Department of Signal Processing and Acoustics

Sonic Gestures and Rhythmic Interaction between the Human and the Computer

Antti Jylhä

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 32/2012

Sonic Gestures and Rhythmic Interaction between the Human and the Computer

Antti Jylhä

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the School of Electrical Engineering for public examination and debate in Auditorium S1 at the Aalto University School of Electrical Engineering (Espoo, Finland) on the 20th of April 2012 at 12 noon (at 12 o'clock).

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics

Supervisor
Prof. Vesa Välimäki

Instructor
Dr. Cumhur Erkut

Preliminary examiners
Prof. Petri Toiviainen, University of Jyväskylä, Finland
Dr. Adam J. Sporka, Czech Technical University in Prague, Czech Republic

Opponent
Assoc. Prof. Stefania Serafin, Aalborg University Copenhagen, Denmark

Aalto University publication series
DOCTORAL DISSERTATIONS 32/2012

ISBN 978-952-60-4553-5 (printed)
ISBN 978-952-60-4554-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Unigrafia Oy
Helsinki 2012

Finland

The dissertation can be read at http://lib.tkk.fi/Diss/

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Antti Jylhä

Name of the doctoral dissertation
Sonic Gestures and Rhythmic Interaction between the Human and the Computer

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL DISSERTATIONS 32/2012

Field of research Acoustics and audio signal processing

Manuscript submitted 15 December 2011 **Manuscript revised** 15 March 2012

Date of the defence 20 April 2012 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

This thesis addresses the use of sonic gestures as input in human-computer interaction with a special applicability focus on rhythmic interactive systems and their design and evaluation. Sonic gestures are defined as human-generated sounding actions which convey information to a computational system. Examples of such gestures are impulsive sounding actions such as hand claps and finger snaps, sustained actions such as humming and blowing, and iterative actions, such as tapping a table to the beat of music.

The use of sonic gestures at the interface requires analysis algorithms that are capable of tracking the desired information from an audio stream containing the human-generated sounds. In interactive systems, these algorithms must be capable of real-time processing. In this thesis, the focus is on percussive sonic gestures, which can be seen to be analogous to the sounds of percussive instruments. Therefore, it is reasonable to assume that the same tools that are applied for retrieving information from drums and percussion in music can be deployed for sonic gesture analysis. This work presents algorithms for the classification of different percussive sounds, such as different types of hand claps.

To demonstrate the use of sonic gestures, a hand clap interface capable of recognizing different hand clap types and extracting continuous information, such as the tempo, from a clapping sequence has been developed. This interface has been utilized in the development of various rhythmic prototype applications, most importantly a system called iPalmas, an interactive Flamenco rhythm tutor. The iPalmas system can produce realistic-sounding synthetic Flamenco hand clapping patterns to the user, listen to the clapping of the user, and give audiovisual feedback on the learning and performance.

The iPalmas system was evaluated in a subjective experiment, resulting in qualitative and quantitative findings related to the system design, the human capabilities, and the interaction. In conjunction with this evaluation, a structured framework for evaluating this kind of systems has been proposed. Based on the evaluation results, the system has undergone iterative development of the audiovisual feedback elements.

The main outcomes of the thesis are a novel definition of sonic gestures in human-computer interaction and a taxonomy of the information they can convey to computational systems and the interactive iPalmas system, resulting in several relevant findings that can be generalized in the design and evaluation of rhythmic interactive systems.

Keywords Sonic interaction design, rhythmic interaction, audio input

ISBN (printed) 978-952-60-4553-5 **ISBN (pdf)** 978-952-60-4554-2

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Espoo **Location of printing** Helsinki **Year** 2012

Pages 167 **The dissertation can be read at** http://lib.tkk.fi/Diss/

A! Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Antti Jylhä

Väitöskirjan nimi
Äänieleet ja rytminen vuorovaikutus ihmisen ja tietokoneen välillä

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Signaalinkäsittelyn ja akustiikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 32/2012

Tutkimusala Akustiikka ja äänenkäsittelytekniikka

Käsikirjoituksen pvm 15.12.2011 **Korjatun käsikirjoituksen pvm** 15.03.2012

Väitöspäivä 20.04.2012 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Tämä väitöskirja käsittelee äänieleiden käyttöä ihmisen ja tietokoneen välisessä vuorovaikutuksessa, keskittyen erityisesti rytmisten vuorovaikutusjärjestelmien suunnitteluun ja arviointiin. Äänieleet määritellään ihmisen tuottamiksi äänitapahtumiksi, jotka välittävät informaatiota tietokonejärjestelmälle. Äänieleet voivat olla impulsiivisia, kuten taputukset ja sormien napsutukset, jatkuvia, kuten vihellys ja puhallus, sekä iteratiivisia, kuten pöydän naputtaminen musiikin tahdissa.

Äänieleiden käyttö vuorovaikutusrajapinnassa vaatii algoritmeja, jotka pystyvät erottamaan halutun äänieleiden sisältämän informaation äänisyötevirrasta. Vuorovaikutteisissa järjestelmissä näiden algoritmien on kyettävä reaaliaikaiseen äänenkäsittelyyn. Tässä väitöskirjassa keskitytään perkussiivisiin äänieleisiin, jotka ovat rinnastettavissa perkussiosoittimiin; niinpä näiden äänieleiden analyysiin voidaan soveltaa samoja menetelmiä kuin perkussiosoittimien äänten analyysiin. Tässä työssä esitetään algoritmeja erilaisten perkussiivisten äänieletapahtumien tunnistamiseen.

Äänieleiden käyttöä vuorovaikutteisissa järjestelmissä havainnollistetaan kehitetyllä vuorovaikutusrajapinnalla, joka tunnistaa erityyppisiä käsien taputuksia ja jatkuvan taputuksen sisältämää jatkuva-aikaista informaatiota, kuten taputuksen tempo. Rajapintaa käyttäen on kehitetty erilaisia prototyyppisovelluksia, joista merkittävin on iPalmas, vuorovaikutteinen flamenco-opettaja. iPalmas osaa tuottaa realistisen kuuloisia flamenco rytmejä taputussynteesin avulla, kuunnella käyttäjän taputuksia ja analysoida niitä, sekä antaa audiovisuaalista palautetta taputussarjojen oppimisesta ja suorittamisesta.

iPalmas -järjestelmälle tehtiin käyttäjäkoe, jonka perusteella saatiin paljon kvalitatiivista ja kvantitatiivista informaatiota liittyen sekä järjestelmään itseensä että käyttäjien rytmisiin taipumuksiin ja vuorovaikutukseen järjestelmän kanssa. Arvioinnin yhteydessä kehitettiin myös arvioinnin viitekehys tämänkaltaisten vuorovaikutusjärjestelmien arviointiin. Arvioinnin perusteella järjestelmän audiovisuaalinen palaute suunniteltiin uudelleen.

Väitöskirjan keskeiset tulokset ovat uusi määritelmä äänieleille ja luokittelu näiden välittämälle informaatiolle sekä iPalmas -järjestelmä ja sen arvioinnin yhteydessä saavutetut löydökset, joita voidaan yleistää rytmisten vuorovaikutusjärjestelmien suunnitteluun ja arviointiin.

Avainsanat Ääniperusteinen vuorovaikutussuunnittelu, rytminen vuorovaikutus, äänisyöte

ISBN (painettu) 978-952-60-4553-5 **ISBN (pdf)** 978-952-60-4554-2

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Espoo **Painopaikka** Helsinki **Vuosi** 2012

Sivumäärä 167 **Luettavissa verkossa osoitteessa** http://lib.tkk.fi/Diss/

Preface

Looking back at the past few years spent working on this thesis I must say that this journey has been most interesting and inspiring. When it all started four or five years ago, I had little idea of all the places it would take me. In the end, all the pieces have fallen into place surprisingly well.

This work was carried out at the Department of Signal Processing and Acoustics at Aalto University School of Electrical Engineering (formerly the Laboratory of Acoustics and Audio Signal Processing at Helsinki University of Technology) during the years 2008-2012. The work was financially supported by the Graduate School at Aalto University School of Electrical Engineering, the Academy of Finland (project 120583 SCHEMA-SID), Emil Aaltonen Foundation, and the Finnish Foundation for Technology Promotion.

I am eternally grateful to my instructor, Dr. Cumhur Erkut, for hiring me in his project from the beginning on, for all the patience and openness towards my ways of working and my sometimes erratic ideas, and for all the support he has given me over the years. As a living whirlpool of ideas and innovation, it is impossible not to get inspired working with you! Thank you for listening, brainstorming, challenging, and being a friend.

I would also like to thank my supervisor, Professor Vesa Välimäki, for his insightful support throughout the years and all the help with getting this thesis finalized, and for being an excellent boss overall.

I could never have finalized this work without all my brilliant co-authors. I would especially like to thank Inger Ekman for extremely influential collaboration, and for bringing up the idea of working with Flamenco in the first place. Also, Dr. Koray Tahiroğlu, Umut Şimşekli, Professor A. Taylan Cemgil, Matti Pesonen, and Reha Disçioğlu — it has been a great pleasure to collaborate with all of you. I also wish to thank my pre-examiners, Professor Petri Toiviainen and Associate

Professor Adam J. Sporka for their constructive comments regarding the manuscript, and Luis Costa for carefully proof-reading my text.

I have had the honor of being part of the unique work community of the acoustics lab. Despite all the changes that we have experienced during the past few years the working atmosphere has remained supportive, friendly, fun, and inspirational. I want to thank all the aku-people for maintaining the lab as a nice place to be. Heidi, Hynde, Lea, Mara, Markku, Mirja, and Ulla deserve big thanks for maintaining the infrastructure and handling various practical things. Hannu, it has been great sharing the room with you; thanks for all the relevant (and not so) discussions and friendship. Henkka, interacting with you and knowing you all these years has been brilliant. Special thanks also to the Helsinki Mobile Phone Orchestra, Jussi P. for the LaTeX template among other things, Sami for always keeping the spirits high, Tapsa, Olli, Mikkis, and Tuomo for the music and always fun discussions, Akis, Heikku, Henna, Jari, Javier, Juho, Julian, Jussi R., Magge, Marko, Okko, Rafael, Seppo, Symeon, Ville P., and all the current labsters and also the previous ones, including Jykke, Laura, Mairas, Matti (In Memoriam 1946–2010), Miikka, Mikko, Stefano, Tomppa, Ville S., and so many others. I thank all of you for making this workplace the most memorable ever.

I want to thank my family and friends for everything in life, especially my mom and dad, Ilse and Mikko, who have always been there to support me, and my siblings Ville and Ilona. Mummi, thanks for all the great food and fun Canasta sessions. Thanks also to the Kujala family, especially Tea, Roope, and Sara, for being part of my life, and to my in-laws, especially Anita for taking good care of Sohvi. Harry, Sami, and The Mökki Gang, thank you for your lasting friendship and all the extra-curricular activities.

Finally and most importantly, I want to express the deepest thanks to my lovely wife Tuuli for all the love, support, and patience she has given me over the years. I owe you for so much I cannot even describe in words. You make me a better man. I love you.

Espoo, March 15, 2012,

Antti Jylhä

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Antti Jylhä. Sonic Gestures as Input in Human-Computer Interaction: Towards a Systematic Approach. In *Proceedings of Sound and Music Computing Conference*, Padova, Italy, pp. 1–7, July 2011.

II Antti Jylhä and Cumhur Erkut. Inferring the Hand Configuration from Hand Clapping Sounds. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx)*, Espoo, Finland, pp. 301–304, September 2008.

III Umut Şimşekli, Antti Jylhä, Cumhur Erkut, and A. Taylan Cemgil. Real-Time Recognition of Percussive Sounds by a Model-Based Method. *EURASIP Journal of Advances in Signal Processing*, pp. 1–14, January 2011.

IV Antti Jylhä, Cumhur Erkut, Umut Şimşekli, and A. Taylan Cemgil. Sonic Handprints: Person Identification with Hand Clapping Sounds by a Model-Based Method. In *Proceedings of the 45th Conference of the Audio Engineering Society*, Espoo, Finland, pp. 1–6, March 2012.

V Antti Jylhä and Cumhur Erkut. A Hand Clap Interface for Sonic Interaction with the Computer. In *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems*, Boston, MA, USA, pp. 3175–3180, April 2009.

VI Antti Jylhä, Inger Ekman, Cumhur Erkut, and Koray Tahiroğlu. Design and Evaluation of Human-Computer Rhythmic Interaction in a Tutoring System. *Computer Music Journal*, vol. 35, no. 2, pp. 36–48, May 2011.

VII Antti Jylhä, Cumhur Erkut, Matti Pesonen, and Inger Ekman. Simulation of Rhythmic Learning - A Case Study. In *Proceedings of the 5th Audio Mostly Conference*, Glasgow, UK, pp. 146–149, September 2010.

VIII Cumhur Erkut, Antti Jylhä, and Reha Disçioğlu. A Structured Design and Evaluation Model with Application to Rhythmic Interaction Displays. In *Proceedings of International Conference on New Interfaces for Musical Expression (NIME)*, Oslo, Norway, pp. 477–480, May 2011.

IX Antti Jylhä and Cumhur Erkut. Auditory Feedback in an Interactive Rhythmic Tutoring System. In *Proceedings of the 6th Audio Mostly Conference*, Coimbra, Portugal, pp. 109–115, September 2011.

Author's Contribution

Publication I: "Sonic Gestures as Input in Human-Computer Interaction: Towards a Systematic Approach"

The author is solely responsible for writing the article.

Publication II: "Inferring the Hand Configuration from Hand Clapping Sounds"

The author implemented the algorithm for recognizing different hand configurations from their sound, designed the experiment for evaluating the algorithm, participated in the recording of test data, labeled the data and evaluated the algorithm, and wrote Sections 3, 4, and 5.

Publication III: "Real-Time Recognition of Percussive Sounds by a Model-Based Method"

The author had an integral part in planning and conducting the study, contributed in the recordings of test data, and participated in the analysis of the results. Sections 1, 2, 6, 8, and 9 were mainly written by the present author, who also contributed in writing Sections 3 and 5.

Publication IV: "Sonic Handprints: Person Identification with Hand Clapping Sounds by a Model-Based Method"

The author came up with the original idea for the article, performed the recordings for the test data, participated in obtaining the results from the data and their analysis, and wrote all of the article except for Section 3.

Publication V: "A Hand Clap Interface for Sonic Interaction with the Computer"

The author designed and implemented the hand clap interaction interface, designed and implemented the prototype applications except for the hand clap synthesis engine, and wrote the article in collaboration with the second author.

Publication VI: "Design and Evaluation of Human-Computer Rhythmic Interaction in a Tutoring System"

The author designed with co-authors the concept for the system in the paper and implemented the hand clapping analysis, the control logic, and the main interface of the system. The present author along with the co-authors also designed the evaluation experiment and was solely responsible for conducting the evaluation sessions and preparing the test data for statistical analysis. The present author wrote the introduction, the sections concerning implementation, and the conclusions and participated in the writing of the section concerning results and analysis.

Publication VII: "Simulation of Rhythmic Learning - A Case Study"

The author came up with the original idea for the article, designed and implemented the model for a virtual learning palmero except for the rhythmic stability analysis function, and wrote the article excluding Section 2.

Publication VIII: "A Structured Design and Evaluation Model with Application to Rhythmic Interaction Displays"

The author contributed to the evaluation described in the paper and wrote most of Sections 3 and 4.2.

Publication IX: "Auditory Feedback in an Interactive Rhythmic Tutoring System"

The author designed the auditory feedback components based on subjective evaluation of the iPalmas system. The present author also wrote the article, except for Section 4.1.

