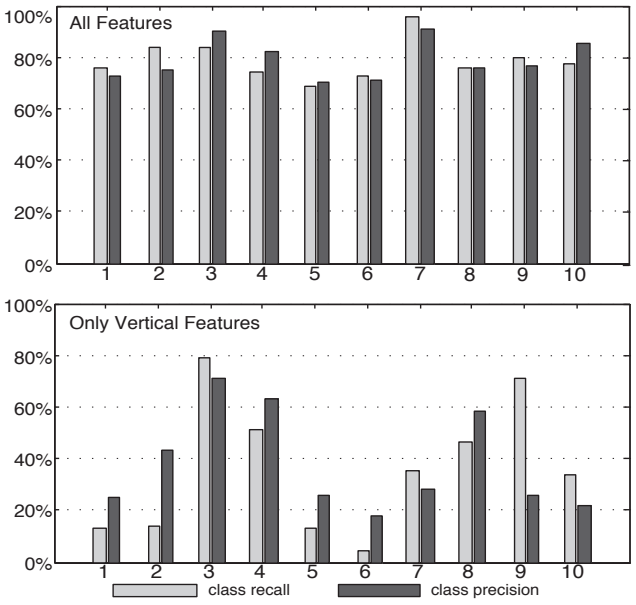


Figure 2: The figure shows the *class precision* and *class recall* for each of the classes (see text for explanation). A class consists of sound-tracings related to the same sound. High scores on both bars indicate that the estimation of this class is satisfactory. In the top chart, all features have been used as input to the classifier, in the lower chart, only the vertical displacement features were used.

71.1% for Sound 9, however the class precision of this sound was as low as 25.8%, indicating that the SVM classifier has made the class too broad for this to be regarded truly significant. The lower plot in Figure 2 shows that the class recall and class precision for all the sounds with changing pitch (3, 4 and 8) have relatively high scores on both accuracy and precision.

5. DISCUSSION

Our goal in this paper is to evaluate the use of a classifier to discover correlations between sound features and movement features. We have evaluated the method by using the data related to Sound 3, where we discovered a relationship between pitch and vertical movement. The fundamental frequency of this sound decreases from 300 Hz to 200 Hz. Figure 3 shows the vertical components of the actions performed to this sound by the subjects. The heavy lines denote the mean value and standard deviation of the vertical positions. Some of the actions do not follow the pitch downwards. This may be because the subject chose to follow the upwards moving spectral centroid. Also, quite a few of the actions make a small trip upwards before moving downwards. Still, there is a clear tendency of downwards movement in most of the performances, so we believe it is safe to conclude that there is a relationship between pitch and vertical position in our dataset. This finding makes it interesting to study the relationship between vertical position and pitch in a larger scale. Would we find similar results in a group that is large enough for statistical significance? Further on, we might ask if this action-sound relationship depends on things like cultural background or musical training.

We have also found similar, although not equally strong, indications of other correlations between sound and movement features. One such correlation is the *shake* feature. With only this as input, the classifier was able to distinguish well between Sounds 9 and 10. These were two rhythmic segments where the only difference was that Sound 10 had decaying slopes after each attack and Sound 9 had simply sound on or sound off with no adjustments in between. This could indicate that for one of the sounds, the subjects performed actions with impulsive attacks, resulting in a rebound effect which has been picked up in the shake feature.

Another relationship is the features *turnMean* and *turnStD* which seem to distinguish between the number of onsets in the sound. Sounds 1 and 2 had three onsets, and were quite well distinguished from the rest, but often confused with each other. The same was the case for Sounds 3, 4, 5, 6, 7 and 8 which had a single onset and Sounds 9 and 10 which had five onsets. A plausible explanation for this is that the subjects tended to repeat the same action for each onset of the sound, implying a somewhat circular movement for each onset. This circular motion is picked up in *TurnMean* and *TurnStD*.

The relationship between pitch and vertical displacement described in this section may seem obvious. But we believe the method is the most interesting. By using a classifier, we get an idea of where to look for cross-individual correlations between sound features and movement features.

6. CONCLUSIONS AND FUTURE WORK

The paper has presented a method for studying how perception of sound and movement is related. We believe that machine learning techniques may provide good indications of cross-individual correlations between sound features and movement features. Our experiments have shown that it is possible to study these relationships by feeding movement data to a classifier and carefully selecting the features used for classification. The paper has mainly focused on evaluating the method itself rather than the results, since a larger selection of subjects would be necessary to draw strong conclusions on the existence of action-sound relationships. Future research plans include experiments with a larger selection of subjects, and to expand the setup to full-body motion capture. In our research, we hope to learn more about how features of movement can be used to develop new intuitive movement-based instruments.

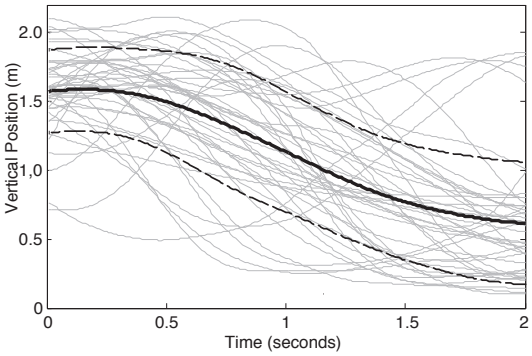


Figure 3: Plot of vertical position of the performances to Sound 3. The heavy lines denote mean value and standard deviation.

7. REFERENCES

[1] F. Bevilacqua, R. Müller, and N. Schnell. MnM: a Max/MSP mapping toolbox. In *Proceedings of NIME 2005*, pages 85–88, Vancouver, BC, 2005.

[2] A. Camurri, B. Mazzarino, and G. Volpe. Analysis of expressive gesture: The Eyesweb expressive gesture processing library. *Lecture Notes in Computer Science*, 2915:460–467, February 2004.

[3] A. Cont, T. Coduys, and C. Henry. Real-time gesture mapping in PD environment using neural networks. In *Proceedings of NIME 2004*, pages 39–42, Singapore, 2004.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

[5] S. Fels and G. Hinton. Glove-talk: A neural network interface between a dataglove and a speech synthesizer. *IEEE Trans. Neural Networks*, 4(1):2–8, 1993.

[6] R. Fiebrink, D. Trueman, and P. R. Cook. A meta-instrument for interactive, on-the-fly machine learning. In *Proceedings of NIME 2009*, Pittsburgh, 2009.

[7] K. Glette. Extracting action-sound features from a sound-tracing study. Tech report, University of Oslo, 2009.

[8] R. I. Godøy. Gestural-sonorous objects: embodied extensions of Schaeffer's conceptual apparatus. *Organised Sound*, 11(02):149–157, 2006.

[9] R. I. Godøy, E. Haga, and A. R. Jensenius. Exploring music-related gestures by sound-tracing. a preliminary study. In *2nd ConGAS Int. Symposium on Gesture Interfaces for Multimedia Systems*, Leeds, UK, 2006.

[10] M. Lee, A. Freed, and D. Wessel. Neural networks for simultaneous classification and parameter estimation in musical instrument control. *Adaptive and Learning Systems*, 1706:244–255, 1992.

[11] M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH/ SCA '06:*, pages 137–146. Eurographics Association, 2006.

[12] N. H. Rasamimanana, E. Fléty, and F. Bevilacqua. Gesture analysis of violin bow strokes. In S. Gibet, N. Courty, and J.-F. Kamp, editors, *Gesture Workshop*, volume 3881 of *LNCS*, pages 145–155. Springer, 2005.

[13] E. Schoonderwaldt and M. Demoucron. Extraction of bowing parameters from violin performance combining motion capture and sensors. *The Journal of the Acoustical Society of America*, 126(5), 2009.

Paper VI

Analyzing sound tracings: a multimodal approach to music information retrieval.
K. Nymoen, B. Caramiaux, M. Kozak, and J. Torresen.
In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 39–44, Association for Computing Machinery (ACM) 2011.

Analyzing Sound Tracings - A Multimodal Approach to Music Information Retrieval

Kristian Nymoen
University of Oslo
Department of Informatics
Postboks 1080 Blindern
0316 Oslo, Norway
krisny@ifi.uio.no

Baptiste Caramiaux
IMTR Team
IRCAM, CNRS
1 Place Igor Stravinsky
75004 Paris, France
Baptiste.Caramiaux@ircam.fr

Mariusz Kozak
University of Chicago
Department of Music
1010 East 59th Street
Chicago, IL 60637, USA
mkozak@uchicago.edu

Jim Torresen
University of Oslo
Department of Informatics
Postboks 1080 Blindern
0316 Oslo, Norway
jimtoer@ifi.uio.no

ABSTRACT

This paper investigates differences in the gestures people relate to *pitched* and *non-pitched* sounds respectively. An experiment has been carried out where participants were asked to move a rod in the air, pretending that moving it would create the sound they heard. By applying and interpreting the results from Canonical Correlation Analysis we are able to determine both simple and more complex correspondences between features of motion and features of sound in our data set. Particularly, the presence of a distinct pitch seems to influence how people relate gesture to sound. This identification of salient relationships between sounds and gestures contributes as a multi-modal approach to music information retrieval.

Categories and Subject Descriptors

H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Signal analysis, synthesis, and processing*

Keywords

Sound Tracing, Cross-Modal Analysis, Canonical Correlation Analysis

1. INTRODUCTION

In recent years, numerous studies have shown that gesture, understood here as voluntary movement of the body produced toward some kind of communicative goal, is an important element of music production and perception. In the case of the former, movement is necessary in performance on acoustic instruments, and is increasingly becoming an important component in the development of new electronic musical interfaces [17]. As regards the latter, movement synchronized with sound has been found to be a universal feature of musical interactions across time and culture [14]. Research has shown both that the auditory and motor regions of the brain are connected at a neural level, and that listening to musical sounds spontaneously activates regions responsible for the planning and execution of movement, regardless of whether or not these movements are eventually carried out [4].

Altogether, this evidence points to an intimate link between sound and gesture in human perception, cognition, and behavior, and highlights that our musical behavior is inherently multimodal. To explain this connection, Godøy [6] has hypothesized the existence of *sonic-gestural objects*, or mental constructs in which auditory and motion elements are correlated in the mind of the listener. Indeed, various experiments have shown that there are correlations between sound characteristics and corresponding motion features.

Godøy et al. [7] analyzed how the morphology of sonic objects was reflected in sketches people made on a digital tablet. These sketches were referred to as *sound tracings*. In the present paper, we adopt this term and expand it to mean a recording of free-air movement imitating the perceptual qualities of a sound. The data from Godøy's experiments was analyzed qualitatively, with a focus on the causality of sound as impulsive, continuous, or iterative, and showed supporting results for the hypothesis of gestural-sonic objects.

Godøy and Jensenius [8] have suggested that body movement could serve as a link between musical score, the acoustic signal and aesthetic perspectives on music, and that body movement could be utilized in search and retrieval of music. For this to be possible, it is essential to identify pertinent motion signal descriptors and their relationship to audio signal descriptors. Several researchers have investigated motion signals in this context. Camurri et al. [1] found strong correlations with the quantity of motion when focusing on recognizing expressivity in the movement of dancers. Fur-

http://doi.acm.org/10.1145/2072529.2072541
MIRUM'11, November 30, 2011, Scottsdale, Arizona, USA.

thermore, Merer et al. [12] have studied how people labeled sounds using causal descriptors like "rotate", "move up", etc., and Eitan and Granot studied how listeners' descriptions of melodic figures in terms of how an imagined animated cartoon would move to the music [5]. Moreover, gesture features like acceleration and velocity have been shown to play an important role in synchronizing movement with sound [10]. Dimensionality reduction methods have also been applied, such as Principal Component Analysis, which was used by MacRitchie et al. to study pianists' gestures [11].

Despite ongoing efforts to explore the exact nature of the mappings between sounds and gestures, the enduring problem has been the dearth of quantitative methods for extracting relevant features from a continuous stream of audio and motion data, and correlating elements from both while avoiding *a priori* assignment of values to either one. In this paper we will expand on one such method, presented previously by the second author [2], namely the Canonical Correlation Analysis (CCA), and report on an experiment in which this method was used to find correlations between features of sound and movement. Importantly, as we will illustrate, CCA offers the possibility of a mathematical approach for selecting and analyzing perceptually salient sonic and gestural features from a continuous stream of data, and for investigating the relationship between them.

By showing the utility of this approach in an experimental setting, our long term goals are to quantitatively examine the relationship between how we listen and how we move, and to highlight the importance of this work toward a perceptually and behaviorally based multimodal approach to music information retrieval. The study presented in the present paper contributes by investigating how people move to sounds with a controlled sound corpus, with an aim to identify one or several sound-gesture mapping strategies, particularly for pitched and non-pitched sounds.

The remainder of this paper will proceed as follows. In Section 2 we will present our experimental design. Section 3 will give an overview of our analytical methods, including a more detailed description of CCA. In Sections 4 and 5 we will present the results of our analysis and a discussion of our findings, respectively. Finally, Section 6 will offer a brief conclusion and directions for future work.

2. EXPERIMENT

We have conducted a free air sound tracing experiment to observe how people relate motion to sound. 15 subjects (11 male and 14 female) participated in the experiment. They were recruited among students and staff at the university. 8 participants had undergone some level of musical training, 7 had not. The participants were presented with short sounds, and given the task of moving a rod in the air as if they were creating the sound that they heard. Subjects first listened to each sound two times (more if requested), then three sound tracing recordings were made to each sound using a motion capture system. The recordings were made simultaneously with sound playback after a countdown, allowing synchronization of sound and motion capture data in the analysis process.

