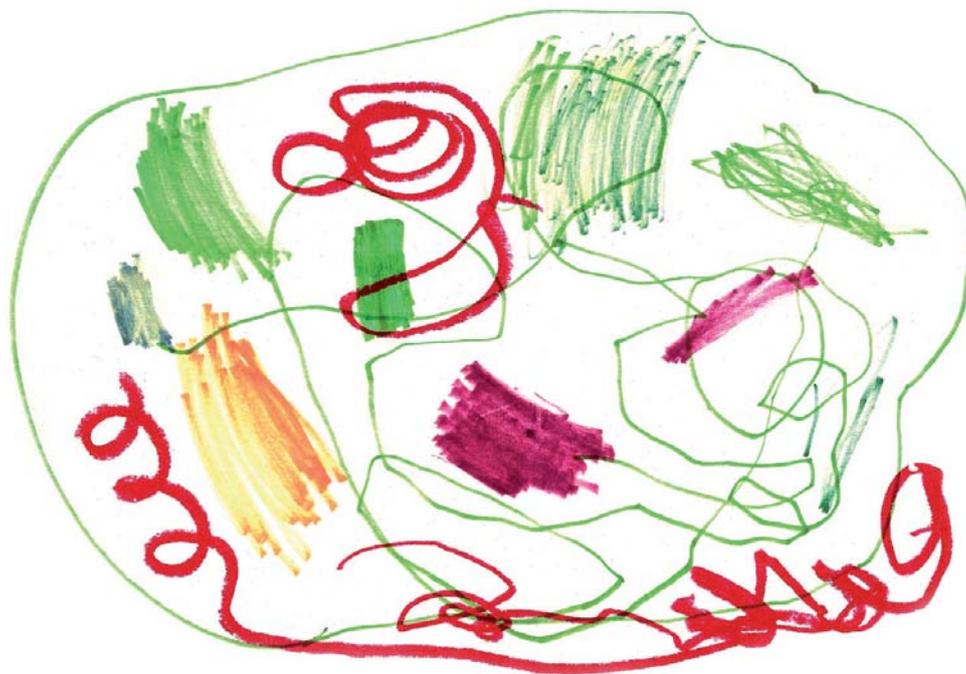


Kai Tuuri

Hearing Gestures

Vocalisations as Embodied Projections
of Intentionality in Designing
Non-Speech Sounds
for Communicative Functions



JYVÄSKYLÄ STUDIES IN HUMANITIES 155

Kai Tuuri

Hearing Gestures

Vocalisations as Embodied Projections of Intentionality in Designing Non-Speech Sounds for Communicative Functions

Esitetään Jyväskylän yliopiston humanistisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Juomatehtaan auditoriossa (JT120)
kesäkuun 21. päivänä 2011 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Humanities of the University of Jyväskylä,
in building Juomatehdas, auditorium (JT120), on June 21, 2011 at 12 o'clock noon.



UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2011

Hearing Gestures

Vocalisations as Embodied Projections of
Intentionality in Designing Non-Speech
Sounds for Communicative Functions

JYVÄSKYLÄ STUDIES IN HUMANITIES 155

Kai Tuuri

Hearing Gestures

Vocalisations as Embodied Projections of
Intentionality in Designing Non-Speech
Sounds for Communicative Functions



UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2011

Editors

Tuomas Eerola

Department of Music, University of Jyväskylä

Pekka Olsbo

Publishing Unit, University Library of Jyväskylä

Jyväskylä Studies in Humanities

Editorial Board

Editor in Chief Heikki Hanka, Department of Art and Culture Studies, University of Jyväskylä

Petri Karonen, Department of History and Ethnology, University of Jyväskylä

Paula Kalaja, Department of Languages, University of Jyväskylä

Petri Toiviainen, Department of Music, University of Jyväskylä

Tarja Nikula, Centre for Applied Language Studies, University of Jyväskylä

Raimo Salokangas, Department of Communication, University of Jyväskylä

Cover picture: Miisa Tuuri

URN:ISBN:978-951-39-4367-7

ISBN 978-951-39-4367-7 (PDF)

ISBN 978-951-39-4290-8 (nid.)

ISSN 1459-4331

Copyright © 2011, by University of Jyväskylä

Jyväskylä University Printing House, Jyväskylä 2011

ABSTRACT

Tuuri, Kai

Hearing Gestures: Vocalisations as Embodied Projections of Intentionality in Designing Non-Speech Sounds for Communicative Functions

Jyväskylä: University of Jyväskylä, 2011, 50 p.(+included articles)

(Jyväskylä Studies in Humanities

ISSN 1459-4331; 155)

ISBN 978-951-39-4290-8 (nid.), 978-951-39-4367-7 (PDF)

Finnish summary

Diss.

It has been argued that we humans understand each other's emotions and intentions on the basis of attuning to the performed actions of others at an embodied level. This dissertation examines the possibilities for utilising such an embodied nature of nonverbal communication in non-speech user interface sound design for human-computer interaction. The approach of this work emphasises situated interactional experiences and the active role of the body and kinaesthesia as a framework of meaning-creation. The approach reciprocally conceives kinaesthetic experience as a mode for performing actions and as well as a mode for understanding them. The central assumption is that the intentionality in gesturally relevant sound events is kinaesthetically graspable through imitative and empathetic involvement in perception.

The focus of this research is on studying the communication of intentionality (i.e., intentional and affective states) through nonverbal gestural projections in vocal expressions. The emphasis is on examining the couplings between intonation structures of vocal expressions and the context-situated communicative intent behind them. As an epistemological background, it is assumed that intonation structures relate to kinaesthetic experience of movement, which in itself is meaningful in terms of the perceiver's *intoned knowledge* – a tacit form of knowledge based on the experiential background of vocal interactions. The proposed action-oriented approach to sonic communication inseparably concerns affective and functional kinaesthetic characteristics involved in sounds.

The empirical studies of this dissertation specifically examined intonation patterns involved in vocal gestures performed for certain interactional communicative functions. In this case, the functions were related to the regulation of physical training activity. Results clearly demonstrated couplings between intonation patterns and communicative functions. The findings also showed that these function-specific intonation patterns communicated the intended meanings when they were implemented in user interface sounds. As a result, this dissertation provides a validated framework for utilising vocal gestures in the design of non-speech sounds for communicative functions of human-computer interaction. According to the experimental results, the proposed design method can be also utilised cross-modally, in the design of vibrotactile user interface elements. Moreover, the underlying design principles, outlined in the dissertation, can be more broadly applied to interaction design and to the development of user interfaces.

Keywords: embodied cognition, sonic interaction, gesture, sound design, user interface, human-computer interaction, intonation, prosody, intoned knowledge

Author	Kai Tuuri Department of Music University of Jyväskylä Finland
Supervisors	Professor Tuomas Eerola Department of Music University of Jyväskylä Finland Senior Researcher Antti Pirhonen Department of Computer Science and Information Systems University of Jyväskylä Finland
Reviewers	Professor Marc Leman Institute for Psychoacoustics and Electronic Music Ghent University Belgium Assistant Professor Dik Hermes Department of Industrial Engineering and Innovation Sciences Eindhoven University of Technology The Netherlands
Opponent	Professor Jaana Parviainen Department of History and Philosophy University of Tampere Finland

PREFACE

One of my favourite toys was a tape recorder. So, I guess I have always been fascinated by sounds and the expressive and communicative powers they possess. Ever since I started my academic studies (in the early 90s), that "childish" attraction has gradually matured into a more formulated interest in sonic phenomena and the ways they become meaningful for us. For instance, I remember how learning about R. Murray Schafer's ideas on studying the soundscape had a huge personal impact on "opening my ears" to acoustic ecology and sonic interactions. This study has its roots in this soil of enthusiasm which has nurtured my growing interest in sound design.

By the time I earned my Master's degree in music education, my professional emphasis was started to shift from music towards sound design. I recognise two aspects that became the main catalysts for this shift. The first was my friendship and collaboration with Jarkko Tornberg. In many ways he was my mentor, who essentially guided me into the mentality of a professional sound designer. The second aspect was the impact of having multimedia studies as part of my university curriculum. Back then, these studies provided a valuable opportunity for carrying out and developing practical design work. They also allowed me to combine sound design with my other long-term area of interest, which is computing.

I was fortunate enough to end up working as a lecturer in the very same multimedia studies programme. Those years as a teacher provided me with an ideal environment for in-depth learning about design for interactive digital media. For a large part, studies were informally embedded into practical design projects which brought together students from several different departments of the university. I cannot overemphasise how much I myself learned through the collaboration with our students and with my colleagues during that time. Also, through numerous insightful discussions, I was introduced to a rich variety of approaches relating to interaction design and content creation. These approaches were characterised by different academic disciplines, such as studies in arts, culture, communication, information systems, education and cognitive science. Relevantly to the present study, my time with the multimedia studies programme shed light on many of the challenges of user interface design, and the related sound design.

Here I must also refer to Simo Alitalo's insightful radio-feature *Mitä kuulemalla tietää?* (*What do we know by hearing?*), which explores the primordial basis of knowing through sounds and first made me explicitly aware of the idea of action-oriented ontology of meanings.

My deepest gratitude goes to Antti Pirhonen, not only for being a supportive colleague for the past years, but also for giving me the opportunity to work on the GEAR2 and GEAR3 research projects (funded by TEKES). After a few years' hiatus from sound design, these projects brought me back to the subject and helped me to outline the goals for my PhD research. The GEAR projects also

brought me into contact with the research field of human-computer interaction. I warmly thank all my project colleagues at Jyväskylä for inspiring collaboration, especially Manne-Sakari Mustonen for sharing a workroom for two years, and Wesley Hatch for the short but effectual stint he worked in the project. Equally warm thanks go to the research partners of the GEAR projects; Pasi Välikkynen from VTT (Technical Research Centre of Finland), Eve Hoggan from the University of Glasgow and Harri Rantala, Markku Turunen and Roope Raisamo from TAUCHI (University of Tampere). I also thank all our industrial partners for providing motivating real-life design cases for our studies. Special gratitude goes to the people at Suunto Ltd. for all the cooperation and support they gave to my work.

My dissertation would never have been finished without the constant support offered by my supervisors Tuomas Eerola and Antti Pirhonen. I can honestly say that it has been a luxury to work with such outstanding advisors with complementing expertises and orientations. Tuomas, in particular, has my sincere appreciation for repeatedly encouraging me to go on with my research ideas – always making me feel that my plans are doable and worthwhile. Furthermore, the help and guidance he has provided in statistical analysis has been invaluable.

My visit to the ISSSM 2007 summer school at Ghent University greatly influenced the theoretical framework of the dissertation, by bringing the views of embodied cognition and gestural views in music research to my attention. It also provided me with encouragement for studying vocalisations with empirical methods. My gratitude goes to all the great teachers and fellow students of ISSSM 2007 for making the intensive course a truly memorable and fun experience.

I am grateful to Marc Leman and Dik Hermes for reviewing the manuscript for this dissertation, and to Jaana Parviainen for being an opponent at the public examination. I thank the Publishing Unit of Jyväskylä University Library and the editors of the Studies in Humanities series for accepting this work for publication. I also warmly thank Glyn Hughes and Steve Legrand for their help in writing English. For funding my work, I am thankful to TEKES along with the Department of Computer Science and Information Systems, Department of Music and Faculty of Humanities at the University of Jyväskylä.

Without doubt, my warmest thanks go to my dear wife Kirsti and to my precious daughters Elina and Miisa. The latter also deserves special thanks for drawing the cover picture, which illustrates "thoughts in the brain". I dedicate this book to my mum and dad.

Jyväskylä, May 2011

A handwritten signature in black ink, appearing to be 'H. O.', written in a cursive style.

CONTENTS

ABSTRACT

PREFACE

CONTENTS

LIST OF INCLUDED ARTICLES

1	INTRODUCTION	11
1.1	Meanings Created through Interaction	12
1.2	Communicative Potential of Vocal Gestures	13
1.3	Structure of the Dissertation	14
2	THEORETICAL POSITION.....	15
2.1	Epistemological Issues	15
2.1.1	The Ongoing Paradigm Shift in Cognitive Science	15
2.1.2	Motoric Involvement in Perception	18
2.1.3	Embodied Communication of Intentionality.....	19
2.2	What is Sound Design?	20
2.2.1	Communication of Mental Images.....	20
2.2.2	Three Aspects of Design	21
2.2.3	Issues of Design Methodology	23
2.3	Vital Meanings of Melody	24
2.3.1	Knowing through Vocal Involvement	24
2.3.2	Couplings between Communicative Functions and Intonation	25
3	STUDIES	28
3.1	Research Aims	28
3.2	Study 1: Modes of Listening	28
3.3	Study 2: Bodily Engagement in Multimodal Interaction	31
3.4	Study 3: Gesture-Based Approach to UI Sound Design.....	33
3.5	Study 4: Outlining and Evaluating the Method of Prosody-Based Sound Design.....	34
3.6	Study 5: Testing the Cross-Modal Functionality of F_0 Contours	36
4	CONCLUSIONS	38

REFERENCES

YHTEENVETO (FINNISH SUMMARY)

INCLUDED ARTICLES

LIST OF INCLUDED ARTICLES

- PI Kai Tuuri, Manne-Sakari Mustonen & Antti Pirhonen. Same Sound – Different Meanings: A Novel Scheme for Modes of Listening. In Proceedings of Audio Mostly 2007, 2nd Conference on Interaction with Sound, Fraunhofer IDMT, Ilmenau, Germany, pp. 13–18, 2007.
- PII Kai Tuuri & Tuomas Eerola. Formulating a Revised Taxonomy for Modes of Listening. Submitted for consideration in the Journal of New Music Research.
- PIII Kai Tuuri, Antti Pirhonen & Pasi Välikkynen. Bodily Engagement in Multimodal Interaction: A Basis for a New Design Paradigm?. In S. Kurkovsky (Ed.), *Multimodality in Mobile Computing and Mobile Devices: Methods for Adaptable Usability*, IGI Global, Hershey, pp. 137–165, 2009.
- PIV Kai Tuuri. Gestural Attributions as Semantics in User Interface Sound Design. In Kopp, S. & Wachsmuth, I. (Eds.), *Gesture in Embodied Communication and Human-Computer Interaction (LNAI 5934)*, Revised Selected Papers of 8th International Gesture Workshop GW2009, Bielefeld, Springer-Verlag, Heidelberg, pp. 257–268, 2010.
- PV Kai Tuuri & Tuomas Eerola. Could function-specific prosodic cues be used as a basis for non-speech user interface sound design?. In P. Susini & O. Warusfel (Eds.), *Proceedings of the 14th International Conference on Auditory Display, IRCAM, Paris, France, 2008*.
- PVI Kai Tuuri & Tuomas Eerola. Identifying function-specific prosodic cues for non-speech user interface sound design. In J. Pakarinen, C. Erkut, H. Penttinen & V. Välimäki (Eds.), *Proceedings of the 11th International Conference on Digital Audio Effects, Helsinki University of Technology, Espoo, pp. 185–188, 2008*.
- PVII Kai Tuuri, Tuomas Eerola & Antti Pirhonen. Design and Evaluation of Prosody Based Non-Speech Audio Feedback for Physical Training Application. *International Journal of Human-Computer Studies*, Accepted pending revision.
- PVIII Kai Tuuri, Tuomas Eerola & Antti Pirhonen. Leaping Across Modalities: Speed Regulation Messages in Audio and Tactile Domains. In Nordahl, R., Serafin, S., Fontana, F. & Brewster, S. (Eds.), *Haptic and Audio Interaction Design (LNCS 6306)*, Proceedings of 5th International Workshop HAID 2010, Copenhagen, Springer-Verlag, Heidelberg, pp. 10–19, 2010.

Author's contribution to the included articles: In PI, the author was responsible for the theoretical contribution and co-contributed to empirical data collection and discussion on observations. In PII, the author was mainly responsible for the theoretical contribution and discussions. The author carried out the bulk of the work in PIII, with the second author co-contributing to the theoretical analysis and the third author conducting and reporting the evaluation experiment. In PIV, the author was the sole author. In PV and PVI, the author performed the majority of the work, including the experimental design, data collection and reporting. Statistical analysis was mainly carried out by the second author of the article. The author was responsible for the major part of the work in PVII. Statistical analysis was mainly executed by the second author of the article, and the third author co-contributed to the experimental design and data collection in the evaluation experiment (sub-study 3). In PVIII, the author performed most of the work, including the experimental design, data collection and reporting, with the second author mainly handling statistical analysis.

1 INTRODUCTION

Technology has become a more and more prominent aspect of our everyday environment and modern culture. The ways in which we use and experience technology are therefore not insignificant. Indeed, for this very reason, "human aspects" have already merited a growing amount of attention in the design of technological products (e.g., Norman, 1988, 2004). Arguably, human sciences should largely be responsible for unveiling these aspects and also for ensuring that they are integrated into designs. In an effort to contribute to such an agenda, this dissertation puts its emphasis on sonic interactions – the ways of knowing through sound. It aims at contributing to the design of interactive products by focusing on sonic experiences, interpersonal communicative attributes of sound, and, in particular, the usage of vocal expressions in the context of sound design.

Due to the history of computing, the field of Human-Computer Interaction (HCI) has its roots deep in information theory (Shannon and Weaver, 1949) and system-centered design perspectives (Czaja, 1997). HCI research was also vastly influenced by the mainstream cognitive science of those days, which portrayed the human mind as an information-processing system analogous to computers (Card et al., 1983). As a consequence, the elements of a user interface (UI) have been easily seen in terms of information processing and transmission between input and output "modules" of a computer and a user (Card et al., 1983), but not so much in terms of interactional experience or interpersonal communication. More recently, the focus has shifted towards acknowledging the importance of human experience in designing interactive technologies (e.g., Norman, 2004; Dourish, 2001; Buxton, 2007). But the legacy of conceptualising HCI in mechanistic terms of computer science and psychology, at least to some degree, still makes it difficult to truly appreciate and understand these so-called human factors as fundamental design principles. For example, different products may have very similar interfaces in measurable terms of usability (see Nielsen, 1994) and functionality, but they could still differ dramatically in terms of use experience (Buxton, 2007).

In general, the development of UIs has strongly focused on textual and graphical forms of presentation. However, as humans, we have evolved to naturally gather information through listening, in order to meet our needs for in-

interacting with the environment. Sonic interaction has also played an equally important role in maintaining social relationships and in the development of culture. From the moment of birth – and even before that – we are capable of interpersonal communication via sound, mostly involving expressive vocalisations (Gibson and Pick, 2000; Fernald, 1992). In all, our evolutionary and experiential background of sonic interactions should provide a rich variety of communicative potential to be utilised in HCI. Obvious strengths of UI sounds lie, for example, in applications with small or ubiquitous devices where the visual attention of the user cannot be taken for granted. The specific focus of this study is on the design of so-called *non-speech sounds* (Brewster, 2003), which do not rely on linguistic attributes of sonic communication. This study, however, does not exclusively contribute to HCI as the present subject is also relevant to non-speech communication in music or in other domains of sound design.

1.1 Meanings Created through Interaction

When it comes to actual design of sounds for interaction, we must answer the following fundamental questions:

- How to couple meanings with sound events?
- How to design sounds that support interaction?

Concerning the first question, it would be tempting to argue that the coupling of sound and meaning is a straightforward process. Such a semantic reference could be made, for example, by using some kind of symbolic coding, like Morse code (typical of the design of *earcons*, see Blattner et al., 1989). Alternatively, one can imitate some "naturally meaningful"¹ acoustic phenomena in the design, like the sound of an opening door (*auditory icon* approach, see Gaver, 1989). Musical expressions may also use analogies which refer to phenomena outside the musical domain (see, e.g., Tagg, 1992; Chion, 1993; Tarasti, 2002; Clarke, 2005). However, it is the second question which makes the issue more complex.

Paul Dourish has argued (2001) that couplings between a UI artefact (such as a sound event) and meaning are not something that the designer can absolutely determine. Rather, these couplings are intentional connections that ultimately arise in the course of interaction. It is essential to understand such a connection between a sound and its interactional context, especially when doing sound design for interaction. And any sound design method should derive its design principles from the nature of interaction.

Ultimately we thus attribute meanings to the sounds through interaction. According to the views of ecological perception (Gibson, 1979) and embodied cognition (Varela et al., 1991), meanings are interactionally structured as (en-active) couplings of action and sound. Perception and action are thus intertwined together since (1) the meanings of an environment are structured through

¹ Such "natural" meanings refer to ecologically structured couplings.

embodied subject-environment interactions and (2) meaning-structures, such as action-sound couplings, are organised in terms of directly perceivable action-relevant cues, i.e., *affordances* (Gibson, 1979; Varela et al., 1991). Our experiential background therefore provides an action-oriented perceptual basis for knowing-about-the-world. From this action-oriented perspective, listening is essentially making sense of the world through actions, as is implied by action-relevant cues of listening experience resonating with the well-structured background of sonic experiences (Gaver, 1993; Leman, 2008). Contemporary trends in HCI sound research acknowledge these views, hence preferring to talk about *sonic interaction design* (Rocchesso et al., 2008).

Following the line of thought of Edmund Husserl (1997), I will argue that kinaesthetic experience is fundamental to the coupling of action and perception. After all, considering our existence as corporeal subjects through which we conceptualise the world, kinaesthesia should indeed outline our basic ontology of actions. It is interesting to notice that also in the contemporary theory of *enactive* perception (Noë, 2004), the pivotal emphasis has been put on the sensorimotor capabilities of a perceiver, underlining the role of action and kinaesthesia in perception.

Some may argue that UI sounds could be directly recycled from application to application since they are designed to convey a certain "message". Even if this strategy may work to a some degree, such an absolute view of semantics takes interactional context and its relation to a sound as self-evident, and thus overlooks the communicative potential that a sound can have when it is tailor-designed *as* activity – and *for* activity.

1.2 Communicative Potential of Vocal Gestures

The main motivation of this study is in exploiting the power that vocalisations possess as interpersonal communication. Vocal actions surely represent a highly familiar, expressive and intuitive means of interacting with sound, suggesting that action-sound couplings of vocal interactions should be taken into an account in sonic interaction design for HCI. In this study, nonverbal meanings conveyed by various *prosodic* features (such as melodic contours) of vocalisations are conceived in the framework of gestural communication – and ultimately in terms of *kinaesthetic* experience.

When we express something to someone, our whole body participates in the act of communication (e.g., Kendon, 2004). Unlike the linguistic elements of an expression, gestures are not detached from the direct sensorimotor basis of social interaction. Movement-based communication, the "body language", is best seen as operating in a specific kinaesthetic field (Parviainen, 2006, p. 35) – an interpersonal subset of kinaesthetic habituation, which still maintains its relation to the rich continuum of more general kinaesthetic meanings. "Gestures form the basis of mutual adaptive behavioral resonances that create shared attention and

are responsible for the feeling of being unified with other people" (Leman, 2008, p. 21).

This dissertation aims at explaining the communicative ability of vocal gestures by acknowledging the kinaesthetic foundation from which the gestural meanings arise. Kinaesthesia is not confined to the experiences relating to the internal domain of our bodies; rather, it is essentially a mediator between the internal and external (Husserl, 1997). Therefore the "sense" of kinaesthesia intrinsically taps into the domain of anticipatory, affective and intentional stance of one's bodily existence in-the-world (Merleau-Ponty, 1962). Hence, kinaesthetic experience can be conceived as bodily projection of *intentionality*, the physio-mental engagement with the world. Within this study, a particular interest lies in intentionality being gesturally manifested and communicated in vocal behaviour within an interactional context.

The intuitivity of gestural communication justifies the aim of utilising it in design. However, this matter also has its flip side. The way that gestures are intuitively meaningful is often invisible to us, and camouflaged as common sense. One can argue that designers might have already implicitly utilised gestural communication in their work. But if knowledge about gestural communication, such as the use of voice intonation, is deliberately left implicit, it cannot be utilised in any consistent fashion. It is therefore a challenge to explicate this intuitive, situation-dependent communicative behaviour to be utilised in the sound design process.

1.3 Structure of the Dissertation

The following chapter outlines the theoretical position of this dissertation. Discussions on epistemological issues and the nature of sound design thus aim at providing a unified framework for the included articles. Chapter three summarises the research aims and the five sub-studies that are reported in the compilation of eight articles. In chapter four, conclusions are drawn from the results. Implications and limitations of the results are also discussed.

2 THEORETICAL POSITION

2.1 Epistemological Issues

2.1.1 The Ongoing Paradigm Shift in Cognitive Science

In his theory of scientific revolutions, Thomas Kuhn (1970) has defined a paradigm shift as a change in basic assumptions within the prevailing theory of science. Hence, the ruling set of methods and concepts characterises a scientific paradigm – essentially defining the framework in which science is done. Therefore an established paradigm may prevent scientists from seeing something essential about the subject they are studying. Paradigm shifts usually occur in the need to explain anomalies which are not accounted for by the most general scientific approach.

After the invention of the digital computer, the emerging discipline of cognitive science was largely developed upon assumptions which draw parallels between functionalities of a computer and a human mind (Broadbent, 1958; Gardner, 1987). Jerry Fodor's work (1975; 1983) exemplifies maybe the most extreme result of this so-called cognitivist approach. According to this influential theory, cognitive processes are explicitly divided into three hierarchical classes: (1) *transducers*, such as sensory organs, which transform the external stimuli into neurological signals, (2) innately specified *input-modules*, which process only a specific type of information distinctively from each other in an automated manner. The primary purpose of these modules is in providing representations for abstracted (symbolic) processes of (3) *central systems*, which are responsible for all higher cognitive functions such as consciousness and thinking. The most prominent point in this representationalist account is that thinking is seen as separated from sensorimotor processes. The "language of thought" (as suggested by Fodor, 1975) would therefore utilise sensory independent, *amodal* representations of the external world. Also, motor actions upon the world are often seen in terms of output-modules – hence conforming to the scheme which consists of sequential input, processing and output stages as a general portrayal of human mind (Broadbent, 1958).

Over the last three decades, the prevailing cognitivism and its computer metaphor for mind has received criticism. For example, connectionism was a research programme which argued for the existence of nonsymbolic, emergent forms of knowledge representation (McClelland, 1988). Some researchers, however, have opted to completely reject the old paradigm, in favour of a wholly alternative approach (e.g., Gibson, 1979; Lakoff, 1987; Johnson, 1987; Varela et al., 1991).

One of the fundamental shortcomings of the input-output scheme has been that it implies a clear disassociation between the capacities for perception and action. As perception also requires muscle activity and body movement, it is not easy anymore, or even appropriate, to make a distinction as to which is input and which is output. Gibson (1966) saw the involvement of muscle activity as a perceptual subsystem, as it participates in the exploratory effort of obtaining the information. His seminal studies on perception made him admit that "the simple, neat easily-remembered contrast between receptors and effectors, between sensory and motor, will have to be abandoned" (Gibson, 1966, p. 45). Gibson's theory (1966; 1979) was also revolutionary in suggesting that perceptual meanings are organised in terms of actions. Such an action-oriented ontology of perception emphasises a fundamental coupling between action and perception. It assumes that, due to the experiential background of interactions, the meanings of an environment are directly perceivable as action-relevant values (affordances). In our purposeful interactions with the environment, there is thus a tendency to attribute ecologically relevant activity to external stimuli. Along with this perspective, recent accounts of enactive perception (Noë, 2004) and embodied simulations (Gallese and Lakoff, 2005) have further blurred the boundaries between input and output (i.e., perception and action), by arguing that all perception is being intrinsically "acted out", i.e., the acquiring of perceptual experience requires sensorimotor knowledge and skills. Recent findings in neuroscience support these approaches by highlighting the sensorimotor dependencies of cognition, and the integration of perception and motor action at the neural level (Gallese and Lakoff, 2005).

The above developments in the theory of perception can be seen under the umbrella of embodied cognition (Varela et al., 1991), a growing philosophical approach seeking to reveal how body-based experience is involved in our thinking. It generally opposes the assumptions about the disembodied nature of cognition, suggested by the traditional paradigm of cognitive science. Embodied cognition has not yet been established into a unified or ruling paradigm; rather, it may be best described as a common philosophical stance on a variety of theories and perspectives. The main characteristics of traditional and embodied positions are summarised in Table 1. The embodied perspective is influenced by phenomenological philosophy (Husserl, 1997; Merleau-Ponty, 1962), which has emphasised the inevitable role of bodily experience as a framework for understanding the processes of perception and thinking, and has strongly argued against the legacy of Cartesian mind-body dichotomy of Western culture, which promote ideas about disembodied mind.

TABLE 1 A comparison between characteristics of cognitivist and embodied approaches to human mind.

	Cognitivist approach	Embodied approach
Cognition is seen in terms of ...	information processing as symbolic computation (Fodor, 1975).	a history of structural (sensorimotor) coupling that brings forth a world (Varela et al., 1991).
Ontology of meanings	Representational	Action-oriented
Perception involves ...	input mechanisms which analyse and convert sensory information for cognitive processes (Broadbent, 1958; Fodor, 1983).	1) active role of a perceiver in an interactional situation, 2) simultaneous usage of sensory and motor processes (Gibson, 1966; Noë, 2004).
Mind vs. body/world	Separable (cognition can work through any device that can process its functional elements, see Fodor, 1975).	Phenomenologically inseparable (embodied coupling of world and mind, even abstract concepts are built upon sensorimotor experiences, see Varela et al., 1991; Lakoff and Johnson, 1999)

It is worth noting, that recent developments in the philosophy of mind have never fully superseded the temptingly simple and easy-to-understand computer metaphor for mind in HCI studies (see Card et al., 1983). In today's vocabulary of HCI practitioners, interaction is largely conceptualised in terms of technical devices which represent input and output modalities of interaction. The term *multimodality* thus often refers to different technical instrumentations used as op-

tional channels of information transmission, in the design of HCI applications (e.g., Sarter, 2006). For example, visual displays, speakers and motion actuators as output channels, or keyboards, microphones and motion sensors as input channels – implicitly suggesting that we possess corresponding modules for each type of user-interface technology. The appeal of such an oversimplified approach is in making the analysis and development of HCI applications straightforward.

2.1.2 Motoric Involvement in Perception

Motoric processes are traditionally seen in terms of effectors we use to produce movement with our body (i.e., as output). These processes, however, arguably also play a crucial role in perception. For example, when we listen to music we can intuitively notice patterns of movement, which we can also quite effortlessly express with our body. Such a kinaesthetic element in perception is a well documented phenomenon in the music research literature (e.g., Eitan and Rothschild, 2010; Godøy, 2009).

From the embodied point-of-view, we experience and schematise our being in relation to embodied space (Merleau-Ponty, 1962). As we essentially make sense of the world around us through the kinaesthetic element, our experiential background, which we utilise in perception and thinking, is inevitably structured on the basis of recurrent sensorimotor experiences. Geroge Lakoff (1987) and Mark Johnson (1987) call these kinaesthetically characterised schematic structures image schemas, and have proposed that they form the very basis of both our imagination and reasoning (see also Lakoff and Johnson, 1999). From the perspective of enactive perception, image schemas provide structured patterns of sensorimotor contingency¹, within which we can perceive the idiom of possibilities of movement (Noë, 2004). These above approaches follow Husserl's idea (1997) of kinaesthesia operating as a core mediator between our experiencing and the external world, on which all senses are also dependent.

Motor theory of speech perception (Liberman and Mattingly, 1985) already hypothesised that we understand what we hear, because we ideomotorically "resonate" to the corresponding vocal action, i.e., the way the sound is produced in the vocal tract. Motor theory thus suggests a common code for perception and action, following the *ideomotor framework* originally proposed by Hermann Lotze and William James (Iacoboni, 2009). Later it was found that related motor areas of the brain indeed activate in the course of speech perception (Rizzolatti and Arbib, 1998). Contemporary neurostudies have revealed that there is an ideomotoric element involved in perception that supports the understanding of actions. This element has been evidenced as responses in premotor neurons, often referred to as *mirror neurons*, which seem to transform sensory information into knowledge that agrees with the motor repertoire of the observer (Rizzolatti and Craighero, 2004). In other words, the observer understands the performed action as he/she has performed it him/herself. There is also evidence suggesting that these pro-

¹ Contingency of movement may also be seen in terms of kinaesthetic fields (Parviainen, 2006).

cesses yield similar motor-based understanding of actions, regardless of the sensory domain in which the action was presented (Kohler et al., 2002). The encoding of action in the neural processes thus seems to integrate different sensory modalities, similarly to amodal perception (see Stern, 1985). However, in the light of neurostudies, amodality does not seem to refer to independence from modalities (as the cognitivist view suggests), but rather, it refers to close and early interconnections between widely integrated sensory and motor aspects of perception (Gallese and Lakoff, 2005). In sum, we can assume that ideomotoric "resonances" of perception play a fundamental role in experiencing movement and conceiving actions. This dissertation will also examine the intriguing possibility that the experience of action, as a contour of kinaesthetic patterns, might actually provide a key element in explaining phenomena such as cross-modal interactions (e.g., McGurk and MacDonald, 1976) and amodality in perception.

2.1.3 Embodied Communication of Intentionality

Providing that body and kinaesthesia work as a mediator between the world and our intentional stances, bodily existence and behaviour intrinsically communicate this "aboutness" between the world and a subject (Merleau-Ponty, 1962; Almäng, 2007). Bodily actions arise from motivations, emotions and other intentional determinants which are linked to the interactional context. Therefore, when we engage with other people within such contexts, the mentality of others is manifested in their actions. This corporeal projection of intentionality can occur both involuntarily or in a voluntarily regulated manner (e.g., push and pull effects in vocalisations, see Banse and Scherer, 1996). Since the physical constitution of the human body is universal, the other person's behaviour can be mirrored (due to ideomotoric processes) in the perceiver's embodied experience, and empathetically understood both in terms of context and the perceiver's action-oriented ontology of intentions and emotions (Almäng, 2007; Gallese et al., 2007; Leman, 2008; Iacoboni, 2009). The attribution of intentionality is highly observer-dependent (Searle, 1983) and its understanding is always bound up with the situational context.

There is also a bias to perceive sound events as being intentional, if they suggest biologically relevant movement patterns (Leman, 2008). Arguably, the suggested capacity for motor-related action understanding allows us to primordially attune to intentionality being reflected in acoustic patterns of human-caused sounds. But intentionality can be attributed to even artificially made sound, as long as it is able to evoke ecologically valid motor-mimetic experiences (see Godøy, 2009). This dissertation is motivated by the possible existence of stable couplings between intentionality and acoustic structures (i.e., sonic movement). Such couplings are expected to be based on the background of vocal interactions, being developed within a kinaesthetic field of vocal gestures. Any regularities between *function* (i.e., the motivational stance behind the expression) and *form* (e.g., melodic or rhythmic patterns), would provide valuable design principles for implementing attributes of intentionality in artificially structured sounds. This issue

is further discussed later in the text, as it also formulates the central theme for the studies of this dissertation.

2.2 What is Sound Design?

2.2.1 Communication of Mental Images

Sound design is about communication. It is seen as a means for conveying meaningful experiences within a given medium – be it a film, radio-play, tv-show, theatre-play, video game or user-interface of a cash machine. Within a medium and the context of its use, sounds have certain communicative functions, which they are designed to fulfil. The designer is therefore required to recognise appropriate sound-meaning relationships for implementing sounds for their context-situated functions.

But where do these sound-meaning couplings come from? From the perspective of embodied cognition (see Varela et al., 1991), they are based on structured correlations between subjective experiences and engagements with sounds in certain interactional situations. Couplings are constantly structured through interactions between both natural and cultural constraints (Leman, 2008). Within cultural constraints, sound-meaning relationships also adapt to novel uses and contexts of sound, following Wittgenstein's (1953) idea of a *language game*, potentially creating new meanings through our interactions with sound.

Meaning as a subjective experience is sensorimotoric and imaginative in nature (Lakoff, 1987; Johnson, 1987). Therefore, by reflecting his or her own background of sonic experiences, the designer essentially seeks couplings of sounds and (sensorimotor) mental images attributed to them (see Sonnenschein, 2001). This imaginative attribution of an experience often involves the use of metaphors (Lakoff, 1987). With a functionally appropriate image in his or her mind, through imitation of the reflected sensorimotor phenomena, the designer aims at articulating it into sonic form. For example, if one wants to support the feel of "a movie character becoming exhausted" with a sound, by reflecting corresponding kinaesthetic experiences, the designer may try to implement attributes of gradually slowing down and increasingly wearisome movement in sound effects or in background music. Sound design basically is a form of expression where a "landscape" of mental imagery is physically articulated in sound. Even though these mental images are highly dependent on the personal experiences of the designer, for the sake of communication, the emphasis is on recognising and utilising culturally or even universally shared couplings between sounds and experiences.

The overall communicative goal of the designer is to project certain intended subjective experiences or intentional stances through sound. Following the ideas discussed in earlier sections, in such projection, the sensorimotor domain operates as a mediator through which expression and its embodied meanings are intertwined together. In this dissertation, an action-oriented approach is

applied to sound design, emphasising the understanding of sound as an action, and focusing on action as an experiential phenomenon. But rather than implementing sounds in terms of their mere semantic/iconic reference to an action, the focus of this approach is on the temporal form of activity and movement in sounds. Phenomenologically, sound is basically considered as a temporal contour of sensorimotor experience (and the imagery involved), much in a similar manner as contoured patterns of vitality affects (Stern, 1985) are conceived.

2.2.2 Three Aspects of Design

To better illustrate the nature of sound design as communication, I have identified three main aspects to be accounted for, both when either performing or studying sound design: user, sound-meaning structures and medium. These aspects and their relations are also illustrated in Figure 1. (1) *User*, i.e., a subject interacting with a system or medium. The user is seen here as a purpose-driven active agent, who fundamentally makes sense of sounds in terms of the action-oriented ontology of the interaction (see 2.1). (2) *Sound-meaning structures* refer to structured couplings of world and mind required for attributing meanings in listening. Couplings are based both on an experiential and an evolutionary history of interactions with the world (see Varela et al., 1991). They form a continuum of prior knowledge (see, e.g., Blackler and Hurtienne, 2007; Stefani, 1987), which range from universal abilities to more specialised knowledge. Roughly, these structures can be distinguished as innate (couplings due to genetic constraints), ecological (couplings due to natural constraints) or conventional (couplings due to cultural constraints). For the sake of intuitivity, the design should aim at utilising prior knowledge that would presumably be possessed by the target group of users. Finally, sonic communication is never independent of the characteristics of a given (3) *medium*. This third aspect concerns the functioning artefact that ultimately mediates the sound design for the user to contextually experience it.

For traditional media (such as radio, TV and film), the role of medium in sound communication may seem quite transparent and trivial. But emerging types of interactive media, such as games and other forms of applications and UIs (in devices ranging from smart phones to safety-critical hospital technology), force us to see the medium itself in a new, more performative role (see Aarseth, 1997). In other words, the actual presentation of sounds is dynamically generated: various situational conditions, including user participation and the performative functions in the "mechanics" of a medium, are potentially involved in the audible manifestation. The designer thus must be well aware of these factors, in order to appropriately utilise the functionality of a medium in the actual communication.

In the user-medium interaction, sonic experience gets fused with the overall interactional experience. This applies also to traditional types of media. Michel Chion (1990) has proposed that the narrative context of sound perception in film results in a fused experience of the audio and the visual. This "audio-visual contract" is demonstrated, for example, when sounds of "hammering" are per-

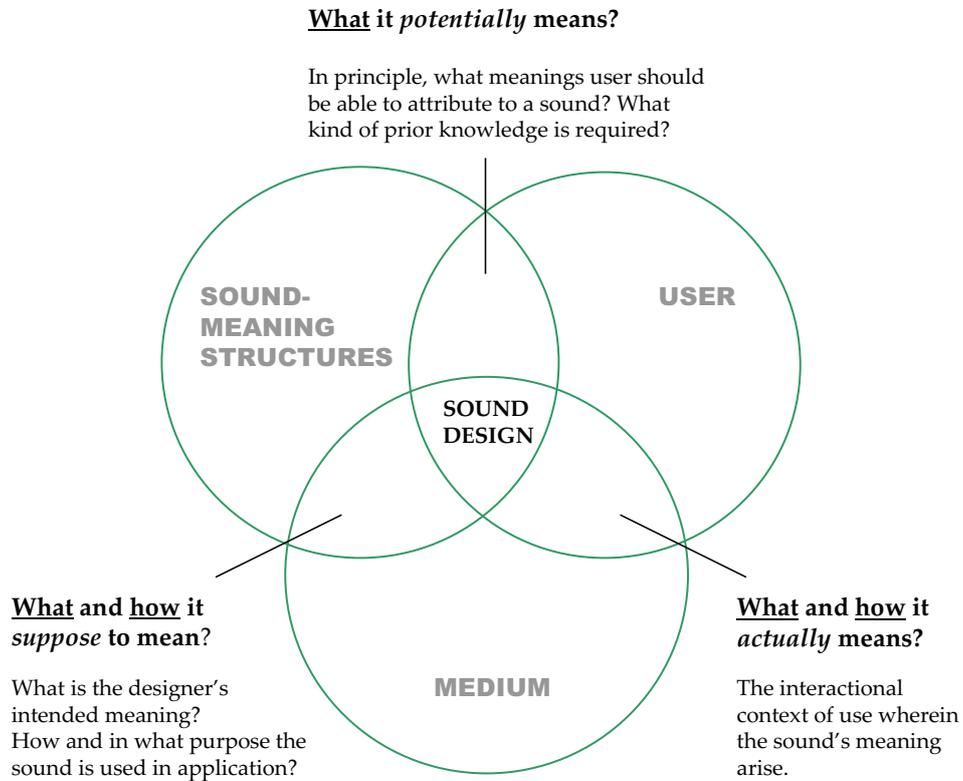


FIGURE 1 Three main aspects of sound design and their interrelations.

ceived as "footsteps" when synchronised with the visual presentation of walking. Chion's notion is important, but from the action-oriented perspective, it could as well be called "audio-action contract" where both the sounds and the visuals are perceptually fused in terms of contingent experience of action (see Noë, 2004). When sounds are being experienced in the context of interactive media and UIs, the enactive fusion of action and sound arguably carries dependencies also for the actions performed by the user. For example, it has been found that the user's motor-movements in interaction and the related sonic/visual feedback can result in pseudo-haptic (Lecuyer et al., 2000) illusions. Thus, auditory or visual sensations, as direct feedback for actions of the user, modulate touch-related perceptions about the physical features of an object, like surface properties (Jousmäki and Hari, 1998; Guest et al., 2002; Mensvoort et al., 2008). User-medium interaction therefore involves a "contract" of action-oriented perceptual fusion of sensory and motor components, which guides the way that sounds are contextually perceived and interpreted. In order to take such experiential complexities into account, designers should be aware of sensorimotor *interaction gestalts* (Svanæs, 2000) to which sounds are intended to relate.