List of Symbols

$p(\theta)$	probability distribution of θ
T_{ooi}	onset-to-onset interval [seconds]
y	observation data in a probabilistic model
θ	hidden quantities in a probabilistic model
θ_{bpm}	tempo [beats per minute]

List of Abbreviations

bpm	beats per minute
FFT	fast Fourier transform
FIR	finite impulse response
GMM	Gaussian mixture model
HCI	human-computer interaction
HMM	hidden Markov model
MFCC	Mel-frequency cepstral coefficients
MIR	music information retrieval
OOI	onset-to-onset interval
PCA	principal component analysis
PD	Pure Data
SG	sonic gesture
SID	sonic interaction design
SMC	sound and music computing
STFT	short-term Fourier transform
Tai-Chi	Tangible Acoustic Interfaces for Computer-Human Interaction
UML	unified modeling language

1. Introduction

Humans are experts at using sound as a key modality in communication. In addition to speech, we are able to produce sounds conveying meaningful information in a multitude of contexts, including, for example, applause to indicate our appreciation on a performance, non-speech utterances to indicate confirmation, negation, or surprise, and tapping or clapping a beat to accompany a musical piece, just to name a few. Furthermore, humans can successfully interpret the information contained in the sound produced by others. For example, musicians can listen to the starting count of a drummer to be able to start their performance synchronously with a shared tempo, and in general people can recognize the identity of familiar people based on the sound of their footsteps.

The use of sound in interactive computational systems is studied in the field of sonic interaction design (SID)[1] (Rocchesso and Serafin, 2009). SID as a discipline considers sound as the conveyor of information, emotions, and aesthetics. SID is by nature multi-disciplinary, as it combines the fields of interaction design, audio signal processing, arts, and to some degree also humanities. The multi-faceted questions regarding sound and interaction indeed require examination from many angles.

In general, interaction comes in two flavors: discrete and continuous. As noted by Rocchesso, Polotti, and Delle Monache (2009), most pre-industrial human actions in the world were continuous, but technology has brought us more and more interfaces that operate in a discrete manner. As an example of differentiating the two, consider getting water from a well. Modern pumps are often operated by pushing a button to switch the pump on or off, whereas older pumps were operated often by a lever, which had to be constantly pushed up and down manually.

[1] http://sid.soundobject.org/

The main feedback in both systems is water running from the outlet; however, the lever-operated pump gives the operating person much more control over the water flow and also gives immediate haptic and auditory feedback on every stroke of the lever. A more modern example of discrete and continuous interaction is operating a visual computer interface with a mouse. A click on the mouse button to invoke an application is discrete, while moving the mouse to move the cursor is continuous.

Until the late eighties, it seemed that the research on informational sounds in computer interfaces concentrated on speech output (Brewster, 2003). Non-speech sounds were mainly utilized as warning signals. However, since the first attempts of designing human-computer interfaces around non-speech sound output (Buxton, Gaver, and Bly, 1994), their advantages over speech have become apparent. This development was catalyzed by the work of Gaver on auditory icons (Gaver, 1986) and everyday sounds (Gaver, 1993a,b), and the emergence of earcons (Blattner, Sumikawa, and Greenberg, 1989) as synthetic informational interface sounds. As discussed by Brewster (2003), non-speech sounds as output do not require visual display, although there are also many benefits in fusing the two modalities in multimodal interaction. Non-speech sounds reduce the visual load and need for visual attention, are typically faster in information representation than speech, and offer better temporal resolution than the visual sense.

Considering computational systems other than speech interfaces to date, the use of sound as an information conveyor has largely concentrated on one-directional use of sound: from the computer to the human. Also in the field of SID, although multisensoriality and sonic output have been an important point of study, for example, in the design of continuous sonic interaction (Rocchesso and Polotti, 2008; Rocchesso et al., 2009) and interface design (Dix, Finlay, and Abowd, 2004), the use of sonic input has not received similar attention. One of the main claims in this dissertation is that if sound output can be informational in auditory interfaces, sound can be informational to the other direction as well, that is from the human to the computer. Sound as input in computational applications has been studied most extensively in the context of speech recognition, often neglecting the other informational sounds we are able to produce. As a notable exception of a large-scale project considering other kinds of sound input, the Tai-Chi [2] (Tangible Acoustic Interfaces for Computer-

[2] http://www.mec.cf.ac.uk/research/pubs/taichi.html

Human Interaction) project has utilized acoustic localization of tangible objects by a microphone array on user interface surfaces. Also, as will be discussed in the next section, several individual studies have proposed implementations utilizing sonic input; however, a systematic approach to this aspect of study has been lacking.

A key requirement for using sound input in interactive applications is an algorithm, which is capable of reliably tracking information from the sounds of the user. The task of information retrieval may concentrate on detecting events, such as a certain temporal pattern of hand claps, or continuous information, such as the evolving tempo of a rhythmic sounding action. In interactive systems, these algorithms need to fulfill the requirements of real-time interaction. In addition to an information retrieval algorithm, an essential component in the interface design is the mapping of retrieved information to system functionalities. These issues, both the algorithms and the mappings, are often largely application-dependent. A general overview of the data flow in interactive systems utilizing sonic gestural input is presented in Figure 1.1.

The computational devices of today, most prominently the computer and the mobile phone, are already equipped with a microphone and sufficient computational power to enable the use of sound as an input modality. Therefore, in general, the use of sound as input requires no specialized hardware. In addition, sonic input can be utilized in eyes-busy situations and requires no physical contact with the device. Sonic input is also appealing as it can suit people with visual and motor impairments, enabling them to interact with applications otherwise unaccessible. Therefore, sonic input shares most of the advantages of non-speech sound output.

This dissertation examines the use of non-speech sound input, namely the use of sonic gestures (defined in Section 2), in interactive applications. Specific focus is on percussive gestures, utilizing hand clapping sounds as the main case example. In addition, as a specific application group, the study focuses on rhythmic interactive systems, which are also used as a tool to understand at a deeper level the phenomena related to rhythmic interaction.

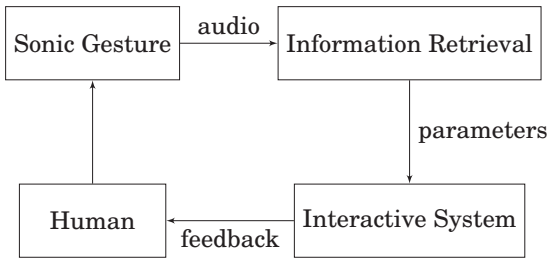

Figure 1.1. Information flow illustration of interactive systems utilizing sonic gestures as input.

1.1 Aims of the Thesis

This thesis discusses the use of sound as input in interactive applications. The problem is approached from two perspectives: identifying the information potential conveyed by non-speech sounds and providing tools for their use in real-time systems and as a specific focus area utilizing these informational sounds to design interfaces and applications that realize rhythmic interaction between the human and the computer. The sounding actions conveying information are called sonic gestures in this thesis. As the study concentrates on interactive applications, a key requirement to process the sounds is that the algorithms and techniques must function in real-time.

The research questions of the thesis can be outlined as follows:

- What information can sonic gestures provide in human-computer interaction?

- What kind of real-time techniques are applicable to capture the information from the sounds?

- How can continuous interaction be facilitated by sonic gestures, for example, in rhythmic applications?

- How do humans respond to multimodal rhythmic interaction systems and can this behavior be modeled?

- How can rhythmic interaction systems be used as a tool for understanding our capabilities of rhythmic perception and production? What kind

of design guidelines can be drawn for these systems from the sound perspective?

1.2 Organization of the Thesis

This thesis comprises an introductory part and nine articles published or accepted for publication in peer-reviewed journals and conference proceedings. The introductory part begins with an overview of sonic gestures in the field of human-computer interaction (HCI), summarizing the past studies related to the matter and the new theoretical outlines of Publication I in Section 2. In Section 3, the focus is on computational techniques and algorithms for identifying percussive sonic gestures based on both the state-of-the-art techniques in the field and Publication II and Publication III. In Section 4, rhythmic interaction with computational systems is discussed, based on previous studies from the fields of HCI, SID, and sound and music computing (SMC), including results and findings from Publications V-IX. Section 5 summarizes the novel results presented in the publications, and, finally, conclusions are drawn in Section 6.

1.3 Main Contributions of the Thesis

The scientific merits of the thesis can be summarized as follows (corresponding publications in parentheses):

- A novel definition for sonic gestures and a taxonomy portraying the retrievable information from different gesture categories (PI).

- Sound recognition techniques specifically adapted for identifying different percussive sonic gestures in real-time (PII, PIII).

- An experimentally verified novel finding that hand clap sounds contain personal information and have potential in person identification (PIV).

- A novel audio user interface, which enables rich use of percussive sonic gestures as input in human-computer interaction applications for both discrete and continuous interaction (PV).

- An interactive Flamenco hand clapping tutor capable of producing synthetic Flamenco hand clap patterns, listening to the clapping of the user, and giving feedback on the performance, including novel techniques for multimodal feedback (PVI, PVIII, PIX).

- Novel findings on rhythmic interaction between the human and the computer (PVI) and, based on these, a system simulating the rhythmic behavior of a human learning new rhythms (PVII).

- New evaluation methodology of rhythmic interaction with computational systems (PVI, PVIII).

2. Sonic Gestures and Human-Computer Interaction

To define sonic gestures and their properties, a look into past studies on gestures as well as sound morphologies is in order. The term gesture itself has been defined in numerous ways in different contexts and, therefore, the vocabulary and understanding of its meaning is somewhat ambiguous (Jensenius, 2008). This section starts with an overview of gesture definitions and focuses on its use in the fields of HCI, SID, and SMC. The relationship of gesture and sound is briefly discussed, followed by the novel definition of sonic gestures based on Publication I.

According to Cadoz (1988), a gesture is non-vocal physical behavior that humans use to inform or transform their environment. He approaches the relationship between gesture and sound from a musical perspective, stating that there can be no music without gesture, and further introduces the concept of instrumental gesture, which can be defined as physical interaction with a concrete object (Cadoz and Wanderley, 2000). Thus, in instrumental gestures there is always an energy transfer between the human actor and an external object that is being excited or manipulated. This is in relation to the notion of manipulative gesture (Quek, McNeill, Bryll, Duncan, Ma, Kirbas, McCullough, and Ansari, 2002) and can be contrasted with empty-handed gestures (Miranda and Wanderley, 2006), which the human performs without contact with an external object. Empty-handed gestures are sometimes also referred to as semaphoric or non-contact gestures and have been studied, for example, as accompaniment to speech communication (Kendon, 2004) or musical performance (Godøy and Leman, 2009), although their communicative value as a stand-alone source of information has also been underlined (McNeill, 2005).

Godøy and Leman (2009) have collected studies on the concept of musical gesture and the way people tend to physically react to the music

they perceive. A tight relationship between bodily movements in both creation and imitation of musically meaningful actions is portrayed.

Jensenius (2008) has discussed gestures from different perspectives in the context of music-related body movement. His tripartite approach to gesture research presents gesture as means for communication, control, or as mental imagery. In communication, gestures are considered related to objects of language, always conveying a meaning or intention of the person making them and often accompanying speech to emphasize the message. Gestures for control, on the other hand, are more or less the traditional HCI viewpoint on gesture, considering them as actions of intentionally inciting or modifying system behavior by meaningful actions. Mental imagery refers to higher-level, metaphorical meanings the gestures can be associated with.

Gestures can be considered as objects characterized by their morphology. Gesture phrases have been characterized as consisting of preparation, nucleus, and retraction (Kendon, 1972). Following Cadoz and Wanderley (2000) and Van Nort (2009), gestures can be divided into different types based on their macro-level morphology. Impulsive gestures are short in time, defined by a near discrete identifiable point in time when they occur. A single tap on a table is an impulsive gesture, for example. What is essential is that the excitation energy is impulse-like. Sustained gestures, on the other hand, have an arbitrary duration and are characterized by continuous excitation energy. Whistling a note for a period of time can be considered a sustained gesture. In addition, there are iterative gestures that are formed by sequentially producing impulsive or sustained gestures. Clapping to a beat is an example of an iterative gesture. Following Schaefferian principles of the morphologies of sound objects (Schaeffer, 1966, 1998) and relating these to gestures, dwelling deeper in morphological levels is possible, considering, for example, the mass, timbre, motion, and dynamic profile of the sound or gesture object (Van Nort, 2009).

Despite the extensive studies on action-sound relationship and musical gesture, to date the mainstream of research in the field seems to have focused on utilizing a physical gesture as input to sounding (musical) systems to generate and/or modify the output sound. Accelerometers, haptic controllers, cameras, and touch sensors have become the tools of the trade of new control input mechanisms. While this body of research is solid and has been utilized in the development of many HCI interfaces

seen in the recent past, the research does not directly consider the use of sounding actions, captured with a microphone, as their own group of gestures on a larger scale.

2.1 Sonic Gestures as Input in HCI

In this thesis and Publication I, sonic gestures are defined as human-generated sound-producing actions that convey information. In the context of this thesis, the information is conveyed to a computational system, imposing interactions between the human and the computer. While the mainstream of gesture-sound and action-sound related research mainly utilizes physical gestures, tracked by motion sensors or video cameras, to inform the sound generated by the system, in this work the sound always occurs prior to the computation. In other words, the primary interest is not necessarily in tracking the gesture creating the sound, but the sound parameters that are used to inform and elicit the interaction.

In contrast to the abovementioned gesture definition of Cadoz (1988), sonic gestures can be vocally produced, as long as they are not speech. Non-speech utterances and other vocal sounds are a powerful conveyor of information, as has been demonstrated, for example, by Ekman and Rinott (2010) in the context of using vocal sounds as a sketching tool for sonic interactions and Tuuri (2011) in studies on the communicative potential of vocal gestures.

While gestures have been stated to be always dynamic (Mulder, 2000), one can argue that a gesture can also have static properties. For sonic gestures, consider humming as an example. Humming with a constant pitch and loudness, apart from having an identifiable start and end, has a static body in terms of spectral properties and energy. Humming with a varying pitch, by contrast, has a dynamic body.

Everything stated above on gesture morphologies applies also to sonic gestures, which may be impulsive, iterative, or sustained. One can argue that impulsive gestures are better suited for discrete interactions and sustained gestures for continuous interactions, but as has been demonstrated, for example, by Sporka (2008, 2009), sustained pitched sonic gestures are also applicable for conveying discrete or event-based information, for example, by mapping certain pitch patterns of humming or whistling to keyboard commands.

In the design of sonic interactions, sonic gestures serve as an interesting way of providing dual sonic feedback. In addition to the sonic feedback displayed by the interactive system, the sonic gesture itself is typically audible to the person performing it, which yields immediate feedback on the action. This implies also that the person interacting with the system can monitor and learn from the sounds he or she is making.

As discussed in Publication I, sonic gestures can be a very rich conveyor of information. For example, the type of gesture, the volume and timbre, and the direction of occurrence can be inferred from isolated sonic gestures. Also continuous information, such as pitch and tempo and their temporal variations, and rhythmic patterns or sequences of gestures can be extracted from iterative and sustained gestures.

While the variety and potential information that sonic gestures can convey is unquestionably rich, one main challenge for the interaction designer or interface implementer is to find suitable computational means of tracking the information. However, as will be discussed later in this thesis, several tools for reliable real-time information acquisition from these sounds actually already exist, and the task at hand is more a matter of choosing the proper tool.