2.1 Sounds

For the analysis presented in this paper, we have chosen to focus on 6 sounds that had a single, non-impulsive onset. We make our analysis with respect to the sound features *pitch*, *loudness* and *brightness*. These features are not independent from each other, but were chosen because they are related to different musical domains (melody, dynamics, and timbre, respectively); we thus suspected that even participants without much musical experience would be able to detect changes in all three variables, even if the changes occurred simultaneously. The features have also been shown to be pertinent in sound perception [13, 16]. Three of the sounds had a distinct pitch, with continuously rising or falling envelopes. The loudness envelopes of the sounds varied between a bell-shaped curve and a curve with a faster decay, and also with and without tremolo. Brightness envelopes of the sounds were varied in a similar manner.

The sounds were synthesized in Max/MSP, using subtractive synthesis in addition to amplitude and frequency modulation. The duration of the sounds were between 2 and 4 seconds. All sounds are available at the project website [1]

2.2 Motion Capture

A NaturalPoint Optitrack optical marker-based motion capture system was used to measure the position of one end of the rod. The system included 8 Flex V-100 cameras, operating at a rate of 100 frames per second. The rod was approximately 120 cm long and 4 cm in diameter, and weighed roughly 400 grams. It was equipped with 4 reflective markers in one end, and participants were instructed to hold the rod with both hands at the other end. The position of interest was defined as the geometric center of the markers. This position was streamed as OSC data over a gigabit ethernet connection to another computer, which recorded the data and controlled sound playback. Max/MSP was used to record motion capture data and the trigger point of the sound file into the same text file. This allowed good synchronization between motion capture data and sound data in the analysis process.

3. ANALYSIS METHOD

3.1 Data Processing

The sound files were analyzed using the MIR toolbox for Matlab by Lartillot et al.[2] We extracted feature vectors describing *loudness*, *brightness* and *pitch*. Loudness is here simplified to the RMS energy of the sound file. Brightness is calculated as the amount of spectral energy corresponding to frequencies above 1500 Hz. Pitch is calculated based on autocorrelation. As an example, sound descriptors for a pitched sound is shown in Figure 1.

The position data from the OptiTrack motion capture system contained some noise; it was therefore filtered with a sliding mean filter over 10 frames. Because of the big inertia of the rod (due to its size), the subjects did not make very abrupt or jerky motion, thus the 10 frame filter should only have the effect of removing noise.

From the position data, we calculated the vector magnitude of the 3D velocity data, and the vector magnitude of the 3D acceleration data. These features are interpreted as the velocity independent from direction, and the acceleration independent from direction, meaning the combination of tangential and normal acceleration. Furthermore, the ver-

[1] http://folk.uio.no/krisny/mirum2011

[2] http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

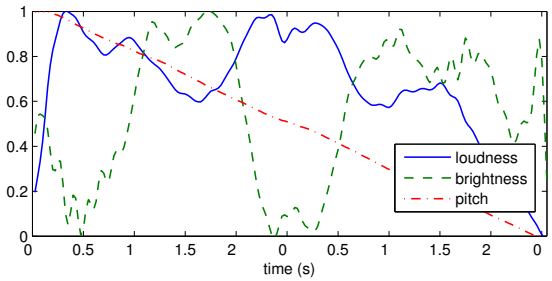


Figure 1: Sound descriptors for a sound with falling pitch (normalized).

tical position was used as a feature vector, since gravity and the distance to the floor act as references for axis direction and scale of this variable. The horizontal position axes, on the other hand, do not have the same type of positional reference. The subjects were not instructed in which direction to face, nor was the coordinate system of the motion capture system calibrated to have the same origin or the same direction throughout all the recording sessions, so distinguishing between the X and Y axes would be inaccurate. Hence, we calculated the mean horizontal position for each recording, and used the distance from the mean position as a one-dimensional feature describing horizontal position. All in all, this resulted in four motion features: *horizontal position*, *vertical position*, *velocity*, and *acceleration*.

3.2 Canonical Correlation Analysis

CCA is a common tool for investigating the linear relationships between two sets of variables in multidimensional reduction. If we let X and Y denote two datasets, CCA finds the coefficients of the linear combination of variables in X and the coefficients of the linear combination of variables from Y that are maximally correlated. The coefficients of both linear combinations are called *canonical weights* and operate as projection vectors. The projected variables are called *canonical components*. The correlation strength between canonical components is given by a correlation coefficient ρ. CCA operates similarly to Principal Component Analysis in the sense that it reduces the dimension of both datasets by returning N canonical components for both datasets where N is equal to the minimum of dimensions in X and Y. The components are usually ordered such that their respective correlation coefficient is decreasing. A more complete description of CCA can be found in [9]. A preliminary study by the second author [2] has shown its pertinent use for gesture-sound cross-modal analysis.

As presented in Section 3.1, we describe sound by three specific audio descriptors[3] and gestures by a set of four kinematic parameters. Gesture is performed synchronously to sound playback, resulting in datasets that are inherently synchronized. The goal is to apply CCA to find the linear relationships between kinematic variables and audio descriptors. If we consider uniformly sampled datastreams, and denote $\mathbf{X}$ the set of m_1 gesture parameters ($m_1 = 4$) and $\mathbf{Y}$ the set of m_2 audio descriptors ($m_2 = 3$), CCA finds two projection matrices $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_N] \in (\mathcal{R}^{m_1})^N$ and $\mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_N] \in (\mathcal{R}^{m_2})^N$ such that $\forall h \in 1..N$, the correlation coefficients $\rho_h = correlation(\mathbf{X}\mathbf{a}_h, \mathbf{Y}\mathbf{b}_h)$ are maximized and ordered such that $\rho_1 > \dots > \rho_N$ (where $N = \min(m_1,m_2)$).

A closer look at the projection matrices allows us to interpret the mapping. The widely used interpretation methods are either by inspecting the canonical weights, or by computing the canonical loadings. In our approach, we interpret the analysis by looking at the canonical loadings. Canonical loadings measure the contribution of the original variables in the canonical components by computing the correlation between gesture parameters $\mathbf{X}$ (or audio descriptors $\mathbf{Y}$) and its corresponding canonical components $\mathbf{XA}$ (or $\mathbf{YB}$). In other words, we compute the gesture parameter loadings $\mathbf{l}^x_{i,h} = (\text{corr}(\mathbf{x}_i, \mathbf{u}_h))$ for $1 \leq i \leq m_1, 1 \leq h \leq N$ (and similarly $\mathbf{l}^y_{i,h}$ for audio descriptors). High values in $\mathbf{l}^x_{i,h}$ or $\mathbf{l}^y_{i,h}$ indicate high correlation between realizations of the i-th kinematic parameter $\mathbf{x}_i$ and the h-th canonical component $\mathbf{u}_h$. Here we mainly focused on the first loading coefficients $h = 1, 2$ that explain most of the covariance. The corresponding ρ_h is the strength of the relationship between the canonical components $\mathbf{u}_h$ and $\mathbf{v}_h$ and informs us on how relevant the interpretation of the corresponding loadings is.

The motion capture recordings in our experiment started 0.5 seconds before the sound, allowing for the capture of any preparatory motion by the subject. The CCA requires feature vectors of equal length; accordingly, the motion features were cropped to the range between when the sound started and ended, and the sound feature vectors were upsampled to the same number of samples as the motion feature vectors.

4. RESULTS

We will present the results from our analysis starting with looking at results from pitched sounds and then move on to the non-pitched sounds. The results from each sound tracing are displayed in the form of statistical analysis of all the results related to the two separate groups (pitched and non-pitched). In Figures 2 and 3, statistics are shown in box plots, displaying the median and the population between the first and third quartile. The rows in the plots show statistics for the first, second and third canonical component, respectively. The leftmost column displays the overall correlation strength for the particular canonical component (ρ_h), the middle column displays the sound feature loadings ($\mathbf{l}^y_{i,h}$), and the rightmost column displays the motion feature loadings ($\mathbf{l}^x_{i,h}$). The + marks denote examples which are considered outliers compared with the rest of the data. A high value in the leftmost column indicates that the relationship between the sound features and gesture features described by this canonical component is strong. Furthermore, high values for the sound features *loudness* (Lo), *brightness* (Br), or *pitch* (Pi), and the gesture features *horizontal position* (HP), *vertical position* (VP), *velocity* (Ve), or *acceleration* (Ac) indicates a high impact from these on the respective canonical component. This is an indication of the strength of the relationships between the sound features and motion features.

4.1 Pitched Sounds

The results for three sounds with distinct pitch envelopes are shown in Figure 2. In the top row, we see that the median overall correlation strength of the first canonical components is 0.994, the median canonical loading for *pitch* is

[3]As will be explained later, for non-pitched sounds we omit the *pitch* feature, leaving only two audio descriptors.

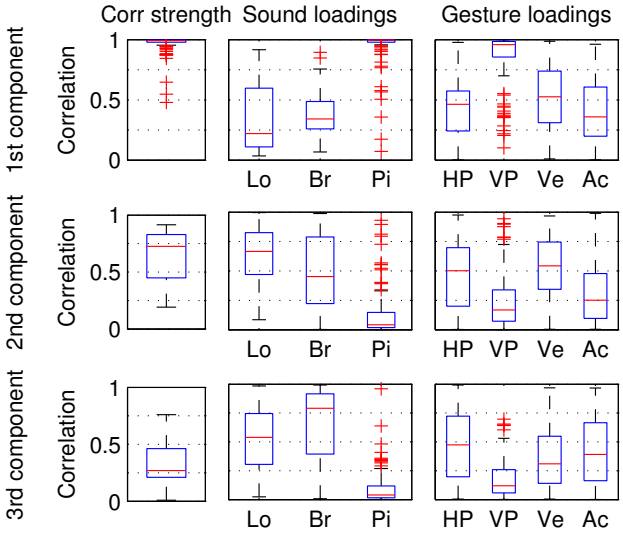


Figure 2: Box plots of the correlation strength and canonical loadings for three pitched sounds. *Pitch* (Pi) and *vertical position* (VP) have a significantly higher impact on the first canonical component than the other parameters. This indicates a strong correlation between pitch and vertical position for pitched sounds. The remaining parameters are: *loudness* (Lo), *brightness* (Br), *horizontal position* (HP), *velocity* (Ve) and *acceleration* (Ac).

0.997 and for *vertical position* 0.959. This indicates a strong correlation between pitch and vertical position in almost all the sound tracings for pitched sounds. The overall correlation strength for the second canonical component (middle row) is 0.726, and this canonical function suggests a certain correlation between the sound feature *loudness* and motion features *horizontal position* and *velocity*. The high variances that exist for some of the sound and motion features may be due to two factors: If some of these are indeed strong correlations, they may be less strong than the pitch-vertical position correlation For this reason, some might be pertinent to the 2nd component while others are pertinent to the 1st component. The second, and maybe the most plausible, reason for this is that these correlations may exist in some recordings while not in others. This is a natural consequence of the subjectivity in the experiment.

4.2 Non-pitched Sounds

Figure 3 displays the canonical loadings for three non-pitched sounds. The analysis presented in this figure was performed on the sound features loudness and brightness, disregarding pitch. With only two sound features, we are left with two canonical components. This figure shows no clear distinction between the different features, so we will need to look at this relationship in more detail to be able to find correlations between sound and motion features for these sound tracings.

For a more detailed analysis of the sounds without distinct pitch we investigated the individual sound tracings performed to non-pitched sounds. Altogether, we recorded 122 sound tracings to the non-pitched sounds; considering the first and second canonical component of these results gives

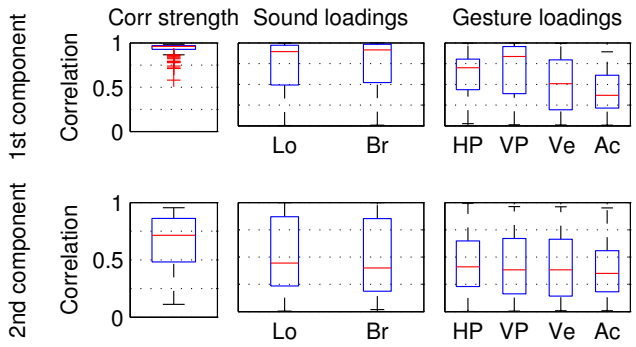


Figure 3: Box plots of the correlation and canonical loadings for three sounds without distinct pitch.

a total of 244 canonical components. We wanted to analyze only the components which show a high correlation between the sound features and motion features, and for this reason we selected the subset of the components which had an overall correlation strength (ρ) higher than the lower quartile,[4] which in this case means a value ≤ 0.927. This gave us a total of 99 components.

These 99 components all have high ρ-values, which signifies that they all describe some action-sound relationship well; however, since the results from Figure 3 did not show clearly which sound features they describe, we have analyzed the brightness and loudness loadings for all the recordings. As shown in Figure 4, some of these canonical components describe loudness, some describe brightness, and some describe both. We applied k-means clustering to identify the three classes which are shown by different symbols in Figure 4. Of the 99 canonical components, 32 describe loudness, 30 components describe brightness, and 37 components showed high loadings for both brightness and loudness.

Having identified the sound parameters' contribution to the canonical components, we can further inspect how the three classes of components relate to gestural features. Figure 5 shows the distribution of the gesture loadings for *horizontal position*, *vertical position* and *velocity* for the 99 canonical components. *Acceleration* has been left out of this plot, since, on average, the acceleration loading was lowest both in the first and second component for all sounds. In the upper part of the plot, we find the canonical components that are described by vertical position. The right part of the plot contains the canonical components that are described by horizontal position. Finally the color of each mark denotes the correlation to velocity ranging from black (0) to white (1). The three different symbols (triangles, squares and circles) refer to the same classes as in Figure 4.

From Figure 5 we can infer the following:

- For almost every component where the canonical loadings for both horizontal and vertical positions are high (cf. the upper right of the plot), the velocity loading is quite low (the marks are dark). This means that in the instances where horizontal and vertical position are correlated with a sound feature, velocity usually is not.