2.2.3 Issues of Design Methodology

From the action-oriented perspective, sound-meaning couplings are fundamentally structured in terms of action-sound dependencies. In this sense, meanings of sounds always relate to the interactional situation and its contingent experiences of action. From this theoretical stance, sound design may be conceived as *sketching of interactional situations*, where sound is intended to evoke or support a certain interactionally relevant experience in the user. As already discussed in the introductory chapter, the perceived involvement of a sound in such a situation is inevitably part of the sound's meaning. In a way, designers have to mentally "foresee" these situations from the user's point-of-view in order to find out appropriate, interaction-derived principles for design. Therefore, the methodological starting point of sound design should be in understanding the interactional context of the sound's application.

The utilisation of *use scenarios* is quite an established practice in the design of UIs. They can provide help in outlining the overall concepts and functionalities of an application (Carroll, 2000) or they can put the emphasis on the user's experience of an application in its context (Cooper, 2004; Pirhonen et al., 2007) and participatory sketching of design ideas and the related experiences (Oulasvirta et al., 2003; Buxton, 2007; Ozenc et al., 2010). Such approaches provide designers with the means to immerse themselves in the situation of use, and from that perspective, to outline appropriate interaction gestalts and tentative functions and ideas for sounds. Participatory approaches should provide designers with embodied knowledge, which is essential for understanding the ways in which sounds fuse with the overall sensorimotor experience.

A fundamental aspect in exploring the context of use lies in recognising the communicative roles of sounds in the course of interaction. The communicative function of a sound, one of the central concepts of this dissertation, should be based on tacit understanding of an interactional situation and its needs and motivations. The well-defined communicative intent of a sound already implies ideas for the intended sonic experience. These may include propositional meanings that sound should convey to fulfil its communicative purpose. But ideas are also evoked in an experiential manner, i.e., in a (non-propositional) form of "feelings" and mental images. In terms of design method, the aim of this dissertation is specifically in finding ways to capture sensorimotor attributes of these tacit, hard-to-be-verbalised design ideas. I will suggest that the capturing could be done by using *corporeal articulations* (Leman, 2008) of such an idea or communicative intent. The resulting physical movement can then be ultimately analysed and utilised in kinaesthetic terms of meaning (i.e., in terms of corporeally projected intentionality).

In practice, UI sound design has both analytic and holistic characteristics, in varying proportions. Analytic approaches tend to emphasise components, such as acoustic parameters of sound (as in the design of alarm sounds, see Patterson, 1982; Edworthy et al., 1991) or some kind of formal coding (as in the design of earcons, see Brewster et al., 1995), rather than the whole. The use of analytic tools

provides consistency and systematics to the design methodology. Purely analytic sound design methods, however, are likely to remain inadequate, since the design process is not something that could be completely reduced to a well-defined subject of analysis. Sound design is essentially a creative and expressive task, rather than just consisting of engineering of acoustic features. Holistic and heuristic approaches are therefore needed to establish a methodological balance between the creative aspects of the design process and the analytic methods required.

2.3 Vital Meanings of Melody

2.3.1 Knowing through Vocal Involvement

In the course of evolution, vocalisations are developed for agonistic² and communicative behaviour both within and between different species (Ploog, 1992; Morton, 1977). Also for humans, nonverbal elements in vocalisation are arguably adapted to serve communicative functions that precede the use of linguistic means of communication (Ohala, 1984; Fernald, 1989). These functions are demonstrated, for instance, in vocal interaction with pre-linguistic infants (Fernald, 1989; Bergeson and Trehub, 2007). Viewed from the theoretical position of this dissertation, vocal involvement as gestural behaviour reflects the bodily intentionality involved in an interaction. The epistemological basis of nonverbal meanings of vocal patterns are therefore found in vocal actions themselves and in the embodied ontology of the intentional stance it projects. These meanings are basically accessible through ideomotoric processes in perception (see 2.1.2). The interest of this dissertation is in utilising this kinaesthetic mode of "knowing through vocal actions" as a domain of sound-meaning structures that are utilisable in sound design.

The proposed ideomotoric processes in perception are thought to be probabilistic in nature (see included articles PII and PIV). In perceptual experience, cues of action-relevancy are thus treated in terms of sensorimotor contingency – in an uncertain and speculative manner. Perceived idioms of movement relate to vocal actions whenever sound (in its context) is able to suggest vocally relevant involvement. But this emergence of kinaesthetic affordances is not deterministic. As the embodied projection of mentality itself is arguably multimodal (i.e., involves multiple parallel bodily expressions³, see Huron et al., 2009; Ohala, 2009), there is reason to assume that vocally produced gestures may evoke correspondingly graded patterns of kinaesthetic experience in perception. Conversely, this would also suggest that any parallel modality of a gesture (e.g., facial, vocal or hand gesture) might incorporate a shared general-level kinaesthetic "coding". If

² Agonistic behaviour is an ethological term that relates to aggressive or submissive stances in interaction.

³ The multimodal nature in producing vocal gestures was also evident in experimental conditions reported in PV, as the whole upper body was mostly involved in vocal expressions.

this is the case, the perceptual involvement of this shared code would potentially permit action-relevant ideomotoric affordances in any of the related modes.

To some extent, approaches of corporeally signified musical meaning have been applied also within the discipline of semiotics. For example, corporeal signs may refer to indexical meanings of musical performance that corporeally emanates the performer's emotional state and the will to express something (Tarasti, 2002). Although semiotics has traditionally had a tendency for structural approaches in studying music, some semioticians has seen that the essence of musical meaning lies in its temporally experienced continuum of "kinetic energy", thus usually emphasising the melodic element (Tarasti, 2003, pp. 172–174). Boris Asafjev's theory of intonation (Zak, 1982) is a prominent example of this line of thought. Bearing similarities to the action-oriented approach, his theory aimed at conceiving music in terms of human voice, focusing on *intoned ideas* captured in music. Some descriptions of such type of musical ideas arguably are demonstrated in studies which use vocal or graphical attuning techniques (see a review in Leman, 2008, pp. 117–121). These studies also point out that it is relatively effortless for us to imitatively express musical (melodic and rhythmic) experience with our voice and body. In general, Asafjev's interesting theory puts weight on the assumption that musical and vocal communication share the same fundamental code (Juslin and Laukka, 2003). I would argue that this code is based on shared kinaesthetic understanding, which is at least partly manifested through intonation.

2.3.2 Couplings between Communicative Functions and Intonation

Intonation, i.e., the "vocal melody", might be the most prominent nonverbal feature of vocal communication. Many studies (e.g., Morton, 1977; Ohala, 1984; Fernald, 1989; Banse and Scherer, 1996; Scherer, 2003) have documented that attribution of affective states and communicative intent is often coupled with melodic characteristics, i.e., temporal structures in contours of fundamental frequency (F_0). Based on regularities found in vocalisations of mammals and birds, Morton (1977) has suggested a set of cross-species *motivation-structural* rules (MS) that describe general correlations between the sound structures (i.e., form) and the situational motivation to communicate (i.e., function). Figure 2 illustrates sound structures associated with varying levels of motivational contexts. MS rules have the following main dimensions:

1. The degree of fear or appeasement. In general, lower F_0 s correlate with dominant/confident position and the image of "bigger" utterer, while higher F_0 s correlate with submissive, fearful/friendly position and the motivation to portray itself as "smaller", appeasing or approachable.
2. The degree of aggression. In general, harsher sound quality with wider spectrum correlates with aggression and hostility, while tonal sounds correlate with non-aggressive motivations. Harsh, low-pitched vocalisations generally indicate imminent attack.

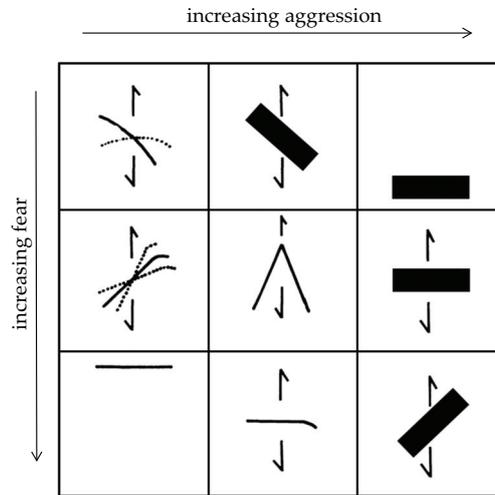


FIGURE 2 Motivation-structural rules of Morton (1977). In each block, the figure's height indicates the frequency. A thick line indicates a harsh tone. Arrows indicate the potential for frequency change and the dotted lines indicate the alternatives in figure's slope.

It has been argued that MS rules would also provide a framework for explaining cross-cultural regularities in speech prosody of humans (Ohala, 1984). A few years ago, I participated in conducting an experiment (Kornysheva et al., 2007) which demonstrated regularities within short infant-directed utterances for *warning* and for *comforting*. The linguistic background of the subjects (10 females, 12 males) was heterogeneous. The found regularities in sound structure can indeed be interpreted on the basis of MS rules (see Figure 2): Warning utterances (high-pitched F_0 contours with steep rising or rise-fall structures) demonstrated increased fear, reduced control and signs of increased aggression, while comforting utterances (low-pitched, falling F_0 structures) correlated with the endpoint of confident/calm position, thus presumably indicating security being provided by a "big" caregiver.

MS rules have been suggested to be predominantly involved with agonistic encounters (Morton, 1977). We might assume that the greater complexity of social interactions of humans has evolved the use of vocal melodies towards more detailed communicative functions that are more socially and culturally oriented. In the framework of human speech communication, it is easily conceived that communicative functions of intonation serve linguistic needs (see Xu, 2005), rather than also seeing these functions in the light of a more primordial continuum that precedes linguistic communication. Such non-linguistic communicative functions of intonation structures have been studied relatively little. Studies of this area of research mostly concern the context of adult-infant interaction. In cross-linguistic studies of infant-directed speech (Fernald, 1992), four distinctive func-

tions (attention bid, approval, prohibition and comfort) have been recognised. All of them demonstrated function-specific F_0 characteristics, despite different wordings and languages used in utterances. In an earlier study (Fernald, 1989) it was found that such stereotyped intonation patterns (with verbal content filtered out) also communicated the intended functions in a recognition test.

Studies presented above make a convincing suggestion that couplings between forms of intonation and communicative functions do exist. These couplings pose themselves as a promising source of sound-meaning structures, as melodic characteristics of vocalisations can be easily implemented in virtually any kind of UI sounds. In sound design applications, pitch-based features of sound are also expected to resist masking effects (see Bregman, 1990) of noisy conditions more robustly when compared to features such as timbre or intensity. But what is the nature of the F_0 coding? Is it innate, or is it learned through interaction? Ohala (1984) argues that it is not appropriate to make clear-cut distinctions between these two: innate behaviour usually involves some learned component, and all learned behaviour requires innate components. But how much is this coding dependent on culturally adapted usage of intonation? It may be best to assume that the use of intonation as gestural behaviour relies on cultural habits to a varying degree. Intoned gestures can thus be placed in a continuum between *affect bursts* and *affect emblems*, referring to a conceptual tool proposed by Klaus Scherer (1994). Affect bursts here stand for a raw vocal display of affective or intentional states. Affect emblems, at the other extreme of the continuum, refer to conventionalised gestures. Intoned acoustic patterns can thus be seen as consisting of varying proportions of both components. In this dissertation, the term "affect" does not exclusively refer to emotion categories. Rather, it more generally refers to a temporally manifested element of intentionality, the purposes and functions of which are characterised in terms of worldly interactions. In many ways, this is close to the conception of vitality affects (Stern, 1985).

3 STUDIES

In all, this dissertation includes eight articles, but their content will be handled within five different sub-studies which are presented in this chapter. Study 1 consists of two articles (PI and PII), studies 2 and 3 correspond to articles PIII and PIV respectively, study 4 comprises three articles (PV, PVI and PVII) and finally, study 5 corresponds to article PVIII.

3.1 Research Aims

The overall aims of this dissertation are threefold. They relate to (1) outlining a theoretical framework, (2) outlining a design methodology and (3) empirical testing of the theoretical and design methodological approaches. The corresponding aims of the five studies are summarised in Table 2. Although all studies have theoretical goals, studies 1-3 have an emphasis on theoretical analysis. Conversely, in studies 4 and 5 the emphasis is on empirical experiments and their results. All studies have a common general-level agenda for applying both theoretical and empirical efforts to a design perspective. Each of these studies also possesses explicit aims relating to design methodology.

3.2 Study 1: Modes of Listening

This study lays a theoretical foundation for understanding sound-based communication. It was motivated by the basic question: what do we know by listening? Listening is an active process, through which we gather information to meet our needs for interacting with the environment. It is not a homogeneous process of grabbing meanings from the information provided by the auditory system, although it is sometimes portrayed as such. The embodied perspective embraces the listener as an action-oriented intentional being making sense of the world.

TABLE 2 Summary of the research aims of the studies.

	Theoretical Aims	Aims Relating to Design Methodology	Empirical Aims
S1	Formulating an epistemological framework of "knowing by listening".	Providing the designers with the means for understanding multifaceted characteristics of a listening experience.	–
S2	On the basis of an action-oriented epistemological perspective, outlining a conceptual framework for designing multimodal UIs.	Conceptualising ways for outlining and utilising experiential "interaction gestalts" for UI element design.	Illustrating the proposed approach with a brief sound design case.
S3	Justifying a gesture-based approach to UI sound design which utilises a kinaesthetic ontology of action-oriented meanings.	Outlining a general process of implementing gestural cues in sound design.	–
S4	Formulating an interpersonal approach to sonic interaction design which justifies the use of gestural invariants of vocal interaction.	Outlining a design method for prosody-based non-speech sounds.	Testing the methodological approach with experiments comprising production and recognition tasks, in the context of a sound design case.
S5	Hypothesising that the kinaesthetic ontology of meanings permits cross-modal communication of the intended meaning.	Proposing an action-oriented approach for conceiving the concept of amodality for the needs of cross-modal design of UI elements.	Evaluating the F_0 contour-based stimuli with supposedly kinaesthetic characteristics in both audio and tactile domains.

Listening experience thus involves active participation (both mental and physical), and it is this intentionality of the listener that ultimately enforces the couplings between sound and meanings. The theoretical stance of this study promotes inherent heterogeneity in ways of knowing by listening. Previous literature has suggested that such a multifaceted nature of meaning-creation could be outlined in terms of distinct listening modes (see below). Modes refer to different constituents of meaning-creation in the process of listening.

The central aim of this study was to deepen understanding of the epistemology of listening by formulating a coherent typology of different listening modes. This approach proposes that each mode of listening is related to its own sources of meanings evoked by sound and its involvement in the interactional context. In the course of evolution, our capacities for different ways of meaning-creation have presumably been shaped to serve different aspects of coping with the world. The taxonomical work on listening modes was done on the basis of a review of the literature and theoretical synthesis.

Previous accounts of Schaeffer (1966), Gaver (1989) and Chion (1990) have made separate but mutually consistent attempts at dividing the ways of attributing meanings in listening into separate processes, broadly divisible into causal, semantic and reduced listening. PI suggests multiple additional categories of listening modes, also loosely incorporating Huron's idea (2002) of activating systems. In this first scheme, eight modes of listening (*reflexive, connotative, causal, empathetic, functional, semantic, critical, and reduced*) were organised into a hierarchical structure, corresponding to the different levels of cognitive operations. PI also includes some empirical observations which, in the form of freely expressed verbal descriptions of the listening experience, demonstrate the involvement of different modes of listening. In addition, this article discusses the relevance of the listening modes approach in the context of UI sound design. Most prominently, this discussion emphasises that the prevalent distinction between auditory icons (see Gaver, 1989) and earcons (see Blattner et al., 1989) is not intended as a distinction between UI-sounds themselves. Rather, it seems to be more related to which modes of listening are being emphasised within the *design paradigm* in question. It is thus argued that different design paradigms, or different methods for sound design, do not necessarily exclude each other because, in the process of design, the involvement of different listening modes can be considered in parallel.

Building upon the first scheme, PII proposes a revised taxonomy of nine listening modes (*reflexive, kinaesthetic, connotative, causal, empathetic, functional, semantic, reduced and critical* listening), with reduced hierarchy between modes and the notable addition of kinaesthetic mode. Kinaesthetic meanings evoked in listening were not covered clearly in the previous taxonomy. In the revised scheme, all modes are re-arranged into domains of *experiential, denotative and reflective* meaning-creation. This arrangement is loosely inspired by Peirce's (1998) scheme of firstness, secondness and thirdness. Modes referring to the experiential domain are reflexive, kinaesthetic and connotative listening. This domain, however, operates in tight interaction with more interpretative processes of denotative domain (causal, empathetic, functional and semantic listening) and re-

flective domain (reduced and critical listening). In all, the revised scheme focuses in much more detail on the experiential basis of meaning in listening. Experiential meanings are theoretically conceived of as emerging resonances between (1) experiential patterns of sensations, (2) well-structured patterns of recurrent sensorimotor experiences (action-sound couplings) and (3) the projection of action-relevant mental images in thinking.

In sum, the taxonomy of listening modes provides a conceptual tool for outlining meaningful attributes of the listening experience and for implementing corresponding features in sound design. Considering the aim of utilising intonation patterns in UI sound design, the most relevant modes of listening are (1) kinaesthetic listening (refers to ideomotoric body movements sensed in the listening experience), (2) empathetic listening (refers to intentionality being empathetically perceived through the resonances between kinaesthetic sensations and the perceiver's action-oriented ontology of intentionality), and (3) functional listening (refers to context-oriented interpretations of a sound's communicative function). According to the revised taxonomy, empathetic and functional perceptions also relate to – or even rely on – a variety of connotations with intersubjective and sociocultural characteristics. For the designer, it is important to acknowledge the involvement of these modes; firstly, how these are positioned in a "big picture" of meaning-making processes and, secondly, how they potentially interact with other ways of knowing by listening (for example, through connotations).

3.3 Study 2: Bodily Engagement in Multimodal Interaction

This study, presented in PIII, was motivated by the implications of the embodied cognition perspective on multimodal interaction in the context of HCI. Theoretical discussion in this article is fundamentally tied into four main arguments:

1. Due to the action-oriented ontology of embodied cognition, *interaction is always multimodal in nature.*
2. *Design arises from mental images and results in mental images.*
3. *Action-relevant mental imagery can be communicated by manifesting action-relevant attributes in the UI.*
4. *Acknowledging the bodily nature of interaction as a basis in design inevitably results in support for multimodal interaction.*

The first argument stresses the integration of sensory and motor components in the interaction experience, and opposes the perspectives which predominantly conceive multimodality in terms of selectable and separable means of interaction (or optional communication channels). Latter views are often being suggested in HCI studies (see, e.g., Sarter, 2006; Naumann et al., 2010). The second and third arguments refer to ideas already introduced in 2.2.1 from the perspective of sound

design. The discussion on the third argument specifically refers to methodological challenges involved in transforming mental images into physical attributes of the design implementation. The fourth argument sums up the overall point of the article.

This study proposes a "multimodal paradigm" for interaction design, where the emphasis of design is in a sensorimotor perceptual coherence instead of, for example, mere sound perception. Because multimodality is conceived as an inevitable experiential quality of interaction, it is not a quality that designers implement in applications – even though multiple different interaction technologies are used. Rather, multimodality is the nature of interaction which designers must take into an account. Hence, even mere sound design should predominantly aim at providing support for a certain, intended multimodal experience. This essentially requires emphasis on how each sound would perceptually fuse into the sensorimotor experience in the user-medium interaction (see 2.2.2 and 2.2.3). Article PIII suggests the exploration of action-relevant mental images for outlining essential sensorimotor qualities of the intended interaction experience. Acknowledging such interaction gestalts (see Svanæs, 2000) would permit seeing potential couplings between sound and the interaction experience intended for the design. Interaction gestalts incorporate schematic structures, comparable to image schemas (Lakoff, 1987; Johnson, 1987), that are based on recurrent (i.e., general type of) sensorimotor experiences. Within the design, interaction gestalts can be based on metaphors which are appropriate to the user-medium interaction.

Article PIII also presents a sound design case which illustrates how design of an unimodal (i.e., sonic) feedback element can acknowledge the couplings of sensory and motor as a perceptual whole. The aim of the case was to sonify an act of "virtual touch" for a so-called physical browsing application (Välkkynen, 2007). Virtual touch here refers to reading of a Radio-Frequency Identification (RFID) tag with a handheld device, without actually causing physical contact between the two. The interactional situation was imaginarily explored by using the Rich Use Scenario method (Pirhonen et al., 2007) and by "bodystorming" (Oulasvirta et al., 2003) the actual physical experience. This process resulted in three interaction gestalts, i.e., mental *action models* that appropriately relate to the interactional context of virtual touch. All of them have slightly different functions. The first one, referring to "connecting", was based on the analogy of real touching instead of virtual. The second one referred metaphorically to "picking up" the content. The third action model was based on a person-to-person interaction metaphor, referring to vocal feedback with polite characteristics for asking "is this the one you want?". Action models are proposed as tools for outlining sensorimotor regularities that are ecologically valid for the intended interaction experience. These regularities (i.e., gestalts) are conceptualised as a certain general type of action which characterises the essential sensorimotor qualities for the design. In addition to UI feedback, action models may be used for conceptualising design ideas for UI input elements such as motion-based control.

Each of the outlined action models was manifested in sonic form. An appropriate click sound was used as a basis for the illustration of a physical "connecting". The gestural image of "picking up" was musically articulated as a rising two-note interval. For the third action model, a short vocalisation with a communicative intent of "asking confirmation" was produced spontaneously with a metaphorically appropriate situation in mind. Contours of F_0 and the first formant (F_1) were extracted from the vocalisation and they were used as a basis for sound synthesis. All three sound elements were appropriately adjusted and mixed together in the final sound implementation. The contextual evaluation test of the implementation yielded positive results. The sound was characterised by qualities such as "friendly", "soft", "moving" and corresponding well with the tag-reading action. Despite the unimodal implementation, the design arguably had multimodal characteristics: Firstly, the sound was intended to be coupled with certain hand movements of the user resulting in a pseudo-haptic illusion of touch. Secondly, the sound itself was intended to support the user in attuning to "movement" in sound. Both the musical and vocal articulations were able to suggest a motor-mimetic image of movement. Also, vocal intonation patterns presumably conveyed kinaesthetic cues of intentionality, interpreted as friendliness by the participants in the evaluation.

The design task of the reported case was deliberately simple – even trivial. Its main purpose was to clearly illustrate the theoretical and methodological approach proposed by the article. The case especially demonstrates the methodological importance of "getting into" the embodied experience of interaction, and seeing interaction gestalts as a starting point of sound design.

3.4 Study 3: Gesture-Based Approach to UI Sound Design

Presented entirely in the article PIV, this study promotes an idea of using gestural projections of body movement as semantics for UI sound design. The general emphasis of the study is on "sensibility for movement", i.e., on the kinaesthetic foundation of bodily mediated action-oriented understanding which forms a basis for interpersonal nonverbal communication. This theoretical background has already been introduced and partly discussed in section 2.1. To justify the design focus on kinaesthetic attributions of *agency* projected through sounds, PIV first provides a review of the literature on interpersonal couplings that permit an understanding of the actions of others. Communication of gestural cues through sounds is then discussed, and finally, a design process is outlined in which gestural cues are implemented in sound design.

To sum up the results, gestural communication relies on two parallel aspects of perception: mimetic involvement (i.e., "mirrored", ideomotor experiences in terms of the perceiver's kinaesthetic schemata) and empathetic involvement (permits understanding of the underlying intentionality in terms of the perceiver's own embodied ontology). It is concluded that such action understanding is not

dependent on which sensory modality is utilised. As documented widely in the music research literature, cues of gestural movement can be perceived through sound. Lens model perspective (Brunswik, 1956; Scherer and Bänziger, 2004) is applied in order to provide an overview on communication where body is acknowledged as a mediator. The applied model simultaneously considers context-situated *encoding* and *decoding* of gestural cues, along with the natural and socio-cultural determinants relating to both processes. The lens model outlines successful communication of gestures by means of multiple, partly redundant action-relevant cues which contribute in a probabilistic and additive manner to the perception. From the perspective of sound design, this can be interpreted as an advantage, because in principle, it allows the designer to select appropriate cues for design and discard others. Therefore gesturally encoded acoustic patterns could be simplified or reduced to conform, for example, to technical or aesthetic restrictions.

For a sound designer, communication with gestural projection essentially means (1) defining a contextually appropriate gesture for a communicative purpose in a UI and (2) articulating it for the design. The first aspect refers to a *modelling phase*, where the designer puts himself/herself into a role as a person who is communicating with the user. The approach proposed in this study elaborates upon the idea of utilising person-to-person interaction gestalts (well-structured patterns of gestural interaction) introduced in study 2. The second aspect mentioned above refers to a *performing phase*, where gestures are spontaneously produced while immersed in an interaction scenario. In a *utilising phase*, recordings of gesturally encoded patterns are implemented in design. Finally, an *evaluating phase* is required for testing the communicative attributes of implemented sounds. These phases (modelling, performing, implementing and evaluating) outline a general design process for utilising interactionally meaningful gestural cues. The emphasis is especially on discovering *stereotypical cues*, which should communicate the intended meaning robustly (i.e., would provide some resistance to contextual variations) without being too strongly dependent on cultural conventions. The presented design phases form a basic framework for the prosody-based design method that will be developed and evaluated in study 4.

3.5 Study 4: Outlining and Evaluating the Method of Prosody-Based Sound Design

This study presents a complete sound design case which outlines a prosody-based method for designing non-speech UI sounds. All stages of the design process are justified with a theoretical analysis and three empirical sub-studies, which comprise production and recognition tasks for four communicative functions involved in the design. Article PVII covers the study as a whole and articles PV and PVI are detailed reports of the aspects involved in the first sub-study. The design principles of the method are founded on nonverbal communicative

functions of prosody – specifically on functions of intonation. Based on the approach outlined in study 3, design aims at utilising gestalts of interpersonal interaction. Gestalts relevant to this study specifically refer to interactional, gesturally encoded intonation patterns. As the discussion in 2.3.2 implies, invariant intonation patterns could be discovered for specific communicative functions, which are coupled with a situated motivation and communicative intent. This study is motivated by the possible existence of such couplings.

The assignment was to design non-speech UI sounds to support interaction between a user and a physical training application in a wrist computer. In the modelling phase, a trainer-runner interaction metaphor was utilised. As a result, four communicative functions for the interaction were outlined: *Slow down* (pace decrease is needed), *Urge* (pace increase is needed), *Ok* (the pace is fine) and *Reward* (for praising the runner on the performance). In the performing phase, interaction-derived vocal gestures were produced for each function. This was done within a controlled experiment incorporating the production task. After the production phase, prosodic features of utterances were extracted and analysed. The results of the analysis widely demonstrated function-specific characteristic in features relating F_0 and intensity. Couplings between function and intonation are also evident by observing within-a-function similarities across different participants in the visualised F_0 contours of utterances (see PV and PVII). Repeated utterances of each participant (for the same function) yielded no statistically significant differences. Also, analysis showed that function-specificity taps into such properties of an expression that are relatively simple and robust to statistically classify even by combination of two basic F_0 descriptors. These observations are interpreted as indications of the robustness of prosodic communication. The half-way results of this study (sub-study 1) were thus very encouraging for the utilisation of these intonation patterns in sound design for these communicative functions.

For the utilisation phase, the best utterances for each function needed to be selected. Therefore the goodness of each utterance for each function was assessed by a listening experiment (sub-study 2). Utterances were then ranked on the basis of this assessment. One utterance for each function was then "designerly" (i.e., on a subjective basis) selected from the shortlist of top-ranked utterances. Intonation contours of the selected utterances are visualised in Figure 3. Reduced renditions of these F_0 contours were then implemented in the final sounds, which conformed to the very limited sound producing (i.e., beeper) capabilities of a wrist computer device.

In the final sub-study, four device-reduced sounds were contextually evaluated. An experiment was conducted which assessed the intuitive recognition of the four meaning categories. The experimental design resorted to a laboratory setting in which the interactional context of the application was simulated to a certain degree. The recognition of the intended meanings was fairly robust across functions. Participants' verbal characterisations of each sound (i.e., descriptions of the evoked mental images) also mostly matched the interaction gestalt outlined by the intended function. In all, on the basis of the results, it could be argued

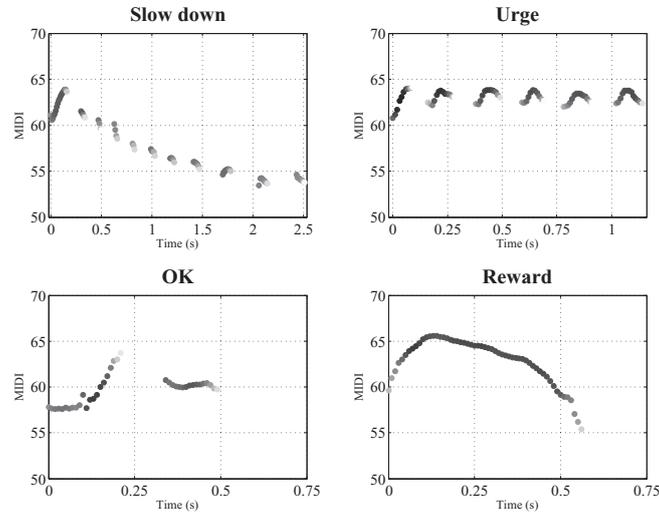


FIGURE 3 Visualisations of the chosen intonation contours for each communicative function.

that all tested sounds possessed communicative attributes (derived from the intentional stance of a vocal source) that facilitated intuitive recognition of their intended purpose. Therefore, in the real-world use of the application, the process of getting accustomed to these sounds should be quite effortless.

3.6 Study 5: Testing the Cross-Modal Functionality of F_0 Contours

This last study is a follow-up to study 4, and it is reported in its entirety in article PVIII. The motivation of this study is in the hypothesis that F_0 contours might function across the domains of auditory and tactile senses. Physical training applications are used in varying contexts (e.g., on a running track or in a gym environment), and therefore it would be beneficial to have options between sounds and more intimately felt vibrations for the UI presentation. According to the ideomotor framework (see Iacoboni, 2009) and the action-oriented approach of this dissertation, perception should involve a common motoric code for sensory experience and action understanding. There is thus reason to believe that all sensory domains integrate with ideomotor action understanding which ultimately operates in the domain of kinaesthetic experience. It is hypothesised that, regardless of the domain of presentation, prosody-based F_0 contours would evoke sensorimotor *experience contours* that resonate with kinaesthetically characterised schematic gestalts. These gestalts arguably relate to gestural signatures of an interpersonal affect, for example, "calming down" for Slow down function, "agitating" or "rousing" for Urge function, "approving" for Ok function and "praising" for Reward function.

The functionality of F_0 contours in audio and tactile domains was tested in a controlled experiment. In the experimental design, only three communicative functions relating to speed regulation (Slow Down, Urge and Ok) were considered. Three different design bases, i.e., forms of F_0 contours for each function, were tested. The first design base consisted of the selected F_0 contours (of Slow Down, Urge and Ok) from study 4 (see Figure 3). The second design base was also prosody-based, originating from the material gathered in study 4. Contours for each function were selected by means of the statistical classification model reported in PVI and PVII. For the third design base, a simple analogy between changes in frequency and the corresponding function was applied. In the F_0 contours of this design base, vibration rate decelerates for Slow down, accelerates for Urge and remains even for Ok function. Such a direct analogy had been previously utilised for creating vibrotactile messages for speed regulation (Lyykangas et al., 2009). Arguably, these contours also refer to schematic gestalts with a kinaesthetic basis. One may suspect that these gestalts relate to vitality affects (Stern, 1985), such as, "reducing/falling/decreasing" for Slow down function, "growing/rising/increasing" for Urge function and "even/steady" for Ok function. Conceived as intoned structures, they bear some resemblance to corresponding prosody-based contours and can also be interpreted with motivation-structural rules (see Figure 2).

The experiment comprised a counterbalanced within-subjects design. In a separate task for each domain, all three design bases were assessed for their ability to intuitively communicate the intended function. Statistically, the results indicated no significant effect of sensory domain. Many participants also expressed that "... understanding was easy to "catch" in both domains", or that "... both domains felt comprehensive". The results of the experiment, together with the verbal comments of the participants, suggest that stimuli were understood similarly in both domains. Domains therefore seemed to function in an interchangeable manner, giving support to the hypothesis. All three design bases performed well in communicating the intended meanings, and thus any of them could potentially be used as a basis for designing "intuitive" messages. However, across functions, there were some differences in the effectiveness of prosody-based and direct analogy stimuli designs. It is thus suggested that best functioning F_0 contour characteristics from different design bases could be combined in the design for an optimal set of stimuli. In particular, according to the results, prosody-based contours for Urge and Ok functions might benefit from generally more ascending or more flat F_0 structures respectively.

Within the HCI research, the concept of cross-modal design in HCI (see, e.g., Hoggan and Brewster, 2007) is based on an assumed existence of amodal content, which can be presented more or less interchangeably in different sensory domains. Instead of conceptualising amodality in the framework of a traditional cognitivist view (see discussion in 2.1.1), the present study suggests that amodal meanings could be better conceived in terms of sensorimotor gestalts, which have a certain kinaesthetic character that integrates information from different senses.

4 CONCLUSIONS

The main subject of this dissertation is about utilising a vocally relevant kinaesthetic "mode of knowing" as a domain of sound-meaning structures in sound design. This is a tacit form of interaction-coupled knowledge which is involved simultaneously in performing vocal expressions as well as in understanding them in terms of action-oriented ontology of meanings. I propose that this enactive and tacit form of knowledge be called *intoned knowledge*. "Intoned" here refers to the enactive nature of knowing through the kinaesthetic experience of intoning. Through this concept, theoretical ideas will be summarised and empirical results can be interpreted within a unified framework. Intoned knowledge can be characterised through its three main features: *ideomotor framework*, *interaction gestalts* and an *embodied linkage to intentionality* (see Figure 4).

The ideomotor framework of human action (see Iacoboni, 2009) assumes a common "code" for action and perception. As discussed earlier in this dissertation, there is an assumed motoric involvement in perception. Sounds (or tactile vibrations) involved in an interactional context thus become perceptually fused in terms of a contextually contingent experience of action. The confirmed cross-modal functionality of intonation contours indeed suggest that understanding predominantly relates to the experience of action rather than experience in a specific sensory domain.

Ideomotor processes are activated in probabilistic terms of sensorimotor contingency, based on the background of experienced actions. It is proposed that such an experiential background is organised into interaction gestalts, which refer to interactional, well-structured sensorimotor experiences. Interaction gestalts referring to intoning as an interpersonal behaviour therefore outline a potential for directly meaningful kinaesthetic (i.e., intoned) experiences. The results of this dissertation suggest that this potential is utilisable in sonic interaction design, even when implemented sounds themselves do not resemble vocal acts. And although the gestalts of vocal interaction were used in design cases, it has been argued (see 2.3.1) that the "idioms of movement", evoked by the F_0 cues in the implemented sounds, are not confined exclusively to a kinaesthetic field of vocalisations. Rather, they can also refer to a continuum of related fields of gestural in-

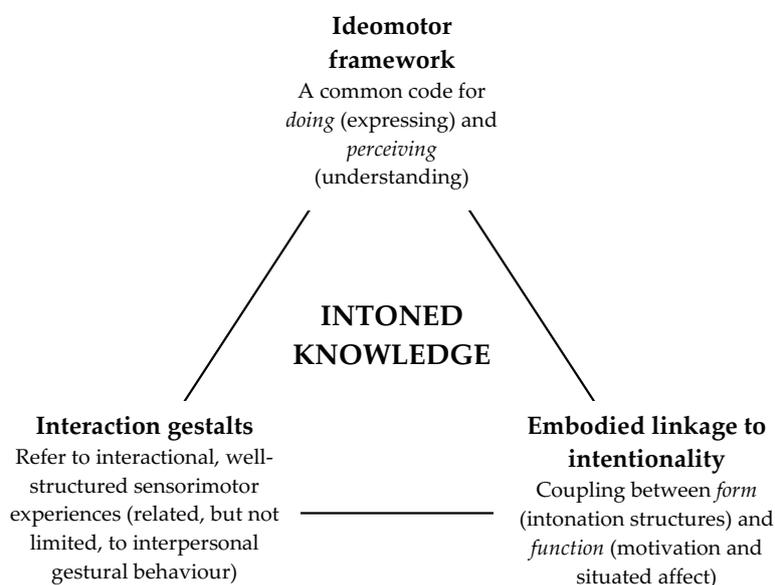


FIGURE 4 Overview of intoned knowledge.

teractions – or even to more general kinaesthetic gestalts of coping with the world (such as vitality affects). It is tempting to hypothesise that intoned knowledge, as a subset of kinaesthetic knowledge, interacts with these related kinaesthetic gestalts.

The third feature in the concept of intoned knowledge concerns the bodily mediated linkage between intentionality and the acts of intoning. The emphasis of this dissertation’s empirical studies has been on this aspect. The empirical results strengthened the already well-justified theory that assumes a coupling between function (motivational and affective states) and the form (intonation structures) of vocalisations. Arguably, these couplings are structured sensorimotor experiences being schematised in interaction gestalts. The communicative value of any intonation pattern is therefore dependent on its relation to the interactional experience. As a consequence, the results of prosody-based design, tailored for certain interaction gestalts (as outlined in communicative functions), would not automatically possess similar functionality in a different interactional situation. For example, the agitative nature of the Urge sound is well suited for physical training situations but might not be fully appropriate for all directive functions referring to acceleration.

With respect to sound design, the dissertation provides conceptual tools, such as a listening modes taxonomy, which can facilitate the human-centred approach to HCI design. Secondly, the action-oriented epistemological approach interestingly puts the emphasis on performative aspects of sound design, i.e., manual ways of producing sounds (or perceiving sounds as being manually caused). This is an important notion, because there is also pressure for developing rule-

based automated processes for HCI sound design (e.g., Leplatre and Brewster, 1998; Bernsen, 1997). However, most importantly, this dissertation proposes a validated method for putting intoned knowledge into the service of sonic interaction design. The proposed method should be inherently scalable to different application contexts, due to the way it arises from the specific demands of the particular application and its interaction gestalts. Prosody as a design principle should potentially lend itself to a host of communicative functions. However, further studies are needed for assessing the validity of the prosody-based method in the UI sound design for different types of interaction and communicative functions.

The proposed methodic framework for utilising intonation patterns of vocalisations is not intended to be taken as an algorithmic guideline for a single "optimal" design process. Rather, the idea is to provide a heuristically adaptable methodic framework. As the prosody-based design in study 2 illustrated, the producing and choosing of vocalisations can also rely on the designer's intuition, rather than a controlled experiment. I would even encourage designers and researchers to explore for more "agile" and less analytic approaches to the utilisation of intoned knowledge. Hence, the sound designer can exploit the intuitivity of vocal expressions in generating, refining and communicating design ideas. In this sense, prosody-based design can be seen in the light of vocal sketching (see Ekman and Rinott, 2010). In early phases of design, vocal sketching would be a logical extension to bodystorming activities (Oulasvirta et al., 2003). Thus, it can be used for vocally expressing the experiential aspects of kinaesthetic thinking (Svanæs, 1997) being involved in the bodystorming process – and simultaneously producing audible sketches that outline ideas for sounds. According to the evidence of using the human voice as a sketching tool (see, e.g., Ekman and Rinott, 2010; Vogt and Höldrich, 2010; Pirhonen and Tuuri, 2009), subjective ideas of actions and events can be sonified quite intuitively with the voice. This stems from the assumption that intoned knowledge integrates with other (often metaphorically projected) aspects of kinaesthetic knowledge. It is also important to note that intoned knowing is not necessarily confined to F_0 -related aspects of sound, as the evidence suggests that F_0 codes are coupled with the mouth shape, which affects the spectral resonances in vocal expressions (Ohala, 1984). Future studies should take these above aspects into account in the development of sound design methods that utilise intoned knowledge.

In future studies, it would also be important to put more emphasis on examining the communicative functionality of more general and graded forms of intonation, rather than focusing strictly on stereotypical patterns. These non-stereotypical "weak cues" (as coined in the included article PIV) are presumably more context-dependent. Indeed, in animal communication, non-stereotypical patterns that conform to MS rules in a graded manner are usually involved in close proximity encounters (Morton, 1977). As the results of the study 5 demonstrated, F_0 patterns with extremely general and nondistinctive characteristics (such as frequency sweeps) can work effectively within certain fixed interactional conditions. Future studies could also examine the possible cross-cultural differences

in encoding and decoding intoned (prosodic) information. Assuming that there is an evolutionary continuity in the formulation of intoned knowledge, at least some universality might be expected. However, it remains to be seen to what extent the results of prosody-based design rely on conventionalised uses of intonation (such as emblems or linguistic dependencies) that are specific, for example, to Finnish or Western culture.

The epistemological background, outlined in this dissertation, potentially has significant implications for the values and the general principles from which the design practises of HCI arise. One of those fundamental principles concerns acknowledging multimodality as an action-oriented, experiential nature of interaction, not as something which designers implement in a technological application. Although there is nowadays a clear striving to understand the user's experience involved in interaction, there is an ever growing need for "phenomenological design tools" through which the experiential domain of interaction design would be accounted for. This dissertation offers a partial answer to such a need, by providing tools for outlining and conceptualising experiential aspects of the interaction design, and also for outlining essential sensorimotor qualities that relate to those aspects.