2.2 Applications of Sonic Gestures in HCI

While the use of sonic gestures as input is still marginal compared to the use of traditional physical gestures, several interesting examples of their utility have emerged during the past decade. These examples serve to exhibit the rich interaction affordances of sonic gestures as well as concretize the taxonomical dimensions discussed above. These examples show that both discrete and continuous interactions are feasible using sonic gestures.

Probably the best-known application of sonic gestures, turning electronic devices on and off with sound, dates all the way back to the 1960s and has since been popularized as a concept mainly by movies. This is a textbook example of discrete interaction with sonic gestures. The earliest devices have not only utilized hand claps and short hand-clap patterns, but also dog whistles as a means for sonic input[1].

Another example of discrete interactions by sonic gestures has been presented by Vesa and Lokki (2005), who developed a finger-snap interface

[1] http://www.time.com/time/magazine/article/0,9171,941481,00.html

for controlling a music player application. They use two microphones attached to the headset of the user to track the position of the finger snap around the head (left/right/center) and map this position to music player functions (previous track/next track/play or pause).

Apart from utilizing a single discrete impulsive gesture as input, interfaces can be developed to distinguish between different gesture types. This has been demonstrated in Publication V with a hand-clap interface that can be taught to recognize different types of hand claps of the user. This information can then be mapped to desired system functions. Furthermore, the hand-clap interface supports tempo tracking, making tracking continuous information from iterative hand clapping gestures possible. While the interface has been labeled as "hand clap interface", it is fully capable of functioning with other sharp impulsive sonic gestures as well. As example applications of the utility of the interface, the authors first presented a sampler driven by sonic gestures, a music tempo control application controlled by continuous hand clapping, and a virtual audience application, in which the user can entrain a simulated crowd of clappers to synchronize with the tempo of the user. These applications are presented in more detail in Publication V. The same interface was used as the backbone of the virtual Flamenco tutor application of Publication VI.

Hanahara, Tada, and Muroi (2007) presented the idea of a hand clapping language for human-robot communication. The aim of the research was to provide a shared language among humans and robots instead of having a separate human-to-robot language and a robot-to-robot messaging system, thus increasing the social dimensions of interaction.

The use of impulsive vocal sounds as input has been utilized, for example, by Hazan (2004), who has developed an application of controlling a drum sampler with beatboxing sounds, that is vocally imitated drum sounds. This research later resulted in the launch of the BoomClap application for the iOS devices. Kapur, Benning, and Tzanetakis (2004) also studied beat-boxing as input, but for music query application.

Considering sustained sonic gestures, many of the presented applications and interfaces are based on tracking the pitch of humming or whistling. Sporka (2008) has studied a number of ways of using pitched non-speech sounds in different applications, including cursor control,

a humming alphabet for keyboard emulation, and control of computer games. The latter aspect was also studied by Hämäläinen (2007).

Vocally produced sounds are also applied to control the Vocal Joystick (Bilmes, Malkin, Li, Harada, Kilanski, Kirchhoffi, Wright, Subramanya, Landay, Dowden, et al., 2006), in which different vowels are mapped to different cursor movement directions in a continuous two-dimensional mapping. The amplitude of the sound is mapped to the movement speed and selection can be indicated by a short hissing sound.

Scratch Input (Harrison and Hudson, 2008) is an interface utilizing instrumental sonic gestures, namely scratching surfaces. With a piezo-electric contact microphone placed on a wall or tabletop, for example, the sounds resulting from human actions on these surfaces can be captured from a relatively large distance without the danger of environmental airborne sounds interfering in the process. Scratch Input can be taught a large number of sonic gestures by means of machine learning and is capable of both discrete (for example, a gesture dictionary) and continuous (for example, a circular scratching motion to control ramping up or down the volume) interactions.

Another instrumental and unpitched sonic gesture has been innovatively utilized by Wang (2009), who has developed the iPhone Ocarina application. The user blows on the microphone of the iPhone to excite a physical model of an ocarina (based on the flute model of Välimäki, Karjalainen, Janosy, and Laine (1992)), while fingers on the touch screen are used to control the pitch output. The interaction closely relates to that of playing a real ocarina.

3. Acquisition of Information from Percussive Sonic Gestures

There is lots of potential information that sonic gestures can convey. For humans, the task of decoding this information relies mainly on tacit knowledge developed from past experience and exposure to the sounds in context, but for the computer the intelligence of tracking the essential parameters must be developed by the engineer or the interaction designer. This often involves the development of specialized algorithms capable of listening to the sounds, a problem for which the literature is vast in the fields of speech recognition and music information retrieval (MIR). The MIR Toolbox (Lartillot, Toiviainen, and Eerola, 2008), for example, contains a set of MIR analysis techniques developed for Matlab. This kind of algorithms can also be seen as a pre-requisite for developing interactive systems utilizing sound as input, in which case they also are required to function in real-time.

In this thesis, the main focus in sound parameter recognition is on percussive sounds, such as the sound of hand clapping or knocking on a table. While these are impulsive gestures, they are also the building blocks of several iterative gestures. Therefore, of special interest are both sound event recognition techniques and algorithms for tracking continuous information, such as tempo and beat, or their higher-level derivatives. The overall procedure of information retrieval from percussive sounds as presented in the context of this thesis is depicted in Figure 3.1.

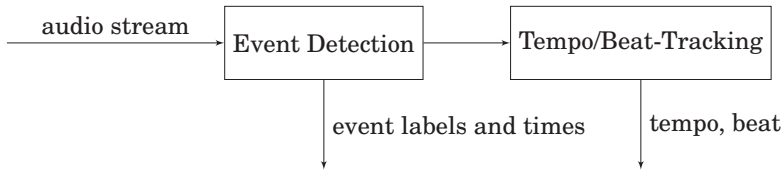

Figure 3.1. Information retrieval from a percussive sound stream.

3.1 Percussive Event Recognition

Percussive sound event recognition has been most intensively studied in the field of MIR, which has produced several techniques and studies on drum hit classification, for example. Often the aim is to apply these techniques in automatic transcription systems, database queries, or musical editing software. Due to the nature of these application areas, offline methods are typically considered sufficient. However, also real-time techniques have started to emerge to be used in, for example, automatic accompaniment systems, for which the real-time requirements are rather strict.

The recognition of percussive sounds typically consists of three steps, onset detection, feature extraction, and classification, although in some techniques the steps are combined (FitzGerald and Paulus, 2006). The steps are visualized in Figure 3.2. In the onset detection step, the potential occurrences of the target sounds are tracked from the audio signal (Bello, Daudet, Abdallah, Duxbury, Davies, and Sandler, 2005; Dixon, 2006) followed by computing desired features around these onsets from the audio signal. Classification is then performed based on these features, typically by a chosen machine learning technique. There are both supervised and unsupervised classification techniques; in the former, the classifier is first taught by a labeled corpus of relevant sounds to identify their differences and in the latter the classifier learns the labels autonomously.

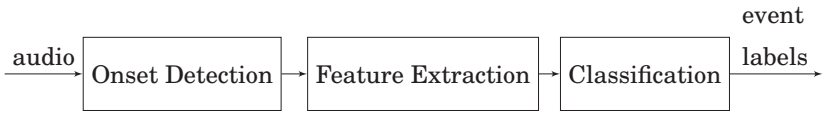

Figure 3.2. Steps of percussive event recognition (typical approach). In an alternative scheme, a hybrid method can combine the three steps into a single algorithm.

The requirements for event recognition algorithms depend not only on the target sound events to be tracked, but also on other signal properties. A very important distinction is that of monophonic and polyphonic signals. In monophonic signals, the only sound events that appear can be basically assumed to be target events, in which case the detection is relatively simple. However, in practice the signals are often polyphonic at least to some degree, as is definitely the case, for example, with musical signals, which complicates the event recognition tremendously as there can be numerous simultaneously sounding events

present at the signal at any time. Percussive sounds, however, are usually quite distinctive in the mix, as they are typically unpitched and have a pronounced transient around the onset.

In the remainder of this section, a summary of applicable onset detection, feature extraction, and classification techniques for percussive sounds is presented based on previous studies.

3.1.1 Onset detection

Bello et al. (2005) describe onset detection in general as a three-stage process of pre-processing, reduction, and peak-picking. After a possible initial step of pre-processing the audio, e.g, by digital filtering to suppress uninteresting frequency bands, the audio signal is reduced, that is, subsampled, yielding a so-called detection function (Bello et al., 2005; Dixon, 2006). The detection function is then processed by a peak-picking algorithm, looking for the onset occurrences.

The reduction step often aims at approximating the amplitude envelope of the signal. There are several ways of achieving this, the most straightforward ones being half-wave rectification of the signal followed by low-pass filtering and the computation of signal energy. These two approaches are illustrated in Figure 3.3. Squaring the signal has the advantage of suppressing the low-amplitude portions of the signal with respect to the high-amplitude ones, making the detection function more robust against disturbances and background noise.

Often the energy function is not applied directly as the detection function; differentiating the energy function provides a function with pronounced peaks at points of rapidly increasing energy. This is especially useful in the context of percussive sounds, which typically have a very pronounced initial transient. Alternatively, the envelope can be followed non-linearly as suggested by Cook (2002), by making the envelope follower react faster to increasing than decreasing signal energy. This can be especially useful with percussive sounds due to their fast attack. If the signal has been pre-processed by a filter bank, the envelope can be estimated for all the individual bands or the ones that are considered relevant for the target events. A filterbank may also be used to account for the psychoacoustics in onset detection (Klapuri, 1999).

Envelope followers are not the only means for percussive event detection, though. An alternative is to compute cross-correlation between the signal and a reference sample representing the interesting events

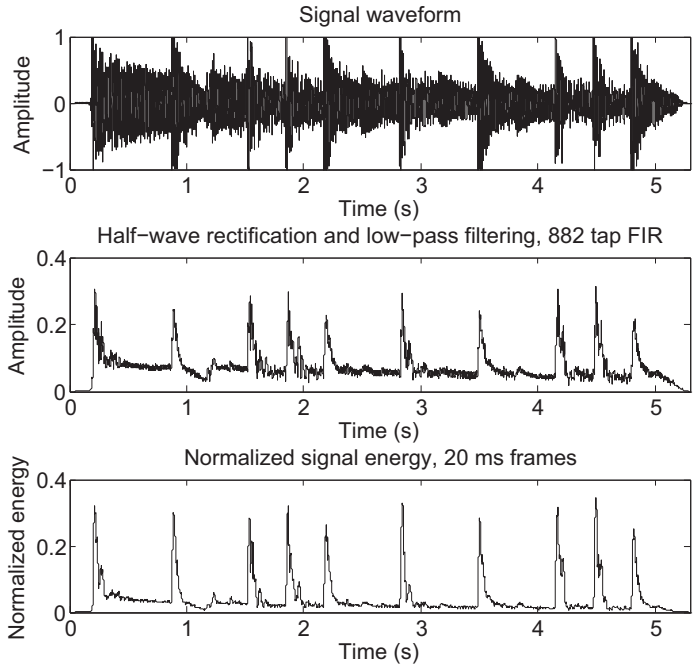

Figure 3.3. Two examples of detection functions by signal reduction for a musical signal waveform (top figure, sampling frequency 44100 Hz). The sharp peaks correspond to percussive onsets. The lowpass filter in the first technique (middle figure) is an averaging finite impulse response (FIR) filter with 882 taps and the energy (bottom figure) is computed in the second technique in 20 ms frames.

(Zils, Pachet, Delerue, and Gouyon, 2002) or utilizing temporal changes in the audio signal features, such as spectral flux or phase deviation (Dixon, 2006). Also, it is possible to separate the steady-state signal from the transient components by a sinusoids+noise decomposition (Duxbury, Davies, and Sandler, 2001) or other signal modeling techniques. The separated transient signal often effectively marks the percussive onsets, which stand out from the steady-state signal. This separation can be effective as such, but mostly relates to pre-processing the signal prior to the reduction step and is essential in the case of polyphonic signals.

Yet another group of techniques is probabilistic reduction, which infers onset occurrences based on a statistical model of some signal properties (Bello et al., 2005). The aim can be to detect change-points or surprises in the signal, which are emphasized around sharp onsets. It is also possible to directly model the interesting event instances and detect them from the audio stream, as will be discussed in the next section.

After the detection function has been computed, the detected signal is processed by a peak-picking algorithm, that is decision criteria are applied to find the onset occurrences (Bello et al., 2005; Dixon, 2006). The most basic form of peak-picking is to use a fixed threshold value, marking as onsets all the time instances where the detection function gets a higher value than the threshold. This approach can be sufficient for signals that have low background noise and consist mostly of desired sound events. In practical circumstances, however, signals tend to have both noise and other interfering sound events, and as a remedy an adaptive threshold value can provide better results. The adaptive threshold is often computed as a smoothed and scaled version of the detection function, for example by low-pass filtering or median filtering the detection function and multiplying with a constant (Bello et al., 2005).

A problem related to peak-picking and onset detection in general is that in the case of some drum hits, for example, the onset itself may be noisy due to double hits (drum stick bouncing several times on the membrane) or the detection function might provide more than one sharp peak per event. A means to overcome these issues is to utilize a temporal mask after the initial detection function peak to eliminate the accidental detection of peaks immediately following the first one.

3.1.2 Feature extraction and classification

In a typical scenario of percussive event recognition, after initial onset detection the audio around an onset is processed to infer the class of that event. First, a set of features is extracted from the audio data, followed by a classification of the event based on those features. Numerous techniques of varying complexity for both feature computation and classification have been proposed in the past. In general, features are computed in short time frames, especially for percussive sounds for which the event type-specific information often is found immediately after the onset within a relatively short time frame.

Audio signal features can be roughly grouped into temporal, spectral, and cepstral features (Xu, Zhu, and Tian, 2002). Temporal features are derived from the time-domain representation of the signal. Examples of such temporal features are signal energy, signal power, attack time, decay time, and zero-crossing rate (Gouyon, Pachet, and Delerue, 2000), temporal centroid and crest factor (Tanghe, Degroeve, and Baets, 2005)

and temporal variations of subsequent frame-level features (Herrera, Dehamel, and Gouyon, 2003).

Spectral features, on the other hand, are derived in the frequency domain, typically after computing the fast Fourier transform (FFT) for the signal (Herrera, Yeterian, and Gouyon, 2002; Herrera et al., 2003). While the spectral bins themselves can be utilized as features, it is more common to compute more parametric representations for the sounds, such as the spectral centroid, spectral tilt, spectral kurtosis, spectral strong peak, and spectral flatness (Herrera et al., 2002, 2003); dissonance (Uhle, Dittmar, and Sporer, 2003); and band-energy ratios (Tanghe et al., 2005). Gillet and Richard (2003) modeled the spectral peaks of different tabla strokes by fitting a Gaussian distribution to the computed spectra.

Cepstral features are basically the cepstral coefficients of the signal computed as the Fourier transform of the logarithmic spectrum of the signal. The most commonly used variant of cepstral coefficients are the Mel-frequency cepstral coefficients (MFCCs), in which the frequencies are mapped to the Mel scale prior to computing the logarithm and making a discrete cosine transform on the resulting Mel spectrum (Tanghe et al., 2005). The main advantage of MFCCs is that the Mel scale provides a better match to the frequency resolution of human hearing than a linear scale.

It is also possible to compute higher-level features not strictly fitting the categorization above. Features such as percussiveness and noise likeness (Uhle et al., 2003) have been shown to be relevant to detect drum sounds in polyphonic signals.