[4]The upper and lower quartiles in the figures are given by the rectangular boxes

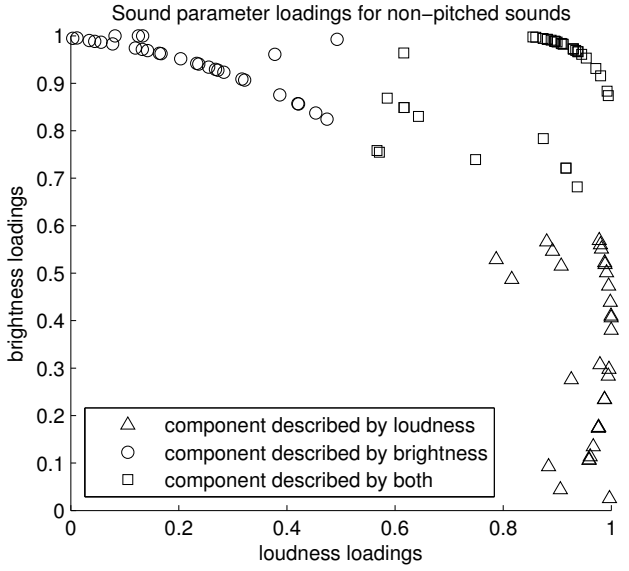


Figure 4: Scatter plot showing the distribution of the sound feature loadings for brightness and loudness. Three distinct clusters with high coefficients for brightness, loudness, or both, are found.

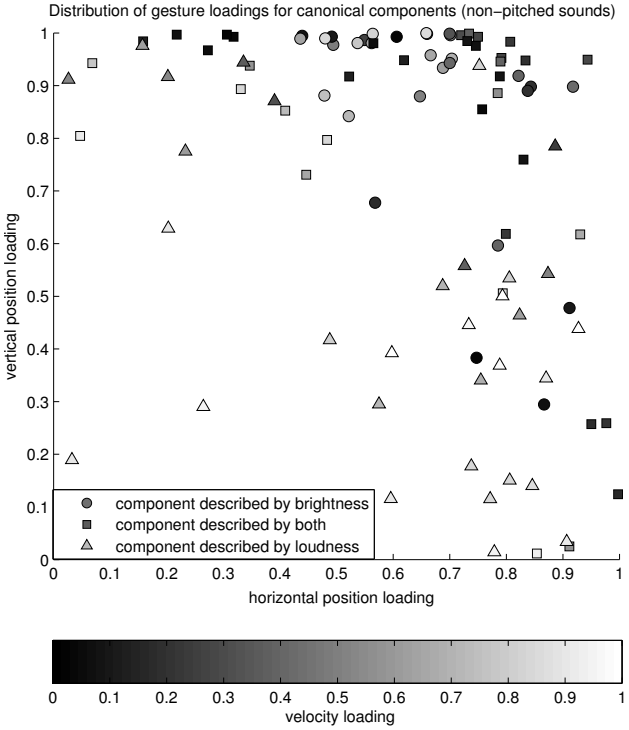


Figure 5: Results for the 99 canonical components that had high ρ-values. X and Y axes show correlation for horizontal position and vertical position, respectively. Velocity correlation is shown as grayscale from black (0) to white (1). The square boxes denote components witch also are highly correlated to brightness.

- The lower left part of the plot displays the components with low correlation between sound features and horizontal/vertical position. Most of these dots are bright, indicating that velocity is an important part in these components.
- Most of the circular marks (canonical components describing brightness) are located in the upper part of the plot, indicating that brightness is related to vertical position.
- The triangular marks (describing loudness) are distributed all over the plot, with a main focus on the right side. This suggests a tendency towards a correlation between horizontal position and loudness. What is even more interesting is that almost all the triangular dots are bright, indicating a relationship between loudness and velocity.
- The square marks (describing both loudness and brightness) are mostly distributed along the upper part of the plot. Vertical position seems to be the most relevant feature when the canonical component describes both of the sound features.

5. DISCUSSION

As we have shown in the previous section, there is a very strong correlation between vertical position and pitch for all the participants in our data set. This relationship was also suggested when the same data set was analyzed using a Support Vector Machine classifier [15], and corresponds well with the results previously presented by Eitan and Granot [5]. In our interpretation, there exists a one-dimensional intrinsic relationship between pitch and vertical position.

For non-pitched sounds, on the other hand, we do not find such prominent one-dimensional mappings for all subjects. The poor discrimination between features for these sounds could be due to several factors, one of which is that there could exist non-linear relationships between the sound and the motion features that the CCA is not able to unveil. Non-linearity is certainly plausible, since several sound features scale logarithmically. The plot in Figure 6, which shows a single sound tracing, also supports this hypothesis, wherein brightness corresponds better with the squared values of the vertical position than with the actual vertical position. We would, however, need a more sophisticated analysis method to unveil non-linear relationships between the sound features for the whole data set.

Furthermore, the scatter plot in Figure 5 shows that there are different strategies for tracing sound. In particular, there are certain clustering tendencies that might indicate that listeners select different mapping strategies. In the majority of cases we have found that loudness is described by velocity, but also quite often by the horizontal position feature. Meanwhile, brightness is often described by vertical position. In one of the sounds used in the experiment the loudness and brightness envelopes were correlated to each other. We believe that the sound tracings performed to this sound were the main contributor to the class of canonical components in Figures 4 and 5 that describe both brightness and loudness. For this class, most components are not significantly distinguished from the components that only describe brightness. The reason for this might be that peo-

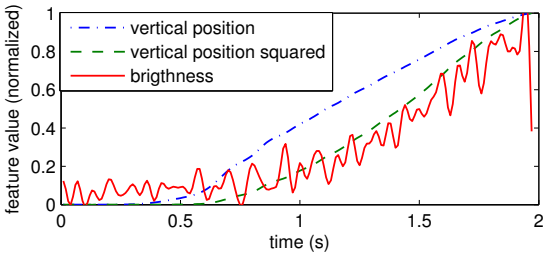


Figure 6: Envelopes of brightness, vertical position, and vertical position squared. The squared value corresponds better with brightness than the non-squared value, suggesting a non-linear relationship.

ple tend to follow brightness more than loudness when the two envelopes are correlated.

In future applications for music information retrieval, we envision that sound is not only described by audio descriptors, but also by lower-level gesture descriptors. We particularly believe that these descriptors will aid to extract higher-level musical features like affect and effort. We also believe that gestures will play an important role in search and retrieval of music. A simple prototype for this has already been prototyped by the second author [3]. Before more sophisticated solutions can be implemented, there is still a need for continued research on relationships between perceptual features of motion and sound.

6. CONCLUSIONS AND FUTURE WORK

The paper has verified and expanded the analysis results from previous work, showing a very strong correlation between pitch and vertical position. Furthermore, other, more complex relationships seem to exist between other sound and motion parameters. Our analysis suggests that there might be non-linear correspondences between these sound features and motion features. Although inter-subjective differences complicate the analysis process for these relationships, we believe some intrinsic action-sound relationships exist, and thus it is important to continue this research towards a cross-modal platform for music information retrieval.

For future directions of this research, we propose to perform this type of analysis on movement to longer segments of music. This implies a need for good segmentation methods, and possibly also methods like Dynamic Time Warping to compensate for any non-synchrony between the sound and people's movement. Furthermore, canonical loadings might be used as input to a classification algorithm, to search for clusters of strategies relating motion to sound.

7. REFERENCES

[1] A. Camurri, I. Lagerlöf, and G. Volpe. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1):213–225, 2003.

[2] B. Caramiaux, F. Bevilacqua, and N. Schnell. Towards a gesture-sound cross-modal analysis. In S. Kopp and I. Wachsmuth, editors, *Gesture in Embodied Communication and Human-Computer Interaction*, volume 5934 of *LNCS*, pages 158–170. Springer Berlin / Heidelberg, 2010.

[3] B. Caramiaux, F. Bevilacqua, and N. Schnell. Sound selection by gestures. In *New Interfaces for Musical Expression*, pages 329–330, Oslo, Norway, 2011.

[4] J. Chen, V. Penhune, and R. Zatorre. Moving on time: Brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training. *Cerebral Cortex*, 18:2844–2854, 2008.

[5] Z. Eitan and R. Y. Granot. How music moves: Musical parameters and listeners' images of motion. *Music Perception*, 23(3):pp. 221–248, 2006.

[6] R. I. Godøy. Gestural-sonorous objects: Embodied extensions of Schaeffer's conceptual apparatus. *Organised Sound*, 11(2):149–157, 2006.

[7] R. I. Godøy, E. Haga, and A. R. Jensenius. Exploring music-related gestures by sound-tracing. A preliminary study. In *2nd ConGAS Int. Symposium on Gesture Interfaces for Multimedia Systems*, Leeds, UK, 2006.

[8] R. I. Godøy and A. R. Jensenius. Body movement in music information retrieval. In *International Society for Music Information Retrieval Conference*, pages 45–50, Kobe, Japan, 2009.

[9] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson. *Multivariate Data Analysis*. Prentice Hall, New Jersey, USA, February, 2009.

[10] G. Luck and P. Toiviainen. Ensemble musicians' synchronization with conductors' gestures: An automated feature-extraction analysis. *Music Perception*, 24(2):189–200, 2006.

[11] J. MacRitchie, B. Buch, and N. J. Bailey. Visualising musical structure through performance gesture. In *International Society for Music Information Retrieval Conference*, pages 237–242, Kobe, Japan, 2009.

[12] A. Merer, S. Ystad, R. Kronland-Martinet, and M. Aramaki. Semiotics of sounds evoking motions: Categorization and acoustic features. In R. Kronland-Martinet, S. Ystad, and K. Jensen, editors, *Computer Music Modeling and Retrieval. Sense of Sounds*, volume 4969 of *LNCS*, pages 139–158. Springer Berlin / Heidelberg, 2008.

[13] N. Misdariis, A. Minard, P. Susini, G. Lemaitre, S. McAdams, and P. Etienne. Environmental sound perception: Metadescription and modeling based on independent primary studies. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[14] B. Nettl. An ethnomusicologist contemplates universals in musical sound and musical culture. In N. Wallin, B. Merker, and S. Brown, editors, *The Origins of Music*, pages 463–472. Cambridge, Mass: MIT Press, 2000.

[15] K. Nymoen, K. Glette, S. A. Skogstad, J. Torresen, and A. R. Jensenius. Searching for cross-individual relationships between sound and movement features using an SVM classifier. In *New Interfaces for Musical Expression*, pages 259–262, Sydney, Australia, 2010.

[16] P. Susini, S. McAdams, S. Winsberg, I. Perry, S. Vieillard, and X. Rodet. Characterizing the sound quality of air-conditioning noise. *Applied Acoustics*, 65(8):763–790, 2004.

[17] D. van Nort. Instrumental listening: Sonic gesture as design principle. *Organised Sound*, 14(02):177–187, 2009.

Paper VII

A Statistical Approach to Analyzing Sound Tracings.
K. Nymoen, J. Torresen, R.I. Godøy, and A.R. Jensenius.
In S. Ystad et al. (eds) *Speech, Sound and Music Processing: Embracing Research in India*, Lecture Notes in Computer Science, volume 7172, pages 120–145, Springer, Heidelberg 2012.

A Statistical Approach to Analyzing Sound Tracings

Kristian Nymoen[1], Jim Torresen[1],
Rolf Inge Godøy[2], and Alexander Refsum Jensenius[2]

[1] University of Oslo, Department of Informatics, Oslo, Norway
{krisny,jimtoer}@ifi.uio.no,
[2] University of Oslo, Department of Musicology, Oslo, Norway
{r.i.godoy,a.r.jensenius}@imv.uio.no

Abstract. This paper presents an experiment on *sound tracing*, meaning an experiment on how people relate motion to sound. 38 participants were presented with 18 short sounds, and instructed to move their hands in the air while acting as though the sound was created by their hand motion. The hand motion of the participants was recorded, and has been analyzed using statistical tests, comparing results between different sounds, between different subjects, and between different sound classes. We have identified several relationships between sound and motion which are present in the majority of the subjects. A clear distinction was found in onset acceleration for motion to sounds with an impulsive dynamic envelope compared to non-impulsive sounds. Furthermore, vertical movement has been shown to be related to sound frequency, both in terms of spectral centroid and pitch. Moreover, a significantly higher amount of overall acceleration was observed for non-pitched sounds as compared to pitched sounds.

1 Introduction

Research on music and motion is an interdisciplinary area with links to a number of other fields of research. In addition to traditional research on music and on kinematics, this area relates to neuropsychology, cognition, linguistics, robotics, computer science, and more [20]. To be able to grasp the complexity of the relationship between music and motion, we need knowledge about how different factors influence the motion and how the musical sound is perceived and processed in the brain. In addition, a certain understanding of experimental and mathematical methods is necessary for analyzing this relationship.

In several fields dealing with sound and motion, it is essential to identify how features of sound relate to events in other modalities. This includes disciplines like auditory display, interaction design, and development of multi-modal interfaces [23, 27]. Furthermore, it has been suggested that this aspect could be utilized in music information retrieval research, for instance by querying sound data bases with body motion [3, 13].

In this paper we present an experiment where motion capture technology was used to measure people's motions to sound. This data has been analyzed in view of perceptual correspondences between lower-level features of sound and motion. The participant's motion was evaluated statistically to find cross-modal relationships that have significance within our data set. There is a need for systematic experiments on sound-action relationships to build a larger corpus of examples on the links between perception of music and motion. The findings in this paper serve as a contribution to this corpus.

We shall begin by introducing the background and motivation for this particular research, including elements from music cognition, accounts of previous experiments on music and motion, as well as our own reflections on the implications of these works. In Section 3 we introduce our experimental setup, followed by a description of the recorded data set, with necessary preprocessing and feature extraction in Section 4. Section 5 presents analysis of the data set, and the results are discussed in Section 6. Conclusions are provided in Section 7.