This work also has implications for music research. Although music is sometimes portrayed as an autonomous and abstract domain, musical meanings are not separable from the everyday aspects of sonic interactions. Indeed, in a historical continuum, musical knowledge has been developed in close interaction with cultural and natural constraints (Leman, 2008). Arguably, musical and intoned forms of knowledge are close relatives, and possess, to at least some degree, common codes for attributing action-oriented meanings to sonic forms (Juslin and Laukka, 2003). Action-oriented ontology of meanings is mutually well-suited for understanding non-speech communication through either music or UI sounds. Both types of sonic communication are able to intuitively evoke meaningful experiences which, for the most part, are not accounted for by the representationalist and semiotic approaches (Johnson, 2007). Similarly to other forms of art, such as dance, sculpture and architecture, it may be concluded that music and the design of non-speech sonic interactions both concern a medium of truly embodied meanings.

REFERENCES

- Aarseth, E. 1997. *Cybertext*. Baltimore, MD: Johns Hopkins University Press.
- Almäng, J. 2007. *Intentionality and intersubjectivity*, Vol. 21. Göteborgs Universitet. *Acta philosophica Gothoburgensia*.
- Banse, R. & Scherer, K. 1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology* 70, 614–636.
- Bergeson, T. R. & Trehub, S. 2007. Signature tunes in mothers' speech to infants. *Infant Behavior & Development* 30 (4), 648–654.
- Bernsen, N. 1997. Defining a taxonomy of output modalities from an HCI perspective. *Computer Standards & Interfaces* 18 (6-7), 537–553.
- Blackler, A. L. & Hurtienne, J. 2007. Towards a unified view of intuitive interaction: definitions, models and tools across the world. *MMI-Interaktiv* 13, 37–55.
- Blattner, M., Sumikawa, D. & Greenberg, R. 1989. Earcons and icons: Their structure and common design principles. *Human-Computer Interaction* 4 (1), 11–44.
- Bregman, A. 1990. *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: The MIT Press.
- Brewster, S., Wright, P. & Edwards, A. 1995. Experimentally derived guidelines for the creation of earcons. In *Adjunct Proceedings of HCI*, Vol. 95. , 155–159.
- Brewster, S. A. 2003. Non-speech auditory output. In J. Jacko & A. Sears (Eds.) *The Human Computer Interaction Handbook*. Hillsdale, NJ: Lawrence Erlbaum Associates, 220–239.
- Broadbent, D. 1958. *Perception and communication*. New York, NY: Pergamon Press.
- Brunswik, E. 1956. *Perception and the representative design of psychological experiments*, Vol. 21. Berkeley, CA: University of California Press.
- Buxton, B. 2007. *Sketching user experiences: getting the design right and the right design*. San Francisco, CA: Morgan Kaufman.
- Card, S., Moran, T. & Newell, A. 1983. *The psychology of human-computer interaction*. Boca Raton, FL: CRC Press.
- Carroll, J. 2000. *Making use: scenario-based design of human-computer interactions*. Cambridge, MA: The MIT Press.
- Chion, M. 1990. *Audio-vision: sound on screen*. New York, NY: Columbia University Press.

- Chion, M. 1993. *Le poème symphonique et la musique à programme*. Paris: Fayard.
- Clarke, E. 2005. *Ways of listening: An ecological approach to the perception of musical meaning*. New York, NY: Oxford University Press.
- Cooper, A. 2004. *The inmates are running the asylum*. Indianapolis, IN: SAMS publishing.
- Czaja, S. 1997. Systems design and evaluation. *Handbook of human factors and ergonomics* 2, 17–40.
- Dourish, P. 2001. *Where the Action Is: The Foundations of Embodied Interaction*. Cambridge, MA: The MIT Press.
- Edworthy, J., Loxley, S. & Dennis, I. 1991. Improving auditory warning design: relationship between warning sound parameters and perceived urgency. *Human factors* 33 (2), 205–231.
- Eitan, Z. & Rothschild, I. 2010. How music touches: Musical parameters and listeners' audiotactile metaphorical mappings. *Psychology of Music*, online Nov 8, 2010.
- Ekman, I. & Rinott, M. 2010. Using vocal sketching for designing sonic interactions. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*. ACM, 123–131.
- Fernald, A. 1989. Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child development*, 1497–1510.
- Fernald, A. 1992a. Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In J. H. Barkow, L. Cosmides & J. Tooby (Eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York, NY: Oxford University Press, 391–428.
- Fernald, A. 1992b. Meaningful melodies in mothers' speech to infants. In H. Papoušek & U. Jürgens (Eds.) *Nonverbal vocal communication: Comparative and developmental approaches*. New York, NY: Cambridge University Press, 262–282.
- Fodor, J. 1975. *The language of thought*. Cambridge, MA: Harvard University Press.
- Fodor, J. 1983. *The modularity of mind*, Vol. 341. Cambridge, MA: The MIT Press.
- Gallese, V., Eagle, M. & Migone, P. 2007. Intentional attunement: Mirror neurons and the neural underpinnings of interpersonal relations. *Journal of the American Psychoanalytic Association* 55 (1), 131–175.
- Gallese, V. & Lakoff, G. 2005. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology* 22 (3), 455–479.

- Gardner, H. 1987. *The mind's new science: A history of the cognitive revolution*. New York, NY: Basic Books.
- Gaver, W. 1989. The SonicFinder: An interface that uses auditory icons. *Human-Computer Interaction* 4 (1), 67–94.
- Gaver, W. 1993. What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology* 5 (1), 1–30.
- Gibson, E. & Pick, A. 2000. *An ecological approach to perceptual learning and development*. New York, NY: Oxford University Press.
- Gibson, J. 1966. *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.
- Gibson, J. 1979. *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.
- Godøy, R. 2009. Gestural Affordances of Musical Sound. In R. Godøy & M. Leman (Eds.) *Musical Gestures: Sound, Movement, and Meaning*. New York, NY: Routledge.
- Guest, S., Catmur, C., Lloyd, D. & Spence, C. 2002. Audiotactile interactions in roughness perception. *Experimental Brain Research* 146 (2), 161–171.
- Hoggan, E. & Brewster, S. 2007. Designing audio and tactile crossmodal icons for mobile devices. In *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 162–169.
- Huron, D., Dahl, S. & Johnson, R. 2009. Facial expression and vocal pitch height: Evidence of an intermodal association. *Empirical Musicology Review* 4, 93–100.
- Huron, D. 2002. A six-component theory of auditory-evoked emotion. In *Proceedings of the 7th International Conference on Music Perception and Cognition*, Sydney, , 673–676.
- Husserl, E. 1997. *Thing and Space: lectures of 1907*. Trans. R. Rojcewicz. Dordrecht: Kluwer Academic Publishers.
- Iacoboni, M. 2009. Imitation, empathy, and mirror neurons. *Annual Review of Psychology* 60 (1), 653–670.
- Johnson, M. 1987. *The body in the mind: The bodily basis of reason and imagination*. Chicago, IL: University of Chicago Press.
- Johnson, M. 2007. *The meaning of the body: Aesthetics of human understanding*. Chicago, IL: University of Chicago Press.
- Jousmäki, V. & Hari, R. 1998. Parchment-skin illusion: sound-biased touch. *Current Biology* 8 (6), 190.

- Juslin, P. & Laukka, P. 2003. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological Bulletin* 129 (5), 770–814.
- Kendon, A. 2004. *Gesture: Visible action as utterance*. Cambridge, UK: Cambridge University Press.
- Kohler, E., Keysers, C., Umiltà, M., Fogassi, L., Gallese, V. & Rizzolatti, G. 2002. Hearing sounds, understanding actions: action representation in mirror neurons. *Science* 297 (5582), 846–848.
- Kornysheva, K., Tuuri, K. & Mustonen, M. 2007. Prosodic Characteristics of Vocal Warning and Comforting Utterances in Humans with and without Interaction Experience with Children. Unpublished raw data.
- Kuhn, T. 1970. *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.
- Lakoff, G. & Johnson, M. 1999. *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York, NY: Basic Books.
- Lakoff, G. 1987. *Women, Fire and Dangerous Things—What Categories Reveal about the Mind*. Chicago, IL: University of Chicago Press.
- Lecuyer, A., Coquillart, S., Kheddar, A., Richard, P. & Coiffet, P. 2000. Pseudo-haptic feedback: Can isometric input devices simulate force feedback? In *Proceedings of Virtual Reality 2000*. IEEE, 83–90.
- Leman, M. 2008. *Embodied Music Cognition and Mediation Technology*. Cambridge, MA: The MIT Press.
- Leplatre, G. & Brewster, S. 1998. Perspectives on the design of musical auditory interfaces. *International Journal of Computing Anticipatory Systems* 4, 227–239.
- Lieberman, A. & Mattingly, I. 1985. The motor theory of speech perception revised. *Cognition* 21 (1), 1–36.
- Lylykangas, J., Surakka, V., Rantala, J., Raisamo, J., Raisamo, R. & Tuulari, E. 2009. Vibrotactile information for intuitive speed regulation. In *Proceedings of the 2009 British Computer Society Conference on Human-Computer Interaction*. British Computer Society, 112–119.
- McClelland, J. 1988. Connectionist models and psychological evidence. *Journal of Memory and Language* 27 (2), 107–123.
- McGurk, H. & MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264 (5588), 746–748.

- Mensvoort, K., Hermes, D. & Montfort, M. 2008. Usability of optically simulated haptic feedback. *International Journal of Human-Computer Studies* 66 (6), 438–451.
- Merleau-Ponty, M. 1962. *Phenomenology of Perception*. Trans. C. Smith. London: Routledge (Original work published 1945).
- Morton, E. 1977. On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist*, 855–869.
- Naumann, A., Wechsung, I. & Hurtienne, J. 2010. Multimodal Interaction: A Suitable Strategy for Including Older Users? *Interacting with Computers* 22 (6), 465–474.
- Nielsen, J. 1994. *Usability Engineering*. San Francisco, CA: Morgan Kaufmann Publishers.
- Norman, D. 1988. *The psychology of everyday things*. New York, NY: Basic Books.
- Norman, D. 2004. *Emotional Design: Why we love (or hate) everyday things*. New York, NY: Basic Books.
- Noë, A. 2004. *Action in perception*. Cambridge, MA: The MIT Press.
- Ohala, J. 1984. An Ethological Perspective on Common Cross-Language Utilization of F_0 of Voice. *Phonetica* 41 (1), 1–16.
- Ohala, J. 2009. Signaling with the Eyebrows—Commentary on Huron, Dahl, and Johnson. *Empirical Musicology Review* 4, 101–102.
- Oulasvirta, A., Kurvinen, E. & Kankainen, T. 2003. Understanding contexts by being there: case studies in bodystorming. *Personal and Ubiquitous Computing* 7 (2), 125–134.
- Ozenc, F., Kim, M., Zimmerman, J., Oney, S. & Myers, B. 2010. How to support designers in getting hold of the immaterial material of software. In *Proceedings of the 28th international conference on Human factors in computing systems*. ACM, 2513–2522.
- Parviainen, J. 2006. *Meduusan liike: Mobiiliajan tiedonmuodostuksen filosofiaa (The Movement of the Medusa: Philosophy of Knowledge Construction in the Mobile Era)*. Helsinki: Gaudeamus.
- Patterson, R. 1982. *Guidelines for auditory warning systems on civil aircraft*. London: Civil Aviation Authority.
- Peirce, C. 1998. Sundry logical conceptions. In P. E. Project (Ed.) *The essential Peirce: selected philosophical writings vol. 2*. Bloomington, IN: Indiana University Press, 267–288. (Original work published 1903).

- Pirhonen, A., Tuuri, K., Mustonen, M. & Murphy, E. 2007. Beyond clicks and beeps: In pursuit of an effective sound design methodology. In Proceedings of the 2nd international conference on Haptic and audio interaction design. Berlin: Springer-Verlag, 133–144.
- Pirhonen, A. & Tuuri, K. 2009. Using Multiple, Role-Related Perspectives in the Design of Alarm Sounds for Safety Critical Context. In Proceedings of the 15th International Conference on Auditory Display. ICAD, 60–63.
- Ploog, D. 1992. The evolution of vocal communication. In H. Papoušek & U. Jürgens (Eds.) *Nonverbal vocal communication: Comparative and developmental approaches*. New York, NY: Cambridge University Press, 6–30.
- Rizzolatti, G. & Arbib, M. 1998. Language within our grasp. *Trends in neurosciences* 21 (5), 188–194.
- Rizzolatti, G. & Craighero, L. 2004. The Mirror-neuron System. *Annual Review of Neuroscience* 27, 169–192.
- Rochesso, D., Serafin, S., Behrendt, F., Bernardini, N., Bresin, R., Eckel, G., Franičević, K., Hermann, T., Pauletto, S., Susini, P. & others 2008. Sonic interaction design: sound, information and experience. In CHI '08 extended abstracts on Human factors in computing systems. New York, NY: ACM, 3969–3972.
- Sarter, N. 2006. Multimodal information presentation: Design guidance and research challenges. *International Journal of Industrial Ergonomics* 36 (5), 439–445.
- Schaeffer, P. 1966. *Traité des objets musicaux*. Paris: Éditions du Seuil.
- Scherer, K. & Bänziger, T. 2004. Emotional expression in prosody: a review and an agenda for future research. In *Proceedings of Speech Prosody 2004, International Conference*, 359–366.
- Scherer, K. 1994. Affect bursts. In S. van Goozen, N. van de Poll & J. Sergeant (Eds.) *Emotions: Essays on emotion theory*. Hillsdale, NJ: Lawrence Erlbaum Associates, 161–193.
- Scherer, K. 2003. Vocal communication of emotion: A review of research paradigms. *Speech communication* 40 (1-2), 227–256.
- Searle, J. 1983. *Intentionality, an essay in the philosophy of mind*. New York, NY: Cambridge University Press.
- Shannon, C. & Weaver, W. 1949. *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.
- Sonnenschein, D. 2001. *Sound design: the expressive power of music, voice, and sound effects in cinema*. Saline, MI: Michael Wiese Productions.

- Stefani, G. 1987. A theory of musical competence. *Semiotica* 66 (1), 7–22.
- Stern, D. 1985. *The interpersonal world of the infant*. New York, NY: Basic Books.
- Svanæs, D. 1997. Kinaesthetic thinking: The tacit dimension of interaction design. *Computers in Human Behavior* 13 (4), 443–463.
- Svanæs, D. 2000. *Understanding interactivity: Steps to a phenomenology of human-computer interaction*. Norges teknisk-naturvitenskapelige universitet. Ph. D. Thesis.
- Tagg, P. 1992. Towards a sign typology of music. *Secondo convegno europeo di analisi musicale*, 369–378.
- Tarasti, E. 2002. *Signs of music: a guide to musical semiotics*. Berlin: Walter de Gruyter.
- Tarasti, E. 2003. *Musiikin todellisuudet*. Helsinki: Helsinki University Press.
- Varela, F., Thompson, E. & Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: The MIT Press.
- Vogt, K. & Höldrich, R. 2010. A metaphoric sonification method – Towards the acoustic standard model of particle physics. In *Proceedings of the 16th International Conference on Auditory Display. ICAD*, 271–278.
- Välkkynen, P. 2007. *Physical selection in ubiquitous computing*. Espoo: VTT Technical Research Centre of Finland.
- Wittgenstein, L. 1953. *Philosophical Investigations*. Trans. G.E.M. Anscombe. Oxford, UK: Blackwell.
- Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions. *Speech Communication* 46 (3-4), 220–251.
- Zak, V. 1982. Asafjev's theory of intonation and the analysis of popular song. *Popular Music* 2, 91–111.

YHTEENVETO (FINNISH SUMMARY)

Puheakteissa kehollisesti välittyvä intentionaalisuus apuna ei-kielellisesti viestivien käyttöliittymä-äänien suunnittelussa

Tietotekniikasta on tullut pysyvä elementti arkipäivän ympäristöömme. Laitteiden muuttuessa pienemmiksi ja niiden sulautuessa elämäämme alati uudellisilla, perinteisestä tietokonekäsituksesta poikkeavilla tavoilla, vuorovaikutussuunnittelun painoarvo niiden kehittämisessä kasvaa. Ääni on hyvin luontainen vuorovaikutuksen keino, mutta sille ei toistaiseksi, puhekäyttöliittymiä lukuun ottamatta, ole tarjottu kovinkaan suurta roolia ihmisten ja koneiden välisessä vuorovaikutuksessa. Koska puhe ei aina ole laitteiden käyttöön sopivin äänellinen tapa viestiä, vuorovaikutussuunnittelun tarpeista katsottuna on tärkeää tutkia äänien välittämiä ei-kielellisiä merkityksiä. Tutkimukseni pyrkii lisäämään ymmärrystä äänien avulla tapahtuvasta ei-kielellisestä vuorovaikutuksesta keskittymällä ihmisten väliseen nonverbaaliseen viestintään.

Meidän on vaivatonta virittäytyä toistemme kehon eleisiin, ja sitä kautta aikomuksiin ja tunnetiloihin. Väitöskirjani tarkastelee, miten tätä toisten ymmärtämistä palvelevaa "kinesteettistä empatiaa" voi ottaa huomioon osana vuorovaikutussuunnittelua ja siihen sisältyvää käyttöliittymän äänisuunnittelua. Vuorovaikutusta korostavassa näkökulmassani kehollinen ja kinesteettinen kokemus hahmotetaan vastavuoroisesti sekä ilmaisun että ymmärryksen viitekehyksenä. Keskeisenä ajatuksena on, että äänitapahtumien intentionaalisuus tulee tilanneyhteydessään perimmiltään ymmärretyksi kehollisen viitekehyksen ja motoristen mielikuvien kautta. Näkökulma mukailee nykyisen musiikintutkimuksen piirissä esitettyä ajatusta, jonka mukaan äänitapahtumien motoris-mielikuvallinen hahmottaminen ja siihen pohjautuva empaattinen havaitseminen olisivat keskeisiä tekijöitä muun muassa musiikin välittämien tunnetilojen ja muiden ei-kielellisten merkitysten syntymisessä kuulijassa.

Tutkimus painottuu tarkastelemaan lyhyiden puheaktien intentionaalisuuden, kuten esimerkiksi viestinnällisten aikomusten ja affektien, eleellistä välittymistä ilmaisun akustisiin piirteisiin. Päähuomio on etenkin puhemelodian (ts. intonaation) ja viestinnällisten aikomusten välisissä yhteyksissä – ja lopulta myös siinä, kyetäänkö puheaktista irrotetuilla melodisilla piirteillä kommunikoimaan niihin heijastunutta intentionaalisuutta. Epistemologisena oletuksena on, että puheaktien melodiset rakenteet ovat kytköksissä keholliseen liikekokemukseen, joka on itsessään merkityksellinen nk. *intonoidun tietämyksen* kautta. Kysymyksessä on keholliseen toimintaan sidotun (enaktiivisen) tietämyksen tyyppi, joka perustuu ennen kaikkea sosiaalisen vuorovaikutuksen vokaaliseen kokemushistoriaan. Selvitykseni ottaa huomioon kokonaisvaltaisesti niin affektiiviset kuin funktionaalisetkin ominaispiirteet äänivuorovaikutukseen liittyvissä kinesteettisissä merkityksissä.

Väitöskirjatutkimuksen empiirinen osa tarkastelee puhemelodian piirteitä ja niiden toimivuutta erityisesti neljän viestinnällisen funktion osalta, jotka liit-

tyvät fyysisen harjoittelusuorituksen säätelyyn – esimerkiksi rannetietokoneen antaman yksinkertaisen äänipalautteen avulla. Empiiriset tulokset tukevat oletusta puhemelodian muotojen ja viestinnällisten aikomusten (eli funktioiden) välisistä yhteyksistä. Tulokset osoittavat, että käyttöliittymä-äännet, jotka perustuvat tiettyä aikomusta kuvastaville puhemelodian muodoille, kommunikoiivat tämän aikomuksen mukaisia merkityksiä onnistuneesti. Näiden tulosten vahvistamana väitöskirja tarjoaa menetelmän vokaalieleisiin perustuvien ei-kielellisten käyttöliittymä-äänien suunnitteluun. Tulokset myös osoittavat, että esitettyä suunnittelumenetelmää voi hyödyntää yli aistipiirien, esimerkiksi tuntoaistiin perustuvien käyttöliittymäviestien suunnittelussa. Lisäksi väitöskirjan hahmottelemat yleiset suunnitteluperiaatteet ovat hyödynnettävissä laajasti vuorovaikutussuunnitteluun ja käyttöliittymien kehittämiseen.

ORIGINAL PAPERS

PI

SAME SOUND – DIFFERENT MEANINGS: A NOVEL SCHEME FOR MODES OF LISTENING

by

Kai Tuuri, Manne-Sakari Mustonen & Antti Pirhonen 2007

In Proceedings of Audio Mostly 2007, 2nd Conference on Interaction with
Sound, Fraunhofer IDMT, Ilmenau, Germany, pp. 13–18

Same sound – Different meanings: A Novel Scheme for Modes of Listening

Kai Tuuri, Manne-Sakari Mustonen, Antti Pirhonen
Department of Computer Science and Information Systems,
P.o. Box 35, FI-40014 University of Jyväskylä, Finland
{krtuuri, msmuston, pianta}@cc.jyu.fi

Abstract. This paper is grounded on the *multimodal listening* hypothesis, which suggests that listening is a multi-focused process based on multiple distinct, environmentally shaped activating systems and listening strategies. Different modes of listening can operate concurrently complementing each other with different perceptual perspectives. In sound design, the potential of this hypothesis lies in its ability to account for the heterogeneity of different, even contradictory levels of interpretations, meanings and emotions evoked by the same listening experience. In this paper the theoretical basis of listening modes is further analysed and reflected upon in the context of sound design. We propose a comprehensive scheme of eight modes of listening (reflexive, connotative, causal, empathetic, functional, semantic, critical, and reduced) accompanied by examples of their significance in sound design for user-interfaces.

1. Introduction

In contrast to hearing, listening is an active process that provides a means to pick out information for our needs from the auditory environment. It is usually associated with voluntary attention and focusing on something. At present, a reasonable amount of studies exist concerning the area of auditory perception [see 1]. However, psychoacoustic models in which perception is built up from low-level perceptual atoms are inadequate for understanding creation of meanings. In most cases we do not perceive sounds as abstract qualities; rather, we denote sound sources and events taking place in a particular environment (e.g. dog barking) or we concentrate on some other level of information. Apart from a few examples in previous literature, modes of listening have received surprisingly little attention.

Listening is highly multimodal activity in nature. Multimodality of listening means that there are several distinct strategies to listen. It is a distinction of listening strategies and perceived experiences, not a distinction between sounds - although some sounds encourage the use of certain modes more strongly than others. Each mode of listening considers its own source of information in the auditory stimuli either with or without its context. The same sound can essentially be listened to with different kinds of attention and with different outcomes. Despite their separate nature, modes can and often do operate concurrently complementing and influencing each other.

How many ways can we listen to the same single sound? The process of listening to speech provides an example of a variety of perceived meanings. In addition to the conventional (linguistic) meaning of speech we can focus on listening to a speaker as a sound source (e.g. gender, age, dialect, emotional state). Or we can concentrate on the interactive nature of speech in a conversational context (e.g. getting attention, approving, encouraging). We can also pay attention to the qualities of speaker's voice (e.g. timbre, melodic contour, rhythm). Such meanings can even be perceived as contradictory e.g. when non-verbal cues in speech do not match with the verbal content.

The exploration of the multimodality of listening experiences promotes a better understanding of how meanings can be conveyed in effective sound design. This is our primary motivation to study this subject. In the following text previous accounts of the modes of listening are reviewed and then a revised account is proposed and detailed with examples.

2. Previous accounts of listening modes

2.1. Everyday listening

Our everyday listening is not focused on sounds. Instead, we usually hear sound sources: actions and events that cause sounds. We hear footsteps on a sidewalk, a car passing by, breaking of glass etc. We might also try to figure out how far and in what circumstances these events happen as we use listening to outline our environment to support our actions. This source-orientated mode of listening seems to be so effortless that we are not conscious of it. Listening studies by Vanderveer, Ballas and Gaver [2] show evidence that subjects indeed tend to describe a sound by its source or an event that caused it. In the case of ambiguous sounds, confusion and misidentifications are argued to be based on similarities in the mechanical structures of source events (such as "hammering" and "walking"). From the perspective of ecological (gibsonian) perception such confusions relate to shared properties in the affordance structures of various sound events. [3] In such cases, additional contextual information is required to confirm the denotation of the sound source.

The automatic nature of source-orientated listening and the phenomenon of ambiguity is frequently exploited in *Foley-tradition* of sound design [4] by framing non-authentic but believable (i.e. affordable) sounds persuasively into a suitable narrative context. As a part of the craftsmanship of a sound designer is required to identify essential components of sound which can convey a desired narrative effect (e.g. denotation of an event in a fictional environment). The freedom to use non-authentic sound sources gives a designer a much wider range of possibilities to suitably enhance an audience's experience. In contemporary audiovisual narration, even source-visualised on-screen sounds are often produced or treated artificially.

2.2. Reduced and musical listening

Possibly the earliest explicit mention of the modes of listening in previous literature can be found in the work of Pierre Schaeffer [5]. He proposes a distinguished *reduced* mode of listening by which we intentionally divorce the phenomena of sound from any functions as a medium of signification. An objective listening perspective was created to manage and handle sounds as abstract and fixed objects (*objet sonore*) for composition purposes of the *musique concrète* tradition. In

Schaeffer's phenomenology, reduced listening is a mode where the sound (and its qualities) is perceived *per se* resisting any claims about the exterior world. Such a totally abstract and "meaningless" perception of sound is of course a purely theoretical concept, and Schaeffer's work has received criticism as such [6].

Gaver's distinction of *musical listening* as opposed to everyday listening [2] shares essential similarities with the definition of reduced listening. Schaeffer's thoughts highlighted the fact that sounds from the everyday world can be used and listened to musically. The mode of musical listening is not intended to be restricted to sounds defined as music. Gaver thus recognises the "cross-referential" nature of different modes; it is possible to listen everyday auditory environments as music (e.g. attending to pitch contours and rhythm) and conversely possible to listen to musical performances in terms of causality and sound sources (e.g. separating different instruments or stems). Therefore as a term, musical listening can be misleading because music can be listened to in various modes – not just as abstract structures. Reduced listening can be referred to as musical listening only when listening process is concerned with musically determined qualities and structures.

However, the reduced mode of listening was later applied to film sound design by Chion [7]. Here the listening experience was objectified, by voluntarily resisting the natural denotation of a sound source or its meaning. By concentrating on the sound itself, sound designers could "open their ears" to the abstract qualities of sound. In this way, more creative or effective ways to utilise sound in narrative or artistic context became possible. The idea of reduced listening stresses the (analytic) perspective in which it is more important to understand what the sound *sounds* like than how it has been produced (see e.g. Foley tradition). Unlike Schaeffer, Chion puts forward the idea that reduced listening is more of a tool for analytic discovery of sound beyond its evident denotative meaning.

Both Schaeffer's and Chion's views are concerned with *acousmatic* situations [5,6,7] where the actual cause of sound is hidden from listener. This is indeed the case with mechanically reproduced and transmitted sounds e.g. telephone, radio and recordings. Acousmatic sounds thus permit more freedom of imagination for a listener to form a sound-only based perception and allow sounds to be composed artificially often by combining a variety of natural or non-natural sources. Acousmatic situations however do not automatically encourage reduced listening. Chion suggests that they can even intensify the motivation for everyday listening (i.e. causal listening, see 2.3.) when the visual support is removed [7].

2.3. Three listening modes of Chion

In his book *Audio-Vision*, Michel Chion was the first to introduce a more comprehensive scheme for modes of listening [7]. It consists of *causal*, *semantic* and *reduced* modes (see Figure 1). Two of them, causal listening (i.e. everyday listening) and reduced listening, were discussed above. The third mode, semantic listening, focuses on conventional meanings that the sound might represent by code, language or habit. As causal (everyday) mode of listening refers to ecologically-orientated evident¹ denotations, there was indeed a place for a mode that addresses socio-culturally shaped and learned (symbolic) codes. Spoken language is the most obvious example of cultural convention, but semantic listening can refer to anything which

creates a meaning for a sound that is not "literally"² there. Examples could range from exact rule-type codes (e.g. Morse code) and natural languages to more passively learned pragmatic habits (e.g. an applause after a performance) or conditioned associations (e.g. an ambulance siren). Many codes are so deep-rooted in the form of dispositions or habits, that they appear almost innate to us.

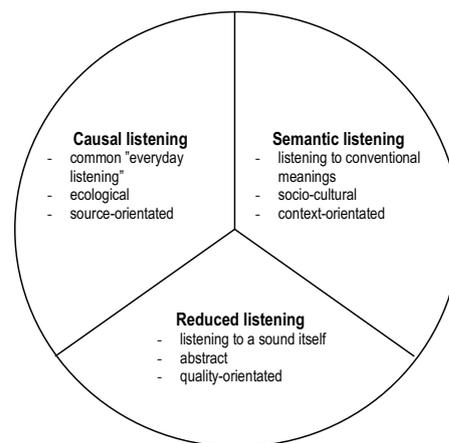


Figure 1. Three listening modes formed on the basis of Chion's classification [7]

Chion's scheme of listening modes is quite well known and it has proven its usefulness for sound designers. It is comprehensive and has broad categories that are easy to understand. However, it fails to capture more refined distinctions between clearly separate modes of listening, which will be covered later in this paper in the form of a revised model. Nevertheless, Chion's scheme forms the basis for our development of a more detailed scheme.

2.4. Activating systems

David Huron has suggested a six-component theory of auditory-evoked emotion. Despite his primary interests in emotions, his theory forms a relevant perspective to the modes of listening. Huron assumes a bio-cultural perspective on emotions as adaptive behaviours expressed within and shaped by an environment. Six *activating systems* are determined (see Table 1). These are evolved to serve specific functions and they all are capable of operating concurrently and evoking various emotional states. [9]

We assume that activating systems are not restricted to evoking emotions only, but in a similar way they can evoke other kind of meanings. Besides the denotative system which is linked to causal listening, the rest of the activating systems concerns novel and complementary perspectives to listening modes.

¹ Evident meanings, as defined by C.S. Peirce, are referred to as *iconic* and *indexical* relations of a sign and the object it refers to [8].

² By literal meanings we refer to ecologically inferred denotations, such as "meow" means a cat object.

Table 1. Brief summary of activating systems [9]

Reflexive system	Fast physiological responses
Denotative system	Processes which allow listener to identify sound sources
Connotative system	Processes that allow listener to infer various physical properties and passively learned associations (e.g. from temporal patterns)
Associative system	Arbitrary learned associations
Empathetic system	Allows the listener to perceive cues that signals someone's state of mind (an agent causing the sound)
Critical system	Reflective self-monitoring concerning the verification of perception and the appropriateness of one's responses

2.5. Functional perspective

Every sound which is intentionally used for some purpose has a specific function. When a listener answers a question “what is the purpose of that sound?” she defines the perceived function of the sound that is used in a particular context. We might become aware of the functions e.g. when the sound is perceived as a fire alarm, or when we feel music suitable for relaxing purposes, or when we perceive a sound effect as a transitional cue in audiovisual narration. The functional perspective of sound was explored in previous literature e.g. by Hermann and Ritter [10], and Jørgensen [11] who has examined the functional aspects of game audio. Roman Jakobson's model of communicative functions [12] is also related to this perspective.

Sound as a function can be seen as a pragmatic frame for meaning. The way that sound appears in a functional context affords a certain perspective to the process of interpreting the meaning of sound. The procedural chain of events, actions and causalities in a situation can give an indicative meaning even to a “meaningless” beep.

Although the sound itself can suggest the purpose of its use, perceiving the function of sound requires an awareness of context. The context concerns equally the situational factors as well as the listener's past experience on similar functionalities. The perception of a function is often related to a certain framework of common habits (e.g. habits of non-verbal interaction, conversation, musical performance, audiovisual narration or different user interfaces). The function of sound is particularly important in practices of interactive communication.

Functional semantics of sound can be seen as a distinct level of meaning which indicates and/or promotes a functional purpose of sound. It is indicated by the situational context but can also be indicated by the sound. For an example; let's consider the spoken word “dad”, whose “pure” verbal meaning of word is carried by the sound. If we shift our attention to the source of sound we can tell that it is a voice of a child. As we concentrate more on the voice and the way the word was spoken (prosodic qualities of voice), we can guess that a child is demanding attention from her dad. She is definitely not just mentioning the word “dad”. The attention-demanding function of this utterance can even be perceived without understanding the verbal content.

3. Hierarchical account of listening modes

Previous accounts of listening modes have been incoherent and limited in their scope. Therefore, we have formulated a new scheme. We have applied various relevant perspectives in order to form a more detailed and comprehensive outline of the listening modes. It is intended that this new outline will be utilised by audio designers and tested by audio researchers.

The basis of our pursuit of new categories is Chion's determination of three modes. One of the obvious shortcomings of the original scheme was its inability to consider connotations of sound. For the sake of interactive communication, we believe that perspective of functions of sounds deserves attention. Furthermore, activating systems introduced by Huron [9] offers relevant perspectives for developing a new scheme.

We propose a novel, *hierarchical* scheme of modes of listening (see Table 2) which consists of two pre-conscious modes (*reflexive* and *connotative*), two source-orientated modes (*causal* and *empathetic*), three context-orientated modes (*functional*, *semantic* and *critical*) and a reduced mode. The order of modes implicates their level of cognitive abstraction from low to high.

3.1. Reflexive mode of listening

In reflexive listening the focus is on automatic and fast audio evoked responses. Huron mentions various reflexes [9] including orientating response, startle response, defence reflex and reflexive responses that relates to expectations, habituation, sensory dissonance and attention. These responses resist conscious mediation, so in a strict sense this category cannot be considered as a pure mode of listening as it is impossible to control or focus on automatic reflexes themselves. In any case, reflexive responses represent clearly an important way by which meanings and emotions can be evoked by sounds.

3.2. Connotative mode of listening

In connotative listening, the focus is on early associations that the process of listening pre-consciously evokes. These associations are references made by similarity to past experiences of a listener, prior to any reasoned denotations. A French semiotician Roland Barthes [13] has concluded: “...denotation is not the first meaning, but pretends to be so...”. He implies that even though it appears that we can reason a denotative meaning instantly, it is only an illusion because we have already made a number of connotative associations. Denotation can then be defined as a “final”, more reasoned connotation. The important thing about connotation to realise is that besides denotative meaning, various connotative meanings can arise from the arsenal of excess associative “building material”.

At the most primitive level, connotative processes permits a listener to infer various physical properties of sound. These properties can indicate perceptual information concerning the sound source and environment: size, material, energy, proximity and excitation. Connotations can also be evoked from certain mechanical structures of source events (e.g. temporal patterns that evokes a “galloping-like” meaning) even if the sound source has nothing to do with what it connotes. [9] Besides the physical and ecological environment, associative cues can also relate to arbitrarily learned cultural experiences. For that reason we propose that the connotative mode of listening is not only linked to processes of connotative but also to processes of associative activating system.

Like the reflexive mode, the connotative mode of listening is involuntary in nature. Therefore focussing on connotations can be somewhat established by mentally exploring free associations and voluntarily resisting denotations.

Table 2. Summary of the revised scheme of listening modes with examples

Type:	Mode:	Questions:	Example:
<i>Pre-conscious modes:</i>	Reflexive	Did you notice any reflexive responses triggered by sound?	<i>Loud sound of train whistle (in a movie)</i> Startle response!! It alarms and grabs an attention.
	Connotative	Can you describe what kind of freely formed associations listening immediately evoked?	Big...strong...lots of power...close proximity ...screaming...air blowing...whistle...scream... ...old steam trains...western movies...
<i>Source-orientated modes:</i>	Causal	What could have caused the sound?	It's a train. <i>Critical second thought:</i> the sound comes from the TV.
	Empathetic	Does it feel that sound signals someone's state of mind or intentions?	Whistle sounds feels desolate and angry.
<i>Context-orientated modes:</i>	Functional	What was the purpose of the sound? What function does the context indicate?	The driver signals train's departure. <i>Critical second thought:</i> sound is used as transitional cue between scenes (just before a visual cut to railway station).
	Semantic	Does the sound seem to represent any symbolic/ conventional meanings?	The whistle represents pain... of a suffering man (by replacing his scream)
	Critical	Was the sound suitable for the situation? Did you understand it correctly?	Ah, no panic. That sound belongs to the movie. It was a cliché but quite effective.
<i>Quality-orientated mode:</i>	Reduced	Can you describe the properties of the sound itself as objectively as possible?	Sound is high-pitched and loud. A big contrast against quiet earlier scene.

3.3. Causal mode of listening

In causal listening the focus is on denotation of the source of sound and determination of an event that caused the perceived sound. This mode of listening is derived from the scheme of Chion (see 2.1 and 2.3.). This mode is also directly linked to denotative activating system. Causal listening is often referred to as a mode of common everyday listening.

3.4. Empathetic mode of listening

In empathetic listening the focus is on cues that could signal someone's state of mind. Empathetic mode of listening is thus directly linked to the empathetic activating system [9]. It is closely related to causal listening, in the sense of considering the possibility of a human or animal as a sound source or cause of sound. It is also related to connotative listening in the sense of potential auditory evoked associations (e.g. from intensity or certain rhythmic pattern) which can refer to emotional states, intentions or even communicative functions. For example listener can recognise a sad or nervous voice. On the other hand, listener can perceive e.g. a loud slamming (lots of energy) of a door as a possible expression of anger.

3.5. Functional mode of listening

In functional listening the focus is on the purpose of a sound in its context (see 2.5.). This mode considers the possibility that the sound is used for some specific function, which is pragmatically indicated by a sound in relation to the context. In the domain of non-speech auditory cues, a perceived function can be e.g. attention-demanding, alarming, orientating, approving, prohibiting, marking, prompting, giving a feedback or noticing.

3.6. Semantic mode of listening

In semantic listening the focus is on denoting any conventional meanings that a sound might represent. This is the second mode of listening which is derived from the scheme of Chion (see 2.3.). By semantic listening lower-level meanings are also reorganised for conventional reasoning to take cultural context (habits, codes) into account.

3.7. Critical mode of listening

In critical listening the focus is on the reflective judgement of auditory perception. As a mode of listening, it applies the idea of

Huron's [9] critical activating system. Critical listening concerns appropriateness or authenticity of sound in a given context. It also considers the appropriateness of one's responses. That includes judgements of possible misunderstanding, deception, false urgency or generally the need to be concerned with the sound. Additionally, at its highest level, critical judgements can be based on aesthetical dispositions.

3.8. Reduced mode of listening

In reduced listening the focus is on the sound itself and its qualities. This is a third mode of listening which is derived from the scheme of Chion (see 2.2. and 2.3.). The examination of sound phenomena itself requires that a listener is consciously resisting any denotations of a sound source or its meaning. This mode of listening is thus exceptionally voluntary and very likely requires high-level cognitive abstraction.

4. Observations from a case study

For a few empirical observations of modes of listening, we present some examples from a group panel discussions of our earlier case study (see Pirhonen et al. [14] which addresses the development of the design panel methodology). Although the panels were not originally conducted for the study of the listening modes, and during the panel sessions the listening modes were not considered, some examples can be pointed out. In the study, the designing of user interface sounds were studied in a series of group panel discussions. A group of panellists carried out design tasks in an iterative fashion; first for idea generation and then for evaluation of the designed outcome. The target of sound design was the user interface of a physical browsing application [15] used in a bicycle exhibition. Each panel session had different tasks and goals. The examples, relevant to this study, consider a warming-up task from the first session and sound evaluating tasks from the third panel session.

Before the actual design tasks, which was the purpose of our panel sessions, the moderator played some soundscape samples to the panellists and after that panellists discussed what they heard. These soundscape tasks were conducted to "open the ears" of the panellists.

The first soundscape sample was recorded in a bird watch tower during morning hours. The panellists described the

environment by the objects causing sounds, such as birds, wind humming in the trees, distant road humming (causal listening) and by connotations and by descriptions such as “trees sound as green as in our summer cottage in the relaxed summer morning, birds sound happy” (connotative and empathetic listening). The sounds of the distant motorway were considered as not suitable to the otherwise relaxed atmosphere (critical listening).

The other listening sample was little walkthrough from an elevator to a silent entrance hall, and from there into a noisy rush hour restaurant. Panellists described the entering from silent room to the noisy restaurant as a defence-reaction evoking event; the sound-mass of the restaurant was described as angry, scary and stressing rush-noise (reflexive, connotative and empathetic listening). The noisy environment was also described as unapproachable and unpleasant (critical listening). The moderator told the panellists before the sample that in the beginning there is a bit noise due to the recording technique. That was actually not true, the recording was clean. We observed that this additional task orientation of considering some inappropriate sounds affected the listening experience of the panellists. On this second task they were generally more critical and analytical listeners and described e.g. the humming and clicking sounds (of the elevator) and considered their appropriateness, whether the sound was original or an error of the recording.

After the ear-opening tasks, the group started the actual design tasks for that panel session. First the panellists familiarised themselves with the application by listening to a use scenario in the form of radio-play narration. Actual events-to-be-sonified (successful activating of a physical link, process of loading, loading ready) were clearly indicated in the story. Secondly, candidate sounds were played alone sequentially for each function and the most promising sounds were voted to go through to the next phase. The third phase of the panel session was that the sounds were played within the radio play, so that the panellists obtained a more holistic experience of the use situation, and heard the nominated UI-sounds connected to the procedure of the use scenario.

When the sounds were played alone, they were mostly judged by the criteria of subjective satisfaction and connotations like “this sound is not good, I do not like it, it sounds too lazy” etc. or “this is good, happy sound, I like it” or “it was good, snappy and attention grabbing sound, suits for the function”. Some sounds were voted to go through to the next phases, others were rejected. One example of the rejected sounds was rattling-like sound which was designed to indicate the loading process. When the sound was judged alone, it was considered as irritating clacking of teeth when feeling cold, and one panellist said that it reminded her of an annoying little boy with a ratchet (connotative, causal, empathetic and critical listening). The sound was rejected as too annoying and not suitable for the context.