The set of possible features to be extracted is potentially very large, which naturally affects the computing time of the subsequent classification step. There are, roughly speaking, two families of techniques to reduce the number of features, namely automatic feature selection and feature space dimensionality reduction. The aim in feature selection is to remove redundant or irrelevant features and find an optimal number and set of features (Herrera et al., 2002). There are different approaches and algorithms for feature selection, which can be categorized as filter models, wrapper models, and hybrid models (Liu and Yu, 2005). For feature space dimensionality reduction, one popular technique is principal component analysis (PCA) (Bishop, 2006), which aims at transforming the possibly correlated features to uncorrelated ones by an orthogonal transformation.

After the features have been computed, they can be used as input to a classification algorithm. There are a multitude of possible approaches to the classification problem; here, the main focus is on real-time techniques[1] that have been utilized in percussive sound classification.

Decision trees are possibly conceptually the simplest classification method. A decision tree is a heuristic set of logical rules that operate on the features, often setting boundary conditions such as "if signal energy is less than X and the spectral centroid is larger than Y, then the event is of class Z". Decision trees have the advantage of being able to utilize any information available on the event properties beforehand, but they can also be taught automatically from data by supervised learning methods (Quinlan, 1986; Gelfand, Ravishankar, and Delp, 1991; Safavian and Landgrebe, 1991). A key challenge is to avoid overfitting the data, as automatic tree induction techniques may be prone to setting the boundary conditions too case-specific (Safavian and Landgrebe, 1991). Decision trees have been applied in sound classification, for example, by Jensen and Arnspang (1999).

Support vector machines (SVM) are a classification technique that maps the feature space into another space, where the features characteristic to different classes are clearly separated (Bishop, 2006). The advantage is that while classifier boundaries may be sometimes impossible to define in the original space, they can be easier to compute in a suitable mapping space. SVMs have been used for percussive sound classification, for example, by Steelant, Tanghe, Degroeve, Baets, Leman, Martens, and Mulder (2004), Tanghe et al. (2005), and Gillet and Richard (2008).

Other "classical" machine learning techniques applied in percussive sound classification are neural networks (Tindale, Kapur, Tzanetakis, and Fujinaga, 2004), K-Nearest Neighbors algorithm (Herrera et al., 2003), and naïve Bayes classifier, which was applied in Publication II to classify different hand clap types, using either spectral bins of a low-order FFT or the filter coefficients of a second-order resonator fitted to the spectrum of each clap type as features. Classification by independent subspace analysis (Uhle et al., 2003), fuzzy logic and self-organizing maps (Eigenfeldt and Pasquier, 2010), and agent-based techniques (Aucouturier and Pachet, 2005) have also been deployed.

[1]Note that the real-time requirements apply to the classification procedure, not necessarily to the learning stage, as it can usually be undertaken beforehand.

3.1.3 Hybrid methods

Hybrid methods for sound event recognition are techniques that do not require a separate step for onset detection, but instead combine this with the feature extraction and classification. Often these techniques treat the audio signal as an unknown sequence of potential events, examine the signal in short segments, and classify these segments into events based on sophisticated machine-learning algorithms.

Hybrid event recognition methods tend to be either template-based, probabilistic, or both. In template-based methods, the sounds to be detected from the signal are modeled by a template of some characteristic properties of those sound events. Zils et al. (2002) approached the drum track extraction problem by an analysis-synthesis technique that uses an initial sound prototype (temporal model) for the bass and snare drum, computes the correlation of these prototypes over the signal, and modifies the prototypes iteratively to closer match the detected, highly correlative instances, thus synthesizing new sounds that should match those in the signal. Yoshii, Goto, and Okuno (2004) have presented another two-step technique based on spectral templates, consisting of template-matching and template-adaptation. In the first step, a seed template is compared to the audio signal frame-by-frame with a distance metric. In the adaptation step, the seed template is refined based on the closest occurrences in the signal, tuning it towards the specific instances present in that particular signal. This is relevant in musical signals, as the drum sounds typically vary between pieces. The technique has been utilized in a drum sound equalizer for volume and timbre control (Yoshii, Goto, and Okuno, 2005).

Puckette, Apel, and Zicarelli (1998) have developed a template-based percussive sound recognition object, `bonk~`, for real-time applications in two popular audio programming languages, Pure Data (PD) and Max/MSP. The algorithm processes the audio signal in frames with a filterbank of second-order band-pass filters derived from the constant-Q transform, and computes the signal power at each frequency band. These power estimates are utilized both for onset detection and for learning templates of different sounds. These templates essentially capture in a compact form the spectral power distribution of sound events. Once learned, the templates are used by `bonk~` to detect and classify sound events in an audio stream.

Non-negative matrix factorization is a template-based technique for non-negative data, aiming to decompose a non-negative matrix of observations into two matrices, a template and an excitation matrix (Cemgil, 2009). The original utility of the technique was in data compression, but it has recently been found useful also in audio classification and drum track extraction problems (Dessein and Lemaitre, 2010; Helén and Virtanen, 2005).

Probabilistic techniques often utilize Hidden Markov Models (HMM) (Bishop, 2006) as part of the inference engine. HMMs are a well-known technique in speech processing, for example, and have been shown to be applicable in real-time techniques (Cappé, Moulines, and Ryden, 2005). In a HMM, the underlying assumption is that the stream of events can be modeled as a Markov chain, where the current state only depends on the previous one. HMMs have been applied in percussive sound recognition with a hybrid approach, for example, by Paulus and Klapuri (2007, 2009), and also after onset detection (and feature extraction) by Paulus and Klapuri (2003) and Gillet and Richard (2003).

An increasingly popular group of probabilistic techniques in MIR problems are those based on Bayesian modeling and inference. Bayesian techniques have been used, for example, in pitch-tracking and tempo estimation, as will be discussed below, but also in percussive sound recognition. The foundation of Bayesian techniques lies in the Bayes' rule of conditional probability. For observed data y and unobservable quantities θ, the joint probability distribution can be written as

$$p(\theta, y) = p(\theta)p(y|\theta), \tag{3.1}$$

where $p(\theta)$ is the prior distribution for θ and $p(y|\theta)$ the sampling distribution (Andrew, Carlin, Stern, and Rubin, 2004). The Bayes' rule states that the conditional probability of θ given observations y is

$$p(y|\theta) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}. \tag{3.2}$$

The objective in Bayesian modeling, when applied as a classification tool, is to construct a probabilistic model for sound generation. A key advantage in Bayesian techniques is that unlike in some other techniques discussed above, it is possible to include any prior information on the problem in the model (Bishop, 2006) in the form of prior distributions. The model can contain one or more hidden variables, that is variables that are not directly observable from the data. One or more of these parameters is then the objective of probabilistic inference.

Bayesian models are conveniently expressed in graphical form (Cemgil, 2004; Barber and Cemgil, 2010). The conditional dependencies of the probability distributions can be drawn as a directed acyclic graph where the nodes correspond to probabilistic variables and edges connecting the nodes to the conditional dependencies. HMM can also be defined by Bayesian principles of conditional independency and is essentially a special case of Bayesian modeling (Bishop, 2006).

3.2 Tracking the Tempo and Beat

Iterative production of percussive sounds can yield a continuous event stream. Most music, for example, inherently consists at least to some degree of sequential events reproduced under certain rules. These repetitive structures lay the foundation for regular rhythms, as experienced in the majority of mainstream music we are exposed to. These repetitive rhythms are not only specific to music, though, but our natural actions such as heartbeat, walking, and breathing, to name a few, are also rhythmic. Such rhythmic streams contain contain higher-level continuous information, such as tempo and beat, as opposed to isolated events. This information can also be computationally retrieved from audio signals, but first a discussion on the key definitions from rhythmic perception studies is in order.

In a stream of events with equal temporal distance, a sense of pulse patterns emerges. Pulses can be defined as equidistant, isochronous time markings (Thaut, 2005), and they can be thought to lay the foundation for rhythm formation. By contrast, beats are perceived and audible pulse markings, which demarcate the pulse locations, but in practical contexts, such as live music, do not always lie strictly on the exact pulse location.

Tempo can be defined as the repetition rate of pulses or beats, that is, with respect to the interval of pulses in a pulse train, consisting of temporally equally spaced pulses. While in music, especially in live situations, beats do not always lie on the exact pulse and the tempo may change during a performance, the tempo can still be at least locally estimated as discussed above. Tempo is most often measured as *beats per minute* (bpm), and for a pulse train this can be mathematically expressed as

$$\theta_{\text{bpm}} = \frac{60}{T_{\text{ooi}}}, \tag{3.3}$$

where θ_{bpm} is the tempo and T_{ooi} is the pulse interval in seconds. In music, tempo defines the occurrence "speed" of rhythmic events. In contrast, beat can be defined as the locations of audible events in the pulse train, that is the "foot-tapping rate" (Thaut, 2005).

The musical meter can be defined with respect to the beat structure (Lerdahl, Jackendoff, and Jackendoff, 1996; Thaut, 2005). It essentially defines how beats are grouped in patterns, in music referred to as measures or bars. In Western music theory, the meter is often utilized synonymously with the time signature of music, such as 3/4, 4/4, or 12/8. Considering the time signature of 3/4, containing three quarter-note beats within a bar, the rhythmic structure is typically defined by placing a metrical accent on the first beat of the bar. This can be achieved by timbral or dynamic modulations of the beat, for example by simply playing the beat louder or with a different kind of drum stroke. However, the relation between grouping and accent is not fixed (Krumhansl, 2000).

Metrical levels are defined by beat divisions (Lerdahl et al., 1996; Gouyon and Dixon, 2005). They can be seen as a layered hierarchy of grids consisting of small time divisions between beats on low levels and longer divisions on higher levels. The metric subdivisions, regardless of the time signature, are often thought of as lying on a temporal grid, the spacing of which is defined by the underlying pulse train. The smallest time division between two musical events is referred to as $tatum$. A beat on a higher level must always align with those on lower levels. The grid-based approach is the foundation of many beat-tracking systems, which rely on rhythmic quantization, that is aligning the observed onsets on the metrical grid. Metrical levels also relate to the definition and measurement of tempo: the tempo depends on the metrical level, as the metrical level dictates the number of beats per measure and, thus, per time unit (Gouyon and Dixon, 2005).

In order to track tempo from audio streams, a trivial solution would be to detect the events of a rhythmic event stream, compute their time difference, and utilize this or its running average to yield an estimate for the tempo. However, in practical contexts this estimate is usually flawed, because the rhythmic patterns do not always have a beat strictly at every pulse location, and even if they do, a missed event detection would immediately lead in the elongation of the computed pulse interval. Offbeat events and irregular metric divisions complicate the task even further. Therefore, more intelligent approaches for tempo-tracking are

required. Several approaches have been proposed in the past for various contexts. In this study, the discussion is restricted mainly to real-time techniques.

Goto and Muraoka (1996) applied a multiple-agent architecture in beat-tracking. The agents, informed by musical knowledge, maintain their own beat-position hypothesis, can evaluate it and adapt their strategy based on the input, and can interact with other agents to co-operatively track the beat. The input to the agents consists of onset times.

The technique of Scheirer (1998) for tracking the tempo and beat directly from an audio signal is based on a network of resonant filters and envelope extraction. The signal is passed through a bank of band-pass filters and banks of parallel comb filters. The resonators are used in order to phase-lock with the beat of the signal and to determine the pulse frequency. Analytically, the technique bears resemblance to autocorrelation methods for tempo-tracking. Klapuri, Eronen, and Astola (2006) later applied a similar use of comb filters to the analysis of the musical meter on multiple hierarchical levels.

Seppänen presented a tempo-tracking technique based on OOI histograms (Seppänen, 2001a,b). The histogram is updated for each detected onset in a weighted manner that gradually de-emphasizes the past observations. By smoothing and using a tatum grid to align the histogram peak on the metric grid, the algorithm is relatively robust for small deviations in tempo. The tatum interval and metric quantization are also the backbones of B-Keeper (Robertson and Plumbley, 2007).

Tracking the tempo and beat can also be approached by statistical techniques. Cemgil, Kappen, Desain, and Honing (2001) have proposed using a tempogram representation and Kalman filtering and Cemgil and Kappen (2003) have utilized a probabilistic switching state-space model and Monte Carlo methods.

For more detailed reviews on tempo tracking and rhythm description systems, see the work of Dixon (2001) and Gouyon and Dixon (2005).

4. Rhythmic Interaction and Computational Systems

In this section, the phenomena related to rhythmic interaction with computational systems are discussed. To ground the discussion, relevant fundamentals of human rhythm perception and production are presented, including definitions for key concepts such as anticipation and prediction, entrainment, and sensorimotor synchronization. This is followed by a discussion on the application domains in HCI, with examples of related systems. Finally, evaluation of rhythmic interaction systems based on the HCI methodology and approaches to evaluating the existing implementations is discussed.

4.1 Human Factors in Rhythm Perception, Production, and Synchronization

The human perception of rhythm as well as the ability for rhythmic production has received relatively widespread attention in the fields of psychology, music cognition, neuroscience, HCI, and SID. Both rhythmic perception and production are suggested to be hard-wired in our brain as biological oscillatory circuits (Thaut, 2005). Also, humans are better at perceiving the pulse from acoustic events than from visual events (Parncutt, 1994), which suggests that sound is the key modality in rhythmic perception and comprehension.

Rhythmic interaction, however, is typically multimodal. While sound often plays a key role in rhythmic perception, it is often accompanied by visual and/or tactile cues. Consider tapping the table to the beat of music, for example, where the rhythmic perception is augmented by the tactile feel of hitting the table. A more complex example is a symphony orchestra, where the movements of the conductor are perceived visually by the instrumentalists, informing them about timing, tempo,

and musical accentuation of the performance. Also, while we are able to perceive a rhythm by hearing it, in order to replicate or synchronize to it we need our motor system. Rhythmic interaction with sound as one of the modalities is one example of continuous sonic interaction (Rocchesso et al., 2009), especially when considered as a cyclic and periodic process.

Several studies have suggested the auditory perception of musical movement and rhythm to be interconnected to physical movement. Perception of a constant beat can lead to tapping the same beat with the foot, for example, portraying embodied cognition related to rhythm perception. Although our natural reactions and movements in synchronizing, for example, dance movements to the beat of the music differ, there are also similarities in the movement trajectories (Toiviainen, Luck, and Thompson, 2009). Periodic movements can also be observed in 2–4 year old children when exposed to familiar music, but synchronization to the tempo is vague at that young age (Eerola, Luck, and Toiviainen, 2006).

The perception of musical rhythm has been presented as a dynamic process (Large and Kolen, 1994), in which the listener gets entrained by the musical events. This has been theorized as a sense of an isochronous pulse being activated in the listener when exposed to (regular) rhythms. If the listener starts to produce rhythms along with the entraining stimulus, the stimulus is often observed to drive the actions of the listener, entraining them to the underlying pulse. Outside of the musical domain, similar entrainment can also be witnessed in walking, when two (or more) people spontaneously and subconsciously synchronize their pace. This is related to the social dimension of synchronization and entrainment (Nagaoka, Komori, and Yoshikawa, 2005).

Synchronization does not necessarily require a fixed external pulse. In fact, emergent synchronization is a phenomenon often found in nature (Strogatz, 2003) and can be witnessed, for example, in concert situations when a crowd of people spontaneously synchronize their applause to call for an encore.

While rhythmicity seems to be hard-wired into humans, it can be argued without question that some people are better at rhythmic production (and perception) than others. Rhythmic expertise can be improved by training, but recent findings suggest that some people may be biologically incapable of some forms of rhythmic perception, which verifies that there can indeed be people who are “rhythm-deaf” when it comes to musical

synchronization (Phillips-Silver, Toiviainen, Gosselin, Piché, Nozaradan, Palmer, and Peretz, 2011).