2 Background

Presently, we have observed an increased popularity of a so-called theory of *embodied cognition*, meaning that bodily sensorimotor processing is understood as an important factor in our cognitive system [28]. Leman [19] put this theory into a musical context in his introduction of *embodied music cognition*. This theory describes how people who interact with music try to understand musical intentions and forms by imitation through corporeal articulations like body motion (e.g. tapping the beat, attuning to a melody or harmony, etc.) and empathy (e.g. attuning to certain feelings or a mood conveyed by the music).

Godøy [9] posits that our understanding of discrete events in music can be explained through *gestural-sonic objects*. These objects are mental constructs that combine the auditory input with gestural parameters, enabling an understanding of the sonic object through its causality (e.g. a perceived sound producing action). The idea of a gestural-sonic object as a discrete perceptual unit, or *chunk*, is based upon Pierre Schaeffer's *sonic object* [26], on Miller's theory of recoding complex sensory information into perceptual chunks [22], and also on the phenomenological understanding of perception as a sequence of now-points introduced by Husserl [14]. According to Godøy, these objects take form at the *meso level* of a musical timescale [10]. In contrast, the *macro level* of a musical timescale could be a whole musical piece, and the *micro level* of the timescale takes place within the sonic object. We believe that *action-sound relationships* [15] are found at all timescale levels, which coexist when a person is involved in a musical experience. Certain musical features like rhythmic complexity or emotional content require a larger timescale perspective than for instance musical features like pitch and timbre which operate in the millisecond range [7].

In a similar manner to the listening experiments Schaeffer performed on sonic objects, we can learn more about gestural-sonic objects by studying lower-level features of sound-related motion. In other words, one can look at the meso

level object from a micro-level perspective. Godøy et al. explored gestural-sonic objects in an experiment they referred to as *sound tracing* [12]. Nine subjects were given the task of making gestures they believed corresponded well with the sounds they heard, by using a pen on a digital tablet. By qualitative comparisons of the sound tracings, the authors found a fair amount of consistency between subjects, and argued that this type of experiment should be done in a larger scale, and include more complex sound objects, to learn more about sound-gesture relationships. The same material was later also analyzed quantitatively by extracting features and classifying the sound tracings using a support vector machine classifier [8]. We shall inherit the term *sound tracing* in the experiment presented in this paper. To be more precise, a *sound tracing* in this sense describes a bodily gesture that has been performed in free air to imitate the perceptual features of a sound object.

Other researchers have also studied how lower-level features of sound objects are related to motion or motion descriptors. Merer et al. [21] asked people to put their own motion-labels on sounds with different sound features. This way they determined which sound parameters were most pertinent in describing motion-labels such as "rotate" and "pass by". Eitan et al. found that for sounds with changing pitch, people imagined the movement of an animated character to follow the pitch up or down, however the authors also argued that changing pitch is related to other dimensions than simply vertical position [5, 6]. This corresponds well with previous research on metaphors and auditory display where increasing pitch has been related to an increase in other dimensions in other modalities, such as temperature [27]. The relationship between pitch and verticality was also found by Nymoen et al. [25] in a sound tracing experiment where participants used a rod to trace the perceptual features of a selection of sounds. In an experiment on synchronization with music, Kozak et al. [17] observed differences for quantity of motion between different lower-level features of sound like pitch, spectral centroid and loudness. Caramiaux et al. [2] applied Canonical Correlation Analysis to a set of sound and motion features derived from sound tracings.[3] This method gave promising results in identifying correlations between features of sound and of motion, and was later applied by Nymoen et al. [24].

The present paper is intended to follow up on the sound tracing experiments presented above. The main idea in this research was to study sound tracings from a more systematic perspective, in particular by using systematically varied sound parameters. This entailed using a number of short sounds, some where only a single sound parameter was varied, and some where multiple sound parameters were varied. In this manner, it was possible to understand the influence of different sound parameters on the sound tracings, which provided knowledge about how these features are reflected in other modalities. Our analytical approach operates at the meso and micro levels of the musical timescale, combining features that describe chunks of sound and motion with continuously varying sound and motion features.

[3] Caramiaux et al. do not refer to them as *sound tracings*, but following the definition presented above, their experiment falls into this category.

3 Experiment

A sound-tracing experiment was designed to be able to systematically distinguish between how people's motion changes and varies in relation to changes in sound features. The data presented in this paper was recorded in Fall 2010.

3.1 Aim

The aim of the experiment was to identify how lower-level features of motion corresponded with features of sound across different participants. By using systematically designed sounds, we can isolate a single sound feature and compare how it relates to motion by itself, or in combination with other sound features.

3.2 Participants

38 people (29 male and 9 female) volunteered to participate in the experiment. They were recruited through mailing lists for students and staff at the University of Oslo and by posting an advertisement on the project website. After participating in the experiment, the participants filled out a questionnaire concerning their level of musical training. 12 people rated their level of musical training as extensive, 11 as medium, and 15 as having little or no musical training. The level of musical training was used in the analysis process to distinguish between experts and non-experts (cf. Section 5). They were also given the opportunity to comment on the experiment. The subjects were not asked for their age, but we estimate the age distribution to be 20–60 years, with most participants aged somewhere between 25 and 35.

3.3 Task

The participants were presented with a selection of short sounds (the sounds will be discussed in Section 3.5). They were instructed to imagine that they could create sound by moving their hands in the air, and move along with the sounds as if their hand motion created the sound. First, each participant was given a pre-listening of all 18 sounds. Following this, the sounds were played one by one in random order. Each sound was played twice: the first time, the participant would only listen, and the second time the participant's hand motion was recorded. A three second countdown was given before each sound, so the participant would know exactly when the sound began.

3.4 Motion Capture

A Qualisys optical infrared marker-based motion capture system was used to record the motion of the people that participated in the experiment. The participants grasped two handles (Figure 1), each one equipped with 5 markers, and the center position of each handle was recorded. There are several advantages to

using this technology for recording motion. The system is very accurate, with a high resolution in both time and space. In our recordings, we used a sampling frequency of 100 Hz. Using several markers on each handle made it possible to uniquely identify the left and right handle, respectively, and enabled tracking of the position and the orientation of each handle.

Fig. 1. One of the two handles that was used for recording the participant's motion.

The main limitation we have experienced with the technology is so-called *marker-dropouts*. This happens when a marker is occluded (e.g. by the body limbs of the participant) or moved out of the calibrated capture space. Marker-dropouts caused a loss of a number of data-frames in several recordings, and it became necessary to perform so-called gap-filling. We will return to how this was done in Section 4. The marker dropouts made it necessary to disregard the orientation data from the handles, although this was initially recorded. This is because gap-filling of the orientation data was more difficult than gap-filling of the position data (interpolation even over small gaps introduces large errors).

3.5 Sounds

A total of 18 short sound objects, each 3 seconds in length, were designed in Max5 using *frequency modulation* (FM) synthesis and digital filters. The design process was to a large extent based on trial and error, to find sounds where the

envelopes of *pitch* (perceived tone height) and *spectral centroid* (here interpreted as perceived brightness) were distinct. *Envelope*, in this sense, is a generic term for a curve describing the development of a sound feature in the time domain. An example of the sound feature envelopes is given in Figure 2. The sound files are available for download at the project website.[4]

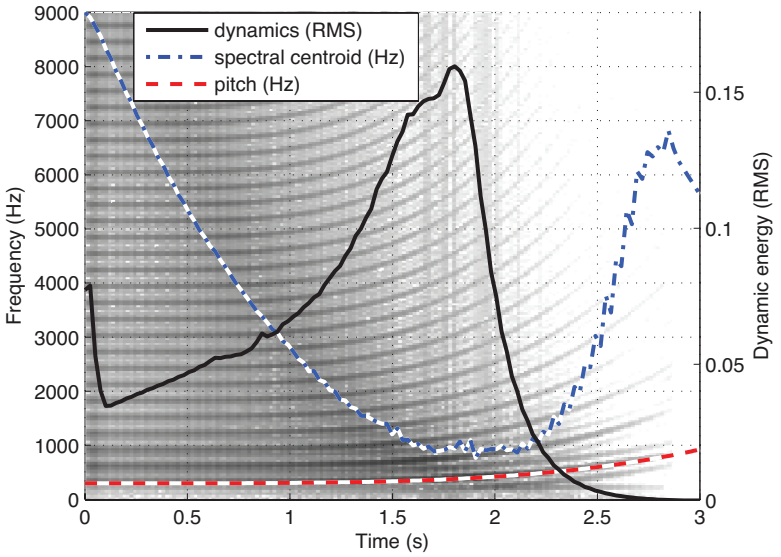


Fig. 2. Spectrogram and corresponding sound features for Sound 15. Pitch and spectral centroid (in Hz) on the left axis. The dynamic envelope scale is on the right axis.

Table 1. Simple description of the sounds used in the experiment. The columns display the pitch envelope, spectral centroid envelope and the dynamic envelope of each sound.

Sound	Pitch	Sp.Centroid	Dyn.Env.	Sound	Pitch	Sp.Centroid	Dyn.Env.
1	Rising	Falling	Bell-shape	10	Noise	Falling	Bell-shape
2	Falling	Rising	Bell-shape	11	Noise	Rising	Increasing
3	Falling	Falling	Bell-shape	12	Noise	Steady	Increasing
4	Rising	Rising	Bell-shape	13	Steady	Rising	Increasing
5	Rising	Steady	Increasing	14	Steady	Falling	Increasing
6	Falling	Steady	Increasing	15	Rising	Falling	Impulsive
7	Steady	Falling	Bell-shape	16	Steady	Steady	Impulsive
8	Steady	Rising	Bell-shape	17	Noise	Steady	Impulsive
9	Steady	Steady	Increasing	18	Noise	Falling	Impulsive

An overview of all of the sounds is presented in Table 1. In the first nine sounds, pitch and spectral centroid were manipulated by controlling the fundamental frequency of the FM sound, and the center frequency of a parametric

[4] `http://folk.uio.no/krisny/cmmr2011`

equalizer which boosted certain parts of the sound spectrum. These sounds were generated by changing the envelopes of pitch between 300 and 1000 Hz (*rising*, *falling* and *steady*) and equalizer center frequency between 50 and 13000 Hz (*rising* and *falling* as well as filter bypass, here interpreted as *steady* spectral centroid). This allowed for an appropriate discrimination between the individual sound parameter changes taking place within the sound. Sounds 10–12 were based on noise rather than a pitched FM sound, and only the filter was adjusted for these sounds. In Sounds 13 and 14, a second parametric equalizer was added. In Sound 13, the center frequencies of the equalizers started at 1000 and 5000 Hz and approached each other towards 3000 Hz, and in Sound 14, the center frequencies started at 3000 Hz, and moved apart to 1000 and 5000 Hz.

The synthesized sounds mentioned in the previous paragraph were multiplied by a window function to control the overall dynamic envelope. Here, we wanted to keep a main focus on the pitch and spectral properties of the whole sound, while influence from onset characteristics of the sounds (changes in sound features during the first part of the sound) was not desired. Therefore, Sounds 1–14 were made with a slow attack and increasing amplitude by applying the amplitude envelope displayed in Figure 3(a).

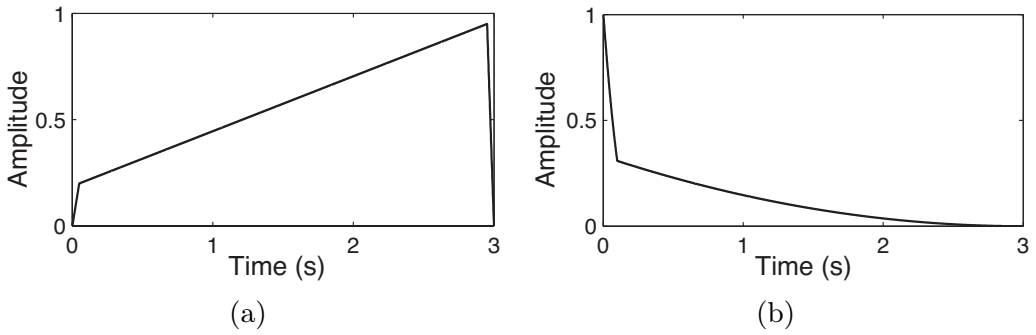


Fig. 3. The envelopes that were used for the amplitude control: (a) Non-impulsive sounds, and (b) Impulsive sounds.

The characteristics of the pitch envelope and the filter frequency also influenced the final dynamic envelope of the sounds. Some of which had a bell-shaped dynamic envelope, displayed in Figure 4(a), while others had a continuously increasing one, displayed in Figure 4(b).

We also wanted to investigate how the onset characteristics of a sound influenced the sound tracings that were performed to it. Therefore, impulsive versions of four of the sounds were made. Sounds 15–18 were versions of Sounds 1, 9, 10 and 12, the only difference was that instead of the slowly increasing dynamic envelope, we applied the impulsive envelope shown in Figure 3(b). It should be noted that the dynamic envelope of Sound 15 was different compared to the other impulsive sounds, because the varying pitch and filter frequency influenced the

dynamics. This resulted in a dynamic envelope which was a combination of the impulsive and bell-shaped envelopes, as shown in Figure 4(c).