In the next phase the panellists heard the radio play again, this time with the selected UI-sounds played within the story. Now, as the panellists were more immersed in the functional context and able to experience the whole situation procedurally, some of the earlier judgements changed. Some sounds that were judged as effective in the second phase were no longer considered suitable for the context, and some rejected sounds were asked to be elevated again. A more important factor than the subjective effect of the sounds was the match with sonic functions and events. During this last phase, the rejected rattling sound was now judged as the top-rated sound as an indicator for the loading process. Now the same sound was described as sound of small cogwheels, indicating the function of something happening, rather than the annoying ratchet or teeth chattering.

This example indicates how the functional context provides a crucial part of understanding the sound.

5. Listening modes and the current design paradigms of user interface sounds

In the research field of auditory cues in user interfaces (UI) there has been only little discussion concerning the multi-faceted meaning construction described in this paper. In 1994, the workshop report of CHI'94 discusses that “A more central concern was how to effectively convey information with non-speech auditory cues. Sounds can be interpreted at several levels...Current user interfaces have not yet addressed this in deeper expressive level in their use of sounds.” [16]. Despite the early recognition of the problem, it has not been widely considered within the research paradigm of UI-sounds.

In this paper we proposed a novel scheme for the modes of listening, which comprehensively binds together somewhat scattered discussions of earlier literature that concerns the issue. Listening modes play an important role as a tool for sound design. As we can listen to a sound with various forms of attention, the process of sound design must then concern auditory signs from various perceptual perspectives, in order to ensure consistent support for common communicative goals. The worst case design scenario would be that different listening modes evoke contradictory meanings (e.g. function implies an alarm sound, but the major chord invokes positive or happy connotations), or when the sound is experienced as annoying despite its perceived relevance in the context.

To demonstrate the relevance between sound design and perspectives of different listening modes, let's first examine two seemingly opposite design paradigms of user interface sounds: *earcons* and *auditory icons*. Earcons are defined “...as abstract, synthetic tones that can be used in structured combinations to create sound messages to represent parts of an interface.” [17]. The symbolic relationship between the sound and its meaning is seen beneficial as sounds do not have to correspond to what they represent [18]. Meanings are thus arbitrarily coded and therefore learning of specific codes is required prior to effective understanding. It seems that the philosophy behind current earcon design is related to information theory [19] with implicit assumption of the role of sound as a carrier of coded information. Current earcon design guidelines consider sound by emphasising psychoacoustic phenomenon on how sound may be masked or how sound stream can be segregated (judgements on timbre, register, rhythm, concurrent sounds etc.) [20]. These channel-orientated perspectives and considerations of channel-noise factors (e.g. masking) further emphasises the information theory based view of communication.

The paradigm of auditory icons, conversely, relies on iconicity and the ecological perspective of auditory perception [21, 3]. This essentially means that when listening we naturally pick up recognisable parts from the auditory stimuli. Relations between sound and its meaning are based on similarities with familiar aspects of our everyday environment. They can be denotations of sound sources or partial indicators that point to some mechanical properties of a sound causing event. In sound design, similarities can also be used in a metaphorical way. The most important difference to the earcon-paradigm is that the design of auditory icons is more focused on how the sound itself, by resemblances, motivates the meaning-creation. Just as in the traditional film sound design, meanings appear to be motivated by the sound.

We can conclude that the earcon-paradigm is concentrated mainly on two modal perspectives: semantic mode (extreme requirement of code) and reduced mode (sound is supposed to be heard as musical parameters). The design of alarm sounds

additionally concerns reflexive mode of listening. On the other hand the auditory icon paradigm is determined on perspectives of causal mode (source recognising), connotative mode (physical property indicators of an event) and in some sense also functional mode (meanings of sounds shares iconic similarities with the event it represents in application). We thus find that the distinction between earcons and auditory icons is not intended as a distinction between UI-sounds themselves. In fact, that categorisation seems to be more related to which modes of listening are adopted for the paradigm in question. In light of listening modes, earcons and auditory icons are to be considered as *design paradigms* – not as necessary distinct types of UI-sounds. The cross-related nature of listening modes allows for the consideration of different design paradigms in tandem.

Indeed, an optimally designed earcon can also utilise its expression with e.g. iconic and affective levels of meaning with cues to some familiar qualities or habits of the experienced world - even when an abstract form of expression is chosen.

6. Conclusions

We feel that our own main contribution of this paper is the systematic review and synthesis of listening modes, and within the proposed scheme the inclusion of an explicit functional mode of listening. We argue that the purpose of sound in a functional context is an important factor in *interaction design* within user interfaces. Firstly, every user interface design must address the role of a sound in interaction. Secondly, the perceived function of sound in a situational context represents itself as an important class of meanings. A user can get context-derived indications to suitably interpret even a “meaningless” beep, not to mention sounds that convey some additional semantic support for the appropriate interpretation of meaning. We can find that the two classic design paradigms discussed above (earcons & auditory icons) are deterministically more concerned with the semantic aspects of UI-sound element itself – not the aspects of how the sound is used in the functional context of UI. The complementary perspective we propose is more procedural in nature; in this approach it is more important *how* meaning is created in a given context than *what* the meaning is *per se*. Functional semantics of sound is based on tacit reasoning and pragmatically evoked semantics.

The general process of sound design for a user interface, at least implicitly, should begin with exploring the relevant communicative purposes of sound in UI-interaction. The outcome of that meta-design process will be a list of functions of sounds referring to various events and processes taking place when user tasks are performed. By analysing those functions within those scenarios and situations, a designer can find associative ideas for relevant functional semantics for the actual sound design.

The scheme for modes of listening, which is presented in this paper, is intended to open a discussion concerning the topic. Also this study is to be seen as compiling a review of various aspects concerning the complex scheme of meanings inferred from sound. The new perspective, the functional mode of listening, is the most prominent contribution from the perspective of audio interaction design and research. The observations from our case study support our assumptions. Nevertheless, more empirical evidence should be gathered to validate the appropriateness of modes of listening for user interface design. Sound design cannot afford to overlook the diversity of meanings and the affective responses that the sound evokes in the context of its use. As a conceptual model, the proposed scheme of modes of listening can guide the designer to find answers to that challenge.

Acknowledgements

We thank Professor Tuomas Eerola for his support and encouragement for our work. Many thanks to Emma Murphy for her comments and help in writing English.

This work is funded by Finnish Funding Agency for Technology and Innovation, and the following partners: Nokia Ltd., GE Healthcare Finland Ltd., Sunit Ltd., Suunto Ltd., and Tampere City Council.

References

- [1] Bregman, A. Auditory Scene Analysis. Cambridge, MA: MIT Press (1990)
- [2] Gaver, W. The SonicFinder: An Interface That Uses Auditory Icons. *Human-Computer Interaction* 4, 1 (1989), 67-94
- [3] Casey, M. Auditory Group Theory: with Applications to Statistical Basis Methods for Structured Audio, Ph.D. Thesis, MIT Media Lab (1998)
- [4] Mott, R. Sound Effects: radio, TV, and film. Boston, MA: Focal press (1990)
- [5] Schaeffer, P. *Traité des objets musicaux*. Paris: Editions du Seuil (1968)
- [6] Kane, B. L'Objet Sonore Maintenant: Pierre Schaeffer, sound objects and the phenomenological reduction. *Organised Sound* 12, 1 (2007), 15-24
- [7] Chion, M. *Audio-Vision: sound on screen*. New York, NY: Columbia University press (1994)
- [8] Peirce, C.S. What is a sign? In *Essential Peirce: Selected philosophical writings* vol. 2. Bloomington, IN: Indiana University Press (1998), 4-10
- [9] Huron, D. A six-component theory of auditory-evoked emotion. In *proceedings of ICMP7 (2002)*, 673-676
- [10] Hermann, T & Ritter, H. Sound and Meaning in Auditory Data Display. In *proceedings of the IEEE*, vol. 92, 4 (2004), 730-741
- [11] Jørgensen, K. On the Functional Aspects of Computer Game Audio. In *proceedings of Audio Mostly 2006*, 48-52
- [12] Jakobson, R. Closing Statements: Linguistics and Poetics. In Sebeok, T. A. (ed.) *Style In Language*. Cambridge, MA: MIT Press (1960), 350-377
- [13] Barthes, R. *S/Z*. New York: Hill & Wang (1974)
- [14] Pirhonen A, Tuuri, K., Mustonen, M-S. & Murphy, E. Beyond Clicks and Beeps: In Pursuit of an Effective Sound Design Methodology. In *proceedings of HAID2007* (in press)
- [15] Väikkynen, P. Hovering: Visualising RFID hyperlinks in a mobile phone. In *proceedings of MIRW 2006*, 27-29
- [16] Arons, B. & Mynatt, E. The future of speech and audio in the interface. In *SIGCHI Bulletin* 26, 4 (1994), 44-48
- [17] Brewster, S, Wright, P. & Edwards, A. A detailed investigation into the effectiveness of earcons. In Kramer, G. (ed.) *Auditory display*. Reading, MA: Addison-Wesley (1994), 471-498
- [18] Blattner, M, Papp, A. & Glinert, E. Sonic Enhancement of Two-Dimensional Graphic Displays. In Kramer, G. (ed.) *Auditory display*. Reading, MA: Addison-Wesley (1994), 447-470.
- [19] Shannon, C. & Weaver, W. *The Mathematical Theory of Communication*. Urbana, ILL: University of Illinois Press (1949)
- [20] Brewster, S., Wright, P & Edwards, A. Experimentally derived guidelines for the creation of earcons. In *Adjunct Proceedings of HCI'95*, 155-159
- [21] Gaver, W. Auditory Icons: Using sound in computer interfaces. *Human-Computer Interaction*, 2. (1986), 167-177

PII

**FORMULATING A REVISED TAXONOMY FOR MODES OF
LISTENING**

by

Kai Tuuri & Tuomas Eerola

Submitted for consideration in the Journal of New Music Research

Reproduced with kind permission of Taylor & Francis.

Formulating a Revised Taxonomy for Modes of Listening

Kai Tuuri

*Department of computer science and information systems, University of Jyväskylä,
Finland*

Tuomas Eerola

Department of Music, University of Jyväskylä, Finland

Address of correspondence:

Kai Tuuri

E-mail: krtuuri@jyu.fi

Address: Department of computer science and information systems

B. O. Box 35

FI-40014 University of Jyväskylä, Finland

Formulating a Revised Taxonomy for Modes of Listening

Listening to sounds or music is not a homogeneous act of grasping meanings by hearing. Yet it is often portrayed as such, especially when the intentional stance of a listener is overlooked. This paper distinguishes listening as the action-oriented intentional activity of making sense of the world. It is proposed that the multifaceted and heterogeneous nature of ‘understanding by listening’ can be outlined in terms of distinct modes of listening. Building upon previous accounts, a revised taxonomy of nine listening modes (reflexive, kinaesthetic, connotative, causal, empathetic, functional, semantic, reduced and critical listening) is proposed and illustrated by examples. Modes refer to different constituents of meaning creation in the process of listening. In the taxonomy, they are schematically arranged into three levels (experiential, denotative and reflective). The theoretical framework of this revised taxonomy utilises an embodied cognition paradigm. The experiential basis of meaning in listening is theoretically conceived of as emerging resonances between experiential patterns of sensations, structured patterns of recurrent sensorimotor experiences (action-sound couplings) and the projection of action-relevant mental images. The proposed taxonomy of listening modes is discussed in terms of its implications for perception and cognition research on sounds and music.

Keywords: listening modes; ecological perception; embodied cognition

1. Introduction

Listening is an active process, through which we gather information to meet our needs for interacting with the environment. One can draw a simple distinction between hearing and listening, seeing the former as more or less passive ‘receiving’¹ of a sound and the latter as an intentional and attentional creation of meanings on the basis of the sonic experience. Intuitively, we also acknowledge that this meaningful experience is dependent on the way the sound is involved in the situation and how we see its relevance to the context of interaction. In most cases we do not experience sounds as abstract qualities or musical features; rather, we denotate sound sources and events taking place in a particular environment. This already exemplifies two separate modes of listening, each referring to different ways of making sense of potentially the same sound. The central aim of this article is to deepen understanding about the

¹ The input-metaphor of ‘receiving sound’ is a commonly utilised concept for hearing. In line with the arguments of this article, hearing could be seen rather as ‘exposing oneself to the sonic experience’.

multifaceted nature of listening, and to formulate a coherent typology of different listening modes. As a basic approach to such a taxonomical work, it is crucial to acknowledge that the typology of listening modes applies only as a distinction between listening orientations and perceived experiences, not as a distinction between sounds – although some sounds may induce the activation of certain modes more strongly than others. On the basis of multidisciplinary literature, we will propose that each mode of listening is related to its own sources of information evoked by sound and its involvement in the context. In the course of evolution, our capacities for different ways of meaning-creation have presumably been shaped to serve different aspects of coping with the world.

The initial account of different listening modes was made decades ago by Pierre Schaeffer (1966). Despite this early notion, the interest towards the matter in the literature has been fairly low. With this article, we aim at demonstrating that this subject matter merits more attention. As a conceptual framework, taxonomies of listening modes have primarily been conceived as useful tools in the field of sound design (Gaver 1989; Chion 1990; Sonnenschein 2001; Tuuri, Mustonen & Pirhonen, 2007), since they promote an understanding of how meanings can be conveyed in effective design. Indeed, the potential of a coherent framework of listening modes lies in its ability to account for the heterogeneity of different, even contradictory, levels of interpretations, emotions and other meaningful experiences on the basis of the same physical sound. There is also no reason why such a heterogeneity in meaning-creation would not also apply to involvement with music. Clarke (2005) points out that in Western tradition we tend to conceive of music as an abstract, autonomous domain, which has its own structures – or even language – of musical meanings. Such a focus is also embedded in the research on music perception, as the most studied perceptual

aspects relate to the processing of musical structures in an unimodal fashion (e.g., Deutsch, 1999; see also Peretz & Zatorre, 2003). From the perspective of listening modes, a strictly autonomous approach concerns mostly the convention and structure-oriented ways of listening that are specific to the Western music tradition – effectively leaving out other modes of meaningful sonic involvement with music that operate outside the autonomous domain.

A comprehensive and general account of different listening modes would clarify the existing research on music cognition and perception, by providing a tangible and parsimonious framework to interpret the findings related to, for instance, the influence of background music on mood (e.g., North, Shilcock & Hargreaves, 2003), the effects of attention on evaluating harmonic progressions (e.g., Loui & Wessel, 2007), the apparent similarities between heard and imagined evaluation of sounds (e.g., Halpern et al., 2004), the differences between felt and perceived emotions induced and expressed by music (Schubert, 2007), pre-attentive processing of sounds (e.g., Tervaniemi et al., 1997), and the puzzling disparities in musical enjoyment as a function of prior information (e.g., Margulis, 2010). Within the different areas of sound-related research, there might also be implicit presuppositions about listening which are taken for granted. Methodologies of listening studies might not always appropriately take the listener's intentional stance or the context of listening into account. For example, it has been found that the overwhelming majority of music listening is done as a background to other activities (Sloboda & O'Neill, 2001). For the reasons presented above, we feel that a comprehensive account of listening modes is necessary.

The theoretical approach of this article is strongly based on an ecological approach to perception (Gibson, 1979) under the umbrella of *embodied cognition*

(Varela, Thompson & Rosch, 1991). At present, a reasonable amount of studies exist concerning the psychoacoustics of perception (see, Bregman, 1990). However, the traditional information processing (IP) model in which perception is built up from low-level perceptual atoms seems inadequate for understanding different types of listening and the creation of meanings. The wide 'semantic gap' between elementary perception and perceptual and emotional content has proven to be especially problematic (Leman, 2008b). To better explain perceptual experiences, we rely on an alternative epistemological paradigm with an experience-centred perspective on perception. Contrary to the IP-model, the embodied view sees cognition as the enactment of the world and mind (Varela et al. 1991). According to this paradigm, rather than perceiving structural features of sound first, we are naturally sensitive to action-relevant values of the environment. Our ecological knowledge of action-sound couplings provides a perceptual basis for various action-relevant meanings, such as sound sources, gestural signatures in actions (manifesting affect), and conditioned/learned associations. Perception and action are thus intertwined together since (1) the meanings of an environment are structured through embodied subject-environment interactions and (2) meaning-structures, such as action-sound couplings, are organised in terms of directly perceivable action-relevant values, i.e., *affordances* (Gibson, 1979; Varela et al., 1991). In other words, our cognition, including sound-meaning structures, is integrally coupled with an ongoing world and embodied experiences of interaction. In addition to the 'natural world', these adaptive couplings extend also to interactions with the social and cultural environment (Gibson, 1966; Clarke, 2005; Leman, 2008a). Clarke (2005) and Leman (2008a) have utilised this kind of theoretical approach to explore the ecology of musical meaning. Our interest, however, is specifically in formulating a theoretically sound and generally applicable

taxonomy of modes of listening, which is based on ecological principles but which is also in line with earlier literature on the subject. Specifically, our account will be a revision of a scheme for listening modes, proposed earlier by Tuuri et al. (2007).

2. Previous accounts of listening taxonomy

You don't get a buzzing-noise like that, just buzzing and buzzing, without its meaning something. If there's a buzzing noise, somebody's making a buzzing-noise, and the only reason for making a buzzing noise that *I* know of is because you're a bee.

– A.A. Milne, Winnie-the-Pooh

2.1 The “big three” of listening modes

Previous discussion on the modes of listening has mostly concentrated on three main orientations of making sense of sounds. The following discussion examines this scheme of three listening modes.

2.1.1 Causal listening

Our everyday listening is not focused on sounds as such. Instead, we usually hear sound sources: in particular, actions and events that cause sounds. We hear a dog barking, footsteps on a sidewalk, a car passing by, glass breaking, and so on. We might also try to figure out how far and in what circumstances these events happen as we listen to meet our needs to cope with the environment. In previous literature, Schaeffer (1966) called such type of listening *écouter*, Gaver (1989) called it everyday listening, and Chion (1990) coined the term causal listening for it. This source-orientated mode of listening seems to be so effortless that we are not usually aware of utilising it. Listening studies by Vanderveer (1979), Gaver (1988), Ballas (1993) and Plazak and Huron (in press) show evidence that subjects indeed tend to describe a sound by its source or an event that caused it. In the case of ambiguous sounds, confusion and misidentifications are argued to be based on similarities in the

mechanical structures of source events (such as ‘hammering’ and ‘walking’). From the perspective of ecological perception (Gibson, 1979), such confusions relate to shared properties (perceptual invariants) in the affordance structures of different sound events. In such cases, additional contextual information is required to confirm the sound source. Music can also utilise structures that encourage the use of causal listening (Tarasti, 2002; Chion, 1993; Johnson, 2007). Hence, it is possible to perceive musical cues in terms of everyday events and natural phenomena, such as ‘the swell of the sea’. It is also possible to attend to the actual sound-producing actions of the musical performance.

Our heavy reliance on causal listening and the phenomenon of perceptual ambiguity is frequently exploited in the Foley-tradition of sound design (Mott, 1990). When the context makes us expect a sound, it seems that the designer can use just about any type of sound as long as it provides enough believable acoustic properties for expected or contextually plausible events (i.e. sounds with appropriate affordance structures). Indeed, *acousmatic* situations (Schaeffer, 1966; Chion, 1990), where the actual cause of sound remains hidden to the listener, allow the designer (or composer) a wide range of artistic possibilities. Acousmatic sounds thus permit more freedom of imagination for a listener and allow sounds to be composed artificially often combining a variety of natural and synthetic sources.

2.1.2 Semantic listening

Semantic listening (Chion, 1990) is a mode of perceiving sound as signs that stand for something due to socio-culturally shaped and learned codes. Schaeffer (1966) called this mode *comprendre*. Listening to spoken language is the most obvious example involving a cultural code, but semantic listening can refer to any representational socio-cultural constructions and accustomations. Examples could range from exact

special codes (e.g. Morse code) and natural languages to more passively learned pragmatic habits (e.g. a round of applause after a performance) or conditioned associations (e.g. an ambulance siren). Arguably, many codes are so deep-rooted in the form of habits that they appear almost innate to us. Some codes, however, require expert knowledge of a specific cultural field. For example, a music listener may be able to recognise structural elements in music that represent the conventions or rules of the cultural field. In general, the meanings of semantic listening can be described as affordances of cultural behaviour (i.e., habituation) which depend on the capacities and the needs of the perceiver, as well as the opportunities the environment provides (Clarke, 2005).

2.1.3 *Reduced listening*

Schaeffer (1966) proposed *l'écoute réduite*, a reduced way of listening, by which we intentionally divorce the phenomena of sound from any meanings in the world. Reduced listening resists the search for any logic, structure or objectivity beyond the sound itself. Schaeffer named two subcategories of reduced listening: *Ouïr* refers to passive and inattentive listening, appreciating the experience on the most elementary level of perception. *Entendre*, on the other hand, refers to listening that attends to particular qualities of sounds. The perspective of reduced listening was formulated for the needs of managing and handling sounds as objects (*objet sonore*) to be used in compositions of the *musique concrète* tradition.

Gaver's (1989) distinction of musical listening from everyday listening shares essential similarities with the definition of reduced listening. Indeed, Schaeffer's thoughts highlighted the fact that sounds of the everyday world can be used and listened to as abstract musical sounds. Gaver also recognises the possibility of listening to everyday sounds as music (e.g. attending to pitch contours and rhythm).

But the term ‘musical listening’ can be misleading as music can be listened to in various modes – not just as abstract structures. Moreover, reduced listening can be taken as musical listening only when listening attends to autonomous qualities and structures of music.

Chion (1990) puts forward the idea that reduced listening is first of all a method for a reflective discovery of sound beyond its ‘evident’ meanings. By concentrating on the sound itself, sound designers could ‘open their ears’ to become acquainted with the sound and its qualities. With an objective perspective to sound, more creative or effective ways to utilise sound in a narrative or artistic context become possible.

2.2 Activating systems

David Huron (2002a) has suggested a six-component theory of auditory-evoked emotion. He takes a bio-cultural perspective on emotions as adaptive behaviours expressed within and shaped by an environment. Despite his primary interests in emotions, the theory outlines a more general perspective to auditory-evoked meaning-making on the basis of six distinct source components that he calls activating systems. These are evolved to serve specific functions and they are all capable of operating concurrently and evoking various emotional states, along with other meaningful experiences. The activating systems are:

- *Reflexive system*: Fast and automatic physiological responses.
- *Denotative system*: Processes which allow the listener to identify sound sources.
- *Connotative system*: Processes that allow the listener to infer various physical properties and passively learned associations.
- *Associative system*: Arbitrary learned or conditioned associations.

- *Empathetic system*: Allows the listener to perceive cues that signals someone's state of mind.
- *Critical system*: Reflective self-monitoring concerning the verification of perception and the appropriateness of one's responses.

Besides the denotative system, which seems to map neatly to causal listening, the rest of the activating systems concern novel and complementary perspectives on the taxonomy of listening modes. A recent attempt to collapse the 'big three' account and Huron's activating systems is described next.

2.3 Hierarchical account of listening modes

Tuuri et al. (2007) have proposed the following taxonomy in which listening modes operate within a hierarchical scheme. It takes the earlier three-mode account as a starting point and expands it with five additional modes. Four of these new modes are built upon Huron's activating systems. The hierarchical scheme consists of two *pre-attentive modes* (reflexive and connotative), two *source-orientated modes* (causal and empathetic), three *context-orientated modes* (functional, semantic and critical) and finally a *quality-orientated*, reduced mode of listening. The order of the modes implies their level of cognitive abstraction from low to high. However, levels should be highly interactive. Emphasis is not on bottom-up processes. For example, contextual anticipations among other types of higher level thinking are expected to modulate or activate lower levels.

2.3.1 Pre-attentive modes

The *reflexive* mode of listening maps to the reflexive activating system, as it refers to the fast and most innate audio-evoked physical responses as the most primordial source of affective responses and meaning-making. These include orientating

response, startle response and defence reflex, to name but a few. These responses resist conscious mediation, so in a strict sense this category cannot be considered a 'pure' mode of listening as it is impossible to control or focus on automatic reflexes themselves.

Connotative listening focuses on early associations, mental images and feelings pre-attentively evoked in the listening experience. These associations are based on past experiences of a listener, and they occur prior to any reasoned denotations. Roland Barthes (1974, p. 9) has concluded: "...denotation is not the first meaning, but pretends to be so...", meaning that even though it appears that we can perceive a denotative meaning instantly, it is only an illusion because we have already made a number of immanent connotations. Therefore, besides a denotative meaning, various connotative activations have already vigorously taken place. These activations essentially modulate the experiential background of any inferred denotations but provide a residual effect as well. Conscious focus on connotations may be established to some degree by voluntarily resisting denotations (via reduced listening) and simultaneously reflecting on the listening experience and its mental imagery.

At the most primitive level, connotative processes should concern activation contours of vitality affects (Stern, 1985), as an affective coloration that sound may evoke. According to Huron (2002a), connotations provide cues about the sound source and the physical environment, indicating, e.g., movement, size, materials, energy, proximity and excitation. The ecological approach (Gibson, 1979) suggests that such associations arise due to perceptual invariants, i.e., perceived action-relevant similarities between certain patterns of sensation and the listener's experiential background. Connotative associations can arise even if the sound source has nothing to do with what it connotes (e.g., when rhythmic patterns in the sound of a cellular

phone interfering with audio speakers evoke a ‘galloping-like’ meaning). Besides environmental associations, associative cues can also refer to socio-culturally habituated experiences. In music, for example, certain use of instrumentation (such as the harpsichord) can be associated with a specific historic period (such as Baroque), even if the music in itself does not clearly resemble the style of that period of time.² For that reason it is proposed that the connotative mode of listening is not only mapped to the connotative activating system but it is also at least partly mapped to the associative activating system.

2.3.2. *Source-oriented modes*

In line with the earlier discussion, the *causal* mode of listening is about perceiving and denoting causes and source-events of sound. It maps directly with Huron’s denotative activating system.

Empathetic listening is about perceiving and denoting affective states, hence it focuses on cues that could signal someone’s state of mind (i.e., intentionality). This mode maps directly to the empathetic activating system. As Leman (2008a) points out, there is a bias to perceive sound as being intentional, especially if it suggests biologically relevant movement patterns. The perception of another person’s intentionality is thus based on bodily realised affordances of the movement (gestural signatures), and inferred empathetically in terms of our own ontology of emotions and intentions. This view is supported by recent research on the hypothesised role of mirror-neurons in interpersonal attunement and empathy (Gallese, Eagle & Migone, 2007; Iacoboni, 2009). Empathetic listening is closely related to causal and

² In his book, David Sonnenschein (2001) has proposed a referential mode of listening. This type of listening attends to the emotional and dramatic meanings that sound is able to evoke besides the implication of its source. According to this description and the examples in the book, referential listening is quite well covered within the mode of connotative listening.

connotative listening, in the sense of attuning to the source of the sound while seeking gesturally relevant attributes of affect. For example, a listener can recognise a sad or a nervous voice. On the other hand, a listener can perceive, e.g., the loud slamming (lots of energy) of a door as a possible expression of anger (via the gestural image of a hard slam).

2.3.3 *Context-oriented modes*

Functional listening is about perceiving and denoting functional purposes of sounds. Often these purposes refer to a certain communicative use of a perceived sound. We become aware of functions, for example, when a sound is perceived as a fire alarm, or when we feel music suitable for relaxing purposes, or when we perceive a sound effect as a transitional cue in an audiovisual narration.

The functional mode of listening attends more to the context of a sound, which ultimately provides the affordances of the sound's usage or purpose. For example, even if 'meaningless' beeps were used as sound events in a user interface of an interactive application, they still suggest themselves as being inherently purposive. Perceived purposes arise within an interaction as intentional couplings between the actions of a user and the observable events in the context of an application (see, Dourish, 2001). The attribution of contextually implied function may also change the attitude towards the sound, for example, when an 'annoying ratchet' sound, moved into the context of a user interface, transforms into a likable 'cogwheel' sound which illustrates the ongoing process (Tuuri et al., 2007). On the other hand, sound itself can also suggest its purpose, i.e., structural patterns in a sound can provide cues for it. From this bottom-up perspective, functional listening operates closely with empathetic listening. For example, according to Searle (1979), the context-situated intention to utter a speech act (e.g., for asking or prohibiting) is realised in the act of

vocalisation as a certain *illocutionary force*. Metaphorically, illocutionary force can be conceived in terms of image schematic force gestalts (Johnson, 1987). It is quite safe to say that action-relevant values relating to these gestalts are also projected into the acoustic structures of a sound (as prosodic patterns). And these ‘affective signatures’ of communicative intention *in* utterance can be empathetically perceived. Let us consider the spoken word ‘dad’. If we shift our focus to the voice and the intonation with which the word was uttered, we can see that a child is demanding attention from her dad. She is definitely not just mentioning the word ‘dad’. It has been found that such communicative functions of an utterance can be perceived even when all verbal content is filtered out (Fernald, 1989).

In line with the earlier discussion, the *semantic* mode of listening is about perceiving and denoting any arbitrary meanings that a sound might represent. This mode of listening specifically attends to the context of socio-cultural conventions and constructions. Huron’s associative activating system functions in strong relation to this mode.

Critical listening is about reflective judgement of auditory perception. As a mode of listening, it is mapped to a critical activating system. Critical listening concerns appropriateness or authenticity of a sound in a given context. It also considers the appropriateness of one’s responses, which includes judgements of possible misunderstanding, deception, false urgency or generally the need to be concerned with the sound. Additionally, at its highest level, critical judgments can be based on aesthetical dispositions.

2.3.4 *Quality-oriented mode*

As discussed earlier, *reduced* listening is about focusing on sound itself and its qualities. The examination of sound phenomena on their own terms requires that a

listener is consciously resisting any denotations of a sound source or its meaning. This mode of listening is thus exceptionally premeditated and very likely involves high-level self-conscious cognitive reflection.

2.4 Listening attention

Barry Truax (2001) has proposed a listening taxonomy which concerns three levels of attention in listening (listening-in-search, listening-in-readiness and background listening). He points out that listening attention is active, dynamic and a constantly shifting phenomenon in which we have the ability to extract certain sounds for foreground listening as well as being able to put other sounds into the background. The scope of attention can also shift from focused listening to a more general ‘scan’ of the environment. Truax argues against the assumption that only foreground listening is attentive. That assumption does not take into an account the subtler processes involved even in seemingly non-concentrative background listening, which are demonstrated in the classic ‘cocktail party effect’ (see Bregman, 1990).

Truax calls the top level of listening attention *listening-in-search*. It involves a conscious search of the environment for cues. The general focus on details of an environment and the ability to focus on one sound to the exclusion of others are important elements in this level of listening. *Listening-in-readiness* is an intermediate kind of listening, in which the listener’s attention is in readiness for any significant information, but the attentive focus is directed elsewhere. And at the bottom level of listening attention, *background listening* refers to perceptual processes where sound usually remains in the background of our attention. Such listening occurs when we are not attentively oriented towards particular sounds, but we might nevertheless have an ability to recall those sounds at a later time. Truax draws a parallel between background listening and keynote sound – a concept commonly used in soundscape

terminology (Schafer, 1993). Within audio-visual media such as film, background listening is repeatedly exploited in narrative music in order to, for example, manipulate the moods of film scenes. Narrative music is quite rarely intended to be attentively listened to in itself.

2.5 Listening styles

In addition to listening modes and listening attention, taxonomical focus may also be put on individual differences between listeners. Styles of listening refer to individual, preferred listening behaviours. Arguably, listening styles are closely related to other cognitive styles, such as thinking styles (Sternberg, 1997). The foundation of listening styles therefore likely comprises general factors of behavioural dispositions; personality, thinking styles, socio-cultural *habitus* and the abilities of an individual. For example, Watson et al. (1995) have proposed a four-style taxonomy for listening in conversational context. Listening styles have also been formulated for music listening. For example Kreutz et al. (2008) have distinguished listeners in terms of cognitive styles (as emphatizers and systemizers). Also, Huron (2002b) has listed 21 different music-related listening styles and strategies (although he also calls them listening modes). With these rather loosely organised categories, he interestingly describes different cases of involvement in listening. Besides merely describing listening preferences, Huron's practical case-examples seem to also comprise the utilisation of different listening modes and the levels of attention in listening.

2.6 Summary of listening taxonomies

By way of summary, previous accounts of Schaeffer, Gaver and Chion have made separate but mutually consistent attempts at dividing the ways of attributing meanings in listening into separate processes, broadly divisible into causal, semantic and

reduced listening. A recent refinement of listening modes by Tuuri et al. suggests multiple additional categories of listening modes, also loosely incorporating Huron's idea of activating systems. In this scheme, the modes of listening are organised into a hierarchical structure, corresponding to the different levels of cognitive operations.

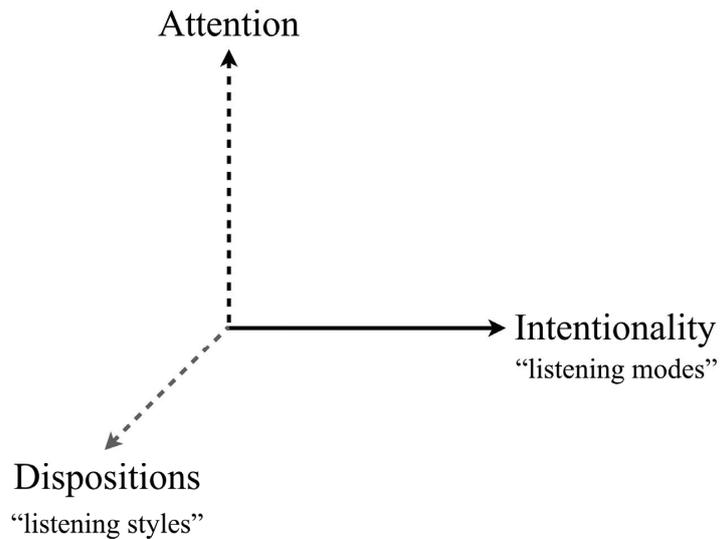


Figure 1. Three dimensions of listening.

The accounts of listening attention and listening styles are relevant to the subject of this article. However, we conclude that these accounts concern dimensions of listening which are separate from the dimension of listening modes. In general, we can characterise these dimensions of listening as being dimensions of *intentionality*, *attention* and *disposition* (see Figure 1). In principle, a separate taxonomy of listening is permitted within each of these dimensions. The dimension of intentionality concerns the aspects of understanding about the world acquired by listening; the main concern of this article. Within the action-oriented ontology of ecological perception, intentionality fundamentally concerns the relationship between the subject and the environment. These meaningful relationships between the two are made effective with

interactionally structured couplings (Varela et al., 1991; Dourish, 2001). The dimension of attention concerns the aspects of focal attention towards the environment, and the third dimension concerns the aspects of individual dispositions in listening behaviour. In this article, our taxonomical focus is on the dimension of intentionality. We view the modes of listening as essentially referring to (1) different types of action-sound couplings relating to schematically structured experiences of interaction, and (2) ways of utilising these couplings (cognitively) in our intentional encounters with the environment.

We acknowledge that the distinction of the dimensions of taxonomy is a theoretical definition for the purpose of simplifying the taxonomical formulation. Therefore, when applying listening modes to any real-life listening experience, the dimensions of attention and listening styles inevitably function as intervening variables.

3. Focusing on the experiential domain of listening

3.1 Revision goals for the taxonomy of listening modes

The listening mode taxonomy of Tuuri et al. (2007) has extended the traditional three-mode scheme with additional relevant perspectives in order to form a comprehensive outline of the listening modes. Overall, we consider that the basic scheme of this taxonomy is still sound. However, one problem has been the rather broad and loose definition of the connotative mode of listening. We also feel that the revised version of the account should better describe the action-oriented basis of causal, empathetic, functional and semantic denotations. In other words, it should better depict meaning-evoking processes on an experiential level and their dynamics between higher-level, interpretative meaning-creation. In particular, a revised version of the taxonomy is made by adopting the embodied cognition and the schematically structured action-

sound couplings as the underlying theoretical stance. We will also reconsider the hierarchical nature of the taxonomy in the revised version. In the earlier scheme, there may have been too much emphasis on the hierarchy between modes.

3.2 Experiential basis of meaning-creation

We can see that in the previous taxonomy of listening modes, pre-attentive modes (reflexive and connotative) are bound up with the experiential domain of meaning-creation, while upper-level modes have an interpretative or conceptual nature.

However, such a distinction should not be taken too literally, as it is very plausible that even the conceptual level of thinking operates inseparably from experiences (Gallese & Lakoff, 2005). Indeed, the embodied theory of meaning (Johnson, 2007) suggests that all creation of meanings has its basis in the situated *flow of experience* that cannot exist without a biological subject engaging its environment. And the meaning of a specific aspect of that experience is that aspect's connections to other parts of past, present and anticipated future experiences. On a higher level, we also utilise conceptually coded meanings and reasoning, but "that is merely the more conscious, selective dimension of a vast, continuous processes of immanent meanings that involve structures, patterns, qualities, feelings and emotions" (Johnson, 2007, p. 10).

Within the embodied approach, the experiential domain of meanings has been characterised as being essentially *sensorimotoric* and *imaginative*. In other words, perceptual experience is closely related to imagination (Gallese and Lakoff 2005; Johnson 2007) and involves synesthetic processes (multisensory integration) as well as kinaesthetic processes (sense of movement via ideomotor processing) (e.g., Leman 2008a). The experiential domain can also be conceived of as *affective*, due to the somatosensory relation between affect and our biological engagement to worldly

interactions (Damasio, 2000). Through corporeal involvement in these interactions, our affective states and other types of intentionality are also vitally manifested and mediated (Merleau-Ponty, 1945; Rosenthal & Bourgeois, 1991).

Lakoff (1987) and Johnson (1987) have put forward an experience-based approach to meaning, imagination and reasoning. According to it, there are basically two sources from which conceptual meaning structures arise: “1) from the structured nature of bodily and social experience, and 2) from our innate capacity to *imaginatively project* from certain well-structured aspects of bodily and interactional experience to abstract conceptual structures” (Lakoff, 1988, p. 121, emphasis added). *Image schemas* are an example of directly meaningful structures (schematic gestalts), which are grounded on regularities of sensorimotor experiences. Imaginative projecting, often involving metaphoric processes, allows the usage of these experiential schemata in the conceptual processes of thinking and language. (Lakoff, 1987; Johnson 1987; Gallese & Lakoff, 2005.) It is important to acknowledge that such schematic structures are based also on experiences of social and cultural constraints in the interaction with the environment (Leman, 2008a; Gibson, 1966). Therefore perceptual experience involves ‘resonances’ to the knowledge of both natural and cultural constraints. In the following sections, we outline a tentative idea of perception which has its emphasis on such resonances involving schematically structured past experiences, conceived in the framework of action-sound couplings.

3.3 Projection of mental images

Our theoretical basis loosely aims at combining two accounts of ecological perception; the direct perception model of Gibson (1979) and the lens model approach of Brunswik (1956). These accounts are often seen as opposed to each other, as Gibson’s view on perception emphasises the direct attunement to action-relevant

values of an environment while the lens model approach emphasises the inferential basis of environment-relevant (i.e., distal) meanings. We do not see these two different modes of knowing the world as necessarily competing with each other. Rather, we see them as dimensions that illustrate the multi-faceted and multi-level nature of meaning creation.

- We accept the embodied sensitivity to action-relevant values (direct perception) as a general principle of making sense of the world. However, depending on level and orientation of meaning-making (e.g., the awareness of context) the same patterns of sensations may lead to different affordances (see, Leman 2008a).
- We accept that perception incorporates inferential processes. Inferred meanings are created in tight interaction with the experiential domain, which works as an *embodied resonator* between patterns of sensation and the speculated action-relevancy of these sensory cues.

Drawing parallels with Brunswik's lens metaphor, the embodied resonator functions as a kind of experiential 'lens-element' which permits patterns of sensation to be inferred in terms of mental imagery being projected from the structured nature of experiences (incorporating action-sound couplings). These projections of action-relevant mental images (I) arise upon emerging resonances of action-relevant values between the other two elements: patterns of sensation (S) and well-structured patterns of recurrent experiences (E). The intention is to portray these elements (I, S, & E) rather as phenomenological entities than physiological objects. In other words, patterns of sensations (S) refer to internal sensory experiences rather than to external physical energies. Figure 2 illustrates the schematic overview of the embodied resonator. The intention is not to describe a flow of information (e.g., Neisser's

perceptual cycle, 1976) between the elements but to illustrate the emergent nature of the resonator. As a consequence, the model is not unidirectional. Mental imagining can also engender imagined patterns of sonorous sensations (S), for example, when we hear sounds of an imagined landscape or listen to music by merely imagining it. The idea of embodied resonator bears some similarities to Godøy's (2006) model of gestural-sonorous interaction.

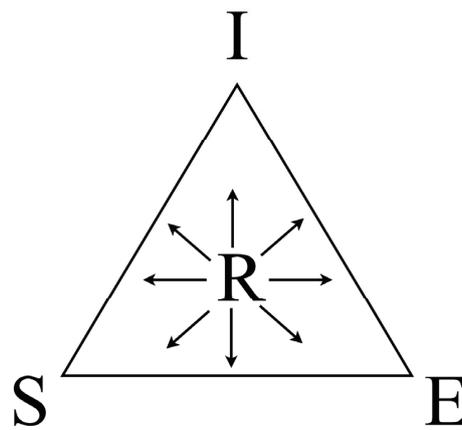


Figure 2. Schematic overview of an embodied resonator (R), and its elements: action-relevant mental images (I), experiential patterns of sensation (S) and well-structured patterns of sensorimotor experiences (E).

The projection of the mental imagery is not deterministic in nature. As in the original lens model, the nature of resonant cues is probabilistic, i.e., cues of action-relevancy have an uncertain and speculative relation to the resulting projections. We thus imagine in terms of what kind of actions we see as a perceptual possibility. The imaginative aspect of perception is best seen as intentional effort and dependent on the contextual constraints of a situation.