When discussing periodic rhythms such as most musical rhythms, rhythmic perception and action are guided by two interrelated concepts, anticipation and prediction (Thaut, 2005). We make predictions of the future and anticipate something to happen. According to Gordon (2000), anticipation in musical patterns relates to being able to foretell the upcoming musical events in familiar music, whereas prediction relates to unfamiliar music involving educated guesses based on the previous exposure to similar music. In other words, prediction is a more conscious process than anticipation.

The phenomenon of being able to produce an action in synchrony with an external event based on predicting or anticipating the occurrence of this event is called sensorimotor synchronization (Repp, 2005). This mechanism is the key to synchronizing movement to music. We can adapt to its rhythm and anticipate the upcoming beat locations, performing coordinated actions "in beat". Also, as most of the popular music we are exposed to tends to be strictly periodic and to have repetitive rhythmic patterns, by experience we have the almost subconscious skill to predict the upcoming rhythm events in a piece of popular music very shortly after the piece has started playing.

Sensorimotor synchronization has been mostly studied in relatively restricted settings with reductionist methods, such as finger tapping to the pulse of a constant tempo (Repp, 2005). This approach can be grounded on the fact that musical rhythms in general can be presented as reduced forms of the actual rhythm, that is as rhythmic prototypes (Lerdahl et al., 1996). Also, as stated by Repp (2005), rhythmic finger movement is relevant to the playing of many instruments. However, although this kind of studies definitely have taught us a lot about our rhythmic capabilities, the process in practical situations of performative music and live ensembles is more complex than replicating a pulse train.

It has been shown that when tempo gets fast, humans (be they trained musicians or not) tend to simplify rhythms in tapping experiments (Snyder, Hannon, Large, and Christiansen, 2006) and that faster tempo leads to rhythmic assimilation (Repp, Luke Windsor, and Desain, 2002). Also, uneven rhythms, which incorporate irregular metric divisions, can be maintained at fast tempi, which do not allow mental subdivision to elementary metrical pulses. However, the intervals corresponding to the

metric divisions tend to be exaggerated (London, Keller, and Repp, 2004). Rhythmic grouping has strong effects on timing, variability, tap force, synchronization, but metrical structure or accents have been found to have little effect (London et al., 2004; Repp et al., 2002).

Ensemble synchronization has been shown to be affected a lot by time delays (Chafe and Gurevich, 2004; Chafe, Cáceres, and Gurevich, 2010). This is the main reason why network-based ensemble performances, in which the players are only connected via a network connection, have been problematic. The same issue can cause problems also in rhythmic interactive systems due to buffering and processing latencies; however, if the rhythms are periodic, the human actor can compensate for a small delay without letting it affect the performance, as long as the delay is constant enough to be anticipated. Furthermore, delaying the decision making in the system regarding the analysis of the user input is possible in some cases, by starting a generic response immediately after detecting an event and modifying the response as soon as the decision (for example, classification of event type) has been made (Stowell and Plumbley, 2010).

Modeling the rhythmic behavior of the human computationally is possible. Thaut (2005) proposed a mathematical model of synchronization and entrainment, while Darabi, Svensson, and Forbord (2010) modeled parametrically the human responses to sudden tempo changes in a tapping experiment with physical oscillatory systems. In Publication VII, the performance of a human clapper learning Flamenco rhythms has been realized as a virtual agent.

Auditory feedback and multisensory fusion have been found to support human performance in tasks involving rhythmic effort, such as walking (Visell, Fontana, Giordano, Nordahl, Serafin, and Bresin, 2009) and rowing (Schaffert, Mattes, and Effenberg, 2009, 2011). The study of Schaffert et al. (2011) involved the use of sonification of boat acceleration in the training of competitive rowing, which involves the synchronization of the movement of many rowers. Comparatively simple sonification, which aligned with the boat motion, was found to improve the performance. In a completely different setting in a basic sonic interaction experiment for giving auditory feedback on cutting vegetables in the kitchen, Rocchesso et al. (2009) found that giving feedback with “upbeat” sonic markers and a tempo adaptive to that of the cutting helped the cutter to maintain a more regular cutting phase and relaxed action than a fixed tempo or downbeat rhythm. On a more general level, Spillers (2008) has identified six design

criteria for entrainment in interactive systems: they should be adjustable, discreet, seamless, receptive to time and timing, responsive (that is, set expectation and give feedback), and incorporate time and tempo to pace the user.

4.2 Applications in HCI

Rhythmic HCI can be seen as an extension and integration of the rhythmicity studies of neurophysiology and psychology on one hand and human-computer interaction and music information retrieval on the other. Rhythmic HCI can be witnessed, for example, in video games, musical accompaniment systems, musical control simulations, and musical education software. Here, a short review of these systems is presented, with a special emphasis on systems utilizing sound as the input modality.

4.2.1 Video games

Even relatively early computer games have incorporated some forms of rhythmic elements. For example, the sports games in the 1980's often required rhythmic manipulation of the game controller in order to make the runner run or the rower row. In the 1990's, more music oriented applications started to emerge, such as the dance games utilizing a special dance mat controller on which the players needed to place their feet at the correct times. While essentially rhythmical, these controllers and interactions actually rely on an illusion of continuous interaction as each control event is typically assessed as a discrete event not utilizing the continuous information, such as tempo in the analysis. Nevertheless, entrainment and synchronization play a key part in these games.

Musical games, such as Guitar Hero and Rock Band, are similar by nature. The player needs to perform actions at the right time to score well. It is noteworthy, however, that while the controller information is utilized discretely, successful performance of these seemingly discrete actions requires the above-discussed sensorimotor synchronization, that is anticipation and prediction of the upcoming events.

Recently, we have witnessed several new interaction paradigms in video games thanks to advances in controller technology. First, Nintendo Wii brought the motion sensing capabilities into mainstream game controllers

and, more recently, Microsoft's Kinect, relying on camera-based movement tracking, stripped the player of external controllers altogether. Kinect also supports audio input by a microphone array, but, to date, applications have made limited use of it.

4.2.2 Musical and rhythmic simulations and control

Simulating rhythmic behavior has been a point of study of Erkut (2006) and Peltola, Erkut, Cook, and Välimäki (2007), who have simulated the behavior of a crowd of clappers falling in and out of sync. The backbone of the simulation is a virtual clapper agent based on a coupled oscillator model, stating the preferred tempo and nonlinear alterations in clap OOI as reactions to the difference with a lead oscillator. The model also affords changing the level of synchronization. This body of work was utilized in Publication V to simulate a virtual audience, synchronizing to the clapping of a human.

Personal Orchestra is an audiovisual system simulating the conducting of an orchestra (Borchers, Lee, Samminger, and Mühlhäuser, 2004). The user can control the tempo and volume of the orchestra with an infrared baton and direct the commands to specific sections. The orchestra is actually a video recording that is synchronized to the motions of the user. The beats on the video have been pre-marked and are synced with the detected events of the user. A sophisticated time-stretching technique is used to avoid audible artifacts.

Another virtual conducting system has been presented by Ilmonen and Takala (1999). Here, the human conductor can conduct a band of virtual musicians instead of a recording. The gestures of the conductor are captured by motion tracking in order to derive the intended musical parameters, such as tempo and beat, which are mapped to the actions of the virtual agents.

Virtual Conductor (Reidsma, Nijholt, and Bos, 2008) approaches the conducting of an orchestra from the opposite perspective, by simulating the actions of a conductor. Virtual Conductor is a virtual agent that can listen to an orchestra playing and react in a corrective way to deviations from the desired tempo. The system follows the principles of mutually coordinated anticipatory multimodal interaction (Nijholt, Reidsma, van Welbergen, op den Akker, and Ruttkay, 2008), which have also been deployed in other applications, namely a virtual dancer and a virtual

sports trainer, all relying on a virtual agent making predictions on the rhythmic actions of the user.

In Publication V, the hand clapping audience synthesis engine ClaPD of Peltola et al. (2007) was utilized in conjunction with a hand clap analysis interface capable of detecting the hand claps of the user and tracking the tempo of continuous clapping. As a result, the user could become a part of the clapping crowd, a lead clapper who could entrain and synchronize the virtual clappers.

A simulation of interaction between a virtual Flamenco tutor and a beginner learning Flamenco hand-clapping patterns was realized in Publication VII. The implementation is based on a subjective experiment where novices interacted with the virtual tutor. The statistical findings from the evaluation were applied to inform a virtual model of a Flamenco learner, which listens to the clapping of the tutor and adjusts the output based on comparison between its own output and that of the tutor.

4.2.3 Musical accompaniment systems

Musical accompaniment systems are in general more sophisticated in their analysis of the players' performance than, for example, video games. In order to produce meaningful accompaniment, musicological knowledge must be incorporated in both the analysis engine and the accompaniment synthesis or production.

B-Keeper by Robertson and Plumbley (2007) is a beat-tracking system designed for live performances in order to synchronize an electronic sequencer with a live drummer, without the need of a click-track. B-Keeper analyzes the tempo of the drummer based on detecting the OOIs of kick drum events and comparing this new estimate to the previous tempo estimate, which is updated if a threshold condition is met. An additional synchronization step, a small tempo adjustment by a simple probabilistic method, is performed in order to synchronize the onsets in the sequencer track and the live music more accurately.

A sophisticated example of a musical accompaniment system is the percussionist robot Haile (Weinberg, Driscoll, and Parry, 2005; Weinberg and Driscoll, 2006, 2007). Haile is capable of listening to the playing of a human percussionist and reacting to this in different ways. The backbone of the analysis engine is the onset detection by bonk~ (Puckette et al., 1998) and pitch and timbre recognition for assessing stroke types. The rhythmic stability model of Desain and Honing (2002) is applied

for higher-level rhythmic analysis (Weinberg et al., 2005). Collaborative interaction is realized based on a model of interconnected musical networks (Weinberg, 2005) and six different interaction modes have been deployed, ranging from call-and-response modes, such as imitation of the playing of the user, and different transformations of the input to perceptual accompaniment, where Haile plays simultaneously with the user, initiating local call-and-response parts.

Ringomatic by Aucouturier and Pachet (2005) is another interactive drummer application, a musical agent, which forms concatenative drum tracks from drum samples based on a constraint-satisfaction mechanism. The system can be controlled by MIDI messages in real-time. The agent analyzes the drum samples based on various descriptors and, based on the analysis of the MIDI messages, adapts the local constraint set.

ZooZBeat (Weinberg, Beck, and Godfrey, 2009) is a gesture-based mobile music studio application and interaction paradigm, which maps physical gestures such as shaking, tilting, tapping, and tossing into the musical output. Using gestural control, the user can produce notes on a backing track loop and expressively modulate the output. ZooZBeat is capable of multi-player musical interaction via wireless data transfer between the devices.

Antescofo (Cont, 2008) is a score following system, which enables the synchronization of music from an electronic score to live music. The system requires both the electronic and instrumental score, and it utilizes audio and tempo agents for anticipatory synchronization of the musical streams. Audio stream observations are based on a probabilistic model that informs the anticipatory agents.

4.2.4 Rhythm education

Interactive systems are an attractive tool for rhythm education, as they ideally can take the responsibility of music tutors at least in some respect. They can also be cheaper and more accessible to people not willing to invest large amounts of time and money in learning new skills. Interactive systems, such as Haile (Weinberg et al., 2005), the Personal Orchestra (Borchers et al., 2004) and the Virtual Conductor (Reidsma et al., 2008), have been outlined as potential tools for rhythm education. However, there are also dedicated educational systems, such as T-RHYTHM, a rhythm education system for school children (Miura and Sugimoto, 2006). T-RHYTHM makes use of tactile actuators that give

rhythmic stimuli to the students in sync with the music. This is especially beneficial in ensemble playing situations, where the tactile stimuli can effectively convey dedicated information to each student on their own rhythmic performance, thus eliminating the potential auditory masking effect from the surrounding musicians.

The iPalmas system of Publication VI, although primarily developed as a testbed for rhythmic HCI, is also a rhythm education system. The virtual Flamenco tutor can present new rhythmic patterns to the user and give audiovisual feedback on the learning and rhythmic performance.

4.3 Evaluation of Rhythmic Interactive Systems

Traditional HCI interfaces and interaction paradigms have a strong foundation in evaluation. Point-and-click interfaces have been around for decades and the evaluation methodology is solid, typically deploying or extending on Fitts's law (Fitts, 1954), which predicts the time needed to point to a target of certain width at a certain distance from the initial hand position. This kind of evaluation methodology, based on simple tasks, has also been proposed for gestural controllers (Wanderley and Orio, 2002); however, it does not apply well, if at all, to rhythmic interactive systems, because the definitions of "target", "distance", and even "time" are not applicable in continuous, cyclic interaction. Stowell, Robertson, Bryan-Kinns, and Plumbley (2009) have also pinpointed the shortcomings of traditional evaluation methodology in musical interactive systems, presenting both qualitative and quantitative alternative approaches to their evaluation.

Three complete constructs can be seen in the task of HCI evaluation: the human, the system, and their meeting point, that is the interaction. Therefore, in order to get a comprehensive evaluation of an interactive system and the interaction, all three should be acknowledged. Related to rhythmic interaction, the sensorimotor synchronization studies (Repp, 2005), albeit carried out in reductionist settings, provide a baseline for human rhythmic capabilities. Systemic factors such as input-output latency and computational complexity, on the other hand, are objectively measurable. The interaction part is the most difficult of the three to tackle.

The evaluation of rhythmic interactive systems has often been studied case-by-case, system-by-system, without a common framework. Qualita-

tive methods, such as making observations on an interactive accompaniment system in a performative context (Robertson and Plumbley, 2007), seem to dominate interaction assessment. ZooZBeat has been evaluated qualitatively based on user feedback on using the application (Weinberg et al., 2009). Haile, the robotic percussionist, has been evaluated by a user study consisting of a perceptual experiment and a written questionnaire with 14 subjects (Weinberg and Driscoll, 2007). The aim was to assess the design, mechanics, interaction, and perception of Haile, resulting in both qualitative and quantitative findings. Virtual Conductor (Reidsma et al., 2008) has been evaluated in several sessions with groups of musicians, resulting in qualitative findings and systemic development, such as refinement of the tempo correction algorithm.

The evaluation of the iPalmas system, reported in Publication VI, fused information from several angles. A subjective, task-based experiment was carried out and applied to gather objective metrics measured by the system and verbal and qualitative data from the comments of the subjects.

Publication VIII reports an attempt to formalize a design and evaluation model for rhythmic interactive systems based on earlier work on multimodal interfaces, implementing the model as real-time Unified Modeling Language (UML) constructs. This framework can be seen as a very promising step towards a common evaluation methodology for interactive systems not falling under the requirements of traditional HCI evaluation.

5. Summary of Publications

This section summarizes the main results presented in the publications.

5.1 Publication I: Sonic Gestures as Input in Human-Computer Interaction: Towards a Systematic Approach

Publication I defines sonic gestures as sound-producing actions that convey information, in this case to a computational system. The definition implies that in contrast to most previous studies on sound and gesture (see Section 2), the information used as input in computational applications lies within the sound itself, that is in its extractable parameters. Sonic gestures are considered as either impulsive, sustained, or iterative based on their macro-level morphology stemming from previous studies on musical gesture (Cadoz and Wanderley, 2000; Miranda and Wanderley, 2006; Godøy and Leman, 2009; Van Nort, 2009). Other taxonomical dimensions are also discussed, namely if the sonic gesture is instrumental or empty-handed, pitched or unpitched, or static or dynamic. The extractable parameters for a set of sonic gestures are presented with respect to the gesture morphology. A novel hierarchical grouping of these parameters for different gesture types is portrayed based on the complexity of the gesture and the required parameter extraction algorithm.