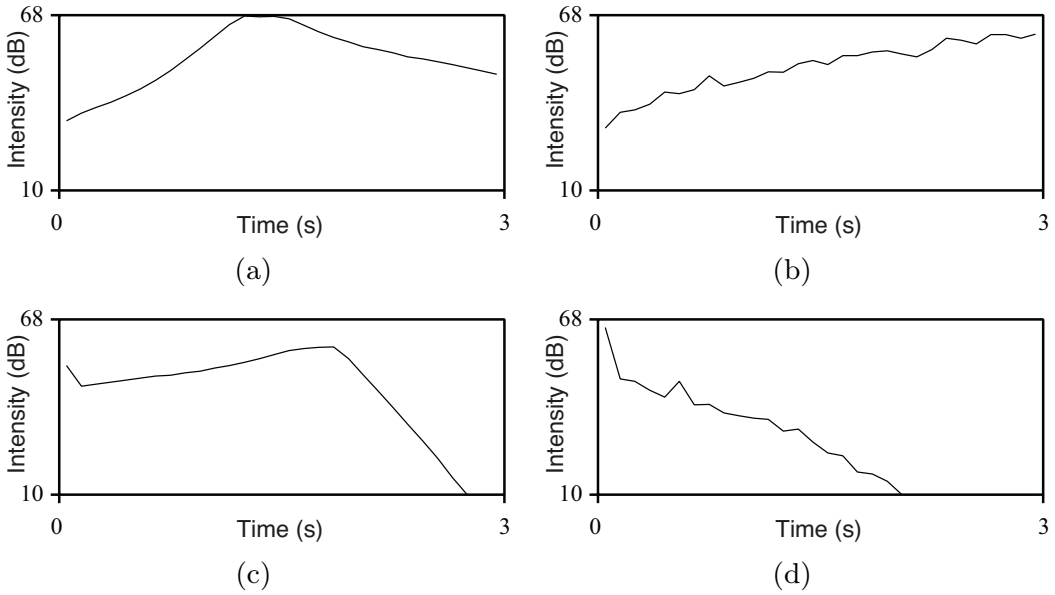


Fig. 4. The figure displays the dynamic envelopes of 4 sounds, analyzed with a perceptual model in the sound analysis software Praat. (a) Bell-shaped envelope (Sound 2), (b) Increasing envelope (Sound 9), (c) Impulsive and Bell-shaped envelope (Sound 15), and (d) Impulsive envelope (Sound 16).

4 Data Processing

In this section, we will describe the processing that was performed on the motion data to prepare it for the analysis process. The position data from the two handles was used, but it was not sufficient for our purpose to use it directly, and hence a number of data processing steps were taken. These steps included gap-filling, smoothing, feature extraction, data reduction and normalization.

4.1 Preprocessing

As mentioned in Section 3.4, some recordings contained missing data frames, and therefore gap-filling was required. We applied gap-filling on small data gaps by interpolating between the first and last missing frame using a piecewise cubic Hermite spline function with the preceding and succeeding frames as reference. A number of gaps were too large for gap-filling to be possible. In these cases, the recordings were discarded.

Certain participants had a large number of discarded recordings, which was due to poor calibration of the system in some sessions, but also because some

participants repeatedly occluded the reflective markers or moved the handles out of the capture space. If a single participant had too many discarded recordings, the consequence would be that this person only influenced a small portion of the data set, and we could risk that one participant only influenced one side of the analysis when two subsets of the dataset were compared. For this reason, we decided to discard the remaining recordings for subjects that had more than 1/3 (i.e. more than six) of their recordings removed.

The datasets from seven subjects were discarded completely, in addition to 30 recordings distributed among the other participants. In total, 156 of the 684 recordings were discarded. After the gap-filling process, a sliding mean filter of 5 samples (i.e. 50 ms) was applied to the position data in all the recordings to reduce measurement noise.

4.2 Motion Features

From the left and right handle position data we calculated a number of features that were used for analysis and comparison of the different sound tracings. Based on the position data, we calculated velocity and acceleration, as these features are related to kinetic energy and change in kinetic energy of the handles. The three axes of the position data cannot all be used directly. Movement in the vertical direction has been used directly as a motion feature, however, as will be explained shortly, the two horizontal axes are conceptually different from the vertical one, and have not been used directly.

The position data describes the position of each handle in relation to the room, or more precisely, in relation to the position of the calibration frame that was used when calibrating the motion capture system. This calibration frame determined the origin of the coordinate system and the direction of the axes. The position of the handles in relation to the calibration frame is not really relevant in light of the task that was given to the participants. The participants could not relate to the position of the calibration frame since it was removed after calibration. Furthermore, the participants were not instructed to face in any particular direction during the experiment, or precisely where to stand in the room. For this reason, we find it misleading to base our analysis directly on the horizontal position data. In contrast, the vertical position of the handles is a reference that was the same for all participants. The floor level remained constant, and was independent of where an individual stood, regardless of the direction he or she faced.

The one thing that varied between the subjects was the height range, as one participant could reach his arms to 2.2 m, while another up to 2.5 m. This was adjusted for by normalization as we will return to in Section 4.4. Based on these considerations, and on experiences regarding which features have proven to be most pertinent in previous experiments [2, 24, 25] the following data series for motion features was calculated:

- Vertical position: The distance to the floor.
- Vertical velocity: The derivative of the vertical position feature.

- Absolute velocity: Euclidean distance between successive position samples.
- Absolute acceleration: Euclidean distance between the successive derivatives of the position data.
- Distance: Euclidean distance between the hands.
- Change in distance: The derivative of the distance feature.

The features mentioned above are all data series, which we shall refer to as *serial features*. From these data series we calculated *single-value features*, meaning features that are given by a single number. These features describe a general tendency for an entire sound tracing. Examples of such features are mean vertical velocity and mean acceleration.

4.3 Data Reduction

To be able to compare the sound tracings, the data representation of each recording should be equal. In our case, this is not the case with the raw data, since some participants varied between using both hands, and only the left or right hand. Out of the 528 recordings, 454 were performed with both hands, 15 with only the left hand, and 59 with only the right hand. The participants were not specifically instructed whether to use one or two hands. In order to achieve equal data representation for all sound tracings, we had to choose between using the separate data streams from both hands in all cases, or reducing the data to fewer data streams keeping only the pertinent information from each sound tracing. Basing the analysis on data from a hand that was clearly not meant to be part of the sound tracing appeared less accurate to us, than to base the analysis on the data streams from only the active hand(s). Therefore, we calculated one serial position feature from each sound tracing, as well as one velocity feature, acceleration feature, and so forth. For the one-handed sound tracings, we used feature vectors of the active hand directly, and for the two-handed sound tracings, we calculated the average of both hands on a sample-by-sample basis. We did not change the distance feature for the single-handed sound tracings.

Admittedly, this difference between single-handed and two-handed performances presents a weakness in our experiment design, and we could have chosen different approaches to dealing with this challenge. We will continue searching for more comprehensive analysis methods which take into account this extra degree of freedom. If new methods for analysis are not found, a solution could be to instruct the participants to always use both hands.

4.4 Normalization

All feature vectors have been normalized for each subject. This means that all the calculated features were scaled to a range between 0 and 1, where the value was determined by the particular subject's maximum value for that feature. For example, if Subject 14 had a maximum vertical position of 2 meters across all of their sound tracings, all of the vertical position data series related to Subject 14 were divided by 2 meters. This type of normalization reduced individual

differences that were due to height, arm length, and so forth. This means that the data displayed in the plots in the coming section will all be scaled between 0 and 1. A similar normalization was performed on all of the single-value features.

5 Analysis and Results

The following sections present comparisons between three different aspects of the sounds. We will start in Section 5.1 by introducing our analysis method. In Section 5.2, the effect of the onset properties of the sounds are presented. In Sections 5.3 and 5.4, we present how the envelopes of pitch and spectral centroid tend to influence the sound tracings. Finally, in Section 5.5, differences between pitched and non-pitched sounds are presented.

5.1 Analysis Method

Our analysis is based on statistical comparisons between the individual data series, both sample-by-sample in serial features, and also on a higher level, comparing single-value features for the whole data series. The analyses of serial features are presented in plots where the individual data series are displayed together with the average vector of the data series. To facilitate the reading of these plots, we include a small example plot in Figure 5. This particular plot displays five data series ranging between 0 (white) and 1 (black). The vertical dashed lines show the beginning and end of the sound file, the motion capture recording began 0.5 seconds before the start of the sound file, and also lasted beyond the entire duration of the sound file. The black solid and dashed lines show the mean and standard deviations across the five data series on a sample-by-sample basis. From this figure, it is difficult to get precise readings of the values of the individual sound tracings, but the horizontal grayscale plots still give some impression of the distribution of this data set. The 0–1 scale on the y-axis is for the mean and standard deviation curves.

When certain tendencies are observed for different groups, we evaluate the statistical significance of the tendencies by applying one-tailed t-tests.[5] Results from the tests are presented in tables, where df denotes the degrees of freedom,[6] and t is the t-value from the t-test. p is calculated based on df and t, and denotes the probability that the two data sets are equally distributed, a p-value of less than 0.05 denotes a statistically significant difference between the groups.

Two subgroups were selected from the data set for analyzing the impact of musical training on the results in the experiment. Because of the somewhat imprecise classification of subjects' level of musical training, we chose to look at only the subjects that labeled themselves as having either no musical training or extensive musical training. This was done to ensure that there was indeed a difference in musical experience between the two groups.

[5] A t-test is a method to estimate the probability that a difference between two data sets is due to chance. See `http://en.wikipedia.org/wiki/T-test` for details.

[6] df is a statistical variable related to the t-test, denoting the size of the data material. It is not to be confused with e.g. 6DOF position-and-orientation data.

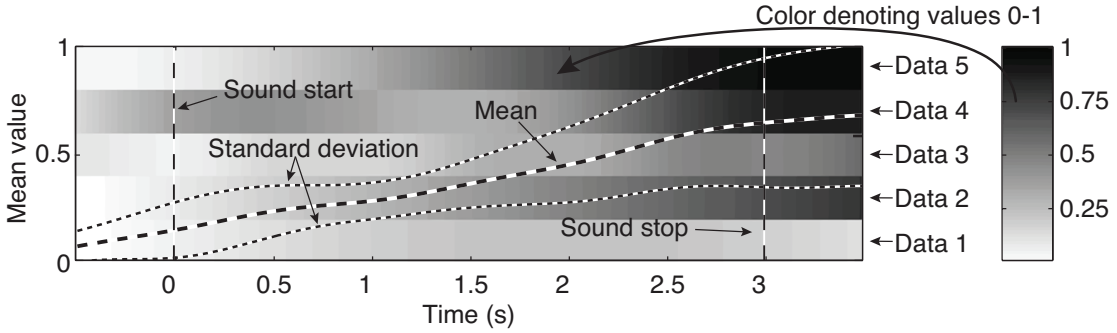


Fig. 5. The figure explains how to read the plots presented below. This is a reduced example with only 5 data series. The data series range between 0 (white) and 1 (black). Please refer to the text for explanation. A single subject is typically associated with multiple data-series, and tick marks on the right Y-axis denote different subjects. The ticks are not equidistant since some recordings were discarded (cf. Section 4.1). Furthermore, some plots display more data series per subject than others, thus the Y-axis resolution differs between plots.

5.2 Onset Acceleration for Impulsive and Non-impulsive Sounds

We evaluated how the onset characteristics of sound influence sound tracings by comparing the acceleration envelopes of impulsive sounds to non-impulsive sounds. We observed a distinct difference in acceleration envelope for sound tracings of the impulsive sounds compared to the rest of the data set, as displayed in Figure 6. To evaluate the significance of this difference, we compared the onset acceleration of the sound tracings. *Onset acceleration* is a single-value feature, which was calculated as the mean acceleration in the beginning of the sound tracing. Figure 6(b) shows that most subjects made an accentuated attack after the start of the sound file. Therefore we used a window from 0.2 seconds (20 samples) before the sound started to 0.5 seconds (50 samples) after the sound started to calculate the onset acceleration. The results of t-tests comparing the onset acceleration for impulsive and non-impulsive sounds are displayed in Table 2. The table shows that onset acceleration values for impulsive sounds are significantly higher than non-impulsive sounds, $t(526) = 13.65$, $p < 0.01$.[7]

Figure 7 displays separate acceleration curves of impulsive sounds for musical experts and non-experts. The figure shows that both groups have similar onset acceleration levels, and a t-test showed no statistical difference between the onset acceleration levels from the two groups, $t(84) = 0.55$, $p = 0.29$. However, the plots do show a difference in timing. By defining time of onset as the time of maximum acceleration within the previously defined onset interval, experts hit on average 163 ms after the start of the sound file, while non-experts hit 238 ms after the start of the sound file, a difference which was statistically significant, $t(63) = 2.51$, $p = 0.007$. This calculation was based only on Sounds 16–18, because several subjects did not perform an accentuated onset for Sound 15.

[7] This is the American Psychological Association style for reporting statistical results. Please refer to [1] for details.

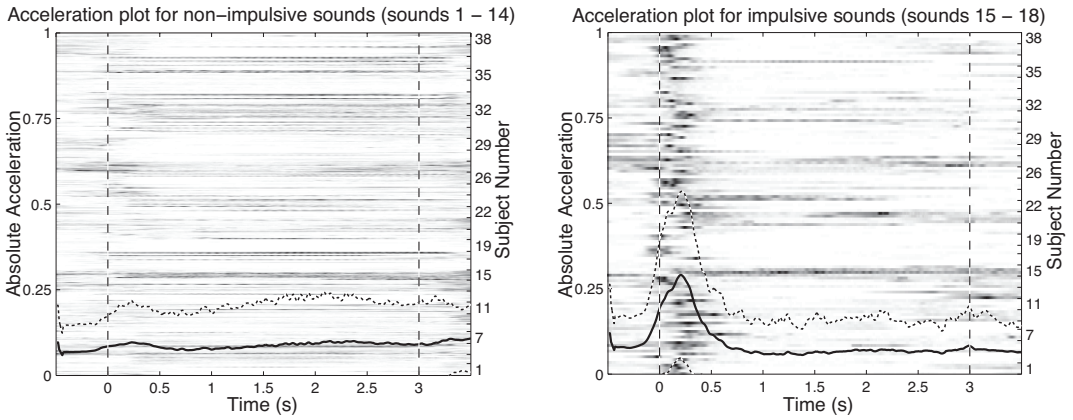


Fig. 6. Acceleration for (a) non-impulsive sounds (406 sound tracings) and (b) impulsive sounds (122 sound tracings). The black solid and dashed lines show the mean value and standard deviation across all sound tracings. Each horizontal line in the image displays the acceleration vector of a single sound tracing ranging between 0 (white) and 1 (black), normalized per subject. See Figure 5 for guidelines on how to read these plots.