3.4 Enactive effort in perception and thinking

We will use the concept of *enactive perception* (Noë, 2004) to further elaborate our idea in order to better understand the imaginative effort to ‘catch’ the perceptual

content by utilising our ecologically shaped experiential background. The central argument of enactive perception is that perception is not something that happens in us, but it is something we *do*. “When we perceive, we perceive in an idiom of possibilities of movement” (Noë 2004, p. 105). And this idiom is mediated by structured patterns of sensorimotor contingency. In the context of embodied resonator, the enactive effort in perception may be seen as a kind of immanent and intentional ‘excitation’ for an imagined action-relevant experience (I), based on the sensorimotor profiles in our possession (E).

The enactive approach provides an interesting point-of-view on the perception of features relating to sound-sources. It may be intuitive to think that we initially find out appearances through perception. Indeed, as the classical view of ecological perception suggests, we find out features of sound-emitting or sound-reflecting objects by utilising our experiential knowledge of perceptual invariants (Michaels & Carello, 1981). However, the enactive approach promotes an idea that, at the time we are able to understand features of a sound-causing object, we have already *acted out* the contingent scenario of a sound event in a simulated manner. It is thus suggested that our sonic perception of the material elements and appearances of the world is initially coupled with (ideomotoric) understanding of *actions* that could be performed to cause sound events³ – including actions performed with a vocal apparatus. In this sense, sound perception essentially has a certain haptic or kinaesthetic character, and requires sensorimotor skills. Perceptual invariants, as properties of action-sound couplings, thus need to be conceived of as referring to co-occurring regularities in both motor and sensory patterns (Mossio & Taraborelli, 2008).

³ This idea is in line with Godøy’s (2001) suggestion that mental images of action would comprise separate components of ‘excitation’ (the images of what we do) and ‘resonances’ (the images of the effects of what we do).

There is evidence suggesting that, in perception and thinking, ideomotoric processing of actions indeed occurs at the neural level, as embodied simulations.

Gallese and Lakoff (2005) outline three different types:

- (1) simulations in *action-location neurons* (for successfully interacting with objects in their spatial position),
- (2) simulations in *canonical neurons* (for encoding the goal of an action and the intrinsic physical features of objects into suitable motor programmes to act on them),
- (3) simulations in *mirror neurons* (for imitative understanding of actions of others).

Embodied simulations are inherently multimodal as they occur in certain functional (neural) clusters, which are shared for both perception and doing and which respond to more than one sensory modality (Gallese & Lakoff, 2005). Such a tight sensorimotor integration also offers an explanation for the synesthetic and kinaesthetic processes of perception which permit multimodal and motoric imagery as a response to sounds or musical cues. As noted in the previous section, the generation of multimodal images does not even need any external stimulus. As proposed by Gallese and Lakoff (2005), embodied simulations are generally involved in imagination and thinking, as well as in perception.

To conclude the discussion on the enactive viewpoint, we hypothesise that there likely exist at least two types of enactive effort; firstly, *imitative enactions*, in which the ideomotoric processes follow or conform to the sonic event itself (as demonstrated in motor-mimetic affordances of movement; see Godøy, 2010), and secondly, *responsive enactions*, in which the ideomotoric processes speculate on responsive counteractions to the sound event (as demonstrated in affordances that, for

example, doorbell sound evokes). In both cases, however, enactive effort may be seen as synonymous to imaginative effort of the embodied resonator, resulting in ecologically relevant mental images of doing.

3.5 Types of action-sound couplings

We refer to action-sound couplings as schematically structured experiences able to project meaningful action-relevant mental images relating both to our body (kinaesthetic/somatic ontology) and the environment (action-oriented ontology of environment). In the literature such schematic structures have been referred to, for example, as body-schemas, motor-schemas or image schemas. As a basis of meaning-creation, different action-sound couplings can provide great varieties of meaningful experiences in listening. For the sake of listening modes taxonomy, we will distinguish three main levels of action-sound couplings involved in imaginative projections: *reflexive couplings*, *kinaesthetic couplings* and *connotative couplings*. The first two mostly involve innate and early developed schemata while the last one likely involves schemata which are more adaptive and learned. For the connotative level, we also suggest a distinction of three sub-types of couplings; *action-sound-object*, *action-sound-intersubjectivity* and *action-sound-habit*.

Reflexive action-sound couplings refer to quickly evoked, phylogenetically developed, innate action-sound-reaction affordances. They are based on automated (or ‘hard-wired’) schemas which are due to the evolutionary adaptation to our ecology.

Kinaesthetic action-sound couplings refer to kinaesthetic affordances of a perceptual experience; an imaginative sense of motor-movements on the basis of sound perception. This gestural character of sound perception is arguably based on ideomotoric processes that manifest innate or early developed structures of kinaesthetic schemata concerning bodily movements, coordination and postures

(Johnson, 1987; Merleau-Ponty, 1945). In the light of vitality affects (Stern, 1985; Johnson, 2007), kinaesthetic perception can also be seen as bodily resonated contours (or patterns) of feeling. These dynamic patterns may concern, for instance, sensitivity to the haptic and tactile feelings relating to movement (e.g., tensions and textures), sensitivity to gestural signatures of an interpersonal affect (see below), and in general, sensitivity of coping with the physical world. It has been suggested that musical involvement in listening strongly comprises different levels of imitative effort, which relate to the experience of corporeal movement in accord with music (Leman, 2008a). The gestural dimension of sound perception is also discussed in the present author's previous article (see, Tuuri, 2010).

Connotations refer to vigorously activating imaginative projections of action-relevant values as resonances of schemata based on interactions with both natural and cultural constraints. All three sub-types of couplings, namely action-sound-object, action-sound-intersubjectivity and action-sound-habit, likely involve the schemata mostly acquired by learning. Action-sound-object couplings refer to sonic experiences that are about actions of encountering and manipulating objects in the environment. Couplings of action-sound-intersubjectivity refer to sonic experiences of interpersonal encounters. These couplings resonate especially with gestural signatures (or motoric invariants) in the patterns of kinaesthetic sensation, and function as *gestalts* for interpersonal understanding. It is suggested that mirror neurons act as a basic mechanism for such an empathetic involvement (Iacoboni, 2009); they permit bodily realised (ideomotoric) affordances of movement, which can be interpreted in terms of the perceiver's own body-based ontology of intentions and emotions (Leman, 2008a; Gallese et al., 2007). As a third category, couplings of action-sound-habit refer to various habituated aspects of cultural ecology that are involved in actions.

Connotative projections are probably highly interactive with contextual orientations and anticipations of the higher, interpretative level of listening. In addition, they are likely to be highly interactive with the kinaesthetic experience of a lower level. We hypothesise that at the connotative level, the meaning-making is elaborated through analogical and metaphorical (Lakoff & Johnson, 1999) processes, organising vigorously into a variety of connotative associations. We propose that these mental projections provide the essential, immanent and experiential basis for making interpretations relating to acting on the sounding objects of the world (causal and empathetic listening), to intersubjective attuning to emotions and intentionality (empathetic and functional listening) or to dealing with the norms of sound usages and sounding artefacts of the socio-cultural ecology (functional and semantic listening).

4. Revised scheme for listening modes

We now present a newly formulated version of the taxonomy of listening modes that conforms to the revision principles and theoretical discussion presented in the previous section. The schematic overview of the new taxonomy is presented in Figure 3. Also, illustrative examples for each mode of listening are presented in Table 1.

The modes mostly referring to the experiential domain of listening are reorganised in accordance with the suggested three bases for imaginative projections of action-relevant meanings: reflexive, kinaesthetic and connotative. Therefore, a *kinaesthetic mode* of listening has been added in between the reflexive and connotative listening modes. Kinaesthetic meanings evoked in listening, a well documented phenomenon in the music research literature (e.g., Eitan & Rothschild, 2010), was not covered clearly in the previous taxonomy. Most importantly, this new mode of listening in general emphasises the bodily basis of meaning-creation

(Johnson, 2007), as it also implies the projection of kinaesthetic character (sense of movement) into the higher-level perceptual processes. Sheet-Johnstone (1999) has defined four qualitative dimensions in the sense of movement, and they are: (1) *tension* (relating to the forces and effort in movement), (2) *linearity* (relating to a path of motion), (3) *amplitude* (relating to the range of motion) and (4) *projection* (relating to a temporal projecting of force). Arguably, these qualities in the experience of movement are integrated with the experience contours of a vitality affect (Stern, 1985). The sense of movement may also become accompanied with tactile sensations of textures. Slowly moving romantic ‘string pads’, for example, can evoke the tactile feel of a smooth cushion (Tagg, 1992).

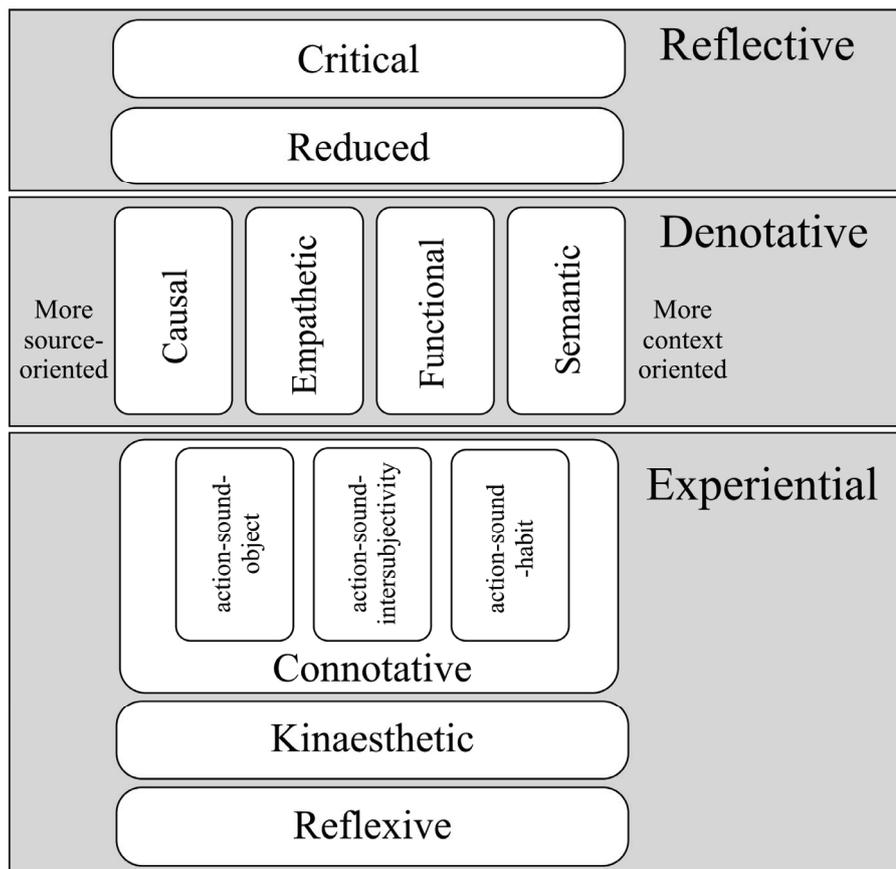


Figure 3. Overview of the revised scheme for modes of listening.

In the revised scheme, connotative listening is divided into three folds, each representing a different kind of action-sound couplings incorporated in connotative projections. These folds are arranged on the same horizontal level, to avoid implying any rigid hierarchy between them. Upper level modes are arranged similarly, thus the hierarchy between causal, empathetic, functional and semantic modes of listening is removed. From left to right, the ordering of different modes within this class of *denotative* listening modes implies their orientation either to sound sources (modes on the left) or to the context of a sound (modes on the right). These modes are in general denotative, because their utilisation results in an ability to conceptualise the interpreted perceptual content, be it an aspect of a sound's cause, an affective state being involved, a sound's purpose or a symbolic meaning. The horizontal order of these modes also indicates their interactional relation to the connotative mode: beneath each denotative mode are the best corresponding categories of connotative action-sound couplings, which provide the experientially projected immanent basis for making conceptual interpretations. Thus, action-sound-object couplings are linked to causal and empathetic modes, action-sound-intersubjectivity couplings are linked to empathetic and functional modes, and finally, action-sound-habit couplings are linked to functional and semantic modes. We see that empathetic listening, while being intersubjectively oriented, also involves 'agency-seeking' attuning to the causal dimension of a sound which also ultimately manifests the cues of intentionality, either directly (via vocal sound) or instrumentally (via sounding objects). Functional listening also mainly attends to the cues of intentionality, albeit in a more context-orientated way. But to some degree, the functional meaning of sounds is arguably also a matter of socio-cultural habit.

Table 1. Summary of the listening modes with examples.

Mode:	Questions:	Example 1: <i>A cell phone rings during a lecture (with a classic 'Nokia-tune' ringtone)</i>	Example 2: <i>Listening to Bruce Springsteen Live Rendition of "The River" on iPod whilst travelling.</i>
Experiential Modes:			
Reflexive	Did you notice any reflexive responses triggered by sound?	Startle and orienting responses! It alarms and grabs attention. Sound was surprising because it was not anticipated in the situation.	No surprises, no reflexive responses. This is a recording I am well familiar with.
Kinaesthetic	How does it physically manifest? In what way does the sound imply movement?	Highly tensed and 'shrilling' qualities in the experienced path of motion, which overall feels descending but has distinct 'wavy' patterns with sharp onsets. Force feels evenly projected in the movement (as there are no changes in sound intensity).	Sense of a 'swaying' motion (driving rhythm), tension building and release across the chord progressions. Vocal attuning (sensations relating to the vocal apparatus) to the voice of the lead singer. An urge to play 'air drums'.
Connotative	What kind of freely formed associations did listening immediately evoke?	...something small and light (high pitch, not much energy)...artificial ...plastic resonances... ...arousing...but feels cold and mechanistic... ...ditties in electronic games of 80's...ringtones of 90's...old Nokia commercials...	... feeling of a projection of a space...live instruments played... presence of large crowd... ...self-confidence in voice...emotionally charged...rock-ballads...Springsteen...stadium concerts... cigarette lighter lightshow...
Denotative Modes:			
Causal	What could have caused the sound?	It is a cell phone ringing on someone's desk.	Live musical performance by the E-street band. Played from a portable music player.
Empathetic	Does it feel as if the sound signals someone's state of mind or intentions?	First, the (tense and sharp) sound is like shouting at you, but when repeated, it starts to feel more like wailing, albeit in a machine-like way.	Melancholic and sad (flat tone of the voice), but nevertheless decisive or aggressive. Also a sense of shared togetherness (an audience singing with the lead singer).
Functional	What was the purpose of the sound? What function does the context indicate?	Somebody is calling; the sound has functions for alarming, identifying and locating. In general, the sound also has a branding function for the manufacturer. In musical terms, the melody manifests tonal functions of chord progression.	In the context of commuting, diverting the attention away from the mundane act of sitting in a bus and transporting the listener to an epic concert among thousands of listeners.
Semantic	Does the sound seem to represent any symbolic or conventional meanings?	Clearly, the sound represents Nokia. For me it also represents Finland. For the owner, it represents his or her choice of ringtone. This is a tonal melodic sequence in ¾ meter. It is a theme from a composition by Francisco Tárrega.	Nostalgic song from the early 80's that epitomises an era of rock ballads. Lyrics convey the feelings of readjustment and hopelessness after unemployment. A classic live stadium rock concert.
Reflective Modes:			
Reduced	Can you describe the properties of the sound itself as objectively as possible?	The sound is a clearly separated, iterative object. It is in high-pitched register and a bit loud. The sharp onsets and tone quality resemble a simple sound synthesis. In overall, a big contrast to the previous quietness.	The trademark sound of the E-street band with harmonica, tingling piano arpeggios, and sparse guitar and bass backing. The vocal sound is rasping and forced, as if produced with great effort with respirations audible in many sections.
Critical	Was the sound appropriate for the situation? Did you understand it correctly?	No panic, this is not my iPhone. How disturbing...it is totally inappropriate to keep your phone switched on! But hey, those classy ringtones are rare nowadays.	This music matches my current mood well. The time of the big stadium concerts has probably passed. How convincingly is Bruce singing about unemployment with his 20 Grammy awards?

The functions of critical and reduced modes of listening are redefined. We conclude that these modes of listening are both *reflective* in nature, and therefore they both operate in the highest-level class of reflective listening. Even though the reduced mode of listening is often conceived of as having an orientation to sound qualities, we

want to emphasise the more general function of this mode. We see that the essence of reduced listening is in self-reflective analysis of one's listening experience and, by resisting any denotations, also intentional manipulation of that experience. This reducing function, of course, permits the orientation towards the qualities of perceived sound. This orientation could be seen as a subcategory of reduced listening, in a manner similar to that originally suggested by Schaeffer (1966). In the previous version of the taxonomy, the critical mode of listening was listed as context-oriented mode. However, in the revised scheme we make a clear distinction between modes of denotation and reflection. Quite intuitively, critical listening essentially functions as a reflective mode, constantly judging the appropriateness of listening-based interpretations. Through these judgments, the critical element in listening evokes new meanings and re-evaluates those already evoked.

Wholly imagined listening experience, evoked completely without any external stimulus, can also be conceived of as a reflective element of listening. However, we argue that such a case does not refer to a distinct mode of listening (e.g., similarity of imagined and heard timbres, see Halpern et al., 2004). We rather see that imagined listening does not fundamentally differ from a 'normal' case of listening. According to the perspective of this article, the pivotal difference in imagined listening concerns the initiation of an enactive effort that excites and evokes the listening experience. In imagined listening, this effort is completely based on mental initiative.

5. Conclusions and discussion

The main purpose of this article was to present a useful taxonomy of listening modes that brings out a justified account of the most relevant constituents of meanings in the process of listening, and situates these on a view of embodied mind. Such a taxonomy

has several important outcomes: it (1) reappraises and reorganises the past accounts of listening in general, (2) frames many central yet contrasting issues of music cognition as questions related to the meaning-creation process within different modes of listening, and finally (3), sets out to explain listening as an experiential and action-oriented process.

The account of listening modes aims at dispelling the notion that listening is a homogeneous and monolithic form of activity, perhaps only divisible in terms of attention (e.g., Chamorro-Premuzic & Furnham, 2007). This kind of mindset prevents us disentangling important distinctions related to experiential (reflexes, kinaesthetic qualities, associative mental images), denotative (causal, empathetic, functional and semantic listening), and reflective modes (reduced and critical listening). This broadening is not only motivated by purely theoretical arguments but by acknowledging that music listening is often construed as a heterogeneous activity where relatively little concentration is submitted specifically to the listening (North et al., 2004; Juslin et al., 2008).

The revised account of the listening modes develops previous accounts in many aspects. First of all, it fundamentally outlines attention, individual dispositions and intentionality as theoretically separable dimensions of listening. The taxonomic formulation of listening modes specifically focuses on the dimension of intentionality, which concerns the aspects of meaning in terms of relationship between the subject and the environment. The revised account clarifies the action-oriented nature of listening by conceiving the aspects of meaning in terms of schematically structured action-sound couplings and the ways of cognitively utilising these couplings in our intentional encounters with the world. Compared to the earlier account of Tuuri et al. (2007), the most prominent changes in the revised taxonomy are the addition of a

kinaesthetic mode of listening and the new, less hierarchic schematic arrangement of modes.

The chosen framework relies on the perspective of embodied cognition, but nevertheless is not dependent on the validity of a single theoretical account (e.g., direct perception, lens model, enactive perception). Listening is considered from experiential, denotative, and reflective perspectives. The main focus of this framework is on intentionality and experience (phenomenological emphasis) rather than on a flow and processing of information in the manner of the traditional IP-paradigm. This is, in our opinion, the most important aspect of the theoretical stance. The embodied perspective embraces the listener as an action-oriented intentional being making sense of the world. This sets the stage for questions that are immediately relevant to listeners, such as: what caused the sound, what are the interpersonal attributes (emotion, intention) in a particular sound or music, and what kind of socio-cultural values may be attributed to them? In the theoretical account, this process is explained by relying on the interaction-based couplings between the sensorimotor processes, and well-structured patterns of these, that all contribute to the imaginative creation of experiential, immanent meanings (via an embodied resonator). This experiential domain operates in tight interaction with more conceptual and interpretative processes of denotative and reflective meaning-creation. The whole process is essentially seen as an emergent property of the active, goal-oriented organism (i.e., listener).

The most immediate, practical implication of the taxonomy of listening modes might exist for the discipline of sound design. This discipline concerns various applications ranging from narrative sound design for media to sonic interaction design for various applications of human-technology interaction, as well as design of sonic

environments and installations. Whether focusing on functionality or expressivity, sound design cannot afford to overlook the diversity of meanings and the affective responses that the sound evokes in the context of its use. As a conceptual apparatus, the proposed taxonomy of listening modes can provide the designer with the means for understanding the attributes of a listening experience.

Another practical implication of the proposed taxonomy is for broadening the scope of music studies to include a wider variety of orientations within the realm of listening. We see this as a change of focus from syntactic and cognitive structures of music (e.g., chord progressions, instrumentation, and melodic archetypes) to components of meaning that have their foundation in more experiential properties of the sounds of music, such as voice quality and motoric processes. These are not only important for theoretical reasons but will facilitate operationalising those research problems more in line with contemporary studies identifying vocal expressivity (e.g., Dibben, 2006), gestures (e.g., Godøy & Leman, 2010) and interactive music technologies (Leman, 2008a) as the crucial challenges for music research.

We hope that this framework will stimulate empirical work on listening typologies by developing appropriate, open-ended data collection methods. Data-driven qualitative methods may be appropriate for avoiding the potential effect of researchers' presuppositions. For instance, free verbalisations about sounds in small panel groups might generate a rich form of information about participants' listening experiences of either complete soundscapes or isolated sounds (see an example in Tuuri et al., 2007). Recorded soundscapes, with either fixed or moving focal position, for example, might be used for permitting more attentional freedom to participants. Researchers may also utilise different orientations to sounds, for example, by orchestrating certain contexts for listening.

Besides developing the methodology for empirical studies on listening, further work could also extend to the other putative listening dimensions of attention and disposition. We also want to bring forward the idea that, to some degree, the sensorimotor integration and the imaginative nature of experiential meaning-creation provide a basis for cross-modal applicability to the presented taxonomy – beyond the auditory domain of listening. We consider that the taxonomy presented could be applied at least to the sense of touch, as many studies have demonstrated perceptual integrations between the auditory and tactile/haptic domains (e.g., Jousmaki & Hari, 1998; Tuuri, Eerola & Pirhonen, 2010). These observations suggest that sensations of both domains could evoke sensorimotor patterns, which resonate with shared schematic gestalt structures for both domains. It also intuitively seems that, for example, causal, empathetic and functional perceptions could be plausible via the sense of touch.

Finally, we see that an ultimate practical implication of the taxonomy of listening modes concerns the awareness of our sonic interactions with an environment at the personal level. By understanding the processes of knowing about the world and how meanings become coupled with sounds, we fundamentally understand more about ourselves as a part of our natural and socio-cultural environment. From an educational perspective, such understanding would also permit a broader repertoire for expressing ourselves either musically or in any other sound-mediated way.

Acknowledgements

This work is funded by Finnish Funding Agency for Technology and Innovation, and the Academy of Finland (Finnish Centre of Excellence in Interdisciplinary Music Research, project number 7118616).

References

- Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance* 19(2), 250–267.
- Barthes, R. (1974). *S/Z: An Essay*. New York: Hill & Wang.
- Bregman, A. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.
- Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.
- Chamorro-Premuzic, T. & Furnham, A. (2007). Personality and music: Can traits explain how people use music in everyday life? *British Journal of Psychology*, 98(2), 175–185.
- Chion, M. (1990). *Audio-vision: Sound on screen*. New York: Columbia University Press.
- Chion, M. (1993). *Le poème symphonique et la musique à programme*. Paris: Fayard.
- Clarke, E. F. (2005). *Ways of Listening: An Ecological Approach to the Perception of Musical Meaning*. Oxford: Oxford University Press.
- Damasio, A. (2000). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Houghton Mifflin Harcourt.
- Deutsch, D. (1999). *The psychology of music*. 2nd ed. USA: Academic Press.
- Dibben, N. (2006). Subjectivity and the Construction of Emotion in the Music of Björk. *Music Analysis*, 25(1-2), 171–197.
- Dourish, P. (2001). *Where the action is: The foundations of embodied interaction*. Cambridge, MA: MIT Press.
- Eitan, Z. & Rothschild, I. (2010). How music touches: Musical parameters and listeners' audiotactile metaphorical mappings. *Psychology of Music*, Online November 8, 2010, doi: 10.1177/0305735610377592.
- Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child development*, 1497–1510.
- Gallese, V. & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in reason and language. *Cognitive Neuropsychology*, 22, 455-479.
- Gallese, V., Eagle, M. & Migone, P. (2007). Intentional attunement: Mirror neurons and the neural underpinnings of interpersonal relations. *Journal of the American Psychoanalytic Association* 55 (1), 131–175.
- Gaver, W. (1988). *Everyday listening and auditory icons*. Doctoral Dissertation, University of California, San Diego.
- Gaver, W. (1989). The SonicFinder: An interface that uses auditory icons. *Human-Computer Interaction* 4 (1), 67–94.
- Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.
- Godøy, R. I. & Leman, M. (ed.) (2010). *Musical Gestures: Sound, Movement, and Meaning*. Routledge.
- Godøy, R. I. (2001). Imagined action, excitation, and resonance. In Godøy, R. I. & Jørgensen, H. (Eds.), *Musical Imagery*, Lisse: Swets and Zeitlinger, 237–250.
- Godøy, R. I. (2006). Gestural-sonorous objects: embodied extensions of Schaeffer's conceptual apparatus. *Organised Sound* 11(2), 149–157.

- Godøy, R. I. (2010). Gestural Affordances of Musical Sound, In Godøy, R. I. & Leman, M. (ed.), *Musical Gestures: Sound, Movement, and Meaning*. Routledge.
- Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, 42(9):1281–1292.
- Huron, D. (2002a). A six-component theory of auditory-evoked emotion. In *Proceedings of ICMPC7*, 673-676.
- Huron, D. (2002b). *Listening Styles and Listening Strategies*. Retrieved Dec 9, 2010 from <http://http://www.musiccog.ohio-state.edu/Huron/Talks/SMT.2002/handout.html>
- Iacoboni, M. (2009). Imitation, Empathy, and Mirror Neurons. *Annual Review of Psychology*, 60(1), 653-670.
- Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. Chicago, IL: University of Chicago.
- Johnson, M. (2007). *The Meaning of the Body: Aesthetics of Human Understanding*. The University of Chicago Press.
- Jousmäki, V. & Hari, R. (1998). Parchment-skin illusion: sound-biased touch, *Curr. Biol.* 8, p. 190.
- Juslin, P., Liljeström, S., Västfjäll, D., Barradas, G., & Silva, A. (2008). An experience sampling study of emotional reactions to music: Listener, music, and situation. *Emotion*, 8(5), 668–683.
- Kreutz, G., Ott, U., Teichmann, D., Osawa, P., & Vaitl, D. (2008). Using music to induce emotions: Influences of musical preference and absorption. *Psychology of Music*, 36(1), 101–126.
- Lakoff, G. & Johnson, M. (1999). *Philosophy In The Flesh: the Embodied Mind and its Challenge to Western Thought*. New York: Basic Books.
- Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press.
- Lakoff, G. (1988). Cognitive Semantics. In Eco, U., Santambrogio, M. & Violi, P. (eds.) *Meaning and Mental Representations*. Indiana University Press, 119-154.
- Leman, M. (2008a). *Embodied music cognition and mediation technology*. Cambridge, MA: MIT Press.
- Leman, M. (2008b). Systematic musicology at the crossroads of modern music research. In Schneider, A., (Ed.), *Systematic and Comparative Musicology: Concepts, Methods, Findings*. Hamburger Jahrbuch für Musikwissenschaft, volume 24, pages 89–115. Peter Lang, Frankfurt am Main.
- Loui, P. & Wessel, D. (2007). Harmonic expectation and affect in Western music: Effects of attention and training. *Perception & Psychophysics*, 69(7), 1084–1092.
- Margulis, H. (2010). When program notes don't help: Music descriptions and enjoyment. *Psychology of Music*, 38(3), published online 24 March 2010 DOI: 10.1177/0305735609351921.
- Merleau-Ponty, M., (1945/1962). *Phenomenology of Perception*. Trans. Colin Smith. London: Routledge.
- Michaels, C. F. and Carello, C. (1981). *Direct Perception*. New York: Prentice-Hall.
- Mossio, M. & Taraborelli, D. (2008). Action-Dependent Perceptual Invariants: From Ecological to Sensorimotor Approaches. *Consciousness and Cognition* 17 (4):1324-1340.

- Mott, R. L. (1990). *Sound effects: Radio, TV and Film*. Boston, MA: Focal Press.
- Neisser, U. (1976). *Cognition and reality: principles and implications of cognitive psychology*. W. H. Freeman.
- Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.
- North, A., Hargreaves, D., & Hargreaves, J. (2004). Uses of music in everyday life. *Music Perception, 22*(1), 41–77.
- North, A., Shilcock, A., & Hargreaves, D. (2003). The effect of musical style on restaurant customers' spending. *Environment and Behavior, 35*(5), 712–718.
- Peretz, I. & Zatorre, R. (2003). *The cognitive neuroscience of music*. Oxford University Press Oxford, UK.
- Plazak, J. & Huron, D. (in press). The First Three Seconds: Listener Knowledge Gained from Brief Musical Excerpts. Accepted by *Musicae Scientiae*.
- Rosenthal, S. B. & Bourgeois, P. L. (1991). *Mead and Merleau-Ponty: Toward a Common Vision*. New York: State University of New York Press.
- Schaeffer, P. (1966). *Traité des objets musicaux*. Paris: Éditions du Seuil.
- Schafer, M. R. (1993). *The Soundscape: Our Sonic Environment and the Tuning of the World*. Destiny Books.
- Schubert, E. (2007). The influence of emotion, locus of emotion and familiarity upon preference in music. *Psychology of Music, 35*(3), 499–515.
- Searle, J. (1979). *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.
- Sheets-Johnstone, M. (1999). *The Primacy of Movement*. Amsterdam: John Benjamins.
- Sloboda, J. A., & O'Neill, S. A. (2001). Emotions in everyday listening to music. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 415-430). Oxford: Oxford University Press.
- Sonnenschein, D. (2001). *Sound design: The expressive power of music, voice and sound effects in cinema*. Saline, MI: Michael Wiese Productions.
- Stern, D. (1985). *The Interpersonal World of the Infant: A View from Psychoanalysis and Developmental Psychology*. New York: Basic Books.
- Sternberg, R. J. (1997). *Thinking Styles*. Cambridge University Press.
- Tagg, P. (1992). Towards a sign typology of music. In R. Dalmonte & M. Baroni (Eds.), *Secondo Convegno Europeo di Analisi Musicale* (pp. 369-378). Trento, Italy: Università Degli Studi di Trento.
- Tarasti, E. (2002). *Signs of Music: A Guide to Musical Semiotics*. Mouton de Gruyter.
- Tervaniemi, M., Schröger, E., & Näätänen, R. (1997). Pre-attentive processing of spectrally complex sounds with asynchronous onsets: an event-related potential study with human subjects. *Neuroscience letters, 227*(3), 197–200.
- Truax, B. (2001). *Acoustic communication*. Ablex Publishing.
- Tuuri, K. (2010). Gestural Attributions as Semantics in User Interface Sound Design. In S. Kopp & I. Wachsmuth (Eds.), *Gesture in Embodied Communication and Human-Computer Interaction*. Lecture Notes in Artificial Intelligence (LNAI 5934). Berlin, Germany: Springer-Verlag. 257-268.
- Tuuri, K., Eerola, T. & Pirhonen, A. (2010). Leaping across Modalities: Speed Regulation Messages in Audio and Tactile Domains. In R. Nordahl, S. Serafin, F. Fontana & S. Brewster (Eds.), *Haptic and Audio Interaction Design*. Lecturer Notes in Computer Science (LNCS 6306). Berlin Heidelberg: Springer-Verlag. 10-19.
- Tuuri, K., Mustonen, M. & Pirhonen, A. (2007). Same sound – Different meanings: A Novel Scheme for Modes of listening. In *Proceedings of Audio Mostly 2007*.

- Ilmenau, Germany: Fraunhofer Institute for Digital Media Technology IDMT. 13-18.
- Vanderveer, N. J. (1979). *Ecological Acoustics: Human Perception of Environmental Sounds*. Ph. D. thesis, Cornell University, Ithaca, NY.
- Varela, F., Thompson, E. & Rosch, E. (1991). *The embodied mind*. Cambridge, MA: MIT Press.
- Watson, K., Barker, L. L., & Weaver, J. B. (1995). The Listening Styles Profile (LSP-16). Development and validation of an instrument to assess four listening styles. *The International Journal of Listening* 9, 1-14.

PIII

**BODILY ENGAGEMENT IN MULTIMODAL INTERACTION: A
BASIS FOR A NEW DESIGN PARADIGM?**

by

Kai Tuuri, Antti Pirhonen & Pasi Väikkynen 2009

In S. Kurkovsky (Ed.), *Multimodality in Mobile Computing and Mobile Devices:
Methods for Adaptable Usability*, IGI Global, Hershey, pp. 137–165

Reproduced with kind permission of IGI Global.

Multimodality in Mobile Computing and Mobile Devices: Methods for Adaptable Usability

Stan Kurkovsky
Central Connecticut State University, USA

Information Science
REFERENCE

INFORMATION SCIENCE REFERENCE

Hershey · New York

Director of Editorial Content: Kristin Klinger
Senior Managing Editor: Jamie Snavelly
Assistant Managing Editor: Michael Brehm
Publishing Assistant: Sean Woznicki
Typesetter: Michael Brehm, Kurt Smith
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: <http://www.igi-global.com/reference>

Copyright © 2010 by IGI Global. All rights reserved. No part of this publication may be reproduced, stored or distributed in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher.

Product or company names used in this set are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI Global of the trademark or registered trademark.

Library of Congress Cataloging-in-Publication Data

Multimodality in mobile computing and mobile devices : methods for adaptable usability / Stan Kurkovsky, editor.

p. cm.

Includes bibliographical references and index.

Summary: "This book offers a variety of perspectives on multimodal user interface design, describes a variety of novel multimodal applications and provides several experience reports with experimental and industry-adopted mobile multimodal applications"--Provided by publisher.

ISBN 978-1-60566-978-6 (hardcover) -- ISBN 978-1-60566-979-3 (ebook) 1.

Mobile computing. I. Kurkovsky, Stan, 1973-

QA76.59.M85 2010

004.167--dc22

2009020551

British Cataloguing in Publication Data

A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

Chapter 6

Bodily Engagement in Multimodal Interaction: A Basis for a New Design Paradigm?

Kai Tuuri

University of Jyväskylä, Finland

Antti Pirhonen

University of Jyväskylä, Finland

Pasi Väikkynen

VTT Technical Research Centre of Finland, Finland

ABSTRACT

The creative processes of interaction design operate in terms we generally use for conceptualising human-computer interaction (HCI). Therefore the prevailing design paradigm provides a framework that essentially affects and guides the design process. We argue that the current mainstream design paradigm for multimodal user-interfaces takes human sensory-motor modalities and the related user-interface technologies as separate channels of communication between user and an application. Within such a conceptualisation, multimodality implies the use of different technical devices in interaction design. This chapter outlines an alternative design paradigm, which is based on an action-oriented perspective on human perception and meaning creation process. The proposed perspective stresses the integrated sensory-motor experience and the active embodied involvement of a subject in perception coupled as a natural part of interaction. The outlined paradigm provides a new conceptual framework for the design of multimodal user interfaces. A key motivation for this new framework is in acknowledging multimodality as an inevitable quality of interaction and interaction design, the existence of which does not depend on, for example, the number of implemented presentation modes in an HCI application. We see that the need for such an interaction- and experience-derived perspective is amplified within the trend for computing to be moving into smaller devices of various forms which are being embedded into our everyday life. As a brief illustration of the proposed framework in practice, one case study of sonic interaction design is presented.

DOI: 10.4018/978-1-60566-978-6.ch006

INTRODUCTION

In early days of human-computer interaction (HCI), the paradigm was mainly seen as a means to “synchronise” the human being and a computer (Card et. al., 1983). While the number of computer users rose rapidly and computers were suddenly in the hands of “the man in the street”, there was an evident need to make computers easier to use than when used by experts in computing. Psychologists were challenged to model the human mind and behaviour for the needs of user-interface design. It was thought that if we knew how the human mind works, user interfaces (UIs) could be designed to be compatible with it. To understand the human mind, computer metaphor was used. Correspondingly, multimodality has often meant that in interaction with a computer, several senses (“input devices”) and several motor systems (“output devices”) are utilised. This kind of *cognitivist* conceptualisation of the human being as a “smart device” with separate systems for input, central (symbolic) processing and motor activity has indeed been appealing from the perspective of HCI practices. However, contemporary trends of cognitive science have drifted away from such computer-based input-output model towards the idea of mind as emergent system which is structurally coupled with the environment as the result of the history of the system itself (Varela et. al, 1991). We thus argue that as a conceptual framework for HCI the traditional cognitivist approach is limited, as it conflicts with the contemporary view of the human mind and also with the common sense knowledge of the way we interact with our everyday environment (see Varela et. al., 1991; Noë, 2004; Lakoff & Johnson, 1999; Clark, 1997; Searle, 2004). One of the shortcomings of the traditional input-output scheme is that it implies that the capacity for perception could be disassociated from the capacities of thought and action (Noë, 2004).

For a long time, the development of UIs had been strongly focused on textual and graphical

forms of presentation and interaction in terms of the traditional desktop setting. However, as computing becomes increasingly embedded into various everyday devices and activities, a clear need has been recognised to learn about the interaction between a user and a technical device when there is no keyboard, large display or mouse available. Therefore, the need to widen the scope of human-computer interaction design to exploit multiple modalities of interaction is generally acknowledged.

Within the mainstream paradigm of HCI design, conceptions of multimodality tend to make clear distinctions between interaction modalities (see, e.g. Bernsen, 1995). There is the fundamental division between perceiving (gaining feedback presentation from the system) and acting/doing (providing input to the system). These, in turn, have been split into several modality categories. Of course traditional distinctions of modalities have proved their usefulness as conceptual tools and have thus served many practical needs, as they make the analysis and development of HCI applications straightforward. However, too analytic and distinctive emphasis on interaction modalities may promote (or reflect) design practices where interaction between a user and an application is conceptualised in terms of technical instrumentation representing different input and output modalities. Such an approach also potentially encourages conceptualising modalities as channels of information transmission (Shannon & Weaver, 1949). Channel-orientation is also related to the ideal that information in interaction could be handled independently from its form and could thus be interchangeably allocated and coded into any technically available “channels”. We see that, in its application to practical design, the traditional paradigm for multimodality may hinder the design potential of truly multimodal interaction. We argue that in the design of HCI, it is not necessarily appropriate to handle interaction modalities in isolation, apart from each other. For instance, the use of haptics and audio in interaction,

Bodily Engagement in Multimodal Interaction

though referring to different perceptual systems, benefits from these modalities being considered together (Cañadas-Quesada & Reyes-Lecuona, 2006; Bresciani et. al., 2005; Lederman et. al., 2002). However, even recent HCI-studies of cross-modal interaction, although concerning the integration of modalities, still seem to possess the information-centric ideal of interchangeable channels (see e.g. Hoggan & Brewster, 2007).

In this study, we are looking for an alternative paradigm for multimodality. Although modalities relating to perceptual awareness and motor activity differ in their low-level qualities, they have interconnections and share properties, which make them highly suitable for study and handling within a shared conceptual framework. This chapter outlines a propositional basis for such a framework, gaining ingredients from the literature of embodied cognition, cross-modal integration, ecological perception and phenomenology. In the *embodied* approach (Varela et. al., 1991), the human mind is inseparable from the sensory-motor experiencing of the physical world and cognition is best described in terms of embodied interaction with the world. The resulting framework is bound up with a concept of physical embodiment, which has been utilised within several scientific disciplines to reveal the role of bodily experience as the core of meaning-creation. In the course of this chapter, we will present arguments to support the following basic claims as cornerstones of the proposed paradigm:

1) *Interaction is always multimodal in nature*

The embodied approach to human cognition implies understanding and meaning as being based on our interactions with the world. Understanding is thus seen as arising inherently from an experiential background of constant encounters and interaction with the world by using our bodies. Although our sensory modalities depend on different perceptual systems, human awareness is not channel-oriented. Instead, it is oriented to actions

in the environment: objects and agents involved in actions and our own action possibilities (see e.g. Gibson, 1979; Varela et. al., 1991; Clark, 1997; Gallese & Lakoff, 2005). In *action-oriented* ontology, multimodality and multimodal experience appear as inseparable characteristic of interaction. This view is supported by recent research in neuroscience; in perception and thinking, the neural linkages of motor control and perception as well as the integration of sensory modalities appear to be extensive (Gallese & Lakoff, 2005).

2) *Design arises from mental images and results in mental images*

We argue that the design of UI elements for human-computer interaction involves the communication of *action-relevant mental imagery*. This imagery, being bonded to embodied experiences, is essentially multimodal in nature. Hence, we argue that the starting point of UI-element design should not be rigidly any specific “channel” – presentation or input modality – but the embodied nature of interaction itself, and the subjective exploration of the “imagery” of its action-related meanings. Of course, “channels” do exist in terms of a technical medium. But they should be primarily responsible for supporting the construction of contextually coherent (action-relevant) mental imagery. This kind of construction of imagery, as a gestalt process, could be called *amodal* completion. But unlike the traditional cognitivist view (Fodor, 1975), we propose that amodality is not symbolic in nature but inseparably bound up with our sensory-motor system. Therefore it makes more sense to call these constructed mental images multimodal, whether they are mental images of the designer or mental images of the user.