The major outcome of the discussion in the article is that there is a rich variety of information that sonic gestures can convey. This outcome is important in the design of interfaces utilizing sonic gestures as an input, since a common framework of these gestures and their parameter affordances has not been previously presented. In order to provide a concrete toolbox of gestures and related information retrieval techniques,

the use of design patterns (Borchers and Mühlhäuser, 1998; Borchers, 2001) is proposed as future work.

5.2 Publication II: Inferring the Hand Configuration from Hand Clapping Sounds

In Publication II, the automatic recognition of a clapper's hand configuration was studied. As discussed by Repp (1987), humans are capable of recognizing their own hand clapping sound from that of others. He also presented an eight-class taxonomy of hand configurations based on the angle and alignment of the hands relative to each other, showing that each of these hand configurations would result in audibly different sounds. This result was later utilized by Peltola et al. (2007), who built the hand clapping synthesis engine ClaPD that modeled the sound of these eight different hand configurations.

The aim of this study was to show that the hand configurations can also be distinguished from each other by audio signal processing. Experiments were performed with both synthetic hand clapping sounds, generated by the hand clap synthesis engine ClaPD (Peltola et al., 2007), and real hand clapping sounds of two subjects. Recordings of the eight hand configurations were performed and the resulting data was used to train and test a naïve Bayes classifier on two different feature sets: the coefficients of a 128 bin FFT and those of a second-order all-pole filter fitted to the clap spectrum under the assumption of a single prominent resonance. The results show that indeed some hand configurations can be easily differentiated from others based on their sound. For synthetic claps, a controlled benchmark data set, the overall classification rate for the filter coefficient features (69.9 %) was only marginally lower than for the full FFT spectrum (71.7 %). Thus, for real hand clapping, the filter coefficients were chosen as the representative features. For one subject, the overall classification rate was 48 % and for the other, 64 %.

Certain hand configurations caused systematic classification errors, suggesting that Repp's taxonomy (Repp, 1987) is not completely unambiguous from the acoustic point of view. The results also are different for the clapping sounds of different subjects, and an interesting finding was that the classifier trained with one person's clapping did not produce good results for the other person, suggesting that there is personal information in the clapping sounds.

5.3 Publication III: Real-Time Recognition of Percussive Sounds by a Model-Based Method

Publication III builds on Publication II and proposes a more sophisticated Bayesian algorithm for percussive sound recognition. While the use of Bayesian techniques has become popular in the field of MIR, the techniques have typically only been suitable for offline processing due to computational issues. The algorithm presented in this paper is capable of real-time processing with a low latency due to buffering observation samples. The probabilistic model of percussive sounds is based on spectral template observations and a HMM. Inference in the model is realized by fixed-lag smoothing instead of accumulating all the data, bringing the algorithm to real-time performance. The well-known Expectation-Maximization algorithm Bishop (2006) is applied as the technique for learning spectral templates.

The technique was evaluated by a hand clapping sound set similar to that of Publication II and by percussion instrument sounds, namely different strokes on two Turkish instruments, the Darbuka (goblet drum) and the Bendir (frame drum). The recordings were performed both in anechoic conditions and in a standard listening room.

For the eight hand clapping modes of three subjects recorded in anechoic conditions, the offline version (accumulating all the data) of the algorithm yields an overall correct classification rate of 66 %, and 61 % for the real-time version. In normal room conditions, the classification rate was 41 % for one subject and 65 % for the other. These results are in line with those of Publication II. Also, the finding that the classifier trained with the clapping of one subject did not perform well with the clapping of another subject was replicated, yielding a classification rate of 30 %.

For the percussion instruments, two types of strokes on each drum were recorded: “düm” strokes (fingers hitting the membrane near the rim) and “tek” strokes (hitting the rim with one finger). The metrics chosen in this experiment, the precision, recall, and latency, are defined as

$$\begin{aligned} \text{precision} &= \frac{\text{no. of correctly recognized events}}{\text{no. of events recognized by the method}}, \\ \text{recall} &= \frac{\text{no. of correctly recognized events}}{\text{no. of true events}}, \\ \text{latency} &= \text{estimated onset time} - \text{true onset time}. \end{aligned}$$

The objective was not only to find the classification precision and recall rates, but also an optimal lag value for the fixed-lag smoothing.

In the overall results, the lag value of 20-30 ms is noticed to be a good optimum with respect to the accuracy. Combined with the computational latency this yields a total latency of around 50 ms. The overall precision is over 85 % and recall over 80 %. However, there were differences in the results depending on the room conditions, the results being better for anechoic conditions (approximately 85 % precision and recall at 20 ms lag for both cases as opposed to 80 % in normal room conditions). Also, interestingly the results were better for the bendir, although its sound has a significantly more resonant body than that of the darbuka.

The algorithm is implemented by the authors as a real-time deployable module that can first be taught a number of percussive events and can then, in real-time, detect and classify the events from an audio stream.

5.4 Publication IV: Sonic Handprints: Person Identification with Hand Clapping Sounds by a Model-Based Method

Publication IV is based on the finding in Publication II and Publication III that a hand clap recognition technique trained on the clapping of one person does not perform well with that of another. Therefore, it is reasonable to assume that the hand clapping sounds contain personal information. The hypothesis in this study is that hand clap sounds could be used as one form of person identification in, for example, multi-user interactive systems, such as multiplayer console games.

The hand clapping of 16 people was recorded in normal room conditions in order to find out how well the recognition algorithm of Publication III is able to distinguish between different people. The subjects were instructed to clap freely to any constant tempo they liked for a minimum of 30 seconds, yielding a data set of 78 claps per subject on average. The sound files of each individual subject were divided into four equally long segments, then two of these segments were randomly assigned as training data and two as test data. The algorithm of Publication III was trained with the data, learning the spectral templates of all 16 subjects, and then evaluated.

The overall correct classification rate of recognizing the subject based on hand clapping was 64 %, which is a convincing result. There were, however, large differences in individual classification rates. Some system-

atic misclassification is observable for people whose spectral templates are similar. Also, as the subjects were not instructed to maintain a fixed hand configuration, the evolution of the clapping sound throughout the recording session made the results worse for few subjects.

Since the algorithm is real-time and the classification rates encouraging, it can be concluded that hand claps indeed have potential in real-time interactive systems as a means for person identification.

5.5 Publication V: A Hand Clap Interface for Sonic Interaction with the Computer

In Publication V, an interface that can track various parameters from hand clapping sounds is presented. The interface makes use of sound event recognition and tempo-tracking algorithms to provide means for extracting rich information from basic sonic gestures. Three prototypes of interactive applications are presented: controlling a sampler, controlling the tempo of music, and entraining a virtual crowd of clappers to the tempo of the user.

The interface was implemented in Pure Data (PD) (Puckette, 1996), a graphical patching language originally developed for audio signal processing. The `bonk~` object was used for event detection and the `rhythm_estimator` (Seppänen, 2001a,b) for tempo tracking. The modules were combined to form an analysis object capable of tracking both event-based (type) and continuous (tempo) information from the audio stream.

The informal evaluation of the prototype applications found that a process of negotiation over the tempo exists between the human and the computer. In practice this means that when the human aims at gradually changing the tempo of, for example, a piece of music, he or she is affected by the perceived tempo of the computer playing back the music, which makes the changing of the tempo challenging. This can be explained by the use getting entrained by the rhythm of the music.

5.6 Publication VI: Design and Evaluation of Rhythmic Interaction with an Interactive Tutoring System

Publication VI introduces iPalmas, an interactive flamenco rhythm tutor application. Flamenco music is rhythmically very rich and differs in many ways from the standard metric structures of traditional Western

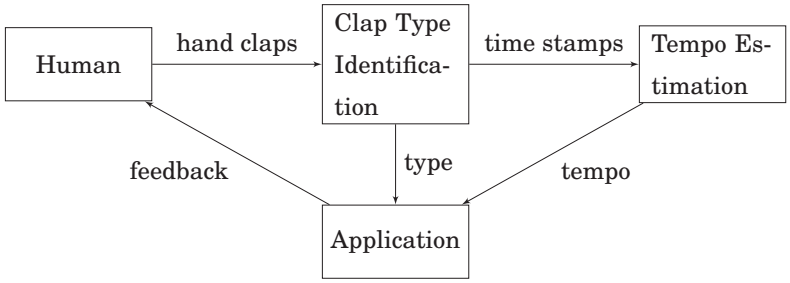

Figure 5.1. Data flow at the hand clap interface.

music; for example, most flamenco rhythms are characterized by a 12-beat rhythmic pattern of accentuated and non-accentuated beats (Maduell and Wing, 2007). This pattern is often performed by hand clapping to provide a rhythmic backbone for the performance. The iPalmas system is capable of reproducing synthetic flamenco hand clap patterns of different flamenco genres, it can simulate an ensemble of multiple clappers, and can listen to the user learning the patterns, giving feedback on the performance. The ClaPD hand clapping engine (Peltola et al., 2007) was utilized as the backbone for hand clap synthesis. The hand clap interface of Publication V was used for analysis, extended with new functionalities of computing the running standard deviation of the tempo of the user, the correctness of accentuation, and an adaptive thresholding mechanism for accent detection.

The iPalmas system measures the performance of the user based on three main metrics: the difference between the tempo of the user and that of the tutor, the internal tempo deviation of the user, and the accentuation correctness. These metrics are displayed to the user both numerically and with sliders and with an abstract representation of two dancing circles reacting to the rhythmicity and measured performance. A screenshot of the visual feedback is presented in Figure 5.2.

In an optional interaction mode, the virtual tutor can also give new challenges to the user once the performance gets better based on the metrics. The tutor gradually speeds up in tempo when the user meets set objectives in the performance.

The system was evaluated in a subjective experiment in which 16 subjects were instructed to learn four different hand clapping patterns using the system. Four different modes of tutoring were applied, that is audio-only tutoring with a fixed tempo, audio-only tutoring with an

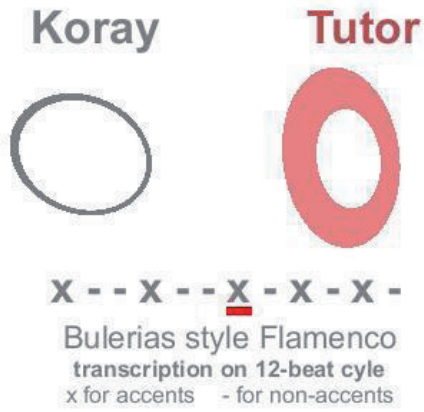

Figure 5.2. Main visual feedback of the iPalmas system. Top: the two circles react rhythmically to the clapping of the user (left) and the tutor (right). Bottom: a transcription of the pattern is presented to the user with a moving red marker indicating the current position. The name of the user is portrayed in the top left corner.

adaptive tempo, audiovisual tutoring with a fixed tempo, and audiovisual tutoring with an adaptive tempo. The initial tempo in all the cases was 175 bpm. The order of clapping patterns and tutoring styles was randomized and balanced throughout the experiment. The subjects were first introduced to the interface and interaction via simple 4-beat patterns and then instructed to learn and perform the patterns one by one in a two-phase manner. The subjects first practiced a pattern with a given tutoring style for as long as they felt necessary, but for the minimum of 10 full pattern cycles. Then the subject was instructed to perform the same pattern for one minute. In this test phase, the hand clapping sounds of the tutor faded away after two full pattern cycles after the subject started clapping along. At the end of the experiment, the subjects were instructed to reproduce as many of the patterns as they remembered.

Throughout the experiment, all the performance metrics were stored in a log file. The subjects' verbal comments were collected by the experiment supervisor both during the experiment and in a post-test interview.

Careful statistical and qualitative analysis of the gathered data lead to several interesting findings regarding the learning and the rhythmic performance of the subjects, but also regarding the system design. While there were subjective differences, the easiest pattern to learn and remember was *soleas*. Two subjects remembered all the patterns at the end of the test and on average two patterns were remembered. Interestingly, two subjects had difficulties in producing consistent accentuated claps,

that is they experienced difficulties in the control of clap loudness. As an explanation, the subjects noted that they usually clap in a crowd of clappers and therefore are not used to listening to their own clapping.

In rhythmic performance in the training phase, the average time to listen to the pattern before starting to clap along was nearly 10 full pattern cycles (roughly 40 seconds), but varied significantly among the subjects. The tempo-adaptation of the tutor helped the subjects to perform the accents more consistently (74.2 % correct with tempo adaptation versus 67.6 % without), as did the visual feedback (73.7 % versus 68.1 %). With the adaptive tutor, the OOI before an accentuated beat was slightly longer (344.7 ms on average) than before non-accentuated beats (343.0 ms). This finding was not true for the fixed-tempo tutor. The pattern itself affected the rhythmic performance significantly.

In the test phase, once the clapping of the tutor faded away, the subjects tended to increase the tempo. If the visual feedback was not present, the tempo increased gradually towards the end of the one-minute test. With visual feedback, however, the subjects were able to see that they got ahead of the target tempo, which led them to slow down. This resulted in a "pumping" tempo for some of the subjects.

The sound of the tutor's clapping was the key element in learning the patterns. The most useful visual element was the transcription of the pattern, which the subjects considered helped them understanding the accent locations. The circles, on the other hand, were considered attractive but uninformative and the numeric feedback was mostly beneficial in the initial stages of learning how to perform accents in general.

This evaluation is revisited in Publication VIII.

5.7 Publication VII: Simulation of Rhythmic Learning - A Case Study

Based on the results obtained from Publication VI, this study addresses the modeling of the learning Flamenco beginner using the iPalmas system. The aim of the study is to provide a means to assess rhythmic interaction by simulation. The learning clapper is implemented as a virtual agent, listening to the clapping sounds of the virtual tutor. The behavior of the learning clapper utilizes the results from Publication VI to inform the model and the rhythmic stability model of Desain and Honing (2002) is used to assess pattern difficulty. The initial listening

time is based on the statistics from the experiment by modeling it with a Gaussian distribution. A slight variation is added to the tempo of the virtual clapper with another Gaussian distribution based on the statistics from the experiment. The virtual clapper learns accentuation gradually, constantly assessing the similarity of the produced pattern with that of the tutor. The initial probability of mistakes is informed by the rhythmic stability mode. As the virtual clapper becomes more successful, the probability of accentuation mistakes decreases during the interaction.

Since both the virtual learner and the tutor listen to audio streams, either component can be replaced by a human clapper. This enables a person to become a tutor in the system, too, potentially giving further insight into rhythmic interaction design.

5.8 Publication VIII: A Structured Design and Evaluation Model with Application to Rhythmic Interaction Displays

Studies on interactive systems that incorporate non-visual modalities lack a common methodology for design and evaluation. This study presents a structured model for sonic and rhythmic interactions, basing the model on a structured approach to multimodal interfaces and formalizing it similarly to real-time Unified Modeling Language (UML) constructs (Larman, 2004). The design model is concerned with modalities, framing them as simple or complex and event-based or streaming. The guiding principle in model construction is to focus on the effect to be produced on the user. The evaluation model is based on defining user constraints and external constraints.

The study uses the iPalmas system, a rhythmic interface, to illustrate both the design and the evaluation model. The primary input modality is an audio stream, that is the hand claps of the user, which is converted to an event-based modality by event detection. The primary output modality is complex and audiovisual, presenting multiple streams of information to the user. The visual feedback of the circles can be considered a dynamic output modality.

This study reinterprets the experiment reported in Publication VI and Section 5.6 and associates its outcomes to system and display components in the next design iteration according to the proposed model. In other words, no new experiments were conducted, and the setup, procedure, and tasks are the same as in Publication VI.