Table 2. Results from t-tests comparing onset acceleration for impulsive sounds to non-impulsive sounds. There was a significant difference between the groups for both expert and non-expert subjects. See the text for explanation of the variables.

Onset acceleration, impulsive and non-impulsive sounds

Test description	***df***	***t***	***p***
Impulsive vs non-impulsive, all subjects	526	13.65	< 0.01
Impulsive vs non-impulsive, expert subjects	182	8.65	< 0.01
Impulsive vs non-impulsive, non-expert subjects	183	7.86	< 0.01
Onset acceleration level, experts vs non-experts	84	0.55	0.29
Onset time, expert vs non-expert subjects	63	2.51	< 0.01

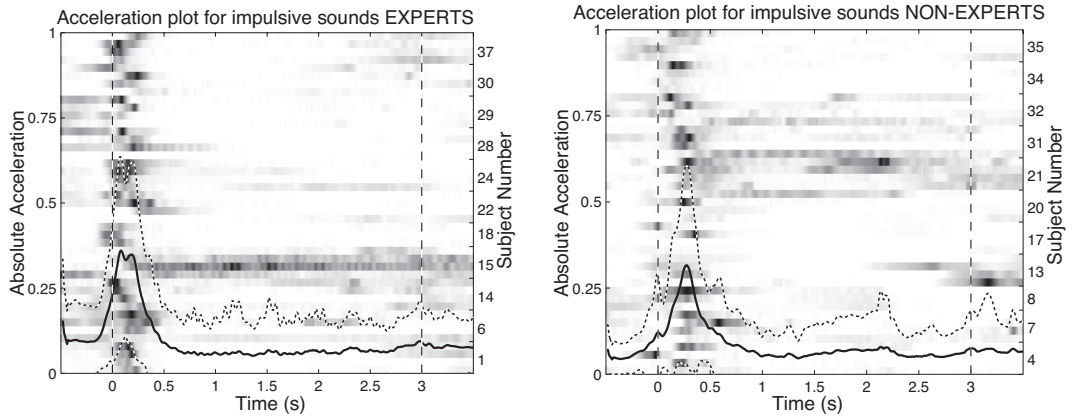


Fig. 7. The plot from Figure 6(b) separated into (a) experts and (b) non-experts.

5.3 Vertical Position and Sound Frequency Features

As mentioned in Section 2, other researchers have documented a relationship between vertical position and pitch. Not surprisingly, this relationship was also found in the data set presented in this paper. In addition to pitch, we observed that the frequency of the spectral centroid is relevant to the vertical position.

Sounds 1, 4 and 5 all had rising pitch envelopes, and Sounds 8 and 11 had rising spectral centroids combined with stable pitch and noise respectively. For the sound tracings of these sounds, there was a clear tendency of upward movement. Similarly, for the sounds with falling pitch, or with falling spectral centroid, there was a clear tendency of downward movement. t-tests comparing the average vertical velocity of the "rising" sounds to the "falling" sounds showed highly significant distinctions between the groups, as shown in Table 3. The mean normalized vertical velocity for the first group was 0.74, and for the second group 0.28 (a value of 0.5 indicates no vertical motion). This is shown in Figure 8.

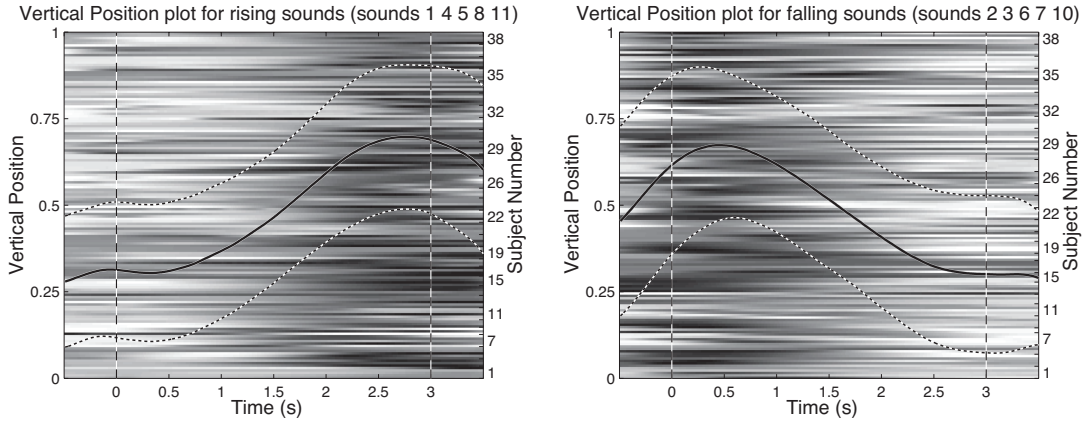


Fig. 8. Vertical position for (a) rising sounds (142 sound tracings) and (b) falling sounds (144 sound tracings). The black line shows the mean value across all the data series, each horizontal line in the image displays the vertical position of a single sound tracing normalized per subject between 0 (lower position, white) and 1 (higher position, black).

Table 3. T-tests comparing the average vertical velocity of rising and falling sounds.

Average vertical velocity, rising and falling sounds

Test description	***df***	***t***	***p***
Rising vs falling, all subjects	284	18.89	< 0.01
Rising vs falling, non-expert subjects	98	8.86	< 0.01
Rising vs falling, expert subjects	97	11.69	< 0.01
Rising, experts vs non-experts	98	0.58	0.28
Falling, experts vs non-experts	97	1.79	0.04

There was no significant difference between the average vertical velocity for experts and non-experts for the rising sounds, however, for the falling sounds there was some difference between the two groups. Experts had a higher extent of downward motion than non-experts, $t(97) = 1.7982$, $p = 0.04$.

It is worth noting that even though Sounds 1 and 2 had increasing and decreasing pitch envelopes, respectively, they had opposing spectral centroid envelopes. When the spectral centroid envelope and the pitch envelope moved in opposite directions, most subjects in our data set chose to let the vertical motion follow the direction of the pitch envelope. The direction of vertical motion seems to be more strongly related to pitch than to spectral centroid.

The observed difference between sounds with varying pitch and sounds with only varying spectral centroid makes it interesting to take a more in depth look at the individual sounds in the *rising* and *falling* classes. Since subjects tended to follow pitch more than spectral centroid in the sounds where the two feature envelopes moved in opposite directions, it is natural to assume that subjects would move more to sounds where the pitch was varied, than to sounds where only the spectral centroid was varied. Figure 9 displays box plots of the average vertical velocities for rising and falling sounds. In Figures 9(a) and 9(b), we observed that the difference between the sounds is larger for falling than for rising sounds. We can also see that Sounds 7 and 8, which are sounds where the pitch is constant but spectral centroid is moving, show less extreme values than the rest of the sounds. Figures 9(c) and 9(d) suggest that the difference between the sounds is larger for expert subjects than for non-expert subjects. There also seems to be more inter-subjective similarities among experts than non-experts, as the variances among experts are lower.

Table 4 shows the results of one-way analyses of variance (ANOVAs) applied to the sound tracings in the rising and falling class, respectively. The table shows that on the one hand, the difference in vertical velocity between the falling sounds was statistically significant $F(4, 139) = 7.76$, $p < 0.01$. On the other hand, the corresponding difference between the rising sounds was not statistically significant $F(4, 137) = 1.53$, $p = 0.20$. The table also reveals that the significant difference between the groups was only present for expert subjects, $F(4, 44) = 4.92$, $p < 0.01$, and not for non-experts, $F(4, 45) = 1.52$, $p = 0.21$.

Table 4. Results from one-way ANOVAs of the vertical velocity for sound tracings within the *rising* and *falling* classes. There is a significant difference between the five falling sounds for expert subjects. *df* are the degrees of freedom (between groups, within groups), F is the F-value with the associated f-test, p is the probability that the null-hypothesis is true.

Subjects	**Rising sounds**			**Falling sounds**		
	df	***F***	***p***	***df***	***F***	***p***
All subjects	(4, 137)	1.53	0.20	(4, 139)	7.76	< 0.01
Expert subjects	(4, 45)	0.39	0.81	(4, 44)	4.92	< 0.01
Non-expert subjects	(4, 45)	0.25	0.90	(4, 45)	1.52	0.21

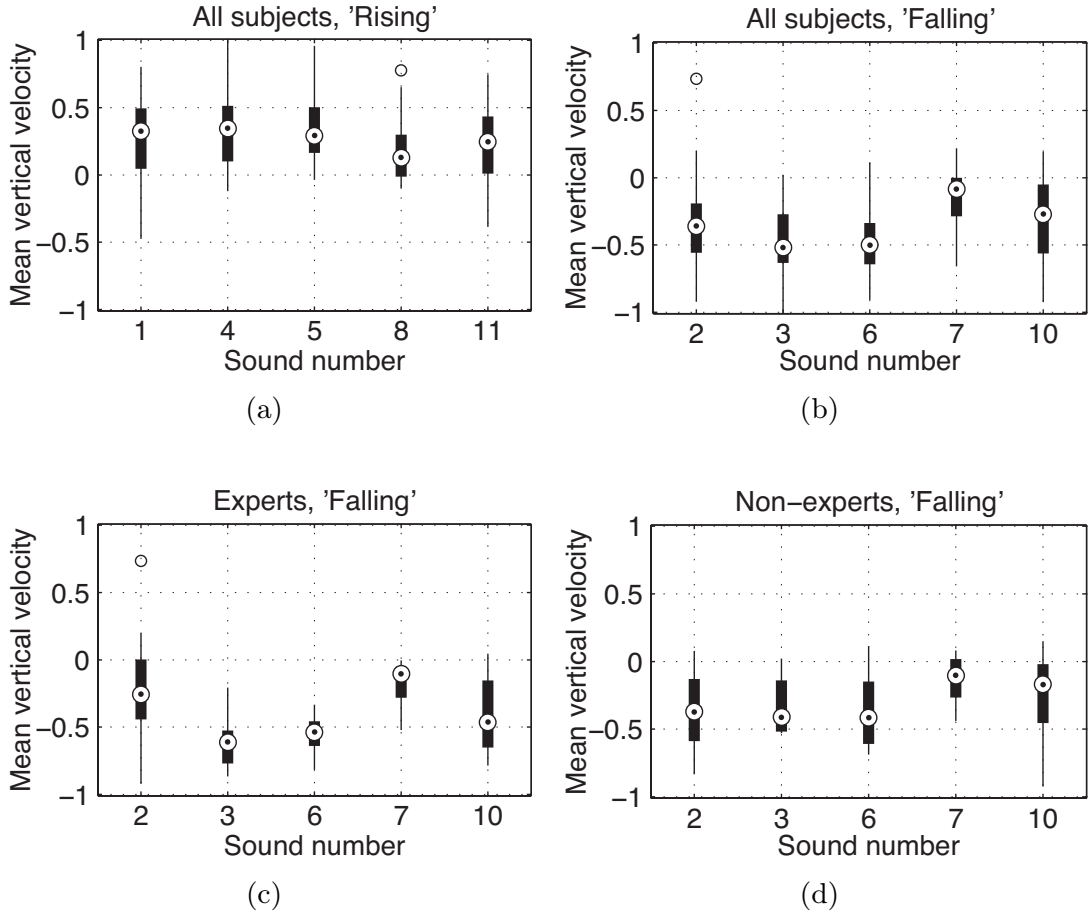
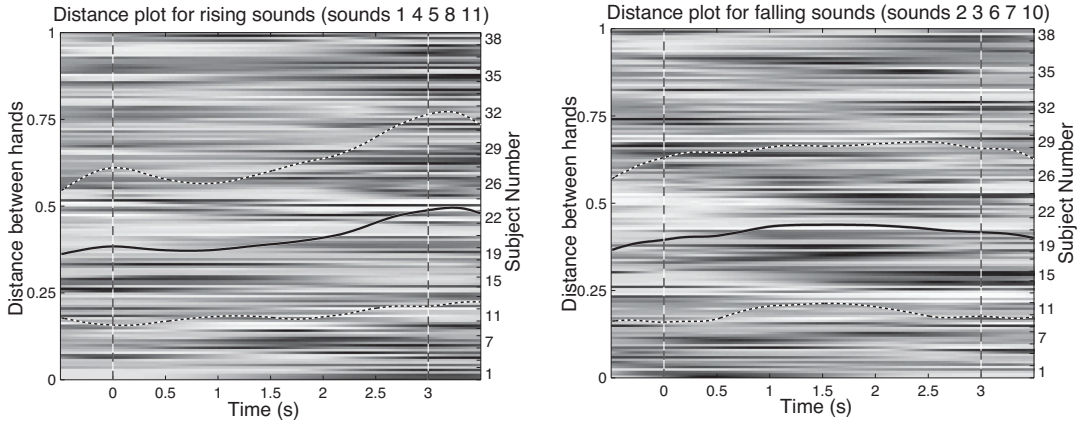


Fig. 9. Box plots of average vertical velocities for (a) rising sounds and (b) falling sounds. The difference between the sounds is greater for falling sounds than for rising sounds, and greater for (c) experts than for (d) non-experts. Note also that the sounds where the pitch is constant and only the spectral centroid is manipulated (Sounds 7 and 8) have the least extreme values in all the plots.

5.4 Pitch and Distance Between Hands

Eitan and Timmers pointed out that the relationship between pitch and motion features may be more complex than mapping pitch to vertical position [6]. For this reason, we have also analyzed how the distance between the hands corresponds to pitch frequency.