3) *Action-relevant mental imagery can be communicated by manifesting action-relevant attributions in the contextual appearance of UI*

As already implied above, we propose that the appearance of elements of a UI can convey attributions of a certain type of mental imagery. These attributions are action-relevant because they propose the occurrence or afforded potential of some activity, and are also indexed to contextual activity. Because of the situated nature of UI appearances, the interpretation of action-relevant attributes is likely to be highly context dependent. Unlike traditional linguistic approaches, our approach to the semantic content of perception is embodied, i.e., it stresses a) the sensory-motoric experiential background/skills of a perceiver b) the action-oriented bias of perception, c) the perceiver's own activity as an integral part of the perception and d) the situated context of perception. Semiotics of a linguistic tradition tend to consider the relation between appearance and meaning as arbitrary or symbolic (Saussure, 1983). However, the peircean school of semiotics has also acknowledged non-symbolic (iconic, indexical) meaning relations (Peirce, 1998), which clearly are related to the action-oriented perspective because both views imply the existence of "the world" (its appearances and laws) as a familiar semantic reference.

- 4) *Acknowledging the bodily nature of interaction as a basis in design inevitably results in support for multimodal interaction*

If designers take into account our naturally multimodal and action-oriented bias to perceive the world in terms of situationally embodied meanings, it would ultimately provide at least a fraction of the "easiness" of everyday interactions to the design of human-computer interaction. However, such meanings are often invisible to us, and camouflaged as common sense. The challenge for our conceptual analysis is to explicate them.

The claims above define a paradigm for *multimodal interaction design*. The outlined paradigm provides a new conceptual framework for the design of multimodal UIs, which is based on a

sound theoretical foundation. Within the framework, this chapter also explores how to utilise the concept of multimodal mental imagery in UI design. We thus aim to provide conceptual tools to understand the relations between meaningful subjective action-related experiences and concrete physical properties of a UI. In addition to theoretical discussion we expose this suggested paradigm to a real-world design case. The brief case study of sound design for mobile application is meant to illustrate how the new framework is realised in design practices and in the resulting design.

CONCEPTUALISING MULTIMODALITY

The term multimodal and the related, more technical and presentation-oriented concept multimedia have been conceptualised in numerous ways. This section provides a comprehensive summary of how multimodality has been handled in the literature. In addition to these previous accounts, we analyse the concept of multimodality in terms of some recent research on embodied cognition and discuss how multimodality could be conceptualised within that framework. As will be discussed, much of the previous conceptualisations is still relevant in the framework of embodied cognition, but some aspects deserve a critical look.

Perspectives on Multimodality

Multimodality as a Technical Opportunity

In the development of information and communication technology (ICT) products, a typical driver is the emergence of new technical opportunities. Visual displays and the related display processors, for instance, have rapidly developed. However, the development of visual display technology is mainly due to the investments in the research and development of that technology, not the needs of

Bodily Engagement in Multimodal Interaction

multimodal interaction. While the amount of data presented in contemporary displays is several hundred times greater than its 20 year old ancestor, the user of a personal computer has pretty much the same typewriter-derived means to control an application as her parents had in the 70's. Thus it can be seen that the development of technology for multimodal interaction has not been ruled by the needs to enhance human-computer interaction, but the merely commercial assumptions about what consumers want to buy.

Once the technology is there, whatever motivated its development, we have to find uses for it. Much of what is marketed as multimedia are products resulting from this kind of approach. Especially in the early stages of multimedia, the producers were under pressure to show their technical sophistication by supporting all available means of interaction – which, as discussed above, mainly meant ever fancier screen layouts.

As soon as a critical mass had been reached in sales, multimedia products can be argued to have become part of our everyday life. The next step was to elaborate the multimedia technology. An essential part of the elaboration was to legitimate the technology in terms of human-computer interaction. Advantages were sought from multimodal interaction. The models of human cognition, on which the multimodality conception was based, were very simple. A typical example is an idea of a free cognitive resource; for instance when information was presented via a visual display, other sensory systems were thought of as free resources. When this claim was empirically found unsustainable, human ability to process information from multiple sources and in multiple modalities became a central issue. An important source of information was attention studies. In them, human ability to process information had been under intensive research since the 1950's, when the rapidly growing air traffic made the cognitive capacity of air-traffic controllers the bottle-neck of fluent flight organisation. These studies resulted in models which either modelled

the structure of those mechanisms which define attention (Broadbent, 1958; Deutsch & Deutsch, 1963), or models which analysed human capacity (Wickens, 1984).

In mobile applications, the technical challenges for multimodal interaction differ from what they used to be in the static context. However, we argue, even in mobile computing it is the new technical opportunities which are the driving force for developing mobile multimedia. A good example is the current addition of accelerometers to various mobile devices. The popularity of the Wii gaming console with its innovative control methods might have something to do with recently grown interest in accelerometer-based gestural control. Now that we have similar technology included in our mobile phones, interaction designers of mobile applications have been challenged to utilise it. In the near future, we will see whether the application of accelerometer technology in mobile phones turns out to be just another technology driven craze or a useful opportunity resulting in novel ways of interacting with mobile devices.

Multimodality Provides Options

The use of multimedia or designing multimodal applications is often thought of as a selection of means of interaction. For instance, it is easy to find texts which give an impression that a given piece of information can be presented in various forms; e.g., text, speech, picture or video (e.g. Waterworth & Chignell, 1997). The underlying idea is that there is the content and there is the form that is independent of it. However, this notion has been found untenable in various disciplines. In the context of information presentation in UIs, it has been found that paralleling sound and an image, for instance, is extremely complicated. When trying to trace meaning creation on the basis of non-speech sound by asking the participants in an experiment to pair sounds and images, it was found that conclusions could be made only when the images were simple symbols indicating

a clearly identifiable piece of information, such as physical direction (Pirhonen, 2007; Pirhonen & Palomäki, 2008). In other words, the idea that the designer is free to choose in which modality to present certain information is a gross oversimplification and lacks support from the studies concerning semantics. Worn phrases like “the medium is the message” (McLuhan, 1966) or that “a picture is worth a thousand words”, still hold in the multimodal context. Referring to the sub-heading, i.e., multimodality provides options, it should be understood as that technology provides modality-related options but that these options are qualitatively different from each other. The process of choosing an interaction modality is not independent of other design efforts.

Multimodality Provides Redundancy

In mathematical information theory (Shannon & Weaver, 1949), redundancy was introduced as something to get rid of. Redundancy unnecessarily uses the resources of a communication channel, thus lowering the efficiency of an information system. However, in the context of information systems, the concept of redundancy has also more positive connotations; redundancy can be seen as a way of increasing system stability by providing backup. This idea, which originates from the mathematical theory of communication, has been applied to human-computer interaction in the era of multimedia. It has been argued that if information is delivered in multiple formats, the message is more reliably received. A classic example is users with disabilities; if information is provided both in an audio and visual format, for example, the same application can be used by users with vision impairment as well as by those with impairment in hearing (Edwards, 1992).

As discussed in the previous section, the inseparability of form and content of information inevitably questions the endeavour to present “the same” information in multiple formats. However, as the research on cross-modal design of UI feed-

back indicates (e.g., Hoggan & Brewster, 2007), such cross-sensorial information can – to some extent – be defined and thus be interchangeably attributed to multiple formats. Even though these kinds of redundant combinations undoubtedly are beneficial in many applications, they should not be seen as a straightforward, universal solution for multimodal interfaces. But we admit that, in the mobile or ubiquitous context, redundant information presentation can provide valuable flexibility. While the actual context of use is hard or impossible to anticipate, it is important that there are options for interacting with the application. In one situation, visual presentation is the best form, in some other situation, audio or haptics works best.

Natural Interaction is Multimodal

All the approaches discussed above are technically oriented in that they analyse modalities in terms of available technology. However, when conceptualising human-computer interaction from a technical perspective, the essential difference between interaction with the real world and virtual objects has to be noted. When constructing a virtual object we as designers split our mental image of the whole object into its constituents. For instance, when creating a virtual dog we consider separately its appearance, sound and how the user could control it in the application. The division into constituents (in this case image, sound, and control elements) is based on technical facilities. Technically, all constituents are separate entities. Only when linked and synchronised with each other, is the illusion of a virtual dog able to emerge.

In contrast to a virtual dog which is analytically constructed from separate parts, a real world dog is one single physical object. It causes many kinds of perceptions; we can see, hear, and smell it. We can also communicate with it. In other words, we are in multimodal interaction with the dog, even if it doesn't have separate devices to cause stimulus in different sensory modalities or provide control.

Bodily Engagement in Multimodal Interaction

This common sense notion that interaction with the real world is always multimodal by nature, has often been used as a rationale for multimodal UIs in various application areas (see, e.g. Oviatt & Cohen, 2000).

However, the suggested naturalness through multimodality cannot be achieved by burying the application under a heap of visuals, sounds and vibrations. The naturalness can only be achieved by designing objects which are not in the first place “sounds” or “images” or anything else which primarily refers to a certain technology. Multimodality should not be an end in itself. Rather, it should be treated as an inevitable way of interaction. This is important to understand in all human-computer interaction design, because multimodality is a basic quality of our way of interacting with our environment, whether “real” or “virtual”.

An oversimplification, which may result from striving towards natural-like virtual objects, is the mechanical imitation of their real-world counterparts. When creating a multimodal, virtual dog, the best strategy is not necessarily to go with a camera and sound recorder to a real dog – unless you are sure about what kind of recordings you will *exactly* need for your purpose. As will be discussed in the next section, filmmakers have learned long ago that the (action-relevant) expressive qualities of sound are much more important than the “natural” authenticity of the sound source. Moreover, the coincident perception of sound and visual results in an impression which is qualitatively different from a product of its constituents (Basil, 1994). Therefore, the argument that natural interaction is multimodal should not encourage one-sided imitation or modelling of real world objects.

Multimodality Provides a Perceptual Bias

As interaction with the real world always provides multimodal understanding of its objects, actions and environments, in what way, then, would that

affect the “virtual world” where we are technically able to combine different presentation modalities in an artificial manner? Cinema, for example, allows theoretically endless arbitrary defined combinations of sounds and visuals. Thus, an audiovisual relationship in a multimedia presentation is not real but an illusion. Such an option to re-associate images and sounds is essential to the art of filmmaking. A single visual basically affords an infinite number of sounds. For instance, the sound of chopping wood, played in sync with a visual of hitting a baseball, is not perceived as a mistake or as two distinct events. It is heard as a baseball hit – with a particular force. Hence, when the context makes us expect a sound, it seems that the designer can use just about any sound source which produces enough believable acoustic properties for the action.

The phenomenon, where the perceiver is tricked into believing that the artificially made sound effect originates from the source indicated by the context of narration, has a long tradition of exploitation within sound design practices for radio, television and cinema. The French filmmaker and theoretician, Michel Chion (1990), talks about *synchresis* (a combination of words synchronism and synthesis) which refers to the mental fusion of a sound and a visual when these coincide. According to Chion, watching a movie, in general, involves a sort of contract in which we agree to forget that the sound is coming from loudspeakers and the picture from the screen. The spectator considers the elements of sound and image as representing the same entity or phenomenon. The result of such an audio-visual contract is that auditory and visual sensations are reciprocally influenced by each other in perception. From the perspective of the filmmaker, audio-visual contract allows the potential to provide “added value” in the perception of cinema – something bigger than the sum of the technical parts. As a perceptual bias, it induces us to perceive sound and vision as they both fuse into a natural perceptual whole. With his known assertion: “we never see the same thing when we

also hear; we don't hear the same thing when we see as well", Chion (1990, p. xxvi) urges us to go beyond preoccupations such as identifying so-called redundancy between the two presentation modalities and debating which one is the more important modality.

In his account Chion (1990) also proposes the notion of transsensorial perception. He means that it is possible to achieve multimodal or cross-modal perception even with unimodal presentation and via a single sensory path. In music, for example, kinetic, tactile or visual sensations can be transmitted through a sole auditory sensory channel. In cinema, one can "infuse the soundtrack with visuality", and vice versa, images can "inject a sense of the auditory" (Chion, 1990, p. 134). By using both the visual and auditory channels, cinema can also create a wealth of other types of sensations. For example, in the case of kinetic or rhythmic sensations, the decoding should occur "in some region of the brain connected to motor functions" (Chion, 1990, p. 136).

Transsensorial perception as a concept appears to be in accordance with many theories concerning multimodal integration. The classical McGurk effect (McGurck & MacDonald, 1976), for instance, indicates the integration of different sensory modalities in a very early stage of human information processing. Interestingly, the account of audiovisual contract and especially the notions of transsensorial perception can actually be seen as apposite exemplifications of embodied multimodality and the involvement of mental imagery in perception. Next, we will focus on this embodied perspective.

Embodied Meaning

Traditional paradigms of cognitive science have had a tendency to treat the human mind in terms of functional symbol processing. The problem of these computing-oriented views is that they have not been able to explain how experiences of phenomena in the surrounding world are encoded

into symbols of cognition, meaning and into concepts of thinking – not to mention how this is all linked to intentionality and how we are ultimately able to use the cognitive processes in the control of our body (Searle, 2004). As a contrasting top-down approach, scientific perspectives like phenomenology, pragmatism or ecological perception have considered meaning as being based on our interactions with the world rather than as an abstract and separated entity (Dourish, 2001). According to the ecological view of perception (Gibson, 1979), our interaction with the world is full of meanings that we can perceive rapidly without much effort. The perspective of embodied cognition continues such a line of reasoning while stressing the corporeal basis of human cognition and enactive (sensory-motor) coupling between action and thinking (Varela et. al., 1991).

Overall, the perspective of embodied cognition seeks to reveal the role of bodily experience as the core of meaning-creation, i.e., how the body is involved in our thinking. It rejects the traditional Cartesian body-mind separation altogether as the "terms body and mind are simply convenient shorthand ways of identifying aspects of ongoing organism-environment interactions" (Johnson & Rohrer, 2007). Cognition is thus seen as arising inherently from organic processes of organism-environment interaction. So, imagination, meaning, and knowledge are embodied in the way they are structured by our constant encounter and interaction with the world via our body (Lakoff & Johnson, 1999; Gallese & Lakoff, 2005). The perspective of ecological perception is closely related to the embodied one. Both perspectives draw upon the action-oriented bias of the human organism, which considers meanings as action-relevant properties of environment (i.e., affordances) that relate to our experiencing of the world.

We should now see that the embodied point of view defines a coupling between action and perception. So-called motor theories of speech perception (Lieberman & Mattingly, 1985) have suggested that we understand what we hear,

Bodily Engagement in Multimodal Interaction

because we sensory-motorically “resonate” the corresponding vocal action by imaging the way the sound is produced in the vocal tract. It was later found that related motor areas of the brain indeed activate in the course of speech perception (Rizzolatti & Arbib, 1998). Several contemporary studies (see a review in Gallese & Lakoff, 2005) in neuroscience suggest that perception is coupled with action on a neural basis. Discoveries of common neural structures for motor movements and sensory perception have elevated the once speculative approach of motor theories of perception to a more plausible and appealing hypothesis. According to the studies mentioned, all sensory modalities are integrated not only with each other but also together with motor control and control of purposeful actions. As a result, doing something (e.g. grasping or seeing someone grasping) and imaging doing it activates the same parts of the brain.

Gallese and Lakoff (2005) argue that to be able to understand something, one must be able to imagine it, i.e., one understands by mentally *simulating* corresponding activity. In the case of understanding objects of the environment, one simulates how they are involved in actions, or in what way one can afford to act on them for some purpose. In other words, understanding is action-relevant mental imagery, and meaning thus equals the way something is understood in its context. Other authors have also suggested that understanding involves embodied simulation (or mental enactment) of the physical activity we perceive, predict or intend to perform. For example, the account of motor involvement in perception (Wilson & Knoblich, 2005), Godøy’s (2003) account of *motor-mimetic* perception of movement in musical experience, and the theory of enactive perception (Noë, 2004) are all related to the simulation approach. According to Gallese and Lakoff (2005), action simulation seems to occur in relation to 1) motor programmes for successful interaction with objects in locations, 2) intrinsic physical features of objects and motor programmes

(i.e., manners) to act on them to achieve goals, and 3) actions and intentions of others.

The embodied simulation of motor movements of others is found to take place in so-called *mirror neurons*. It is suggested that such an innate mirror system is involved in imitation, as it indeed can explain why it is so natural for us to imitate body movements (Heiser et. al., 2003). Mirror processes, as corporeal attuning to other people, are also hypothesised to act as a basic mechanism for empathy (e.g. Gallese, 2006), i.e., perception of intentionality (mental states such as intentions and emotions). When we observe someone’s action, the understanding of intentionality can be conceived as an emerging effect of the experienced motor resonances through embodied attuning (Leman, 2008). For instance a vocal utterance, as sonic vibrations in the air, “derives” cues of *corporeal intentionality* (or motor intentionality) in the vocal action. Via motor resonances, the perceived cues of the observed action relate to our personal, embodied experiences.

As we can see, the perspective of embodied cognition provides insights at least for the creation and establishing of action-related meanings and concepts. However, it is proposed that even concepts of language could be based on an embodied foundation of action-oriented understanding and multimodal sensory-motor imagery. According to the neural exploitation hypothesis, neural mechanisms originally evolved for sensory-motor integration (which are also present in nonhuman primates) have adapted to serve new roles in reasoning and language while retaining their original functions (Gallese & Lakoff, 2005; Gallese, 2008). The theory of *image schemas* (Johnson, 1987), provides maybe the most prominent perspective on how to characterise our pre-conceptual structures of embodied meaning. Johnson and Rohrer (2007) have summarised image schemas as: “

- 1) recurrent patterns of bodily experience,

- 2) “image”-like in that they preserve the topological structure of the perceptual whole, as evidenced by pattern-completion,
- 3) operating dynamically in and across time,
- 4) realised as activation patterns/contours in and between topologic neural maps,
- 5) structures which link sensory-motor experience to conceptual processes of language, and
- 6) structures which afford “normal” pattern completions that can serve as a basis for inference.”

Image schemas are *directly meaningful structures* of perception and thinking, which are grounded on recurrent experiences: bodily movements in and through space, perceptual interactions and ways of manipulating objects (Hampe, 2005). The shared corporeal nature of the human body, universal properties of the physical environment and common sensory-motor skills for everyday interactions are the providers of a range of experiential invariants.

As any recurrent experiences of everyday interaction can be considered in terms of inference patterns of image schemas, we therefore provide a couple of examples of how these image schemas can be conceived. Image schema can be based on any type of recurrently experienced action, e.g., experience of picking something up. It can thus refer to, for example, forces, kinaesthesia and intentionality experienced in picking up an object in a rising trajectory. Image schema may be based on interpersonal experiences as well. Therefore it can, for example, be related to an experience of illocutionary force (Searle, 1969; see the application in the case study in this chapter) in a vocal utterance – referring to a type of communicative intent in producing that utterance (e.g., asking something).

It is important to remember that image schemas are not equal to concepts nor are they abstract representations of recurrent experiences. Instead, they are “schematic gestalts which capture the

structural contours of sensory-motor experience, integrating information from multiple modalities” (Hampe, 2005, p. 1). Such “experiential gestalts” are “principle means by which we achieve meaning structure. They generate coherence for, establish unity within, and constrain our network of meaning” (Johnson, 1987, p. 41). As such, image schemas operate in between the sensory-motor experience (acted/perceived/imagined) and the conceptualisation of it. However, they are not “lesser” level concepts. On the contrary, being close to the embodied experiences of life, the activation of image schemas in the use of concepts provides an active linkage to the richness of such experiences. Therefore when reading the sentence “Tom took the ball out of the box”, in our mind we can easily complete the “picture” and “see” the instance of someone or ourselves doing just that.

We can now conclude that the creation of embodied meanings is coupled with activity. Perception and imagining of an action is proposed as involving active doing that is dependent on the embodied sensory-motor experiences and skills of a subject. It is suggested that this “internal” activity takes place in embodied simulations that use shared neural substrate with the actual performance of the similar action. It is plausible that these embodied simulations encode both mechanical characteristics of action (properties of objects in locations and motor activity) and structural characteristics of action (gestalt processes such as stream/object segregation or pattern completion) into a sensory-motor neural mechanism. Image schemas are one of the structures by which cognition is linked to sensory-motor neural mechanisms. From that position they allow the embodied experiences to be turned into distinguishable meanings and abstract concepts – and vice versa.

Multimodality is Embodied

The perspective of embodied cognition, which was discussed above, clearly has an effect on the

Bodily Engagement in Multimodal Interaction

conception of multimodality. We should now see that multimodality is a natural and inevitable property of interaction, which emerges from embodied action-oriented processes of doing and perceiving. Theory of embodied simulations suggests that in very early stages of perception, action is being simulated in a corporeal manner. Hence, the perception intrinsically involves participatory simulation of the perceived action, which activates rapidly and is an unconscious process. Such simulation is inherently multimodal because it occurs in the parts of the sensory-motor system which are shared for both perception and doing and which respond to more than one sensory modality. Embodied simulations most likely also involve perceptual gestalt-completions of “typical” and contextually “good” patterns in determining the type of the action and its meanings. Therefore, regardless of the presentation mode, mediated representations—even with limited action-specific cues of, e.g., movement, direction, force, objects properties—can make us simulate/imagine actions that make sense with the current contextual whole. The result of such (amodal) pattern completion is always multimodal experience, even when the stimulus was unimodal – e.g., in the form of written language. Embodied simulations also explain the cross-modal associations based on one presentation modality – for example, why music can create imagery of patterns of movement, body kinaesthesia, force and touch (Godøy, 2004; Tagg, 1992; Chion, 1990). In a similar way, visual presentation can evoke, for example, haptic meanings of textures.

In addition to perception, multimodality takes place in the course of interaction. This is evident in the situations in which one can effortlessly move “across” modalities, for example, when a spoken expression is continued in a subsequent hand gesture, and thus how hand gestures in general are integrated with speech and thought processes (McNeill, 2005). In the context of interaction, embodied simulations of actions should also occur prior to performed actions; when we consider how

we act and imagine what the possible outcomes would be, or when we anticipate actions of others. From the perspective of embodied cognition, it is actually impossible to keep perception fully separated from interaction. We “act out” our perception and we perceive ourselves to guide our actions (Noë, 2004). Thus, our perceptual content is unavoidably affected by the activity of perceiver.

The proposed approach questions some traditional ideas about multimodality and the underlying computational metaphor of human cognition with its separate input and output systems. The embodied view of multimodality clearly denies the existence of separate input/output modules – i.e. separate domains for all senses and motor control which do not integrate until in the presumed higher “associating area” (Gallese & Lakoff, 2005). It is thus inappropriate to consider perception as passive input to the cognitive system, or to consider sensory modalities as channels by which content could be transmitted independently from other senses. Of course, the previous paradigms concerning, for example, attentional capacity and redundant information presentation, should not be completely rejected but they could be understood in a new way in the framework of embodied cognition. Embodied cognition as an approach to design has thus potential in shifting the orientation of traditional, computation-centred models to something which could be called human-centred.

MULTIMODAL INTERACTION DESIGN

What Is “Multimodal” Design?

On the theoretical basis presented earlier in this chapter, we can formulate a new design paradigm which takes into account the natural multimodality of interaction. The proposed design paradigm seeks to support the user in the creation of percep-

tions that would feel natural in the sensory-motor experience of the interaction. In this sense, this paradigm is in accordance with ecological approaches of design (e.g., Norman, 1988). The essence of the multimodal design paradigm, however, lies in its conceptualisation of interaction modalities. Most importantly, this means that multimodality is not something that designers implement in applications. Rather, multimodality is the nature of interaction that designers must take into account. Hence, the multimodal design of UIs does not focus on different presentation or input modalities – nor on any communication technologies in themselves. By acknowledging this, we can begin to appreciate the modally integrated nature of our embodied sensory-motor experience – as a part of natural engagement of action and perceptual content.

What would be the main principles for multimodal design? In fact, the cornerstones of this new paradigm are those proposed already in the introduction of this chapter. Already in the discussion about the ways to conceptualise multimodality, we have concluded that “*interaction is always multimodal in nature*”. As mentioned above, the embodied perspective for multimodality formulates the basic rationale for the paradigm. The second principle, “*design arises from mental images and results in mental images*”, concerns the creation of embodied mental imagery in UI mediated communication. The motivation of this principle is to shift focus from communication channels and symbolic representations of information to embodied meaning-creation based on sensory-motor *attunement* between the user and the UI within the context of use. The starting point of design should be in the flow of interaction itself and the mental exploration of embodied imagery of contextually coherent action-related meanings. Being bonded to embodied experiences, such imagery is inherently multimodal. Therefore design operates on a multimodal foundation – even when its concrete implementation is technically unimodal.

The third principle; “*action-relevant mental imagery can be communicated by manifesting action-relevant attributions in the contextual appearance of UI*”, asserts that a designer can actually utilise action-relevant imagery in finding propositional semantics for UI design (perceptual cues of action including, e.g., cues of movement, objects and intentionality involved in action). The aim of the exploration of mental imagery is ultimately to get ideas for physical cues that associate with the intended action-relevant imagery and thus provide support for the communication of it. For sure, the designer cannot absolutely define such couplings between certain meaning (i.e., mental imagery) and certain characteristics of UI. But these couplings are not completely coincidental either since, due to past interactions with the world, we already possess a rich history of meaningful experiences that can potentially become coupled with the contextual appearance of UI. As these couplings are intentional connections that ultimately arise in the course of interaction (Dourish, 2001), we propose that the designer must immerse herself in (actual or imaginary) participation in interaction in order to find ideas for them. The designer should possess at least a tentative mental model of application-user interaction, from which the need for and intended purpose of a certain UI element has arisen in the first place. By exploring her own relevant experiences, she should be able to pinpoint some presumably “general” action-related associations or their metaphorical extensions which would match with the communicational purpose of a UI element and would support the coherent perceptual whole in the interaction. Such meaningful imagery, as recurrent patterns of experience, is very closely related to the concept of image schema, which we discussed earlier.

The embodied attunement between user and the UI of application is the key in the communication of action-related meanings. When the user encounters the UI element, it should appear as an intentional object which indicates its purpose in

Bodily Engagement in Multimodal Interaction

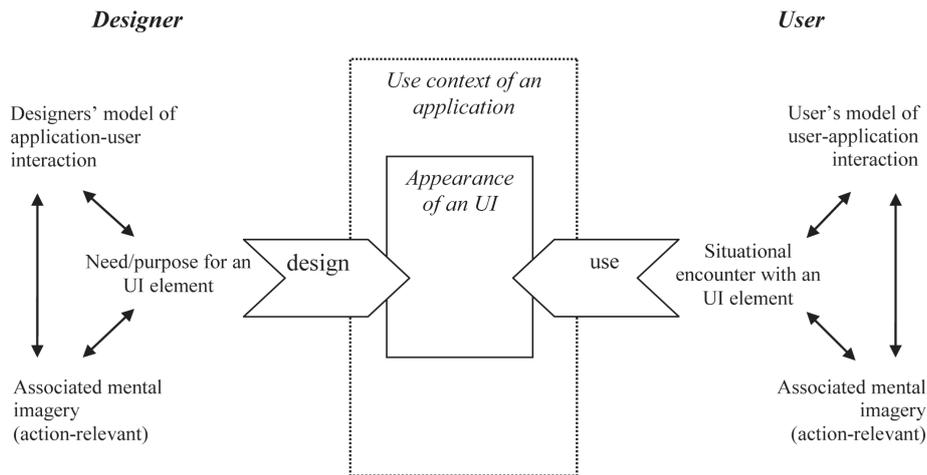
the course of interaction. Action-relevant cues in UI appearances can exploit our natural tendency to attune to environment and its actions. The aim of a designer is to implement such cues that would contextually support the user in creating the intended “feel” and mental imagery. Figure 1 illustrates the UI-mediated communication process of action-relevant mental imagery. The imagery of designer is created in relation to her model of application-user interaction and to the arisen need/purpose of the certain element in UI. Correspondingly, the imagery of a user is created in relation to how she has experientially constructed (i.e., modelled) the interaction and how that certain UI element is encountered in interaction. This is quite analogous with Norman’s (1988) account of system image and the related models, but in our approach, the mental construction covers the whole interaction in the context of use, not the system operation only. Unlike in traditional theory of communication (Shannon & Weaver, 1949), the process is not unidirectional. Both the user and the designer are seen as equally active in creating

mental imagery that contextually makes sense. Even though these two mental constructions would never be precisely the same, the contextual utilisation of common experiential (sensory-motor) invariants can guide the perception of the user be attuned with the designer’s intention.

Mental exploration is a common practice, for instance, in film sound design (Sonnenschein, 2001). The sound designer, having immersed herself in the plot and the narration of the film, reflects her own experiential sensory-motor background in order to mentally “listen” to the sounds in order to determine acoustic invariants that would presumably relate to the preferred subjective experience.

The cues of action-relevant imagery can be encoded into various forms of physical manifestation. Therefore their appearance in UI can, for example, be realised either within a single presentation mode or in co-operative integration of multiple modes. If the propositional mental imagery involved touching a rough-surfaced object, we could equally imagine what kind of

Figure 1. The communication of action-relevant mental imagery via the processes of design and application use



sounds it would produce, what it would feel like (in a haptic sense) and what it would look like. And even if the technical UI element eventually resulted in a unimodal form (e.g., audio feedback), the perception would always be engaged with the multimodal completion of perceptual content. For instance, while sliding our finger on the touch screen, associated audio feedback (indicating a rough surface) can also situationally induce pseudo-haptic perceptions of surface roughness. Radio-plays are also good examples of multimodal completion; they have been called “theatre of the mind” because they allow listeners to use their imagination solely on the basis of sound cues.

However, as in the phenomenon of synchresis, the concurrent contextual framework (e.g. UI appearance of other modes of presentation) will ultimately guide the situated perception of an UI element. Such perceptual modulation is evident, for example, in the findings of Bresciani et. al. (2005) and Guest et. al. (2002). We should therefore pursue integrated design perspectives that take into account all concurrent presentation modalities in the situational context of use of an application. The design of different UI elements should be focused on supporting the construction of coherent imagery as a perceptual whole – not on evoking competing denotations.

One of the advantages of the proposed design paradigm lies in bringing the essence of interaction – action and intentionality – to its central role in perception and meaning creation. If our perceptual content is integrated into our own actions, goal-oriented participation in interaction, we cannot distinctively talk about input and output semantics without acknowledging the embodied situation as a whole. Moreover, if our perceptual content is based on enactive simulation of activity, we should learn to conceive the semantics of HCI as embodied meanings of intentional activity – not just as representations in different channels. The fourth principle of multimodal design sums up the point: “*Acknowledging the bodily nature of*

interaction as a basis in design inevitably results in support for multimodal interaction.”

The proposed design paradigm is nothing completely new; multimodal interaction design is in line with the current development in the philosophical foundation of the HCI-field, emphasising interaction design, user- and context-centred views, user experience, ecological perception, situated actions, and the ideas of tangible interaction. It is largely overlapping with the design principles for embodied interaction, introduced by Paul Dourish (2001). Recent accounts concerning the design of enactive interfaces (see e.g. Götzen et. al., 2008) are also related to the proposed design paradigm. In the following sections, we will further discuss the utilisation of mental imagery in the design of UI elements. The following design case will also demonstrate how our conceptual framework is realised in design. It should be noticed that in the following sections the discussions and examples are biased towards the design of UI presentation/feedback elements. However, it should be seen as a practical restriction of this study and not the restriction of the paradigm itself.

Designing Action-Relevant User Interface Elements

Defining Action Models

In the previous section we already suggested that the designer’s mental imagery arises from her conceptual model of interaction. Imagery is also bound up with the purpose of a UI element. This *communicative function* also arises from the designer’s model of interaction. Indeed, keeping interaction design in mind, we suggest that such a communicative function of a UI-element is the first and foremost factor to be considered in its design. Thus, the first questions to be asked are: Why do we need the UI element? What should be the communicational role of the UI element instance in interaction? And in what way is it used in the model/flow of interaction?

Bodily Engagement in Multimodal Interaction

In considering the above questions, we should be able to *imagine* the flow of application-user interaction and explore different associations that are related to the role and the characteristics of the propositional UI element. This can be achieved, for example, with participatory exploration of use scenarios. These associations are what we call action-relevant mental imagery or embodied meanings, but they could as well be called design ideas or design concepts for an element. Associated mental imagery does not always have to refer to actual HCI context. For the sake of creativity, it is actually important to also explore non-HCI associations which arise from the interaction model – or from the interaction in some other, analogous, context. In the latter case, it is a question of utilising metaphors in design.

When starting to design UI elements with certain action-relevance we need to outline the activity to which this relevance refers. Such a conceptualisation of action involves a mental image of generalised and contextually plausible action, which we will call the *action model*. It is an image-like mental model of sensory-motor experience that involves the general type of activity (forces, movement, objects) and its general type of occurrence. In addition, it may involve the general type of intentionality, i.e., causality between mental states and the action. Like image schemas, action models refer to recurrent patterns of experience. Action models can be conceived as design ideas that guide the designer in attributing certain action-relevant semantics to the design. Together with the designer's model of interaction, they work as high-level mental conceptions, against which the design is reflected when considering its action-relevance.

In our design case (reported later in this chapter in more detail) we ended up using three different action models as three different approaches to the design of the same auditory feedback. These models referred to general types of actions in general types of occurrence like “connection of an object with a restraining object”. It is important to notice

that if more than one action model is used, they all should support the construction of the intended perceptual whole, that is, meanings that become fused into the action experience.

When considering UI presentation, it is appropriate to make a distinction between natural and artificial action-presentation relationships, similar to the action-sound distinction made by Jensenius (2007). In fact, every action-presentation relationship in UI is essentially artificial. However, like in Chion's (1990) synchresis, the relationship can be experienced as natural because of UI presentation's plausible action-relevance in the situational context on which it is mapped. Hence the instance of UI presentation can become perceptually fused with concurrent actions and result in an illusion of naturalness.

The continuum from natural to artificial operates in two dimensions. Firstly, it concerns the relationship between the action model and its context in the interaction model (see Figure 2) and secondly, the relationship between the action model and its degree of realisation in presentation (see Figure 3). These relationships in the first dimension can involve direct, indirect or arbitrary mappings. Indirect mappings involve some kind of mental mediator, for example, use of metaphors or analogical thinking. Arbitrary mappings always involve the use of some sort of code.

Let us consider the design task of creating feedback for controlling certain parameter on a touch screen. The change of parameter value needs to be done in discrete steps – either upwards or downwards. For each change operation, the feedback should be robust enough so that the user can be confident of success. One straightforward way to schematise such a control would be based on mappings between upward or downward stroking on a touch screen surface and changes in a parameter value. Based on that, we could adopt the real physical involvement of “stroking with a finger upwards or downwards” directly as an action model for feedback. But in order to support the required robustness and the feel of control, we

Figure 2. Mappings between the action model and the interaction model

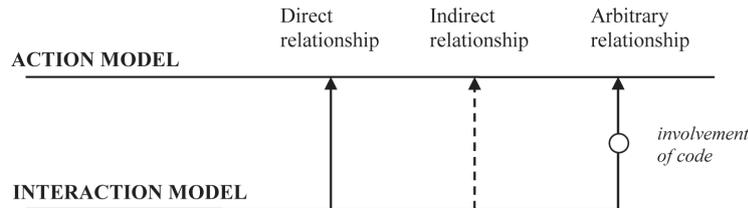
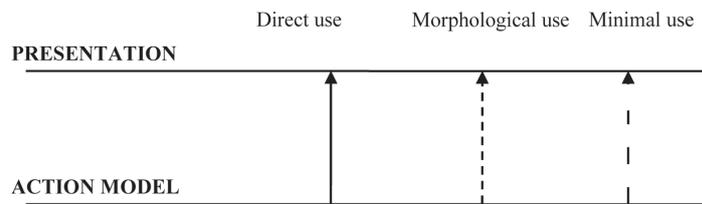


Figure 3. Relationship between the action model and its use in UI presentation



could outline the action model metaphorically as, for example “rotating a resisting wheel upwards or downwards until it snaps into a retainer that holds it”. From such an action model, it would be easy to imagine visual, haptic or sonic feedback that would feel functional and natural (offering the sensations of mass, resistance and restraint) when being coupled with stroking on the screen surface. For the interaction model of discrete changes, we could also adopt a convention of pressing a “plus-button” for increase and pressing a “minus-button” for decrease. Hence the relation of the action model “pressing a button” and the parameter control would be arbitrarily defined.

As will be shown in our design case, for the design of feedback elements it may also be appropriate to explore action models from the perspective of interpersonal interaction metaphor. The active subject would be an imaginary “agent”, i.e., the intentional (2nd person) counterpart of interaction. In feedback element design this can

be a very fruitful approach because it would allow the attributions of 2nd person response.

In general, the use of multiple action models is able to produce more interesting results in design than use of a single model, as it can provide multiple layers of action-based semantics. Because the degree and the means of utilisation of each model in realised presentation varies, it should be possible to implement the manifestations of more than one action model in a single presentation element. When considering multiple action models in design, they can also be allocated for use with separate presentation modalities.

The second dimension, relationships between the action model and its attributes implemented in presentation, involve quite different kinds of relations than the first dimension. Of course, before these general-level action models can be used in design, they have to be manifested in a concrete form. The second dimension operates with considerations of “how much” and “in what

Bodily Engagement in Multimodal Interaction

way” these manifested attributes of action models are utilised in the realised UI element. Roughly defined categories of the second dimension include 1) *direct use*, 2) *morphological use* and 3) *minimal use* of the action model manifestation. These categories are inspired by similar use of model appropriation in the composition of electroacoustic music (Garcia, 2000). Direct use means that the presentation element is a partial or complete reconstruction of the action model manifestation. Morphological use means that the model and its manifestation are taken as a formal structure, i.e., the designer extracts some organisational laws of the model/manifestation and utilises these features morphologically in the design. Minimal use means that only such reduced or general features or cues of the model/manifestation have been utilised in the design which do not appear as resemblances between the presentation and the model/manifestation. In the next section, we will discuss the manifestation of action models.

Manifesting Action-Relevant Mental Imagery

Godøy (2001) has proposed a model of musical imagery where our understanding of (sound-producing) actions is founded on sensory-motor images of *excitation* (imagery about what we do or imagine/mimic doing, e.g., body movement or gestures) and *resonance* (imagery about the effects, i.e. material resonances, of what we do or imagine/mimic doing). Because of the multimodal, holistic nature of such mental images, this model should be easily adaptable to the multimodal design perspective, in which it can be used as a guide to mentally *enact* motor movements and explore the reflecting material/object-related imagery. This kind of “manifesting by enacting” takes the action model as a starting point and aims to elaborate it into a more detailed scenario of action and its resonances. The result of this creative process should include tangible ideas for concrete actions; gestural imagery,

action-relevant mechanical invariants produced by certain materials and objects within interaction and, more generally, a “scheme” of the action that can be consequently articulated in physical form. Physical articulation can aim at analytic realisation of action imagery (e.g., manually or by physical modelling), or it can adopt some structures or ideas from the action imagery and interpret them in a more “artistic” manner. For example, an action model of “letting loose” might be articulated as short drum tremolo, which retains the feeling of “removing a restraint” and a decaying contour of force felt in the imagined action.

To complement Godøy’s model presented above, we suggest an additional type of imagery which provides the motivational basis for the other two. Images of *intentionality* concern goals, motivations and also emotions behind the perceived/imaginary action. The exploration of intentionality behind action should be based on the designer’s considerations of the functional purpose of a UI element in its context (see the earlier discussion about communicative function).

UI elements are intentional objects which, together with their contextual appearance, should indicate their functional purpose. Human perception seems to have a tendency to denote purposes or motivations behind the perceived actions. As perception involves a mirror mechanism – the attunement to the actions of other people – we thus have a corporeal apparatus for “catching” the cues of intentionality. The attribution of intentionality can also be extended to the perception of material movement, providing that we can imagine an intentional subject being involved in the action. (Leman, 2008.) Therefore UI elements can convey intentionality and can be understood as intentional objects, even if they do not directly denote the presence of another person. It is up to the designer to make sure that UI elements comprise suitable attributions of, for example, communicative intent. In other words, the designer must ensure that the imaginary action and its physical articulation is an outcome of intention-in-action (Searle, 1983)

that matches the propositional communicative purpose of a UI element.

Physical articulations that spontaneously occur as the result of direct corporeal involvement in interaction can be called 2nd person descriptions of the subject's intentionality. Such articulations naturally occur, for example, in person-to-person ("from me to you") communication. As they reflect direct involvement in interaction, they are not a result of subjective interpretation but an "experience as articulated". As gestural manifestations of corporeal intentionality, they provide cues of intentionality in the form of physical features and thus offer a way to find correlates between subjective experiences and objective physical properties. (Leman, 2008.) It would be valuable, if stereotypical physical cues of specific intention to communicate could be captured in corporeal articulations. In other words, the use of such bodily determined invariants in the appearances of UI would form an intention-specific semantic basis. Intention-specific cues could be utilised to facilitate HCI – including feedback presentation and gesture recognition. In our design case, we used an action model that was based on a person-to-person interaction metaphor. The action model involved a vocal speech act with a specific communicative intent. Prior to spontaneous vocal articulations, we first explored various mental images of suitable vocal gestures in order to become situationally immersed. From recorded articulations we aimed to capture features of physical energy that would reflect the communicative intent in the speech act (i.e., the illocutionary force). In the design, these acoustic features were thus utilised as cues of corporeal intentionality. A similar type of gestural imagery and its physical articulation has already been utilised in the design of non-speech audio feedback elements (Tuuri & Eerola, 2008).

Despite the fact that the core of design process operates at a highly subjective level, the results of this process must always be transferred into a form of observable features. From the perspective of design methodology, the problem is the

gap between subjective (1st person) ontology of experience and objective (3rd person) ontology of physical appearance, which makes the formulation of systematic methods very difficult. We see that at least one answer to the problem lies in spontaneous physical articulation of design ideas. As an interaction-derived, action-related imagery of the designer is manifested and expressed in physical terms, the structure and features of the physical phenomenon can be analysed and later utilised systematically in design. At the same time, the design idea can more easily be sketched and communicated within a design team.