In the experiment, the users reported on the system properties. In general, out of the visual feedback components consisting of a transcribed rhythmic pattern, dancing circles, and numeric metrics, only the transcription was perceived as useful. A couple of subjects also perceived the reverberation as excessive or unnatural (perceiving two sounds instead of one) and two subjects reported that they were not able to consistently produce accentuated claps. An auditory marker indicating the start of each pattern cycle was also disturbing to some subjects.

Based on the evaluation, the feedback modules, both visual and auditory, have been re-iterated. Essentially, the circles and numeric feedback elements were removed from the interface and a new visual look was designed. The pattern is now represented by a circular display of discs, whose size and color represent accentuated and non-accentuated beats. The re-development of the auditory feedback is discussed in detail in Publication IX.

5.9 Publication IX: Auditory Feedback in an Interactive Rhythmic Tutoring System

This study discusses the re-design of feedback elements in the iPalmas system based on the previous work in Publication VI and Publication VIII. After evaluating the system with both the subjective experiment and the structured model, the system has been enriched with new auditory feedback functionalities. To pinpoint the key targets for auditory feedback, this study further made a task analysis (Benyon, 2010) on the interaction with the system.

As a result of the evaluation, several attributes were pinpointed as design guidelines for the audio feedback. The sounds should be subtle, natural and non-irritating, yet informational and, if fast reaction from the user is desired, urgent. Archetypal sounds were considered favorable to contextualize the system further in the domain of Flamenco.

The correctness of individual accents, the tempo difference (lead/lag) from the tempo of the tutor, and the overall performance based on the metrics in the system were chosen as feedback targets (see Publication VI). The individual accent correctness is indicated by adding a reverberation tail to the hand clap sounds of the user for correct accents to indicate a “hit”. The tempo differences are sonified by a synthetic guitar scale that rises in pitch and shortens in OOI as the tempo increases, and

lowers in pitch and lengthens in OOI if the tempo slows down. The overall performance is indicated by the overall reverberation level in the clapping of the tutor, with the reverberation level gradually rising when the performance improves and diminishing while increasing the dry signal level if the performance deteriorates. This is to convey the feeling of the hall getting larger and the tutor moving further away or, alternatively, that the tutor comes closer to the student to give focused training. Also, with improving performance, additional virtual clappers appear in the mix. In addition, an upcoming tempo speed-up by the tutor is indicated by an archetypal shout (“Olé!”).

This iteration, while yet unevaluated by subjective experiment, closes the loop of the iPalmas system development.

6. Conclusions and Future Directions

This thesis has demonstrated that sonic gestures — sound-producing actions performed by a human to convey information — can successfully be utilized in HCI in numerous ways and can convey a rich amount of information for both continuous and discrete interactions. This was demonstrated by a taxonomy of sonic gestures and their extractable parameters. Even a seemingly simple gesture such as a hand clap can be utilized as a building block for a variety of applications, ranging from rhythmic systems to person identification.

The specific focus in this study was on percussive sonic gestures, for which recognition algorithms were proposed. As has been noted throughout the publications and in Section 3, numerous applicable real-time techniques exist in other fields such as speech recognition and MIR, and these can be harnessed for sonic gesture tracking as well. In addition to the existing techniques, new real-time techniques for analyzing percussive sonic gestures were proposed. Based on the use of these techniques, a hand clap interface was developed and further deployed in the development of a virtual Flamenco tutor application. These examples tangibly exemplify the use of sonic gestures for continuous interactions.

Evaluation of rhythmic interactive systems is far from trivial, and this thesis includes novel insights into the problem. The evaluation of the iPalmas application led to several important findings related to the design and evaluation of rhythmic interaction. The evaluation showed that the initial system design may have shortcomings and, therefore, iterative development by prototype evaluation is required. The evaluation needs to take into account not only the system but also the related human factors in order to get a holistic view of the interaction and system attributes. Both qualitative and quantitative measures need to be acknowledged in the evaluation, as in the subjective evaluation of the iPalmas system.

The iPalmas system turned out to be a fruitful tool and testbed both for constructing new evaluation methodology for rhythmic interactive systems and for assessing the human rhythmic capabilities. The subjects in the experiment were, in general, able to learn new and complex rhythms using the audio-visual interface, and considered the audio representation of the rhythm as the key to learning. Realistic, informative, and non-irritating sound design was found to be favorable. Related to the assessment of rhythmic capabilities, a model of a person learning and performing new Flamenco rhythms was constructed. This kind of modeling can be valuable for the design of realistically behaving rhythmic agents, for example, and for deepening our understanding of rhythmic perception and production.

As sound, rhythm, and bodily motions are known to be closely related, the rhythmic interactions between the human and the computer utilizing sonic gesture interfaces can be considered to be a promising direction. The use of sound input is already available in consumer electronics, and the recent interactive systems realizing rhythmic interaction can be considered as advanced and important, not only from a systemic point of view, but also as tools for understanding the human capabilities and attributes related to rhythmicity.

The studies in this thesis mainly concentrate on considering sonic gestures from a control viewpoint. However, the mental imagery and metaphorical dimensions (Jensenius, 2008) of sonic gestures should not be neglected when designing interfaces and interactions around them. This can be seen as the key aspect to consider in the design of intuitive interfaces.

Sonic gestures as input in HCI is still in its infancy and the studies on the topic are limited. To further facilitate the use of sonic gestures, one potential approach could be to develop design patterns (Borchers and Mühlhäuser, 1998; Borchers, 2001) for their utility.

Care must be taken in the design of interfaces making use of sonic gestures in order to keep the interactions fluent and usable. First of all, if the application will use sound as both the input and output, it is important to ensure that the two do not interfere with each other in a negative way. The danger is that the perception of the user of the interface gets blurred if the input and output sounds do not fit together. Furthermore, if the system is designed to be used with loudspeakers while capturing the sounds with microphones, the feedback of the output

sound to the input can cause problems, unless the output sound is computationally cancelled from the input signal or the sonic gestures differ substantially from the sound output.

A final remark and a definitely valuable point of future research is the social acceptability of sonic gesture interfaces. While using sonic input in a private apartment might not be an issue, clapping your hands to your mobile phone in a public space might be. It is also an interesting challenge for interaction designers to come up with sonic gestures and interfaces that have the potential of becoming socially acceptable.

Bibliography

G. Andrew, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*. Boca Raton, FL, USA: Champan and Hall/CRC, 2004.

J. Aucouturier and F. Pachet, "Ringomatic: A real-time interactive drummer using constraint-satisfaction and drum sound descriptors," in *Proc. Intl. Conf. Music Information Retrieval (ISMIR)*, London, UK, Sept. 2005, pp. 412–419.

D. Barber and A. T. Cemgil, "Graphical models for time series," *IEEE Signal Processing Magazine, Special issue on graphical models*, vol. 27, no. 6, pp. 18–28, 2010.

J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, September 2005.

D. Benyon, *Designing Interactive Systems*. Harlow, UK: Pearson Education Ltd, 2010.

J. Bilmes, J. Malkin, X. Li, S. Harada, K. Kilanski, K. Kirchhoffi, R. Wright, A. Subramanya, J. Landay, P. Dowden *et al.*, "The vocal joystick," in *Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Toulouse, France, May 2006, pp. 625–628.

C. Bishop, *Pattern Recognition and Machine Learning*. Singapore: Springer, 2006.

M. Blattner, D. Sumikawa, and R. Greenberg, "Earcons and icons: Their structure and common design principles," *Human-Computer Interaction*, vol. 4, no. 1, pp. 11–44, 1989.

J. Borchers and M. Mühlhäuser, "Design patterns for interactive musical systems," *IEEE Multimedia*, vol. 5, no. 3, pp. 36–46, 1998.

J. Borchers, E. Lee, W. Samminger, and M. Mühlhäuser, "Personal orchestra: A real-time audio/video system for interactive conducting," *Multimedia Systems*, vol. 9, no. 5, pp. 458–465, 2004.

J. Borchers, "A pattern approach to interaction design," *AI and Society Journal of Human-Centred Systems and Machine Intelligence*, vol. 15, no. 4, pp. 359–376, 2001.

S. Brewster, "Nonspeech auditory output," in *The Human-Computer Interaction Handbook*, J. Jacko and A. Sears, Eds. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2003, pp. 220–239.

W. Buxton, W. Gaver, and S. Bly, "Auditory interfaces: the use of non-speech audio at the interface," *Unfinished book, available at http://www.billbuxton.com/Audio.TOC.html*, 1994.

C. Cadoz, "Instrumental gesture and musical composition," in *Proc. Intl. Computer Music Conf.*, Cologne, Germany, Sept. 1988, pp. 1–12.

C. Cadoz and M. Wanderley, "Gesture-music," in *Trends in Gestural Control of Music*. Paris, France: IRCAM - Centre Pompidou, 2000, pp. 71–93.

O. Cappé, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models (Springer Series in Statistics)*. Secaucus, NJ, USA: Springer, 2005.

A. T. Cemgil, "Bayesian music transcription," Ph.D. dissertation, Radboud University, Nijmegen, Netherlands, 2004.

——, "Bayesian inference in non-negative matrix factorisation models," *Computational Intelligence and Neuroscience*, no. 4, pp. 1–17, 2009.

A. T. Cemgil and H. J. Kappen, "Monte Carlo methods for tempo tracking and rhythm quantization," *J. Artificial Intelligence Research*, vol. 18, pp. 45–81, 2003.

A. Cemgil, B. Kappen, P. Desain, and H. Honing, "On tempo tracking: Tempogram representation and Kalman filtering," *J. New Music Research*, vol. 28, no. 4, pp. 259–273, 2001.

C. Chafe and M. Gurevich, "Network time delay and ensemble accuracy: Effects of latency, asymmetry," in *Proc. AES 117th Conv.*, San Francisco, CA, USA, Oct. 2004, preprint no. 6208.

C. Chafe, J. Cáceres, and M. Gurevich, "Effect of temporal separation on synchronization in rhythmic performance," *Perception*, vol. 39, pp. 982–992, 2010.

A. Cont, "Antescofo: Anticipatory synchronization and control of interactive parameters in computer music," in *Proc. Intl. Computer Music Conf.*, Belfast, UK, Aug. 2008, pp. 1–8.

P. R. Cook, "Modeling Bill's gait: Analysis and parametric synthesis of walking sounds," in *AES 22nd Intl. Conf. Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, June 2002, pp. 73–78.

N. Darabi, U. Svensson, and J. Forbord, "Parametric modeling of human response to a sudden tempo change," in *Proc. AES 129th Conv.*, San Francisco, CA, USA, Nov. 2010, preprint no. 8308.

P. Desain and H. Honing, "Rhythmic stability as explanation of category size," in *Intl. Conf. Music Perception and Cognition*, Sydney, Australia, July 2002, pp. 17–21.

A. Dessein and G. Lemaitre, "Free classification of vocal imitations of everyday sounds," in *Proc. Sound and Music Computing Conf.*, Porto, Portugal, July 2010.

A. Dix, J. Finlay, and G. Abowd, *Human-Computer Interaction*. Cambridge, UK: Prentice hall, 2004.

S. Dixon, "An empirical comparison of tempo trackers," in *Proc. 8th Brazilian Symposium on Computer Music*, Fortaleza, Brazil, July 2001, pp. 832–840.

——, "Onset detection revisited," in *Proc. of the 9th Intl. Conf. Digital Audio Effects (DAFx)*, Montreal, Canada, Sept. 2006, pp. 133–137.

C. Duxbury, M. Davies, and M. Sandler, "Extraction of transient content in musical audio using multiresolution analysis techniques," in *Proc. Digital Audio Effects (DAFx)*, Limerick, Ireland, Dec. 2001, pp. 1–4.

T. Eerola, G. Luck, and P. Toiviainen, "An investigation of pre-schoolers' corporeal synchronization with music," in *Proc. 9th Intl. Conf. Music Perception & Cognition*, Bologna, Italy, Aug. 2006, pp. 472–476.

A. Eigenfeldt and P. Pasquier, "Real-time timbral organization: Selecting samples based upon similarity," *Organised Sound*, vol. 15, no. 2, pp. 159–166, 2010.

I. Ekman and M. Rinott, "Using vocal sketching for designing sonic interactions," in *Proc. 8th ACM Conf. Designing Interactive Systems*, Aarhus, Denmark, Aug. 2010, pp. 123–131.

C. Erkut, "Towards physics-based control and sound synthesis of multi-agent systems: Application to synthetic hand clapping," in *Proc. Nordic Music Technology Conf.*, Trondheim, Norway, Oct. 2006, pp. 1–7.

P. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," *J. Experimental Psychology*, vol. 47, pp. 381–391, 1954.

D. FitzGerald and J. Paulus, "Unpitched percussion transcription," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York, NY, USA: Springer, 2006.

W. Gaver, "Auditory icons: Using sound in computer interfaces," *Human-Computer Interaction*, vol. 2, no. 2, pp. 167–177, 1986.

——, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, no. 1, pp. 1–29, 1993.

——, "How do we hear in the world? explorations in ecological acoustics," *Ecological Psychology*, vol. 5, no. 4, pp. 285–313, 1993.

S. Gelfand, C. Ravishankar, and E. Delp, "An iterative growing and pruning algorithm for classification tree design," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 2, pp. 163–174, 1991.

O. Gillet and G. Richard, "Automatic labelling of tabla signals," in *Proc. Intl. Conf. Music Information Retrieval (ISMIR)*, Baltimore, MD, USA, Oct. 2003, pp. 1–8.

——, "Transcription and separation of drum signals from polyphonic music," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 529–540, 2008.

R. Godøy and M. Leman, *Musical Gestures: Sound, Movement, and Meaning*. New York, NY, USA: Taylor and Francis, 2009.

E. Gordon, *Rhythm: Contrasting the Implications of Audiation and Notation*. Chigago, IL, USA: GIA Publications, 2000.

M. Goto and Y. Muraoka, "Beat tracking based on multiple-agent architecture - a real-time beat tracking system for audio signals," in *Proc. 2nd Intl. Conf. Multiagent Systems*, Kyoto, Japan, Dec. 1996, pp. 103–110.

F. Gouyon and S. Dixon, "A review of automatic rhythm description systems," *Computer Music J.*, vol. 29, no. 1, pp. 34–54, 2005.

F. Gouyon, F. Pachet, and O. Delerue, "On the use of zero-crossing rate for an application of classification of percussive sounds," in *Proc. Digital Audio Effects (DAFx)*, Verona, Italy, Dec. 2000, pp. 1–6.

P. Hämäläinen, "Novel applications of real-time audiovisual signal processing technology for art and sports education and entertainment," Ph.D. dissertation, Helsinki University of Technology, Espoo, Finland, 2007.

K. Hanahara, Y. Tada, and T. Muroi, "Human-robot communication by means of hand-clapping (preliminary experiment with hand-clapping language)," *Proc. IEEE Intl. Conf. Systems, Man and Cybernetics*, pp. 2995–3000, Oct. 2007.

C. Harrison and S. E. Hudson, "Scratch input: creating large, inexpensive, unpowered and mobile finger input surfaces," in *Proc. 21st annual ACM Symp. on User Interface Software and Technology*, ser. UIST '08, Monterey, CA, USA, Oct. 2008, pp. 205–208.

A. Hazan, "Performing expressive rhythms with BillaBoop voice-driven drum generator," in *Proc. 7th Intl. Conf. Digital Audio Effects (DAFx)*, Naples, Italy, Oct. 2004, pp. 1–4.

M. Helén and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Proc. EUSIPCO*, Istanbul, Turkey, Sep. 2005, pp. 1–4.

P. Herrera, A. Yeterian, and F. Gouyon, "Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques," *Proc. Intl. Conf. Music and Artificial Intelligence*, pp. 69–80, Sep. 2002.