Figures 10(a) and 10(b) display the distance between the hands for sound tracings to rising and falling sounds, respectively. On the one hand, the black lines displaying the average distance features do not show very clear overall tendencies towards increasing or decreasing distance, but on the other hand, the underlying images, displaying the individual sound tracings, show that there is a substantial amount of change in the distance between the hands, both for the rising and the falling sounds.



(a) (b)

Fig. 10. Distance between hands for (a) rising sounds (142 sound tracings) and (b) falling sounds (144 sound tracings). The black line shows the mean value across all the data series, each horizontal line in the image displays the vertical position of a single sound tracing normalized per subject between 0 (hands close together, white) and 1 (hands far apart, black).

Figure 10 seems to vaguely suggest that participants let the hand distance increase for sounds with increasing pitch. Table 5 compares the change in distance for rising sounds versus falling sounds. For the two sounds where only the pitch is varied, there is a significant difference between the rising and falling sound tracings. The same is true when all of the rising sounds are compared to the falling sounds. On the contrary, we do not observe significant differences in the instances where the spectral centroid is varied.

Table 5. Results from t-tests comparing the average change in distance between hands of rising and falling sounds.

Average change in hand distance, rising versus falling sounds

Sounds	**Varying features**	***df***	***t***	***p***
5 vs. 6	Pitch	52	3.24	< 0.01
4 vs. 3	Pitch and spectral centroid	54	1.50	0.07
1 vs. 2	Pitch, opposing spectral centroid	53	1.09	0.14
8 vs. 7	Spectral centroid, pitched	60	0.48	0.32
11 vs. 10	Spectral centroid, non-pitched	57	-0.44	0.67
All rising sounds vs. all falling sounds		284	2.60	< 0.01

5.5 Acceleration Envelope for Pitched and Non-pitched Sounds

We have evaluated how the presence of a distinct pitch influences the sound tracing by comparing acceleration envelopes of sound tracings performed to pitched sounds and non-pitched sounds.

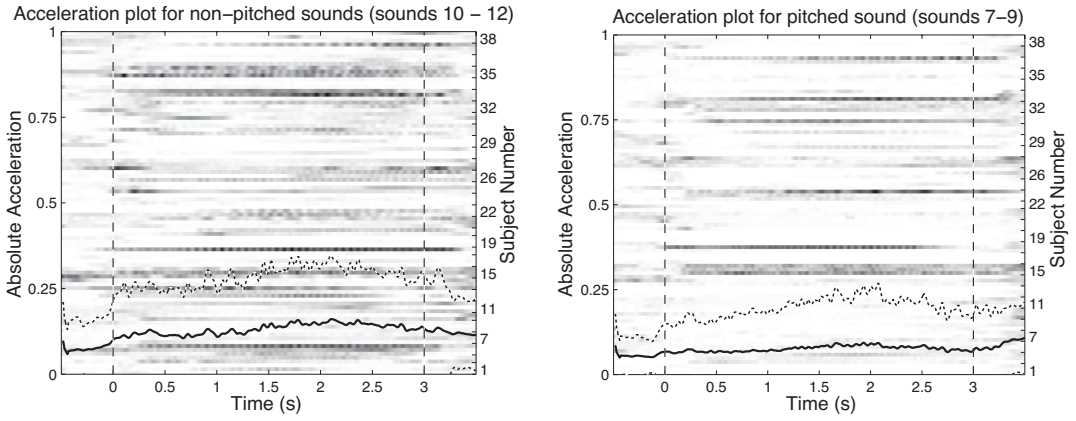
Three of the sounds used in the experiment were based on noise, and three were based on a stable tone with a fundamental frequency of 342 Hz. Within each of these categories, one sound had a falling spectral centroid, one had a rising spectral centroid and one had a stable spectral centroid.[8] Figure 11 shows the acceleration curves from the sound tracings to non-pitched and pitched sounds respectively. The mean acceleration was significantly higher for non-pitched sounds than pitched sounds, $t(179) = 5.53$, $p < 0.01$. For non-pitched sounds the mean normalized acceleration was 0.52, and for pitched sounds it was 0.28.

This significant distinction between acceleration values for pitched and non-pitched sounds was also found when the data from experts and non-experts was analyzed individually. Furthermore, no significant difference was found between the acceleration levels of experts and non-experts, $p = 0.46$ for both pitched and non-pitched sounds, respectively. See Table 6 for statistical results.

Table 6. Results from t-tests comparing acceleration of pitched to non-pitched sounds.

Acceleration, non-pitched and pitched sounds

Test description	***df***	***t***	***p***
Non-pitched vs pitched, all subjects	179	5.53	< 0.01
Non-pitched vs pitched, expert subjects	62	3.31	< 0.01
Non-pitched vs pitched, non-expert subjects	62	3.68	< 0.01
Noise, experts vs non-experts	61	0.10	0.46
Stable tone, experts vs non-experts	63	0.11	0.46



(a) (b)

Fig. 11. Acceleration for (a) non-pitched sounds and (b) pitched sounds. The black line shows the mean value per sample, each horizontal line displays the acceleration of a single sound tracing ranging from 0 (white) to 1 (black). The mean acceleration levels for the non-pitched sounds are generally higher than for pitched sounds.

[8] Sounds based on noise: 10, 11, and 12. Sounds based on a stable tone: 7, 8, and 9.

5.6 Summary of the Results

The results in this section have shown that the participants produced sound tracings with higher onset acceleration for impulsive sounds than for non-impulsive sounds. This was true for experts and non-experts. The onset time of musical experts was on average 75 ms ahead of non-experts. Furthermore, sounds without a distinct pitch seem to induce higher acceleration than pitched sounds.

Vertical displacement of the hands was found to be related to pitch and to spectral centroid. When pitch and spectral centroid moved in opposite directions, most subjects let the vertical position of the hands follow the perceived pitch. When only the spectral centroid was varied, there was less vertical motion than for sounds with varying pitch. This was particularly true for sounds with a stable pitch, as opposed to no perceivable pitch. Overall, falling sounds induced more vertical motion than rising sounds. For the falling sounds, the variance between the vertical velocity of the subjects was low, suggesting that there is more consistency within the expert group than in the non-expert group. Finally, there was significant difference between the change in hand distance for some of the sounds with falling and rising envelopes. We will discuss these findings in the next section.

6 Discussion

The following discussion will address the analysis method, the results from the previous section, as well as how the experiment setup and task may have influenced the results. We will put the results into context in relation to previous research, and in this way try to assess what can be learned from our findings. For certain sound features and specific motion features, we have observed a quite high consistency across the subjects. This supports the claim that there is a relationship between auditory and motor modalities.

The discussion is structured as follows: The statistical approach is discussed in Section 6.1. In Section 6.2, we evaluate the results from Section 5.2. Section 6.3 discusses the results from Sections 5.3 and 5.4, and results from Section 5.5 are discussed in Section 6.4. In Section 6.5, we provide a more general evaluation of the results.

6.1 Statistical Method

Using statistics to evaluate the differences between the groups does provide some indication of the tendencies in our data set. However, it should be noted that the t-test and ANOVA methods assume that the data is normally distributed. The subsets of data in our statistical analyses were tested for normality using a Jaque-Bera test[9] with significance level 0.05. This test revealed that 13 out of the 52 data sets in our experiments do not follow a normal distribution, and thus the statistical results can not alone be used to make strong conclusions.

[9] http://en.wikipedia.org/wiki/Jarque–Bera_test

Nevertheless, the results from the statistical tests support the results that are shown in the corresponding feature plots. This gives us reason to believe that the statistical results are trustworthy.

It should be noted that for the sample-by-sample based plots, the standard deviations are quite high, particularly for the acceleration curves shown in Figure 11. Since these plots were derived on a sample-by sample basis, the high standard deviations are not very surprising. In Figure 11, the high standard deviation reflects that even though several subjects had a high overall acceleration, they vary between high and low acceleration throughout the sound tracing. This demonstrates the importance of looking at the individual sound tracings in the plot, not only the mean and standard deviation curves.

6.2 Impulsive Sound Onset

Let us have a look at the results presented in Section 5.2, where the onset acceleration of sound tracings to impulsive sounds was shown to be much higher than for non-impulsive sounds. In our opinion, these results can best be explained from a *causation* perspective. In other words: people link the impulsive characteristics of the sound to some sort of impulsive action that could have generated it. An example of an impulsive action is displayed in Figure 12. The figure shows how the subject performs an accentuated attack, with high acceleration, followed by a slower falling slope down to a resting position. The sound tracing resembles that of crashing two cymbals together.

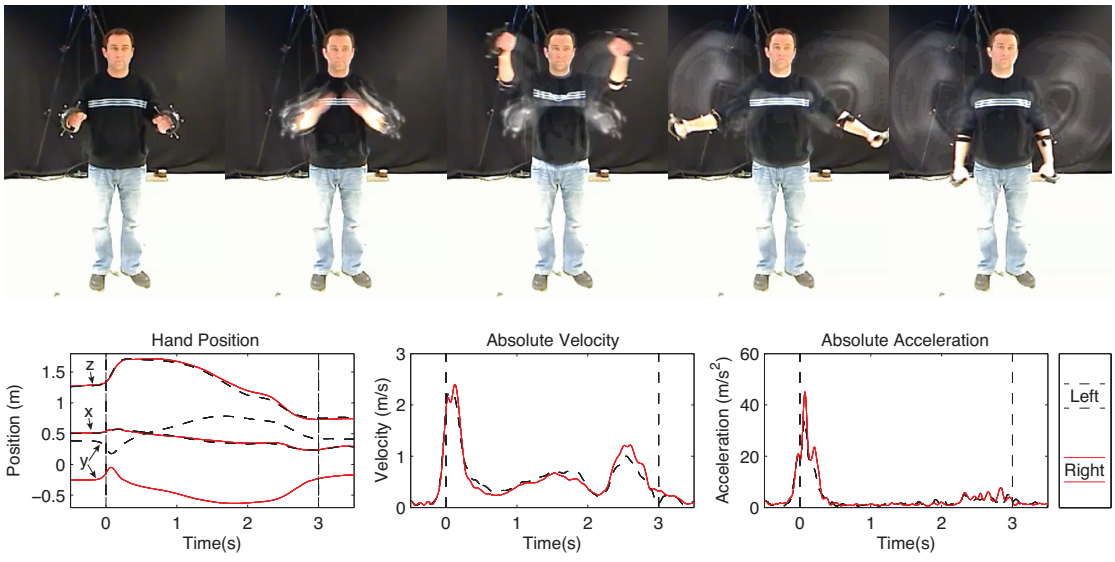


Fig. 12. The figure displays an impulsive action, with motion history images and the calculated motion features for the left (black) and right (red) hand. Motion history images show the current frame and the average frame difference for past video frames [16]. x (back/forth), y (sideways) and z (up/down) are position coordinates from the motion capture system.

Lakoff and Johnson [18] talk about causation as an important element in understanding objects and events in the world. As an example, they mention a paper airplane which we understand primarily as paper, and secondarily as airplane. The making, or cause, of the paper airplane is an essential element in our understanding of this object. It is interesting to compare the *causation* term to Schaeffer's theory of the sonic object [26]. According to Schaeffer, sonic objects are defined by their *causal coherence*, something which Godøy linked to gesture-sensations and chunking in sound [9]. According to Lakoff and Johnson [18], the causation of an object is partly emergent, or something that is present in the object itself, which makes it possible to understand an object as a holistic and metaphorical unit. Again, compared with the sonic object, this emergent property works well with Schaeffer's principle of *stress-articulation*, or natural discontinuities in the continuous sound signal [26].

Following these thoughts, it seems apparent that people link the impulsive onset of sounds to some sort of impulsive or ballistic action or event. Given the constraints of two handles to imitate the sound, some participants imitated the action of crashing two cymbals together, while others imitated a single-handed or two-handed striking action. The discontinuity of stress-articulation in sound has its motor counterpart in the higher derivatives of position data, here shown by a high onset acceleration.

6.3 Sound Frequency Features

In addition to the causation perspective, Lakoff and Johnson also introduced the *metaphor* perspective. Metaphors are crucial to our understanding of events and objects. We understand some event or object by using a metaphor to describe it. According to Eitan and Timmers [6], Cox [4] has linked the metaphor "more is up" to a perceptual relationship between vertical position and pitch. In our experiment, the results show clearly that there is a relationship between these features, and that most subjects follow rising pitch with upward motion, and falling pitch with downward motion.

However, Eitan and Timmers have shown that for pitch, up is not always the best metaphor [6]. In their experiments, low pitch has also been associated with with metaphors like "heavy" or "big". Also, Walker [27] described rising pitch to be a good descriptor for increasing temperature. For this exact reason, we also investigated if the motion feature *hand distance* was related to the rising and falling envelopes. Our results show that when all the rising sounds were compared to the falling ones, there was a significant difference in the average change in hand distance. However, a closer look at the results revealed that on average, for sounds with a rising envelope and with a falling envelope alike, the distance between the hands increased. The significant difference was therefore only due to a faster increase in distance for rising sounds than for falling sounds. In addition, a significant difference between rising and falling envelopes occurred when only the parameter *pitch* was varied. In this case, the average hand distance decreased for falling pitch and increased for rising pitch, and thus to some extent, defined a relationship between these features. Nevertheless, even though several

subjects did change the distance between their hands for these sounds, there was much less similarity among the subjects compared to the vertical position feature. Some subjects moved their hands apart while other moved them towards each other. So to conclude, the "more-is-up" metaphor for pitch seems to be the best metaphor to describe the results in our data set.