UI Element Design as a Creative Process

To sum up the design process and utilisation of action relevant mental images at a general level, we decided to categorise the different design phases by conforming to the classic model of the creative process. Traditionally the process is described in five phases: 1) preparation, where the problem or goal is acknowledged and studied, 2) incubation, where the ideas are unconsciously processed, 3) insight, where the idea of a plausible solution emerges, 4) evaluation, where the idea is somehow concretised and exposed to criticism and 5) elaboration, where the idea is refined and implemented (Csikszentmihalyi, 1996). Our modified process includes four phases: 1) *defining the communicative function for the UI element*, which matches fairly well the classic goal defining and preparation phase, 2) *defining action models and exploring action-relevant mental images*, which comprises recursive incubation and insight phases with the aim of elaborating action-relevant imagery, 3) *articulation*, which refers to concretisation and evaluation of ideas, and finally 4) *implementation of the UI element*, which is the phase where the concretisations of action-related ideas are composed into the final appearance of a UI-element. Of course, this kind of process model is always a crude simplification of the real processes. For

Bodily Engagement in Multimodal Interaction

example, in actual design there is not always a clear boundary between mental exploration and its articulation.

In the sections above, we have introduced some conceptual tools, such as the usage of action models, which we consider essential in applying the action-oriented multimodal perspective in the UI element design. We will next expose this theoretical foundation to the creative process of real-world interaction design by presenting a case report.

Design Case: Physical Browsing Application

Background

The aim of the case was to design an appropriate sound feedback element to complement the experience of “physical selection” in the physical browsing application. Physical browsing is a means of mapping physical objects to digital information about them (Välkkynen, 2007). It is analogous to the World Wide Web: the user can physically select, or “click”, links in the immediate environment by touching or pointing at them with a mobile terminal. The enabling technology for this is tags that contain the information – for example web addresses – related to the objects to which they are attached. The tags and their readers can be implemented with several technologies. In this application, we chose RFID (Radio-Frequency Identification) as the implementation technology. The tags can be placed in any objects and locations in the physical environment, creating a connection between the physical and digital entities (Want et. al., 1999).

Because of the close coupling between physical and digital, the properties and affordances (Norman, 1999) of the physical environment – namely the context, the appearance of the object and the visualisation of the tag – form a connection with the digital content of the physical hyperlink. Each part of this aggregate should complement each

other. The same holds for the interaction with the tags and the links they provide: optimally the user feels as if she is interacting with the physical objects themselves with a “magic wand”. This kind of system is very close to tangible user interfaces (Holmquist et. al., 1999; Ishii et. al., 1997) even while we use a mediating device to interact with the physical objects.

While we can call the short-range selection method “touching”, the reading range is not actually zero, that is, the reading device and the tag do not need to be in physical contact with each other (see Figure 4). The reading range is typically a few centimetres and because it is based on generating an electromagnetic field around the reader, it varies slightly in a way that does not seem deterministic to the user. Typically, the hardware and software combinations only report a tag proximity when the contents of the tag are already read, which means that guiding the user in bringing the reader slightly closer to the tag is not generally possible – we only know if a tag has been read or if it has not.

Because the physical action of selecting a tag is a process that happens in three-dimensional space, during tag reading the mobile device is often at an angle where it is difficult to see what is happening

Figure 4. Selecting an RFID tag with a reading device in physical browsing



on the screen, and an arm's length away from the user's eyes. Therefore, other modalities, such as sounds, are especially important in reporting the successful reading of a tag.

Design Process

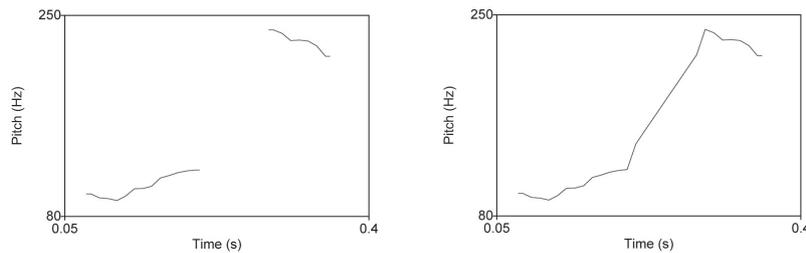
The design goal was fairly straightforward, as the propositional purpose of sound was to provide feedback to the user about successful reading of an RFID tag in the physical selecting. However, in this early phase of design it was important to formulate a more detailed picture of what the actual constitutes of this *feedback function* are in the current context of interaction. Hence we created and, by imaginary participation, explored a contextually rich use scenario about the usage of a physical browsing application (for details see Pirhonen et. al., 2007) and ended up with design principles that determine some important qualitative aspects of the required feedback. Most importantly, sound should clearly illustrate, in real time, the selecting/reading of the tag. For the user, the sound would thus appear as “the” sound of that particular event of virtual touching, providing confirmation of success and a feel of control. As a consequence, the feedback sound would inform the user that physical pointing towards the tag is no longer needed. Feedback should also “answer” the user's intentions in selecting this particular tag (from many) and in wishing to peek at what kind of content the tag holds. To achieve these goals, sound design should evoke confirmatory associations in the user that are contextually suited to such intentions and the user's conception of the tag selecting action. As a secondary purpose, the sound should discriminate different tag types from each other. In this particular application there were four types of RFID tags referring to the type of media content they hold (text, picture, sound or video). However, this second purpose of the feedback sound had much lower priority in actual design considerations than the main purpose presented above.

Relating to a separate study of creative group work in design (Pirhonen et. al., 2007), we started the design by conducting three iterative panel-working sessions with six panellists. In the first panel session the aim was to immerse the panellists in the use of the physical browsing application. Such participatory experience was supported by a detailed story about a person using an application. The story was narrated as a radio-play and was based on the rich use scenario mentioned earlier. Panellists thus got the opportunity to enact contextual instances of physical selecting of RFID tags, and to explore what kind of imaginary sound they heard in association with the experience. Hence, the result of the first panel session was a list of verbally expressed ideas about the sound. These ideas fell roughly in two categories: sounds of connections (such as “clicks” and “dings”) and sounds with an upward gestural motor image (such as a rising two-tone melody). By mental exploration, we interpreted the first category as sound producing actions of “*connecting an object with a restraining object*” and the second category in terms of imagery of “*picking something up (out of the container)*”. These two general types of action experiences were adopted as action models for the design. Both models have an indirect relationship with the actual doing of physical selecting, although the model of “connecting” has a much closer analogy with the physical selecting than the latter model, which is based on a gestural metaphor.

On the basis of both action models and other ideas, we made various types of sounds as recorded articulations of suitable sound-producing actions and as synthetically generated sounds. In the consequent panel sessions, these sounds were evaluated by the same panellists. The most liked one was the sound of a single metronome “click”. A high-pitched two-tone melody of rapid upward 5th interval (played on a metallophone) was also liked, but because of its long decay time it was considered “a little too long”.

Bodily Engagement in Multimodal Interaction

Figure 5. Visualised pitch (F_0) contours of the selected utterance. The original segmented contour is on the left, and the contour with an interpolated gap on the right. The interpolated contour was used in the design.



After the panel sessions, we also came up with the idea of exploring person-to-person interaction metaphors for an additional action model. We made analogies to various interpersonal communication scenarios that involve similar kind of “pointing” to an object, and the related 2nd person utterance as confirmatory feedback. We defined the third action model as a vocal gesture of “asking for confirmation” – with a polite and confidence-supporting intention such as “So, you mean this one?” The essential element of this action model was communicative intent in the speech act (i.e., the illocutionary force), not the verbal content. Suitable action relevant utterances, using a Finnish word meaning “this one?” (as a rhetorical question), were explored and spontaneously articulated the imaginary interaction scenario in mind. Articulated utterances were recorded and analysed acoustically for their prosodic features, namely fundamental frequency (F_0), intensity and first formant (F_1), to be utilised in the design. The presumption was that such prosodic features would carry cues of the corporeal intentionality of the vocal action (Leman, 2008; Banse & Scherer, 1996), allowing its intention to communicate to be implemented in design. The extraction of prosodic features was performed with Praat software (see Boersma, 2001). After some preliminary listening to the synthesised versions of F_0 pitch contours

of the utterances, one utterance was chosen as a basis for the subsequent sound design because we felt it supported the upward gestural imagery similar to the “picking something up” model. Figure 5 illustrates the original and interpolated pitch contours of the selected utterance.

We preferred the interpolated version of the pitch contour, as it provided a solid, undivided basis for sound synthesis. We then shifted all pitch values into a higher register (at the rate of 2.7), before producing renditions of the contour. Four versions were synthesised. Each version used a different type of waveform (triangle, sinusoid, sawtooth, square) resulting in different timbres. This simple feature was intended to support the distinguishing of different tag types. Next, the extracted intensity contour was applied to all four sounds in order to provide them with the dynamics of the original utterance. Likewise, all four sounds were filtered using the extracted information of the first formant (F_1) in the original utterance. It is worth noting that although formant filtering (with one formant) was used, the aim was not to reproduce speech-like sounds. However, filtering did provide the sounds with certain human-like qualities.

Finally, three different sound elements, which originated from corresponding action models, were mixed together in a multi-track audio editor.

Table 1. Summary of action models and their usage in the design of feedback sound

ACTION MODEL:	<i>Connecting an object with a restraining object</i>	<i>Picking something up</i>	<i>Asking confirmation (speech act)</i>
Relationship with an action of physical selecting:	Analogy of real touching instead of virtual	Based on a gestural metaphor (refers to "picking up" the content from the tag)	Based on a person-to-person interaction metaphor
Usage in the UI presentation:	Direct usage (plausible "click" sound for object-object collision)	Minimal usage (action model as a basis for musical articulation, which was partially used)	Morphological usage (acoustic structure of an utterance was used)
Propositional type of semantics:	Object-interaction resonance	Motor-mimetic image of movement (upwards)	Cues of corporeal intentionality (via motor-mimetic image)

These source elements were 1) the metronome click, which provided the proper qualities of physical connection to the beginning of the UI feedback, 2) the two-note metallophone melody, which was shortened by fading out the long-sounding decay and used in the beginning of the sound to provide damped metallic hits, and 3) the synthesised sound we produced on the basis of extracted prosodic information. The third source element provided the body element for the UI feedback sound, also providing the only difference between feedback sounds for different tag types. As a design approach, the use of the acoustic structure of a vocal act in sound design is analogous to parameterised auditory icons (Gaver, 1993). In the editing process, source elements were transformed in various ways (e.g. balanced, filtered and reverberated) in order for the result to be aesthetically pleasing and functional sound. Action models and their use in the design are summarised in Table 1.

Evaluation of the Design

So far we have discussed the physical browsing application without a context. Next we will outline a physical environment in which we evaluated the feedback sound design. We then report some usability issues of the evaluation study we conducted.

The system was used in a bicycle exhibition in a museum. The museum was situated in a former

engineering works¹ thus providing a large open space for exhibition. The physical browsing system consisted of a palmtop computer (Personal Data Assistant, PDA) with wired headphones and an RFID reader integrated into a PDA. The PDA communicates with RFID tags, which are read by touching the tag with the PDA, that is, bringing the RFID reader within a few centimetres of the tag. The contents linked to the RFID tags were videos, sound files, images and short textual descriptions about the artefacts and themes in the exhibition. The software system consists of a web browser, media player, image viewer and RFID reading software. The RFID reading software is responsible for

1. detecting the proximity of the RFID tags,
2. determining the media type corresponding to the tag,
3. playing the sound feedback, and
4. presenting the media in the appropriate media display software.

Ten volunteers participated in the evaluation test. They explored the exhibition just as any other museum visitor carrying a PDA and headphones. During usage, the participants were observed and after the exhibition visit they were interviewed. The participants were asked to rate the suitability and the pleasantness of the sound feedback on a scale from 1 to 5. Then they were interviewed

Bodily Engagement in Multimodal Interaction

about the qualitative aspects of the sound. The participants were also asked whether they noticed the different types of feedback sounds (4), and whether those sounds helped to distinguish the associated media types from each other.

On a scale from 1 to 5, the average rating for suitability was 4.4 and for pleasantness 3.9. The sounds were considered especially clear and intelligible, and several subjects mentioned that the sounds were not irritating. In a positive sense, the sounds were also characterised as “not a typical sound signal” and as “friendlier than a basic beep”. Other qualities of the sound mentioned were “soft” and “moving”, and that the sound “corresponded well with the physical action” of reading the tag.

The sounds seemed to work very well as feedback. Most of the subjects soon stopped paying attention to the sounds as separate user interface entities and took them as a natural part of the physical user interface. After use, the participants considered the sounds to be an important response that made physical browsing easier. One subject summarised the significance of the sound as follows: “because the touching range varied slightly, and the loading of the content took some time, the sounds were important feedback for knowing when touching the tag had succeeded”.

None of the participants noticed any difference between the sounds of the different media types. So, in the use context of an application, slight timbre modification for sounds of different media types evidently did not provide any difference in the feedback experience. This is an interesting observation because, although the sounds are undoubtedly similar, the difference between the sounds is clearly noticeable when comparing them directly to each other.

Illustrating the New Paradigm

The design task of the reported case was deliberately simple, even trivial. This was because a simple case made it possible to illustrate the con-

tribution of the new paradigm at a more detailed level than a complex one. In this chapter, we are not proposing any special design method for any special need nor trying to prove its effectiveness. Rather, we are proposing a new *mental framework* to approach UI design. To illustrate that approach, the case study presented provides a clear and down-to-earth example. The fact that the design task involved a technically unimodal result was also a deliberate choice. In this way we intend to demonstrate that the proposed paradigm sees multimodality as a quality of the design process, not as multiple technical modalities in the resulting design. It can be argued that a professional designer could end up with similar results. We do not deny that. It is true that designers may more or less implicitly follow design practices which are very similar to the ones proposed. However, as mentioned at the beginning of the chapter, one of our challenges is to explicate such tacit knowledge that designers have.

The embodied perspective of multimodal design stresses the coherency of the perceptual whole which the user actively constructs in interaction. In the illustrated sound design case, the goal was to produce the sonic appearance of reading the RFID tag with a mobile device. From the viewpoint of multimodal design, the most essential issue is how this sound would be perceived within the embodied tag-reading experience. In order to achieve a design that would be coherent with the embodied experience of user, the communicative function, action models and design principles in general were built on the enactive exploration of interaction. We demonstrated the explicit usage of action models as mental tools in pinpointing action-relevant imagery to be utilised as propositional perceptual content (semantics). Hence we utilised three types of (non-linguistic) propositional semantics for a simple confirmatory feedback function. Analytic use of conceptual action models allowed these attributions/appearances to be consciously explored and explicitly implemented in the design.

Despite the unimodal presentation of the resulting design, we argue that the result is perceptually multimodal. Firstly, the situated instance of sonic appearance is intended to be coupled with motor movements of the user resulting in a situational pseudo-haptic illusion of a touch. Secondly, the feedback sound itself is designed with the intention of supporting the user in attuning to “movement in sound” and creating action-oriented sensory-motor illusions. Within the traditional conceptualisation of multimodality in HCI, a feedback event would be easily considered merely in terms of transmitting the feedback information concurrently in sonic and visual (tiny PDA screen) channels. Disembodiment from the interaction can also lead to oversimplified considerations of semantic content. The designer could think, for example, that feedback sounds can be directly recycled from application to application since they are designed to convey the same feedback “message”. This kind of “absolute” view on semantics takes a model of interaction and its relation to a UI element as self-evident, and easily dismisses the communicational potential that even a simple feedback sound can have when it is designed *as* activity – and *for* activity.

CHALLENGES AHEAD

As stated already, multimodality should not be understood as a design option, but as a basic quality of human behaviour. Thus the design process is multimodal, as well as interaction between humans, and interaction between a human being and a technical device.

If human-computer interaction has always been multimodal and will always be, why even pay attention to this kind of inevitable phenomenon? Let us think about the current human-computer interaction paradigm. It was largely developed in the era when a computer meant a desktop workstation. The interaction concepts were based on the framework which was defined by the desktop

unit with its software, keyboard - later mouse- and essentially a visual display. The tradition of elaborating interaction design was inevitably connected to the existing hardware. Therefore, the conceptual framework of HCI is still largely related to past technology.

The main contribution of the current work is to provide a new conceptual basis for the multimodal design of multimodal user-interfaces. By using the ideas of Thomas Kuhn (1970), it can be said that we aim at defining a design paradigm for future needs. We argue that the concepts, or perspective, or paradigm, of the past is simply not adequate for future needs. Kuhn argued that the paradigm defines what we see and hear – it is the framework for figuring out our environment. If the paradigm of HCI is inappropriate, we as designers may fail to see something essential about the use of technical devices. Likewise, Lakoff and Johnson (1980), quite some time ago in their famous metaphor book, argued that our metaphors define what we perceive. Metaphors, in this context, mean for instance the terms which we use for conceptualising human-computer interaction. In other words, we aim at providing a conceptual framework which would serve as a highly appropriate and theoretically founded interaction design basis for future technology. The proposed conceptual framework would reveal issues which would otherwise stay hidden. It provides a perspective, which is applicable when shifting from desktop applications to mobile and ubiquitous computing, or any other novel technical setting.

A closely related design issue is the largely evolutionary nature of the development of high technology. Each new generation of technical applications inherits many of the properties of the preceding one. The way we try to interact with a new device is based on our previous experiences. Therefore it is understandable that the industry often tries to soften the introduction of new applications by making them resemble familiar products. However, evolutionary development of technology may also prevent really new ideas from

Bodily Engagement in Multimodal Interaction

being implemented. In the design of multimodal user-interfaces, an evolutionary approach implies the acceptance of the basic structure of past user interfaces. A well-known example is the so-called qwerty-keyboard as the dominating input device. Even if there is no other rationale but tradition to keep it in the form it was developed for mechanical typewriters, not too many manufacturers have dared to make any modifications to it. A similar reluctance to change the form of a technical device can be seen in numerous product categories. When digital photography freed camera designers from the technical constraints of a film roll, the whole organisation of lens, viewfinder and control devices could be rethought afresh, a “tabula rasa”. Some unprejudiced studies of the digital camera were released then. Those products made brand new ways of composing and shooting possible. Unfortunately, a very conservative design has dominated ever since.

It can be argued, that conservative design dominates because consumers are conservative. We would rather not blame only consumers. Sometimes novel ideas have been commercial successes. Apple, with its innovative graphical UI and portable devices with revolutionary interaction concepts, has proved that consumers are ready to accept new ideas as well. This shows that consumers are ready to accept brand new concepts as well as modifications of existing products.

Since computation can nowadays be embedded almost anywhere, the future of computing can be expected to be truly ubiquitous by nature. In ubiquitous computing, our everyday environments are part of the user interface, and there is no clear distinction between different applications within a single physical space. This emerging setting challenges design. Interaction with different ubiquitous applications cannot rely on the models which we adopted in the era of personal computers. In so-called smart environments, technology should support more natural interaction than, for instance, typing. Understanding our naturally multimodal way of interacting with our environ-

ment would be essential in order to design usable ubiquitous applications. The physical browsing case is an early and simple example of embedding the interface – and multimodal interaction – into the environment.

When designing the products of the future, we will need to be able to take a fresh look at what people really need. The elaboration of existing technological conventions (i.e., evolutionary approach) is not necessarily adequate. By understanding the central role of bodily experience in the creation of meaning, designers would be able to get to grips with something very essential in interaction between a human being and a technical device. This, in turn, would help to create product concepts which genuinely utilise the available technology to fulfil human needs.

CONCLUDING STATEMENTS

The prevailing interaction design paradigm inherits its basic concepts from an information processing model of human cognition. In practical design, this approach has proved to be inadequate in conceptualising interaction of a human being and a technical device.

New ideas about interaction design, which challenge the legacy of the mechanical view of human cognition, have already emerged. However, the new approaches have not yet managed to formulate a concise framework. In this chapter, we have proposed a new paradigm for interaction design. It is based on the idea of the human being as an intentional actor, who constantly constructs meanings through inherently multimodal bodily experiences. The case study illustrates that the proposed approach provides a relevant conceptual framework for understanding interaction with a technical device, as well as for understanding the design process. The evaluation of the design case indicates that the implemented sounds reflected essential properties of intentional action in the given context.

The proposed approach does not contain detailed guidelines for successful design. Rather, it is a sound, theoretically founded conceptual framework. As such, it provides designers with concepts which should orientate the design process in a relevant direction in terms of human intentionality in interaction.

ACKNOWLEDGMENT

This work is funded by Finnish Funding Agency for Technology and Innovation, and the following partners: Nokia Ltd., GE Healthcare Finland Ltd., Sunit Ltd., Suunto Ltd., and Tampere City Council.

REFERENCES

- Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*, 614–636. doi:10.1037/0022-3514.70.3.614
- Basil, M. D. (1994). Multiple resource theory I: Application to television viewing. *Communication Research*, *21*(2), 177–207. doi:10.1177/009365094021002003
- Bernsen, N. O. (1995). A Toolbox of output modalities: representing output information in multimodal interfaces. *Esprit Basic Research Action 7040: The Amodeus Project*, document TM/WP21.
- Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glott International*, *5*(9/10), 341–345.
- Bresciani, J.-P., Ernst, M. O., Drewing, K., Boyer, G., Maury, V., & Kheddar, A. (2005). Feeling what you hear: auditory signals can modulate tactile tap perception. *Experimental Brain Research*, *162*(2), 172–180. doi:10.1007/s00221-004-2128-2
- Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon.
- Cañadas-Quesada, F. J., & Reyes-Lecuona, A. (2006). Improvement of perceived stiffness using auditory stimuli in haptic virtual reality. In [Piscataway, NJ: IEEE Mediterranean.]. *Proceedings of Electrotechnical Conference, MELECON, 2006*, 462–465. doi:10.1109/MELCON.2006.1653138
- Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Chion, M. (1990). *Audio-vision: Sound on screen*. New York: Columbia University Press.
- Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.
- Csikszentmihalyi, M. (1996). *Creativity: Flow and the psychology of discovery and invention*. New York: HarperCollins.
- de Götzen, A., Mion, L., Avanzini, F., & Serafin, S. (2008). Multimodal design for enactive toys. In R. Kronland-Martinet, S. Ystad, & K. Jensen (Eds.), *CMMR 2007, LNCS 4969* (pp. 212–222). Berlin, Germany: Springer-Verlag.
- de Saussure, F. (1983). *Course in general linguistics* (R. Harris, ed.). London: Duckworth. (Original work publishes in 1916).
- Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, *70*(1), 80–90. doi:10.1037/h0039515
- Dourish, P. (2001). *Where the action is: The foundations of embodied interaction*. Cambridge, MA: MIT Press.
- Edwards, A. D. N. (1992). Redundancy and adaptability. In A. D. N. Edwards & S. Holland (Eds.), *Multimedia interface design in education* (pp. 145–155). Berlin, Germany: Springer-Verlag.

Bodily Engagement in Multimodal Interaction

- Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.
- Gallese, V. (2006). Embodied simulation: From mirror neuron systems to interpersonal relations. In G. Bock & J. Goode (Eds.), *Empathy and Fairness* (pp. 3-19). Chichester, UK: Wiley.
- Gallese, V. (2008). Mirror neurons and the social nature of language: The neural exploitation hypothesis. *Social Neuroscience*, 3(3), 317–333. doi:10.1080/17470910701563608
- Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in reason and language. *Cognitive Neuropsychology*, 22, 455–479. doi:10.1080/02643290442000310
- Garcia, D. (2000). Sound models, metaphor and mimesis in the composition of electroacoustic music. In *Proceedings of the 7th Brazilian Symposium on Computer Music*. Curitiba, Brasil: Universidade Federal do Paraná.
- Gaver, W. W. (1993). Synthesizing auditory icons. In [New York: ACM.]. *Proceedings of INTERCHI*, 93, 228–235.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.
- Godøy, R. I. (2001). Imagined action, excitation, and resonance. In R. I. Godøy, & H. Jørgensen (Eds.), *Musical imagery* (pp. 237-250). Lisse, The Netherlands: Swets and Zeitlinger.
- Godøy, R. I. (2003). Motor-mimetic music cognition. *Leonardo*, 36(4), 317–319. doi:10.1162/002409403322258781
- Godøy, R. I. (2004). Gestural imagery in the service of musical imagery. In A. Camurri & G. Volpe (Eds.), *Gesture-Based Communication in Human-Computer Interaction: 5th International Gesture Workshop, Volume LNAI 2915* (pp. 55-62). Berlin, Germany: Springer-Verlag.
- Guest, S., Catmur, C., Lloyd, D., & Spence, C. (2002). Audiotactile interactions in roughness perception. *Experimental Brain Research*, 146(2), 161–171. doi:10.1007/s00221-002-1164-z
- Hampe, B. (2005). Image schemas in cognitive linguistics: Introduction. In B. Hampe (Ed.), *From Perception to Meaning: Image Schemas in Cognitive Linguistics* (pp. 1-14). Berlin, Germany: Mouton de Gruyter.
- Heiser, M., Iacoboni, M., Maeda, F., Marcus, J., & Mazziotta, J. C. (2003). The essential role of Broca's area in imitation. *The European Journal of Neuroscience*, 17, 1123–1128. doi:10.1046/j.1460-9568.2003.02530.x
- Hoggan, E., & Brewster, S. (2007). Designing audio and tactile crossmodal icons for mobile devices. In *Proceedings of the 9th International Conference on Multimodal Interfaces* (pp. 162-169). New York: ACM.
- Holmquist, L. E., Redström, J., & Ljungstrand, P. (1999). Token-based access to digital information. In *Proc. 1st International Symposium on Handheld and Ubiquitous Computing* (pp. 234-245). Berlin, Germany: Springer-Verlag.
- Ishii, H., & Ullmer, B. (1997). Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* (pp. 234-241). New York: ACM.
- Jensenius, A. (2007). *Action-Sound: Developing Methods and Tools to Study Music-Related Body Movement*. Ph.D. Thesis. Department of Musicology, University of Oslo.
- Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. Chicago, IL: University of Chicago.

- Johnson, M., & Rohrer, T. (2007). We are live creatures: Embodiment, American pragmatism and the cognitive organism. In: J. Zlatev, T. Ziemke, R. Frank, & R. Dirven (Eds.), *Body, language, and mind, vol. 1* (pp. 17-54). Berlin, Germany: Mouton de Gruyter.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed). Chicago: University of Chicago Press.
- Lakoff, G., & Johnson, M. (1980) *Metaphors we live by*. Chicago: University of Chicago Press.
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books.
- Lederman, S. J., Klatzky, R. L., Morgan, T., & Hamilton, C. (2002). Integrating multimodal information about surface texture via a probe: relative contributions of haptic and touch-produced sound sources. In *Proceedings of the 10th Symposium on Haptic Interfaces for Virtual Environments & Teleoperator Systems* (pp. 97-104). Piscataway, NJ: IEEE.
- Leman, M. (2008). *Embodied music cognition and mediation technology*. Cambridge, MA: MIT Press.
- Lieberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21, 136. doi:10.1016/0010-0277(85)90021-6
- McGurk, H., & MacDonald. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748. doi:10.1038/264746a0
- McLuhan, M. (1966). *Understanding media: the extensions of man*. New York: New American Library.
- McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.
- Noë, A. 2004. *Action in perception*. Cambridge, MA: MIT Press.
- Norman, D.A. (1988). *The psychology of everyday things*. New York: Basic Books.
- Norman, D. A. (1999). Affordance, conventions, and design. *Interactions (New York, N.Y.)*, 6(3), 38-42. doi:10.1145/301153.301168
- Oviatt, S., & Cohen, P. (2000). Multimodal interfaces that process what comes naturally. *Communications of the ACM*, 43(3), 45-53. doi:10.1145/330534.330538
- Peirce, C. S. ([1894] 1998). What is a sign? In Peirce Edition Project (Ed.), *The essential Peirce: selected philosophical writings vol. 2* (pp. 4-10). Bloomington: Indiana University Press.
- Pirhonen, A. (2007). Semantics of sounds and images - can they be paralleled? In W. Martens (Eds.), *Proceedings of the 13th International Conference on Auditory Display* (pp. 319-325). Montreal: Schulich School of Music, McGill University.
- Pirhonen, A., & Palomäki, H. (2008). Sonification of directional and emotional content: Description of design challenges. In P. Susini & O. Warusfel (Eds.), *Proceedings of the 14th International Conference on Auditory Display*. Paris: IRCAM (Institut de Recherche et Coordination Acoustique/ Musique).
- Pirhonen, A., Tuuri, K., Mustonen, M., & Murphy, E. (2007). Beyond clicks and beeps: In pursuit of an effective sound design methodology. In I. Oakley & S. Brewster (Eds.), *Haptic and Audio Interaction Design: Proceedings of Second International Workshop* (pp. 133-144). Berlin, Germany: Springer-Verlag.
- Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences*, 21, 188-194. doi:10.1016/S0166-2236(98)01260-0

Bodily Engagement in Multimodal Interaction

- Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. New York, NY: Cambridge University Press.
- Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. New York: Cambridge University Press.
- Searle, J. R. (2004). *Mind: A brief introduction*. New York: Oxford University Press.
- Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.
- Sonnenschein, D. (2001). *Sound design: The expressive power of music, voice and sound effects in cinema*. Saline, MI: Michael Wiese Productions.
- Tagg, P. (1992). Towards a sign typology of music. In R. Dalmonte & M. Baroni (Eds.), *Secondo Convegno Europeo di Analisi Musicale* (pp. 369-378). Trento, Italy: Università Degli Studi di Trento.
- Tuuri, K., & Eerola, T. (2008). Could function-specific prosodic cues be used as a basis for non-speech user interface sound design? In P. Susini & O. Warusfel (Eds.), *Proceedings of the 14th International Conference on Auditory Display*. Paris: IRCAM (Institut de Recherche et Coordination Acoustique/Musique).
- Välkkynen, P. (2007). *Physical selection in ubiquitous computing*. Helsinki, Edita Prima.
- Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind*. Cambridge, MA: MIT Press.
- Want, R., Fishkin, K. P., Gujar, A., & Harrison, B. L. (1999). Bridging physical and virtual worlds with electronic tags. In *Proc. SIGCHI Conference on Human factors in Computing Systems* (pp. 370-377), New York: ACM Press.
- Waterworth, J. A., & Chignell, M. H. (1997). Multimedia interaction. In Helander, M. G., Landauer, T. K. & Prabhu, P. V. (Eds.): *Handbook of Human-Computer Integration* (pp. 915-946). Amsterdam: Elsevier.
- Wickens, C. D. (1984). Processing resources in attention. In: R. Parasuraman & D.R. Davies (Eds.), *Varieties of attention* (pp. 63-102). Orlando, FL: Academic Press.
- Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 1(3), 460473.

ENDNOTE

¹ www.tampere.fi/english/vapriikki/

PIV

**GESTURAL ATTRIBUTIONS AS SEMANTICS IN USER
INTERFACE SOUND DESIGN**

by

Kai Tuuri 2010

In Kopp, S. & Wachsmuth, I. (Eds.), *Gesture in Embodied Communication and Human-Computer Interaction (LNAI 5934)*, Revised Selected Papers of 8th International Gesture Workshop GW2009, Bielefeld, Springer-Verlag, Heidelberg, pp. 257–268

Reproduced with kind permission of Springer-Verlag.

Gestural Attributions as Semantics in User Interface Sound Design

Kai Tuuri

Department of Computer Science and Information Systems
FI-40014 University of Jyväskylä, Finland
krtuuri@jyu.fi

Abstract. This paper proposes a gesture-based approach to user interface sound design, which utilises projections of body movements in sounds as meaningful attributions. The approach is founded on embodied conceptualisation of human cognition and it is justified through a literature review on the subject of interpersonal action understanding. According to the resulting hypothesis, stereotypical gestural cues, which correlate with, e.g., a certain communicative intention, represent specific non-linguistic meanings. Based on this theoretical framework, a model of a process is also outlined where stereotypical gestural cues are implemented in sound design.

Key words: gestures, user interfaces, sound design, semantics

1 Introduction

Sound-based communication within different kinds of media has a long tradition. Sound design practices for radio-plays from as early as the 1920s have defined the basis for the communicative use of sound effects which is still relevant in today's film and video game sound design [1]. An essential part of the craftsmanship of film sound designers has concerned the creation of sound effects that reflect mental states of the story's characters. The appropriate door knock for film narration, for example, can be *urging*, *gentle* or *angry*, depending on the purpose of that door knock and the feelings and intentions of the person who is knocking. Such focus on *agency* behind sounds is frequently utilised in film sound design. It exploits the perceptual bias towards understanding the human involvement (i.e., *intentionality*) in the sound-causing action [2]. But when it comes to the sound design for human-computer interaction (HCI), such interpersonal focus on interpretation is rarely utilised in a systematic manner. This paper focuses on this agency-orientated perspective.

Due to its history, the HCI field has its roots in information theory [3] and "system-centred" design [4]. Compared to filmmaking tradition, functions of user interface (UI) sounds are more easily conceived in terms of information processing and transmission between machine and user than in terms of interpersonal communication. Such a perspective is well exemplified in the design/research

paradigm of "earcons" [6], which usually refer to abstract user interface sounds with highly arbitrary meanings. It has adopted a linguistically orientated view of semantics which usually sees the semantic content as symbolic units of information essentially separable from its form of expression. The risk of such an approach is that design can become detached from meaningful experiences that get coupled with sounds in the interaction. It should be obvious that the role of sound cannot be as a mere carrier of symbolic information. Indeed, reflecting the ongoing shift towards user-centred design [5], contemporary trends in HCI sound research have preferred to talk about *sonic interaction design* thus emphasising the coupling of sound and its meanings with interaction [7].

The word *gesture* is used here to represent any bodily act that – observable in interaction – operates as a vehicle for interpersonal communication. Such a perspective is not restricted to hand movements, but takes all non-verbal forms of body-related communication into account. In social interaction, we express our mental states with bodily actions which can be either directly perceivable (like in hand/facial gestures or in vocal prosody), and/or indirectly perceivable as reflections of body movements (e.g., in sounds of objects which are acted on). Communication with gestures is primordial [8] and often unconscious for us. The basis for gestural communication – the physical constitution of the human body and our ways to schematise it – is universal. Gestural communication has also been suggested to have a strong phylogenetic background which precedes verbal communication [8]. As Marc Leman has put it: "Gestures form the basis of mutual adaptive behavioral resonances that create shared attention and are responsible for the feeling of being unified with other people" [2].

By suggesting the utilisation of gestural attributions (projections of motor-activity/body movement) as semantics in UI sound design, this paper emphasises bodily mediated action understanding and the role of action-relevant gestural cues in sound as constituents of meaning-creation based on a kinaesthetic foundation. In the scope of interaction design, such tacit "sensibility for movement" also accounts for a "sensibility for responses to movement" [9]. We thus stress the close engagement of interaction and meaning-creation already acknowledged within, for example, the ecological view [10] of perception. Gestural attributions, unlike linguistic ones, are not detached from the direct sensory-motor basis of social interaction. According to the embodied approach to human cognition [11], the human mind is coupled with our environment. That coupling with the environment has emerged in the course of the experiential history of using our bodies for interacting with it. Understanding is thus inseparable from the embodied experiences of the physical world including – most relevantly to this study – interactions with other people.

The aims of this paper are to promote an idea that gestural projections of body movement attributed to sounds could be used as semantics for UI sound design, and to formulate this idea into a justifiable and testable hypothesis. In addition, the aim is to outline a model of the process in which gestural cues are implemented in sound design.

2 Embodied Basis of Understanding Gestures

In the phylogenesis of social mammals, such as humans and non-human primates, it has proved to be beneficial – maybe even essential for survival – to understand the actions of others [12]. Surely, without such an ability, the social and cultural development of humans would have been impossible. In this section we discuss the embodied basis of this attunement to interpersonal relations, which involves shared body-related constituents. First, the neurological foundation of *interpersonal action understanding* is reviewed, which is then applied to the concept of body-schema and empathetic involvement in perception. Lastly, interpersonal action understanding is viewed in terms of the Brunswikian lens model.

2.1 The Human Mirror-Neuron Mechanism

Mirror neurons are a particular class of premotor neurons that discharge both when one performs a specific goal-reaching action and when one observes other individuals executing similar actions [12]. They were originally discovered in the monkey premotor cortex, but there is also evidence for the existence of a similar mirror system in humans [12].

Nowadays there exist two parallel and equally plausible hypotheses about the functional role of mirror neurons. Firstly, they mediate bodily imitation and secondly, they are related to action understanding [12]. The motor representation encoded in mirror neurons thus reflects the understanding of observed action – not object presentation. Via motor representation, mirror neurons transform sensory information into knowledge that agrees with the motor repertoire/skills of the observer [12]. In other words, the observer understands the performed action as she could perform it herself. Action understanding thus involves embodied "resonances" (or embodied simulation [13]) of the observed action.

Experiments have shown that even fragmentary clues about action presented to the observer can trigger the specific response (motor representation) in mirror neurons [12]. Therefore, the audio-visual features of the observed actions seem to be fundamental only to the point where they allow action understanding. For example, the mere sound of action seems to result in a response that matches the responses for the same action observed or executed [14]. The encoding of action in the mirror system thus seems to be highly multimodal in nature. These above aspects underline the possible role of mirror neurons in contributing schematic gestalt processes, which transforms sensory information (like hearing somebody laughing) into preconceptual structures meaningful to the observer (understanding laughing by means of mirrored motor representation of it). It is also suggested that such action understanding operates as an enabling mechanism for empathy [13].

2.2 Interpersonal Body-Schematic Transfer

The concept of *body-schema* refers to a tacit understanding of one's own body in-the-world. As suggested in *Phenomenology of Perception* by Merleau-Ponty [15],

humans possess such specific schemata of our body in relation to embodied space, i.e., space in the environmental setting of our habitual actions. It is reasonable to assume that such body-schemata (or kinaesthetic image schemata [16]) are based on recurrent sensory-motor experiences of bodily interactions with the world.

According to Merleau-Ponty [15], body-schema has a crucial function in the perception of other individuals as human-beings. That function is related to *body-schematic transfer* where the movements of other individuals are perceived as the movements that the observer could imagine executing by her own action repertoire. Therefore the perception of body movements is based on the perceiver's embodied knowledge of body-schema. This theoretical idea is very much in line with the already discussed function of mirror neurons in action understanding. It is thus plausible that the mirror system is a part of the realisation of body-schematic transfer [17].

Jan Almäng [17] has proposed that, at its basic level, body-schematic transfer has at least four characteristic features, which are:

1. The perceiver observes the other as having a body-schema.
2. The perceiver can perceive the action by means of body-schema even when she is unable, e.g. due to its complexity, to perform it herself. Thus, it is sufficient that her body-schema can "read" the movement.
3. Physical similarity between the perceived and perceiver is not required for an apprehension of the movements by body-schematic transfer. Thus, it is sufficient if there appear to be kinematic similarities between observed movement and body-schematic knowledge of how to produce such movement.¹
4. Body-schematic transfer of movements by itself does not imply that the intentionality of the other person is communicated to the perceiver. This is because understanding the mental states of someone, on the basis of physical movement, requires contextual awareness.

Because human actions arise from intentions, emotions and other affective determinants which are linked to the context, we must situate movements in the interaction where it takes place. This broader aspect of body-schematic transfer is discussed in the following section.

2.3 Empathetic Involvement in Perception

We normally perceive other people as engaged in situations that provide meaningful references to their actions. Hence, the crucial element in the Merleau-Pontyan view of understanding others (by means of body-schematic transfer) lies in the ability to automatically re-center our perception of the situation to the perspective of the other. Therefore, together with tacit perception of a person's body movements, the observer re-centers her own primordial perception

¹ This feature is evident in watching animated cartoons, where even objects can indicate such anthropomorphism in movements that they can be perceived by means of body-schematic transfer.

of the surrounding environment – including its action affordances – to become a perception *for* that other person [17, 13]. In this way we can understand motivations and other reflections of intentionality in actions, and we are also able to anticipate possible actions of another person – as if they were actions of our own. Such an attunement to another individual, not only to body movements but also to her intentionality and action affordances, seems to be such a natural part of our interpersonal awareness that it requires no conscious reasoning.

To sum up the discussion on the embodied understanding of movements, we can assume that it is based on two parallel aspects in perception. The first aspect is related to the mimetic involvement in the mirror neuron system, by which perceived movement is understood in terms of body-schema and kinaesthesia. Such corporeal resonances can range from simple synchronisations to more specific motor mimetic attuning. The second aspect is related to empathetic involvement, in which the body-schematic "resonances" (of the first aspect) are associated with the perspective of the other individual engaged in interaction. At the lower level, this refers to mere action-based involvement with the other's movements and thus primordial apprehension of *corporeal intentionality* (i.e., motor intentionality) [2], whereas "genuine" empathy usually refers to more participative, emotional and inferential involvement with the perspective of the other. A similar distinction, between the degrees of motor-system involvement and the degrees of empathetic involvement, has also been suggested in the theory of bodily mediated experiences in music [2].

2.4 Encoding and Decoding of Gestural Cues

We finally take a look at the Brunswikian lens model scheme initially proposed as a framework for understanding how prosodic cues are encoded in a vocal expression of emotion [18]. Based on the original lens model [19], it describes the processes of various *distal cues* being situationally determined in articulation, indicating affective states of a person, and how these acoustically transmitted patterns (as *proximal cues*) play a role in the attribution of an affective state in perception. The lens model simultaneously considers *encoding* (i.e., contextual determination of cues) and *decoding* (i.e., contextual interpretation of cues). It therefore gives a neat overview of communication, where body is acknowledged as a mediator. The model, adapted for gestural communication, is illustrated in Figure 1.

In the encoding of cues, Scherer has emphasised the central role of *push* and *pull effects* [18]. Therefore intentional and affective states of a subject are situationally characterised in gestural articulation by interaction between 1) psychobiological processes, intrinsically related to mental states, that provide a natural influence on body movements (push effects) and 2) interactional processes, which involve voluntary control over body movements and are related to external conditions (pull effects). Conditions of interaction thus often requires a certain strategic display (or hiding) of intentions and other mental states.