P. Herrera, A. Dehamel, and F. Gouyon, "Automatic labeling of unpitched percussion sounds," in *Proc. 114th Conv. Audio Engineering Society*, Amsterdam, Netherlands, Jan. 2003, pp. 1–14.

T. Ilmonen and T. Takala, "Conductor following with artificial neural networks," in *Proc. Intl. Computer Music Conf.*, Beijing, China, Oct. 1999, pp. 367–370.

K. Jensen and J. Arnspang, "Binary decision tree classification of musical sounds," in *Proc. Intl. Computer Music Conf.*, Beijing, China, Oct. 1999, pp. 414–417.

A. Jensenius, "Action-sound: Developing methods and tools to study music-related body movement," Ph.D. dissertation, Faculty of Humanities, U. Oslo, Oslo, Norway, 2008.

A. Kapur, M. Benning, and G. Tzanetakis, "Query-by-beat-boxing: Music retrieval for the DJ," in *Proc. Intl. Conf. Music Information Retrieval*, Barcelona, Spain, Oct. 2004, pp. 170–177.

A. Kendon, “Some relationships between body motion and speech,” *Studies in Dyadic Communication*, vol. 7, pp. 177–210, 1972.

——, *Gesture: Visible Action as Utterance*. Cambridge, UK: Cambridge University Press, 2004.

A. Klapuri, “Sound onset detection by applying psychoacoustic knowledge,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP’99)*, vol. 6, Phoenix, AZ, USA, Mar. 1999, pp. 3089 – 3092.

A. Klapuri, A. J. Eronen, and J. T. Astola, “Analysis of the meter of acoustic musical signals,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 342–355, 2006.

C. Krumhansl, “Rhythm and pitch in music cognition,” *Psychological Bulletin*, vol. 126, no. 1, p. 159, 2000.

E. Large and J. Kolen, “Resonance and the perception of musical meter,” *Connection Science*, vol. 6, no. 1, pp. 177–208, 1994.

C. Larman, *Applying UML and Patterns : An Introduction to Object-Oriented Analysis and Design and Iterative Development (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2004.

O. Lartillot, P. Toiviainen, and T. Eerola, “A Matlab toolbox for music information retrieval,” *Data Analysis, Machine Learning and Applications*, pp. 261–268, 2008.

F. Lerdahl, R. Jackendoff, and R. Jackendoff, *A Generative Theory of Tonal Music*. Cambridge, MA, USA: The MIT Press, 1996.

H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” *IEEE Trans. Knowledge and Data Engineering,*, vol. 17, no. 4, pp. 491 – 502, 2005.

J. London, P. Keller, and B. Repp, “Production and synchronization of uneven rhythms at fast tempi,” in *Proc. 8th Conf. Music Perception and Cognition*, Evanston, IL, USA, Aug. 2004, pp. 223–226.

M. Maduell and A. Wing, “The dynamics of ensemble: the case for flamenco,” *Psychology of Music*, vol. 35, no. 4, pp. 591–627, Oct. 2007.

D. McNeill, *Gesture and Thought*. Chicago, IL, USA: University of Chicago Press, 2005.

E. Miranda and M. Wanderley, *New Digital Musical Instruments: Control and Interaction Beyond the Keyboard*. Madison, WI, USA: AR Editions, Inc., 2006.

S. Miura and M. Sugimoto, “Supporting children’s rhythm learning using vibration devices,” in *Proc. CHI’06 Extended Abstracts on Human Factors in Computing Systems*, Quebec, Canada, Apr. 2006, pp. 1127–1132.

A. Mulder, “Towards a choice of gestural constraints for instrumental performers,” *Trends in Gestural Control of Music*, pp. 315–335, 2000.

C. Nagaoka, M. Komori, and S. Yoshikawa, “Synchrony tendency: interactional synchrony and congruence of nonverbal behavior in social interaction,” in *Proc. Intl. Conf. Active Media Technology*, Kagawa, Japan, May 2005, pp. 529–534.

A. Nijholt, D. Reidsma, H. van Welbergen, R. op den Akker, and Z. Ruttkay, "Mutually coordinated anticipatory multimodal interaction," in *Proc. HH and HM Interaction 2007*. Berlin, Germany: Springer, 2008, vol. LNAI 5042, pp. 70–89.

R. Parncutt, "A perceptual model of pulse salient and metrical accent in musical rhythm," *Music Perception*, no. 11, pp. 409–464, 1994.

J. Paulus and A. Klapuri, "Model-based event labeling in the transcription of percussive audio signals," in *Proc. Digital Audio Effects Workshop (DAFx)*, London, UK, Sep. 2003, pp. 73–77.

——, "Combining temporal and spectral features in HMM-based drum transcription," in *Proc. Intl. Conf. Music Information Retrieval (ISMIR)*, Vienna, Austria, Oct. 2007, pp. 225–228.

——, "Drum sound detection in polyphonic music with hidden Markov models," *EURASIP J. Audio, Speech, and Music Processing*, pp. 1–14, 2009.

L. Peltola, C. Erkut, P. Cook, and V. Välimäki, "Synthesis of hand clapping sounds," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1021–1029, 2007.

J. Phillips-Silver, P. Toiviainen, N. Gosselin, O. Piché, S. Nozaradan, C. Palmer, and I. Peretz, "Born to dance but beat deaf: A new form of congenital amusia," *Neuropsychologia*, vol. 49, pp. 961–969, 2011.

M. Puckette, "Pure data: another integrated computer music environment," in *Proc. Second Intercollege Computer Music Concerts*, Tachikawa, Japan, 1996, pp. 37–41.

M. Puckette, T. Apel, and D. Zicarelli, "Real-time audio analysis tools for pd and msp," in *Proceedings of International Computer Music Conference (ICMC)*, Ann Arbor, MI, USA, Jan. 1998, pp. 109–112.

F. Quek, D. McNeill, R. Bryll, S. Duncan, X. Ma, C. Kirbas, K. McCullough, and R. Ansari, "Multimodal human discourse: gesture and speech," *ACM Trans. Computer-Human Interaction (TOCHI)*, vol. 9, no. 3, pp. 171–193, 2002.

J. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

D. Reidsma, A. Nijholt, and P. Bos, "Temporal interaction between an artificial orchestra conductor and human musicians," *Computers in Entertainment (CIE)*, vol. 6, no. 4, pp. 53:1–53:22, 2008.

B. Repp, "The sound of two hands clapping: An exploratory study," *J. Acoustical Society of America*, vol. 81, no. 4, pp. 1100–1109, 1987.

——, "Sensorimotor synchronization: A review of the tapping literature," *Psychonomic Bulletin & Review*, vol. 12, no. 6, pp. 969–992, 2005.

B. Repp, W. Luke Windsor, and P. Desain, "Effects of tempo on the timing of simple musical rhythms," *Music Perception*, vol. 19, no. 4, pp. 565–593, 2002.

A. Robertson and M. Plumbley, "B-Keeper: A beat-tracker for live performance," in *Proc. 7th Intl Conf. New Interfaces for Musical Expression (NIME)*, New York, NY, USA, June 2007, pp. 234–237.

D. Rocchesso and P. Polotti, “Designing continuous multisensory interaction,” in *Sonic Interaction Design; a Conf. Human Factors in Computing Systems (CHI) Workshop*, Florence, Italy, Apr. 2008, pp. 3–9.

D. Rocchesso and S. Serafin, “Sonic interaction design,” *Intl. J. Human-Computer Studies*, vol. 67, no. 11, pp. 905–906, 2009.

D. Rocchesso, P. Polotti, and S. Delle Monache, “Designing continuous sonic interaction,” *Intl. J. Design*, vol. 3, no. 3, pp. 13–25, 2009.

S. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” *IEEE Trans. Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–673, 1991.

P. Schaeffer, *Traité des objects musicaux*. Paris, France: Seuil, 1966.

——, *Solfège de l’objet sonore*. Paris, France: INA/GRM, 1998, first published in 1967.

N. Schaffert, K. Mattes, and A. Effenberg, “A sound design for the purposes of movement optimisation in elite sport (using the example of rowing),” in *Proc. 15th Intl. Conf. Auditory Display (ICAD)*, Copenhagen, Denmark, May 2009, pp. 18–21.

——, “Examining effects of acoustic feedback on perception and movement adjustments in on-water rowing training,” in *Proc. Audio Mostly*, Coimbra, Portugal, Sept. 2011, pp. 122–129.

E. Scheirer, “Tempo and beat analysis of acoustic signals,” *J. Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

J. Seppänen, “Computational Models of Musical Meter Recognition,” Master’s thesis, Tampere Univ. Technology, Tampere, Finland, 2001.

——, “Tatum grid analysis of musical signals,” in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2001, pp. 131–134.

J. Snyder, E. Hannon, E. Large, and M. Christiansen, “Synchronization and continuation tapping to complex meters,” *Music Perception*, vol. 24, no. 2, pp. 135–146, 2006.

F. Spillers, “"Synch with me": Rhythmic interaction as an emerging principle of experiential design,” in *Proc. 6th Conference on Design and Emotion*, Hong Kong SAR., Oct. 2008, pp. 1–17.

A. Sporka, “Non-Speech Sounds for User Interface Control,” Ph.D. dissertation, Faculty of Electrical Engineering, Czech Technical Univ., Prague, Czech Republic, 2008.

——, “Pitch in non-verbal vocal input,” *ACM SIGACCESS Accessibility and Computing*, no. 94, pp. 9–16, 2009.

D. V. Steelant, K. Tanghe, S. Degroeve, B. D. Baets, M. Leman, J. Martens, and T. D. Mulder, “Classification of percussive sounds using support vector machines,” in *Proc. Annual Machine Learning Conf. of Belgium and The Netherlands (BENELEARN)*, Brussels, Belgium, Jan. 2004, pp. 146–152.

D. Stowell and M. Plumbley, "Delayed decision-making in real-time beatbox percussion classification," *J. New Music Research*, vol. 39, no. 3, pp. 203–213, 2010.

D. Stowell, A. Robertson, N. Bryan-Kinns, and M. Plumbley, "Evaluation of live human-computer music-making: Quantitative and qualitative approaches," *International Journal of Human-Computer Studies*, vol. 67, no. 11, pp. 960 – 975, 2009.

S. Strogatz, *Sync: The Emerging Science of Spontaneous Order*. New York, NY, USA: Penguin Books, 2003.

K. Tanghe, S. Degroeve, and B. D. Baets, "An algorithm for detecting and labeling drum events in polyphonic music," in *Proc. First Annual Music Information Retrieval Evaluation eXchange*, London, UK, Sept. 2005, pp. 11–15.

M. H. Thaut, *Rhythm, Music, and the Brain: Scientific Foundations and Clinical Applications*, ser. Studies on New Music Research. New York, NY, USA: Routledge, 2005.

A. Tindale, A. Kapur, G. Tzanetakis, and I. Fujinaga, "Retrieval of percussion gestures using timbre classification techniques," in *Proc. Intl. Conf. Music Information Retrieval (ISMIR)*, Barcelona, Spain, Oct. 2004, pp. 541–544.

P. Toiviainen, G. Luck, and M. Thompson, "Embodied metre: hierarchical eigenmodes in spontaneous movement to music," *Cognitive processing*, vol. 10, pp. 325–327, 2009.

K. Tuuri, "Hearing gestures - vocalisations as embodied projections of intentionality in designing non-speech sounds for communicative functions," Ph.D. dissertation, University of Jyväskylä, Jyväskylä, Finland, 2011.

C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," in *Proc. 4th Intl. Symp. on Independent Component Analysis and Blind Signal Separation*, Nara, Japan, Apr. 2003, pp. 843–848.

V. Välimäki, M. Karjalainen, Z. Janosy, and U. Laine, "A real-time DSP implementation of a flute model," in *IEEE Intl. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, San Francisco, CA, USA, Mar 1992, pp. 249–252.

D. Van Nort, "Instrumental listening: Sonic gesture as design principle," *Organised Sound*, vol. 14, no. 02, pp. 177–187, 2009.

S. Vesa and T. Lokki, "An eyes-free user interface controlled by finger snaps," in *Proc. 8th Intl. Conf. Digital Audio Effects (DAFx)*, Madrid, Spain, Sept. 2005, pp. 262–265.

Y. Visell, F. Fontana, B. Giordano, R. Nordahl, S. Serafin, and R. Bresin, "Sound design and perception in walking interactions," *Intl. J. Human-Computer Studies*, vol. 67, no. 11, pp. 947–959, 2009.

M. Wanderley and N. Orio, "Evaluation of input devices for musical expression: Borrowing tools from HCI," *Computer Music J.*, vol. 26, no. 3, pp. 62–76, 2002.

G. Wang, "Designing Smule's iPhone Ocarina," in *Proc. Intl. Conf. New Interfaces for Musical Expression*, Pittsburgh, PA, USA, June 2009, pp. 1–5.

G. Weinberg, “Interconnected musical networks - toward a theoretical framework,” *Computer Music J.*, vol. 29, no. 2, pp. 23–39, 2005.

G. Weinberg and S. Driscoll, “Toward robotic musicianship,” *Computer Music J.*, vol. 30, no. 4, pp. 28–45, Jan. 2006.

G. Weinberg, S. Driscoll, and M. Parry, “Musical interactions with a perceptual robotic percussionist,” in *IEEE Intl. Workshop on Robot and Human Interactive Communication*, Nashville, TN, USA, Aug. 2005, pp. 456–461.

G. Weinberg, A. Beck, and M. Godfrey, “ZooZBeat: A gesture-based mobile music studio,” in *Proc. Intl. Conf. New Interfaces of Musical Expression (NIME)*, Pittsburgh, PA, USA, June 2009, pp. 312–315.

G. Weinberg and S. Driscoll, “The interactive robotic percussionist: new developments in form, mechanics, perception and interaction design,” *Proc. ACM/IEEE Intl. Conf. Human-Robot Interaction (HRI ’07)*, pp. 97–104, Mar. 2007.

C. Xu, Y. Zhu, and Q. Tian, “Automatic music summarization based on temporal, spectral and cepstral features,” in *Proc. IEEE Intl. Conf. Multimedia and Expo*, vol. 1, Lausanne, Switzerland, 2002.

K. Yoshii, M. Goto, and H. Okuno, “Automatic drum sound description for real-world music using template adaptation and matching methods,” in *Proc. Intl. Conf. Music Information Retrieval (ISMIR)*, Barcelona, Spain, Oct. 2004, pp. 184–191.

——, “INTER: D: A drum sound equalizer for controlling volume and timbre of drums,” in *Proc. 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies (EWIMT2005)*, London, UK, Nov. 2005, pp. 205–212.

A. Zils, F. Pachet, O. Delerue, and F. Gouyon, “Automatic extraction of drum tracks from polyphonic music signals,” in *Proc. 2nd Intl. Conf. Web Delivering of Music (WEDELMUSIC)*, Darmstadt, Germany, Jan. 2002, pp. 179–183.

Errata

Publication II

The page numbers in reference [2] should read 1100–1109.

To date, sound has been an underutilized modality in human-computer interaction (HCI). Recent advances in sonic interaction design have developed more tools and applications for utilizing sound at the interface, but rarely as the primary input. This dissertation argues that there is lots of unharnessed potential in the use of sonic gestures as input in HCI. Special focus in this research is on percussive sonic gestures, such as hand claps, and the design and evaluation of rhythmic interactive systems utilizing sonic input.

Aalto-DD 32/2012

ISBN 978-952-60-4553-5
ISBN 978-952-60-4554-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
www.aalto.fi