An example of a sound tracing performed to a sound with falling pitch and rising spectral centroid is shown in Figure 13. The motion history images show how the subject prepares for the sound tracing by moving his hands up, then moving them down and out in such a way that the vertical position follows the pitch envelope. At the end of the sound tracing, the subjects increasingly vibrates the right hand, as shown in the acceleration plot. This might be a gesture performed to imitate the increased spectral centroid which is increasingly prominent towards the end of the sound file. As the motion history images show, the hand distance first increases and then decreases in a sound where the pitch is constantly falling and the spectral centroid is constantly rising.

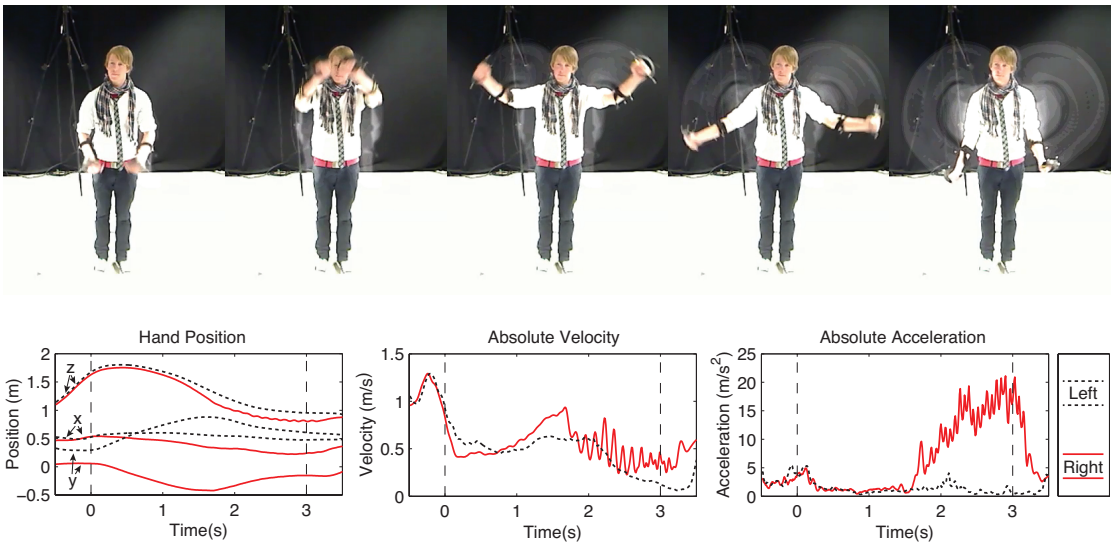


Fig. 13. The figure displays motion history images and feature plots for the left and right hand for a sound tracing performed to a sound with falling pitch, rising spectral centroid, and a bell-shaped dynamic envelope.

An interesting feature regarding the rising and falling sounds, and the correlation to vertical position, is that sound seems to be more descriptive than motion. Our results show that even though the sounds were easy to tell apart, the sound tracings that were performed to the different sounds were similar. This implies that although you can describe certain perceptual features of the sound through an action, it is not necessarily clear which perceptual feature(s) the action imitates. Elevating a hand might refer to increasing pitch or to increasing spectral centroid, or to something else.

6.4 Pitched versus Non-pitched Sounds

We observed a significantly higher amount of acceleration in sound tracings performed to non-pitched sounds than to pitched sounds. This may be explained by participant's associations with this sound property. Sounds based on noise have wind-like properties, which might cause people to move a lot, as if they were blowing with the wind or creating the wind themselves. Pitched sounds, on the other hand, seem to provide something stable for the participants to "hold on to", that is not provided by the non-pitched sounds.

For Sounds 9 and 12, which both had stable spectral centroids, we observed that some participants started shaking or rotating their hands, gradually increasing the frequency or amplitude of the shaking. One example of this is shown in Figure 14. As these sounds had no change in pitch or spectral centroid, the loudness envelope of the sounds seem to have been the main influence in these instances. The increased shaking or rotation intensity may be explained by some sort of engine metaphor: we believe participants wanted to follow the gradually increasing loudness envelope by supplying more and more energy to the sound through their motion.

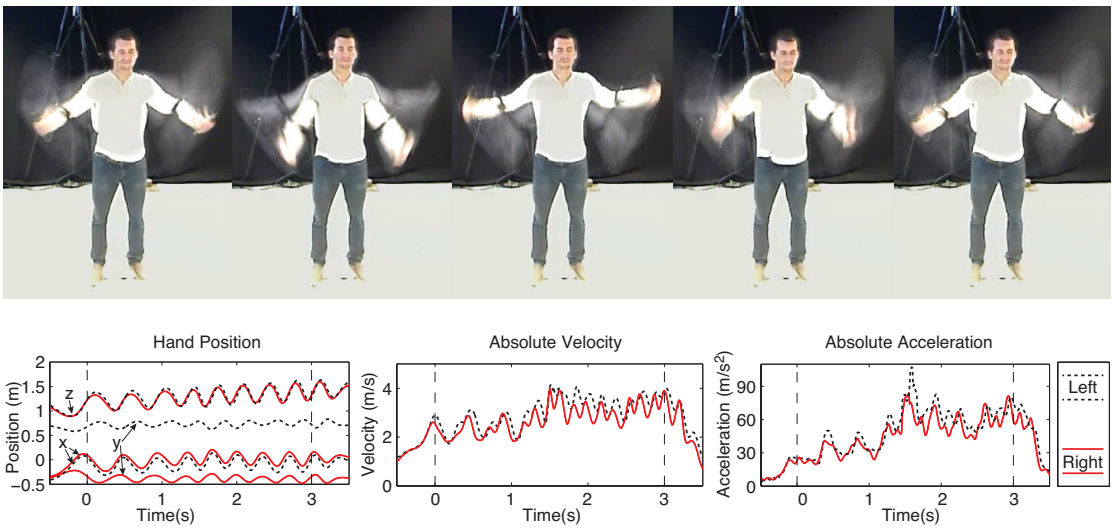


Fig. 14. The figure shows motion history images of a subject moving to a noise-based sound. Plots of xyz-position, velocity, and acceleration are shown below.

6.5 Final Remarks

Admittedly, the results presented in this paper are related to the context of the experiment, and cannot necessarily be claimed to be valid outside this setting. The way in which subjects solved the task may have been influenced by the instructions, which were to imagine that moving the hands in the air created the sound. Even though we did not provide the subjects with an a priori metaphor

connecting upward motion to increasing pitch, the options for movement were limited. Godøy has postulated that our musical motions are goal-directed, and that these motions are guided by goal-postures describing the shape and position of our end-effectors [11]. These positional goal-points may have been more consciously exposed than features describing the relationship between the hands or details of the trajectory between, for instance, a high and low goal point. In light of the experiment task, other pitch relationships like the one between low pitch and "heavy" or "big" [6], may have been less accessible than drawing a trajectory between two positional goal points. Some subjects may have consciously or unconsciously used a feature like hand distance to describe pitch, but as our results show, this was more inconsistent between subjects.

Even though the experimental setup may have prevented the subjects from using other descriptors than spatio-temporal ones, we are confident that the results show some indication of how people relate spatio-temporal features of motion to features of sound. The participants were given an imagined instrument, and they made their own mental model of how this instrument worked. Some aspects of these mental models were similar for the majority of the subjects.

7 Conclusions and Future Work

We have presented an experiment on sound tracing, where motions performed to sound have been analyzed from micro and meso timescale perspectives. Plotting of serial motion features at the micro timescale was used to obtain impressions of general tendencies in the data set, and statistical evaluations of single-value features at the meso timescale indicated the significance of these tendencies.

Rising pitch, and rising spectral centroid correlated strongly with upward motion, and similarly, falling pitch and spectral centroid, correlated strongly with downward motion. When pitch and spectral centroid moved in opposite directions, participants followed the pitch feature. Furthermore, sounds based on noise induced higher overall acceleration than sounds based on a steady pitch, and sounds with an impulsive onset caused a high acceleration peak in the beginning of the sound tracing.

To follow up on this experiment, we are currently starting to evaluate advantages and disadvantages of different methods for analyzing sound tracings. We believe that the different approaches that have been taken to analyze such data provide different types of knowledge, and that the choice of analysis method is important. For instance, some methods may be good at revealing action-sound relationships at a low timescale level, while others may work better at chunk-level or higher. We are also moving towards applying the results from our studies in development of new interfaces for musical expression.

8 Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement n° 257906, Engineering Proprioception in Computer Systems (EPiCS).

References

1. Publication Manual of the American Psychological Association. American Psychological Association, Washington, DC, 6th edn. (2010)
2. Caramiaux, B., Bevilacqua, F., Schnell, N.: Towards a gesture-sound cross-modal analysis. In: Kopp, S., Wachsmuth, I. (eds.) Gesture in Embodied Communication and Human-Computer Interaction, LNCS, vol. 5934, pp. 158–170. Springer, Heidelberg (2010)
3. Caramiaux, B., Bevilacqua, F., Schnell, N.: Sound selection by gestures. In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 329–330. Oslo, Norway (2011)
4. Cox, A.W.: The Metaphoric logic of Musical Motion and Space. Ph.D. thesis, University of Oregon (1999)
5. Eitan, Z., Granot, R.Y.: How music moves: Musical parameters and listeners' images of motion. Music Perception 23(3), 221–248 (2006)
6. Eitan, Z., Timmers, R.: Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. Cognition 114(3), 405–422 (2010)
7. Gjerdingen, R., Perrott, D.: Scanning the dial: The rapid recognition of music genres. Journal of New Music Research 37(2), 93–100 (2008)
8. Glette, K., Jensenius, A.R., Godøy, R.I.: Extracting action-sound features from a sound-tracing study. In: Proceedings of the Second Norwegian Artificial Intelligence Symposium. pp. 63–66 (2010)
9. Godøy, R.I.: Gestural-sonorous objects: Embodied extensions of Schaeffer's conceptual apparatus. Organised Sound 11(2), 149–157 (2006)
10. Godøy, R.I.: Chunking sound for musical analysis. In: Ystad, S., Kronland-Martinet, R., Jensen, K. (eds.) Computer Music Modeling and Retrieval. Genesis of Meaning in Sound and Music, LNCS, vol. 5493, pp. 67–80. Springer, Heidelberg (2009)
11. Godøy, R.I.: Gestural affordances of musical sound. In: Godøy, R.I., Leman, M. (eds.) Musical Gestures: Sound, Movement, and Meaning, chap. 5, pp. 103–125. Routledge (2010)
12. Godøy, R.I., Haga, E., Jensenius, A.R.: Exploring music-related gestures by sound-tracing. A preliminary study. In: Ng, K. (ed.) COST287-ConGAS 2nd Int. Symposium on Gesture Interfaces for Multimedia Systems. pp. 27–33. Leeds (2006)
13. Godøy, R.I., Jensenius, A.R.: Body movement in music information retrieval. In: Proceedings of the 10th International Society for Music Information Retrieval Conference. Kobe, Japan (2009)
14. Husserl, E.: The Phenomenology of Internal Time Consciousness. Indiana University Press, Bloomington (1964, trans JS Churchill)
15. Jensenius, A.R.: Action–Sound : Developing Methods and Tools for Studying Music-Related Bodily Movement. Ph.D. thesis, University of Oslo (2007)

16. Jensenius, A.R.: Some video abstraction techniques for displaying body movement in analysis and performance. Leonardo Music Journal (Forthcoming 2012)
17. Kozak, M., Nymoen, K., Godøy, R.I.: The effects of spectral features of sound on gesture type and timing. In: Efthimiou, E. (ed.) Post-proceedings of Gesture Workshop 2011: Gesture in Embodied Communication and Human-Computer Interaction. LNCS, Springer (Forthcoming 2012)
18. Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago Press, Chicago, IL. (1980)
19. Leman, M.: Embodied Music Cognition and Mediation Technology. MIT Press, Cambridge, Mass. (2007)
20. Leman, M.: Systematic musicology at the crossroads of modern music research. In: Schneider, A. (ed.) Systematic and Comparative Musicology: Concepts, Methods, Findings, Hamburger Jahrbuch für Musikwissenschaft, vol. 24, pp. 89–115. Peter Lang, Frankfurt (2008)
21. Merer, A., Ystad, S., Kronland-Martinet, R., Aramaki, M.: Semiotics of sounds evoking motions: Categorization and acoustic features. In: Kronland-Martinet, R., Ystad, S., Jensen, K. (eds.) Computer Music Modeling and Retrieval. Sense of Sounds, pp. 139–158. No. 4969 in Lecture Notes in Computer Science, Springer-Verlag, Heidelberg (2008)
22. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 63, 81–97 (1956)
23. van Nort, D.: Instrumental listening: sonic gesture as design principle. Organised Sound 14(02), 177–187 (2009)
24. Nymoen, K., Caramiaux, B., Kozak, M., Torresen, J.: Analyzing sound tracings: a multimodal approach to music information retrieval. In: Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies. pp. 39–44. ACM, New York, NY, USA (2011)
25. Nymoen, K., Glette, K., Skogstad, S.A., Torresen, J., Jensenius, A.R.: Searching for cross-individual relationships between sound and movement features using an SVM classifier. In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 259–262. Sydney (2010)
26. Schaeffer, P., Reibel, G.: Solfège de l'objet sonore. ORTF, Paris, France (1967, Reedited by INA-GRM 1998)
27. Walker, B.N., Kramer, G.: Mappings and metaphors in auditory displays: An experimental assessment. ACM Trans. Appl. Percept. 2(4), 407–412 (2005)
28. Wilson, M.: Six views of embodied cognition. Psychonomic Bulletin & Review 9(4), 625–636 (2002)

Paper VIII

Analysing Correspondence Between Sound Objects and Body Motion.
K. Nymoen, R.I. Godøy, A.R. Jensenius and J. Torresen.
ACM Transactions on Applied Perception (to appear),
submitted June 4 2012
accepted December 2 2012

The paper is in press, therefore not included here.