The dominance of push effects is most evident in so-called *affect bursts*, which are mostly a result of physiological arousal. Push effects also have significant role

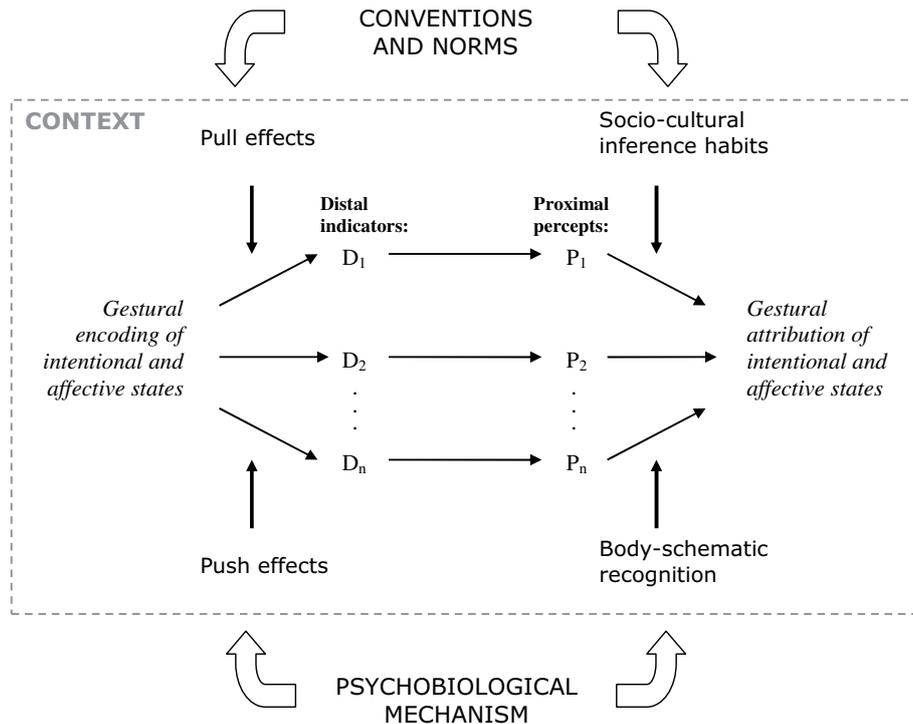


Fig. 1. The Brunswikian lens model with push and pull effects [18].

in spontaneous expressions of, for example, pain or joy, but they also influence even the most premeditated acts of communication. [18] At the other end of the continuum, the dominance of pull effects is evident in expressions of an intended purpose. John R. Searle has suggested three *conditions of satisfaction* that fuel speech acts [20], but they should also apply to any acts which bear the motivation to be understood in interaction. As determinants of gestural articulation they are 1) *articulation intention*; an intention to appropriately produce a certain kind of gesture, 2) *meaning intention*; an intention to mean something with the gesture, and 3) *communicative intention*; an intention to be understood in a certain way – referring to the type of communicative intent such as *asking*. Communicative intention essentially imposes its condition of satisfaction on the conditions of satisfaction of both meaning something and producing a gesture accordingly.

According to the lens model, distal cues cannot be directly perceived. But on the basis of body-schematic transfer, the proximal cues can be perceived in terms of distal gestural movements. Proximal cues are probabilistic, partly redundant and contribute to action understanding in an additive fashion [21]. Gestural action understanding, in turn, provides cues for attributing intentionality [2]. Gestural attributions are thus contextually constructed in interaction between primordial, ecologically developed gestalts and inference based on social habits or cultural norms. These two processes can be seen as perceptual counterparts to push and pull effects, in which proximal percepts of gestural cues resonate with both body-schematic and conventional backgrounds.

3 Gestural Attributions in Sound Design

This section concentrates the discussion around a gesture-based approach to UI sound design – the principle of using sounds as the means of projecting gesturally attributable cues of movement. The basic assumption in such an approach is that human capacity for interpersonal awareness allows body movements of another individual to be understood in terms of the body movements of the perceiver. This seems to apply also in situations where the presented movement is partial or, due to its kinematic similarities with body-schematic reference, merely implies bodily movement. As already discussed earlier in the paper, action understanding is not dependent on which sensory modality is utilised in providing clues of action. Therefore the premise is that sounds can be used for presentation of action-related features.

The lens model perspective (see 2.4) demonstrates how perceptual action understanding is an emerging resonance of the mixture of several parallel cues encoded by the same gesture. In the context of HCI sound design, the strength of the lens model is that it outlines successful communication by means of probabilistic, multiple and redundant cues that allow discarding non-relevant cues for the task. Due to technical or aesthetic reasons, the designer must often conform to a selected set of acoustic cues in attributing intended characteristics to UI-sounds. Limitations should not undermine the utilisation of gestures, as there should be plenty of suitable cue combinations which compensate for the discarded ones. Indeed, it has been found that even simple acoustic cues can communicate the emotions of a musical performer or a speaker (with a general lack of cue interactions) [21]. As interaction is multimodal, the action understanding is ultimately based on the contextual whole, in which the sound instance and its cues become perceptually fused with the other aspects of interaction [22].

3.1 Gestural Articulations Projected in Sounds

From the perspective of the observer/user, gestural projections of sound refer to kinaesthetic imagery of body movement, which arise during the listening experience. The contextual creation of such gestural imagery is based on body-schematic resonances of *motor-mimetic* involvement [23] in listening. It is easiest to assume that gestural imagery is perceived in sounds that are in some way caused by bodily excitation, hence implying sound-producing gestures. But regardless of the type of sound production, motor-mimetic involvement applies to the perception of music, or *any* sound, as long as it is able to imply physical movement [23, 2]. Thus, gestural imagery can even be attributed to abstract and artificially produced sounds.

From the sound designer's perspective, communication with gestural projections means defining a contextually appropriate gesture for the communicative purpose, and then articulating and implementing it in design. Gestural articulation can be a vocal act, musical expression or any physical action that itself produces, or allows its features to be transformed into, acoustic resonances. As discussed in 2.4, the articulation is bound up with the situation. When a gesture

is articulated spontaneously, while being immersed in interaction, articulation is not a subjective interpretation so much as an *experience as articulated*. A sound, caused or modulated by gestural articulation consequently conveys acoustic cues of corporeal intentionality involved *in* the physical articulation.

The basic principle of using gestures as part of sound production in HCI is not new, although the idea has usually been utilised in producing immediate environmental audio feedback on the basis of the user's gestures (e.g. [24]) – not in exploiting gestures as interpersonal communication. It can be argued that traditional film sound design practices have long acknowledged the importance of gestural articulation in creating sound effects. One prominent example of this is a tradition called *foley art* [1]. Despite all the sophisticated audio technology available today, foley art still favours manual ways of producing sounds (in real-time). As noted above, gestural communication also refers to acting on material objects. This is exactly the case with foley art, in which the aim is to express through the sounds of material objects and provide "added value" to the narrative whole. We see that direct bodily involvement during sound creation – often performed concurrently with the related visual narration – enables the intentionality of a performer to be communicated via gestural projection. As illustrated in the introductory section of this paper, even simple sounds like door knocks can have much variety both in their gesturally determined qualities and in how these qualities can affect the contextual interpretations (see also [25]).

3.2 Utilising Stereotypical Gestural Cues in Sound Design

When treating gestural attributions as semantics, the sound designer can approach them from at least two directions: She can use the gesture as a starting point (thus emphasising distal cues), also accounting for the situated articulation and motivation of the gesture. Or she can focus on gesture-specific acoustic characteristics (thus emphasising proximal cues), but only if she has sufficient knowledge about the acoustic correlates of gesture-related understanding. Thus we regard it as easier for the designer to start with communicationally appropriate gestural articulation as the means to acquire and study gestural cues, which in turn can be utilised in sound production. In this way, semantics is always considered as being closely linked to the context.

In order to use gestural semantics systematically, one needs a way to categorise different types of meanings, i.e., gestural patterns that become contextually meaningful. In order to take advantage of the non-linguistic characteristics of gestures, we are especially interested in discovering *stereotypical* gestural cues. They reflect embodied meanings of a specific type of *recurrently experienced* gesture and resonate with primordial gestalts for interpersonal understanding and communication. Stereotypical cues are thus a type of gesturally perceived cues that should communicate a specific meaning robustly, without being too strongly dependent on cultural constraints. They differ from "weak cues" (like indications of direction or force), which are extremely dependent on context. They also differ from "coded cues", which communicate robustly, but are only meaningful because of coding or convention.

But how can the sound designer find gestures that convey such stereotypical meanings? We suggest using the *communicative intention* of gesture as a category of semantics. The assumption is that context-situated articulation – with specific intention to communicate – results in stereotypical physical cues of that intention as an outcome of push and pull effects. There is evidence, for example, that prosodically realised acoustic patterns specific to communicative intention exist [26, 27]. Arguably, in infant-directed speech, such intention-specific prosody (e.g., for alerting or prohibiting) functions as the first regular semantic correspondence to the infant, clearly preceding any linguistically related functions of prosody [26]. Communicative intention of gestural expression thus appears to serve a fundamental and important prelinguistic function.

Prosodic patterns represent a subset of gestural patterns as they are caused by "phonetic gestures"; motor movements of the phonatory apparatus, vocal tract and respiration. Indeed, there are two reasons why prosody of vocal acts promises to be a very important source for gestural cues being utilised in sound design. Firstly, the evolution of the vocal apparatus is related to human communication [26]. Secondly, in prosody, gestural cues are directly realised as acoustic cues which are familiar and ecologically valid to us. Of course, stereotypical cues – specific in communicating different emotions or interpersonal attitudes – exist in other kinds of gestural expressions as well [28]. But in order to be encoded acoustically, non-vocal gestures need to be sonified. In everyday interactions, such sonification is a natural outcome of material resonances of motor excitation when objects and materials are acted on. However, by using physical models (e.g. [24]), the sounds of various material interactions can also be synthesised on the basis of, e.g., kinematic parameters.

Figure 2 specifies the general phases of *modelling*, *performing*, *utilising* and *evaluating* in the process of implementing gestural cues in sound design. In the first phase, the need/purpose of an UI sound element is acknowledged on the basis of the designer's model of application-user interaction. Hence the communicative functions for UI sounds can be determined. The modelling of appropriate gestural action requires mental exploration of interaction (see the discussion about action models in [22]). If the designer puts herself into the dialogue between user and machine, she is able to conceive her role as a person who is communicating with the user. The designer can thus imagine participating in the interaction, which in reality occurs via the mediation of the machine in use. From that perspective, she can mentally explore the patterns of contextual application use and – whenever sonic feedback is required – discover gestural patterns that would feel contextually appropriate for the communicative need. The communicative intention of the modelled gesture thus conforms to the communicative function of the propositional UI sound.

In the performing phase, the specified gesture is articulated. In order to achieve the spontaneity in articulation, the gesture should be performed while being immersed in interaction. To enable such immersion, a suitable scenario can be used providing the situational flow of interaction. This can be done, for example, in terms of metaphorical person-to-person interaction (see example

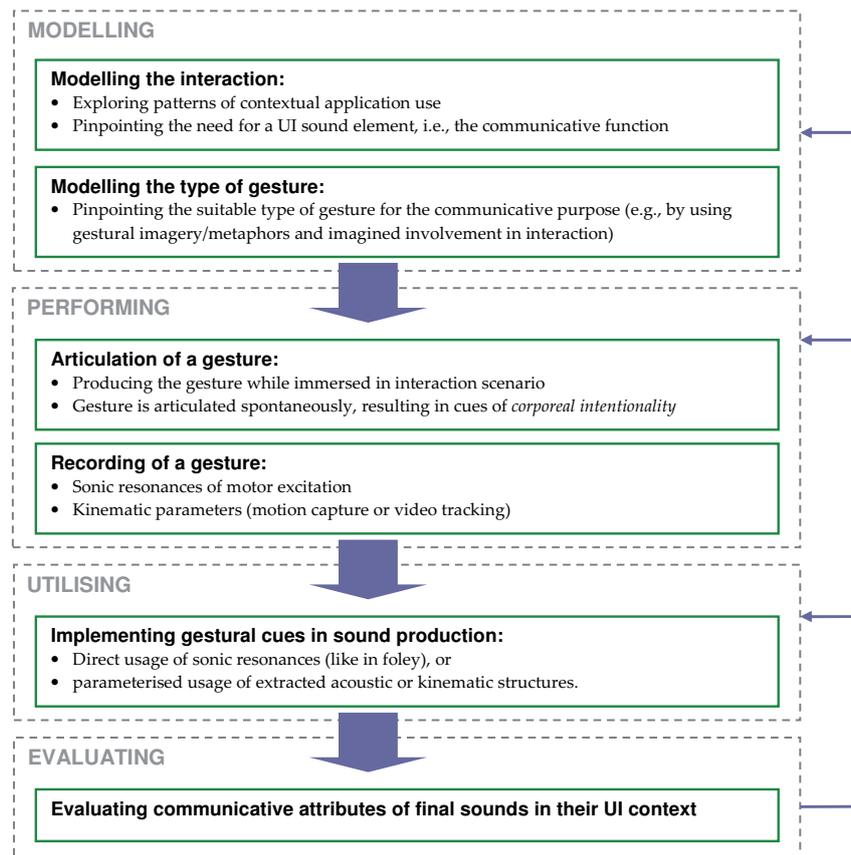


Fig. 2. The process of implementing gestural cues in sound design.

cases in [27, 22]). The articulation of a gesture needs to be recorded, either as a sound recording (of sonic resonances of action) or as a recording of selected kinematic parameters (by motion capture).

The utilisation of recorded gesture in sound production can be based on either "foleyish" direct usage of recorded sound, or – in a more analytical manner – parameterised usage of gestural cues. Parameters can be acquired, for example, by extracting selected acoustic features from sound recording (see example cases in [27, 22]), or by utilising recorded kinematic parameters of gestural articulation. Parameterised gestural cues can be implemented in sounds, for example, as parameters for sound synthesis or sound manipulation. Alternatively they can be used indirectly as structural ideas in sound design and production. The last phase in the process is evaluation, where the communicative attributes of the final UI sounds are contextually tested. When required, the sound designer can iterate and re-evaluate the process starting from any of the previous phases.

4 Conclusions

We can now conclude this paper in the form of a hypothesis about using gestural projections of body movement as semantics in UI sound design. The resulting hypothesis includes the following assumptions:

- Sound design can be founded on theory which emphasises embodied cognition and interpersonal action understanding.
- By exploiting the primordial capacity for interpersonal action understanding in humans, sound design can utilise stereotypical, gesturally realised cues of social interaction, which represent non-linguistic categories of meaning.
- On the basis of the theoretical framework presented, sound design can be outlined as a process which is explicated into distinct design phases where action-relevant cues are determined in terms of interaction and gestural articulation.
- The approach presented results in ecologically valid semantics which should be communicated robustly and require less learning than semantic attributions of linguistically orientated design schemes. The ongoing research, based on the theoretical framework presented, has supported this assumption [29].

The gesture-based perspective on sound design provides an important focus on performative aspects of sound design (i.e., direct involvement with the sound creation), and bodily engagement with sonic communication (both in sound design and contextual perception). These aspects merit more explicit focus within sound design research, although they are most likely tacitly acknowledged by many professional sound designers. A linguistically orientated paradigm often considers semantics as "absolute". Such a perspective easily dismisses the communicational potential that even a simple feedback sound can have when it is designed as physical activity – and for physical activity.

Acknowledgments. This work is funded by Finnish Funding Agency for Technology and Innovation, and the following partners: GE Healthcare Finland Ltd., Suunto Ltd., Sandvik Mining and Construction Ltd. and Bronto Skylift Ltd.

References

1. Mott, R. L.: Sound effects: Radio, TV and Film. Boston, MA: Focal Press (1990)
2. Lemay, M.: Embodied Music Cognition and Mediation Technology. Cambridge, MA: MIT Press (2008)
3. Shannon, C. E. and Weaver, W.: The Mathematical Theory of Communication. Urbana, IL: The University of Illinois Press (1949)
4. Czaja, S. J.: Systems design and evaluation. In Salvendy, G. (ed.) Handbook of human factors and ergonomics (2nd ed.). New York, NY: Wiley, 17–40 (1997)
5. Bannon, L. J.: A human-centred perspective on interaction design. In Pirhonen, A., Isomäki, H., Roast, C. and Saariluoma, P. (eds.) Future interaction design. Berlin/Heidelberg: Springer-Verlag, 31–51 (2005)
6. Brewster, S., Wright, P. and Edwards, A.: A detailed investigation into the effectiveness of earcons. In Kramer, G. (ed.) Auditory display. Reading, MA: Addison-Wesley, 471–498 (1994)
7. Rocchesso, D., Serafin, S., Behrendt, F., Bernardini, N., Bresin, R., Eckel, G., Franinovic, K., Hermann, T., Pauletto, S., Susini, P. and Visell, Y.: Sonic interaction design: sound, information and experience. In CHI '08 extended abstracts on Human factors in computing systems. New York, NY: ACM press (2008)

8. Rizzolatti, G. and Arbib, M. A.: Language within our grasp. *Trends in Neurosciences*, 21, 188–194 (1998)
9. Svanaes, D.: Kinaesthetic thinking. *Computers in Human Behavior*, Vol. 13, No. 4, 443–463 (1997)
10. Gibson, J. J.: *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin (1979)
11. Varela, F., Thompson, E. and Rosch, E.: *The Embodied Mind*. Cambridge, MA: MIT Press (1991)
12. Rizzolatti, G. and Craighero, L.: The mirror-neuron system. *Annual Review of Neuroscience* 27, 169–192 (2004)
13. Gallese, V., Eagle, M. N. and Migone, P.: Intentional Attunement: Mirror Neurons and the Neural Underpinnings of Interpersonal Relations. *J Am Psychoanal Assoc*, Vol. 55, No. 1, 131–175 (2007)
14. Kohler E., Keysers C., Umiltá MA., Fogassi L., Gallese V. and Rizzolatti G.: Hearing sounds, understanding actions: action representation in mirror neurons. *Science* 297, 846–848 (2002)
15. Merleau-Ponty, M.: *Phenomenology of Perception*. London: Routledge (1945/1996)
16. Johnson, M.: *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago, IL: University of Chicago (1987)
17. Almäng, J.: *Intentionality and intersubjectivity*. Ph.D. thesis, Goteborg, Sweden: Goteborg Universitet (2007)
18. Scherer, K. R. and Bänziger, T.: Emotional expression in prosody: a review and an agenda for future research. In *proc. of SP-2004*, Nara, Japan, 359–366 (2004)
19. Brunswik, E.: *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press (1956)
20. Searle, J. R.: *Mind, language and society*. New York, NY: Basic Books (1998)
21. Juslin, P. N. and Laukka, P.: Communication of emotions in vocal expression and music performance: Different channels, same code? *Psych. Bull.*, vol. 129, no. 5, 770–814 (2003)
22. Tuuri, K., Pirhonen, A. and Vääkkönen, P.: Bodily Engagement in Multimodal Interaction: A Basis for a New Design Paradigm? In Kurkovsky, S. (ed.) *Multimodality in Mobile Computing and Mobile Devices*. Hershey, PA: IGI Global (2009)
23. Godøy, R. I.: Gestural-sonorous objects: embodied extensions of Schaeffer’s conceptual apparatus. *Organised Sound*, 11(02), 149–157 (2006)
24. Avanzini, F. Serafin S. and Rocchesso, D.: Friction sounds for sensory substitution. In *proc. of ICAD 2004*, Sydney, Australia (2004)
25. Vitale R. and Bresin, R.: Emotional Cues in Knocking Sounds. In *proc. of ICMPC 10* (abstract only), Sapporo, Japan, p. 276 (2008)
26. Fernald, A.: Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In Barkow, J., Cosmides, L. and Tooby J. (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, 391–428 (1992)
27. Tuuri, K. and Eerola, T.: Could function-specific prosodic cues be used as a basis for non-speech user interface sound design? In *proc. of ICAD 2008*, Paris, France (2008)
28. Feyereisen, P. and de Lannoy, J.-D.: *Gestures and speech: Psychological investigations*. New York, NY: Cambridge University Press (1991)
29. Tuuri, K., Pirhonen, A. and Eerola, T.: Design and Evaluation of Prosody Based Non-Speech Audio Feedback for Physical Training Application. (journal submission)

PV

**COULD FUNCTION-SPECIFIC PROSODIC CUES BE USED AS
A BASIS FOR NON-SPEECH USER INTERFACE SOUND
DESIGN?**

by

Kai Tuuri & Tuomas Eerola 2008

In P. Susini & O. Warusfel (Eds.), Proceedings of the 14th International
Conference on Auditory Display, IRCAM, Paris, France

COULD FUNCTION-SPECIFIC PROSODIC CUES BE USED AS A BASIS FOR NON-SPEECH USER INTERFACE SOUND DESIGN?

Kai Tuuri

University of Jyväskylä
Department of Computer Science and Information Systems
P.O.Box 35, FI-40014, Finland
krtuuri@cc.jyu.fi

Tuomas Eerola

University of Jyväskylä
Department of Music
P.O.Box 35, FIN-40014, Finland
tuomas.eerola@campus.jyu.fi

ABSTRACT

It is widely accepted that the nonverbal parts of vocal expression perform very important functions in vocal communication. Certain acoustic qualities in a vocal utterance can effectively communicate one's emotions and intentions to another person. This study examines the possibilities of using such prosodic qualities of vocal expressions (in human interaction) in order to design effective non-speech user interface sounds. In an empirical setting, utterances with four context-situated communicative functions were gathered from 20 participants. Time series of fundamental frequency (F_0) and intensity were extracted from the utterances and analysed statistically. Results show that individual communicative functions have distinct prosodic characteristics in respect of pitch contour and intensity. This implies that function-specific prosodic cues can be imitated in the design of communicative interface sounds for the corresponding functions in human-computer interaction.

Keywords: prosody, communicative functions, non-speech sounds

1. INTRODUCTION

Finding ways to produce intuitively salient and communicative non-speech user interface sounds has been a major challenge in the research paradigm of auditory display. An interface sound can be seen intuitively communicative if the users' unconscious application of knowledge facilitates effective interaction [1]. One way to achieve this utility of existing abilities and knowledge in sound design is to "...mimic the ways we constantly use sound in our natural environments...", as was noted already in the workshop report of CHI'94 [2]. Alongside the linguistic means to express, the human vocal communication contains an important non-verbal channel. This affective content of speech is conveyed by various *prosodic* cues, which refer certain characteristics in intonation, stress, timing and voice quality - or by acoustic terms - in dimensions such as pitch, intensity and spectrum. It is pointed out by several authors [3, 4, 5] that the basis of encoding and decoding these prosodic features in vocal communication has a strong *phylogenetic* background. Such evolutionary perspective is supported, e.g., by the evidence of cross-cultural prosodic similarities in infant-directed speech [6]. It is hardly the case that all codes related to nonverbal vocal expressions are "hard-wired" into the human species. One can assume that several parts of the coding consist of socio-culturally learned habits. But if the feature determinants and nonverbally evoked meanings of vocal patterns have even partial universality, these codes must be considered to be

serving as a source of relevant knowledge in sound design. While many professional sound designers might implicitly mimic various prosodic cues in their work, there is a definitive lack of explicit knowledge of how certain prosodic characteristics are related with the human meaning-creation.

1.1. Vocally communicated emotions and intentions

A wealth of evidence exists that emotional and intentional states are communicated nonverbally through vocal expressions [4]. The ability to catch the emotional and motivational state of mind of other people has been considered as crucial in forming and maintaining social relationships [3]. In social interaction, the emotional communication can also be utilised for manipulation and persuasion.

1.1.1. Formulation and perception of vocal cues

The acoustic form of vocal expression is the result of several determinants. Scherer [7] has made a basic distinction between push and pull effects in those determinants. *Push effects* are caused by physiological processes that are naturally influenced by emotional and motivational state (e.g., nervousness in voice). *Pull effects* involve external conditions and voluntary control over vocalisation. The external situational context thus often requires certain strategic display of intentions or emotions. Voluntarily controlled vocalisations can consist of innate expressions as well as culturally dependent, learned or invented, vocal patterns.

The perception of emotions has been suggested to involve specialised innate *affect programs* [8], which rapidly and autonomously organise perception in terms of affect categories (e.g., basic emotions). Moreover, as Huron [9] has suggested, emotional responses may be caused by multiple distinctive activating systems. In this current study, the empathetic activating system deserves a particular interest. It allows the listener to perceive cues that signal someone's state of mind. The discovery of "mirror neurons" [10] provides further insights concerning the empathy and understanding of other people's intentions via inner imitation or simulated reenactment. It proposes the existence of a common neural structure for motor movements and sensory perception. As a mechanism for imitation, it codes the description and the motor specification of a perceived action (e.g., vocalisation). Interestingly, it seems that the intention or goal of the imitated action is also encoded. This suggests that empathy may function via the mechanism of this "mirrored" *action representation* by modulating our understanding about the emotions and intentions of other people in a corporeal

way. [11, 10] Of course, in addition to processes that take place in autonomic nervous systems, the rationalisations made on a higher cognitive level are also relevant in interpretations of culturally specific nuances or nonverbal semantics of perceived vocalisations.

1.1.2. Communicative functions

Vocal expressions are in many ways dependent on the situational context in which they take place and which they serve. Emotional and motivational states reflect the current situation and provide various effects to the determinants of vocalisation. But vocalisations are not only for revealing the speaker's emotional and motivational states. The speaker also instrumentally uses the expression to convey information to the others and to influence the communicational process.

Communicative functions of vocalisations refer to the communicative intentions of the speaker as well as the vocalisation's pragmatic meaning¹ (or functional semantics) of nonverbal vocal patterns is indicated by the empathetic perception of sound and its indexical relation to the situational context. The dependency to the situational conditions may vary. For example, an infant can perceive mother's vocal patterns as *prohibitive* in many different situations as long as the child is able to associate the utterance with her actions. On the other hand, the perception of certain functions may have more fine-tuned relationships between the vocalisation and its context.

Communicative functions represent particular categories of vocal expression and also certain contexts of interactions. In this study we will use the term essentially to categorise certain context-specific communicative intentions for using sound.

1.2. Transferring prosodic cues into another domain

This study is grounded on the idea that codes of nonverbal vocal communication could be utilised in the design of non-speech user interface sounds. However, can we make an assumption that prosodic characteristics of vocalisations can be extracted and effectively transferred into a different auditory domain?

Vocal expressions and musical performances are often seen as close relatives. In 1857 Spencer already argued that speech and music have notable similarities due to the physiological processes which are linked to both emotions and sound production [14]. On the basis of an extensive meta-analysis, Juslin and Laukka [3] found that, at least to a certain extent, acoustic cues in musical expression of emotions indeed have similarities to those employed in the vocal expression of emotions. They argue that these similarities are due to a habit of musicians' to communicate on the basis of the principles of nonverbal vocal expressions. Using a similar line of reasoning, it can be argued that those principles of vocal affect can also have an influence to sound design. Despite the differences between the essence of vocalisations and non-speech user interface sounds, prosodic cues may evoke similar affective responses in both domains.

We can speculate that, when compared to musical performances, user interface sounds are potentially even closer to the vocal communication. This speculation is supported by two premises: Firstly, the communicational utility value of interface sounds is prioritised (as it is in vocal communication). Secondly, there are several

¹See Tuuri et. al. [12] for a discussion about the levels of sonic meaning-creation and Rosenthal [13] for defining pragmatic meaning.

conveniently matching communicative functions between human-computer interaction and human vocal interaction (e.g., approving and disapproving).

When certain prosodic cues of speech are associated with certain communicative functions², we can presume that these function-specific prosodic qualities can be effectively imitated in the design of new sound objects as a source for its intended functional semantics. Iconic references to the original vocalisations should be considered in two levels: *imitation of prosodic features* and *imitation of a communicative function*. For the sake of the functional match, it is crucial to define the communicative functions (i.e., purposes) for every sound occurring in the interaction. Those considerations should be a natural part of interaction design and the conceptual design of sounds.

1.3. Goals of study

In order to utilise function-specific prosodic cues, one must examine whether such stereotyped cues in certain function-related vocalisations actually exist. The main goal of this study is to address this central issue. The secondary goal is to construct a suitable empirical method for gathering function-specific vocal expressions.

1.3.1. Design case as a background

Many ideas and determinants of this study have emerged from the context of collaborated sound design case with *Suunto Ltd*, which is a Finnish manufacturer of mobile devices for outdoor activities. The aim there is to design user interface sounds for a training application in a wrist computer. One of the main functions of the sounds within that type of interaction is to persuade the user to control her running speed. Therefore the chosen communicative functions for this study were defined as "slow down" (decrease speed), "urge" (increase speed), "keep this / OK" (current speed is fine) and finally "reward" (positive cheer). The first three functions are for speed control and the fourth one is for general encouragement.

Because the sounds in the training application are intended as relatively short auditory cues, the preferred form of the to-be-gathered function-specific vocal material was also determined to be more like short vocal gestures or communicative sound objects than spoken sentences. Also at this point of the study, due to the typical limitations of wrist devices' sound output, the focus of prosodic features is on the frequency and intensity of prosodic contours instead of spectral qualities of the sounds.

1.3.2. Research questions

In context-situated controlled setting of trainer-runner interaction, will participants encode function-specific (communicative functions mentioned above) vocal patterns in their utterances? More specifically, can we find any evidence of such prosodic cues by analysing the patterns of fundamental frequency (F_0) and intensity?

²For example, Fernald [5] has found cross-cultural evidence of stereotyped prosodic patterns associated with four communicative functions in infant-directed maternal speech.

2. METHOD

2.1. Participants

Vocalisations were gathered from a group of 20 Finnish-speaking students and personnel of University of Jyväskylä. Of the participants, 9 were male and 11 were female. The average age in the group was 24.8 years (with SD of 2.8 years).

The participants were recruited from the Department of Computer Science and Information Technology and from the departments of Teacher education and Music. Of these participants, 55% were IT-students, 25% were students of education and 15% were music students. One of the participants belonged to the University staff.

2.2. Experiment

2.2.1. Experimental design

The basic idea of the experiment was to gather context-situated utterances from participants by recording them in a realistic setting. The prosodic content of those vocal expressions is the dependent variable of the study. The primary independent variable is the communicative function, which has been divided into four distinct functions ("slow down", "urge", "keep this/ok" and "reward").

To set different conditions for the usage of nonverbal means in the expression, we also chose to use an additional *moderator variable* which determines two different methods for vocalisations: *Word condition* is a verbal form of expression using specified words for each function³. However, in this condition, words can be used freely and the participant is free to stress the words in the manner she wishes. The chosen set of words were purposely short, and aside from one expression ("pidä tämä"="keep this") words do not have exact linguistic meanings in the Finnish language. Still, they are pragmatically (by habit) considered to be appropriate for the expressions they were associated with. *Vowel condition* is a fully nonverbal form of expression (using "a"-vowel for all the functions). These two forms of expression were selected from three methods that were evaluated in the pilot testing of the experiment. The rejected third method was a free form of expression. The pilot experiment implied that freely expressed vocalisations favour a verbal channel for coding the intended information while the prosody of all expressions remained relatively similar (a bit like a "coach style"-voice with a general urging function).

Because the pragmatic nature of a situational context is assumed to be a determinative factor for the salience of communicative functions and in the actual producing of vocalisations, the control of contextual and situational factors was also taken into account in the experimental design. The context of trainer-runner interaction were brought into the experimental setting by 1) a short written scenario, which provides the background for an imaginary setting, 2) a simplified computer animation, which controls the situational procedure of interaction and, at the same time, provides information about the situational conditions. To make the experiment as natural as possible for the participants, the context created for the experiment was analogous to normal trainer-runner interaction and was not application specific to any extent. Despite that, the intended communicative functions should remain adaptable for application use.

³Finnish and pseudo-Finnish words that were used to express different communicative functions were "top" (for *slow down*), "hop" (for *urge*), "pidä tämä" (for *keep this / OK*) and "jee" (for *reward*).

2.2.2. Apparatus and setting

The experiment was conducted in a sound shielded room that is suitable for audio recording. The participants were seated in the front of a microphone and a computer screen from where they could follow the animation (see Figure 1). They were also able to hear the included environmental sounds from the earphones that were designed to facilitate the immersion into the imaginary setting at the running track. On the other hand, the button-style earphones were not closed so they did not restrict the hearing of ones own voice. The positions of the microphone and the chair along with the other parts of experiment setting remained fixed between sessions. The recording levels also remained fixed during all recordings and between all sessions. Due to the seated position, the distance between a participant and the microphone remained relatively constant (approx. 40-50 cm), although many participants felt the necessity to move their body at the time of their expression. To make the situation a more comfortable and intimate experience for the participant, the researcher and the setting were separated by a screen.

The animation was made with Macromedia Director MX2004. Other equipment used in the experiment was a Shure KSM-32 microphone, a microphone stand, an HNB Portadisc audio recorder, an HP laptop computer (for running the animation), Olympus earphones, and a Samsung 17" LCD display.



Figure 1: The experimental setting showing the computer display with the animation and a participant.

2.2.3. Procedure

The overall duration of the experiment was 10-15 minutes. At the start, the participant was given a general description of the task in a form of a written scenario. Here is the translation of the original Finnish version:

"Imagine the following scenario: You and your friend are running together. Your friend has an objective to achieve as constant lap times as possible on a short running track. You remain at the start/finish line and have promised your friend to control her speed.

As your friend passes you each lap, your task is to vocally express to her if she must increase or decrease the running speed for reaching the ideal lap time. If the speed is constant with the ideal time,

Table 1: The order of the communicative functions.

Lap	Associated communicative function
1 (warm-up)	Slow down
2 (warm-up)	Urge
3	Urge
4	Slow down
5	OK
6	Reward
7	Slow down
8	Urge
9	OK
10	Reward

then you indicate by your expression that the speed is fine. You have also planned to reward your friend with a praising cheer in the middle and at the end of the performance.”

After a moment of undisturbed concentration to the text, the communicative functions were shortly discussed. The participant was then informed that the experiment was to be divided in two similar tasks. The task specific details were explained to the participant at the beginning of each task. The tasks corresponded to Word and Vowel conditions and were otherwise identical. The Word condition task was always done first. Based on feedback from the pilot experiment, more time was needed to get accustomed to “losing the faculty of speech” thus using only “a”-vowel in expression. Therefore, arranging the Vowel condition to take place after the more intuitive Word condition was justified due to the presumed learning effect.

Each task consisted of 10 running laps. A computer animation visualised the running process with a dot moving along a circle. Towards the end of the lap the animation alarmed the participant (with a text and the sound of an approaching runner). A moment later the animation informed textually about the situational condition; i.e., whether the lap time was a) too fast b) too slow c) fine, or d) if the participant was asked to reward the runner with a cheer. In the case of the Word condition task, the corresponding verbal expression for the associated communicative function was also reminded by the animation. After receiving information about the current lap, the participant had a few seconds to respond vocally to the “passing runner” before the animation indicated that the runner had gone too far (with a marker on the circle, and by fading off the sound of the runner).

Before the tasks, the participant was informed that the purpose of the two first laps in the each condition was for warming-up. The remaining 8 laps were allocated evenly for communicative functions, hence the intended number of gathered utterances per task were 8 (2 utterances for each function). The whole structure of communicative functions associated for each running lap is shown in Table 1.

After the participant has completed both tasks, in all, 20 utterances were recorded (including 4 warming-up utterances). The performance was followed by a short spontaneous discussion with the researcher about the experience. Finally, the participant filled a small questionnaire (for performance self-evaluation) and was rewarded with a gift token for cafeteria.

2.3. Participant self-evaluation

In the questionnaire the participants were asked to evaluate their performance in each task (both Word and Vowel condition) by using a 1-5 scale to indicate the success of their vocalisations (1=successful, 5=unsuccessful). In addition, the participants were asked to give a short verbal description about the success of their expressions.

2.4. Pre-processing of audio material

All the audio recordings were first pre-processed in order to enhance their signal quality. Each take was cut out from the recording and these were organised into audio files in a suitable manner. A take here refers to all vocalisations that a participant produced under the single function-specific experimental trial. Files were then imported into the Praat 4.6 software [15] for annotation and acoustic analysis.

Despite the intended training purpose associated with the warm-up takes, it was clear that those takes could not be automatically rejected from the analysis. Because the number of utterances must be equal in all the function categories, the least affective take (out of the three) from “slow down” and “urge”-categories was rejected from both conditions for each participant.

The selection of the most relevant utterance from each take was made by automatically marking out any undivided vocalisations in the material and then choosing and labelling the most prominent vocalisation of each take. The resulting utterance should be perceived as a coherent and distinct entity in relation to its original context. For this, an automatic marking was successfully implemented by using the sound intensity based annotation feature in Praat. In 4% of all the chosen utterances, the automatically trimmed segments proved to be perceptually incoherent, and the markings had to be manually altered.

2.5. Acoustic analysis

The preprocessing of the prosodic features from audio was carried out using Praat software [15]. The fundamental frequency (F_0) and the voice intensity (energy in dBs) was obtained for each utterance using a 10 ms time-window. Even though the autocorrelation based pitch extraction generally yielded reliable estimation of F_0 , some utterances contained minor inaccuracies, mostly unwanted jumps (octaves or fifths). These errors were corrected in Praat using its pitch editor and re-evaluated by playing back the synthesised pitch contours simultaneously with the original utterances.

For all utterances, F_0 s (in Hz) were converted into linear scale by

$$P = 69 + 12 \times \log_2 \left(\frac{F_0}{440} \right), \quad (1)$$

where P represents the pitch numbering convention used in the MIDI standard ($C_4 = 60$). Note that this scaling does not alter the resolution of the F_0 as they were not reduced to the integers of the MIDI note standard. Next, the F_0 contours were centred to MIDI note 60 (261.6 Hz) within each participant to remove the obvious F_0 differences between the participants caused by gender, size, etc. For intensity, a similar operation was carried out (centred to 70 dB). The examples of the resulting frequency and intensity contours are visualised in Figure 2. In the figure, the intensity is indicated by the colour of the marker (darker colour for higher intensity). Attached sound examples are also available portraying

the utterances and synthetic renditions of the original frequency and intensity contours (see Figure 2).

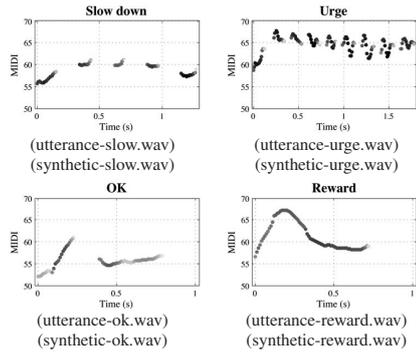


Figure 2: Examples of the F_0 and intensity contour for each four functions from a single participant (Word condition). Darker colour indicates higher dB (intensity) value. Recordings of the utterances and synthetic renditions of the original prosodic contours can be triggered by clicking the corresponding file name.

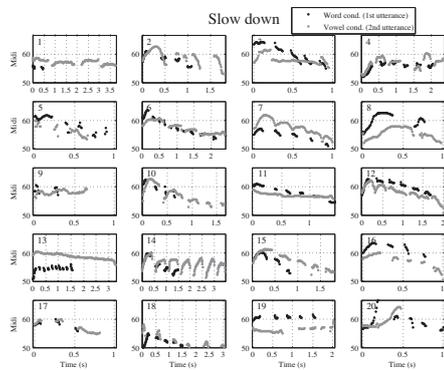


Figure 3: The F_0 contours of two utterances by all the participants for the *Slow down* communicative function.

The utterances were then summarised by 8 simple descriptors: mean frequency, F_0 (M), frequency variation, F_0 (SD), voice intensity, $VolInt$ (M), intensity variation, $VolInt$ (SD), the length of the utterances, $Length$, proportion of pauses within utterances, $Pause prop.$, and the trend of the F_0 and intensity. More sophisticated descriptors such as the attack slope, brightness or formant measures could be viable additions but there is ample evidence that relatively simple measures such as the ones outlined above are able to account for most of the differences in, for example, vocal expressions of emotions [3, 16]. Also, we wanted to focus on F_0

and intensity rather than spectral measures, as F_0 and intensity are easily manipulated in applications with limited audio generating capacities.

In order to visualise the raw data, two utterances for all the participants are displayed for two communicative functions in Figures 3 and 4. The overall patterns within the functions are visible. For example, the *Urge* function seems to have a higher frequency, shorter segments and ascending and level pitch contour. For the *Slow down* function, the segments within the utterances are longer, less variable in frequency compared to the urge segments and the pitch contour is mostly descending. What is also worth of pointing

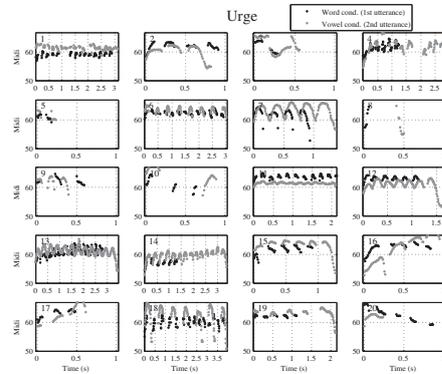


Figure 4: The F_0 contours of two utterances by all the participants for the *Urge* communicative function.

out is that the utterances representing different conditions (Word and Vowel) are remarkably similar within and for the participants, although they were given at separate experimental trials. The extent of this similarity is encouraging when thinking about the possible uses of prosodic information. Nevertheless, this issue will be later examined in detail.

3. RESULTS

3.1. Results of self-evaluation

The participants gave ratings of how well they themselves succeeded in the task. The mean values (Word cond.: 2.2 and Vowel cond.: 2.95, scalar values from 1-5 where low numbers denote a success in conveying the function, $n=20$) indicate that the utterances produced in the Word condition were evaluated as marginally more successful than utterances in the Vowel condition. Up to 85% of participants used the positive end of the scale (answers 1 or 2) to indicate the success with the Word condition, whereas only 25% of participants used similar answers in the case of the Vowel condition. Also, 8 participants described in their free verbal reports of the experiment that the Vowel condition was the harder of the two tasks. Conversely, the Word condition was described as the harder task by only 2 participants. These results imply that the Vowel condition might have been more ambiguous as an experience, and the participants were not quite sure about their own success when using only the vowel in their expressions.

3.2. Differences between repeated utterances, conditions and functions

We first investigated whether there were differences between the repeated utterances each participant gave for each function and condition. One-way ANOVA yielded no statistically significant differences in the mean F_0 s ($F[1,158]=1.22, p=n.s.$) or in mean intensities ($F[1,158]=0.04, p=n.s.$) and hence both utterances are retained in the following analyses. This also suggests that prosodic information is robust in communicating these functions and minimally altered across repetitions in the experiment.

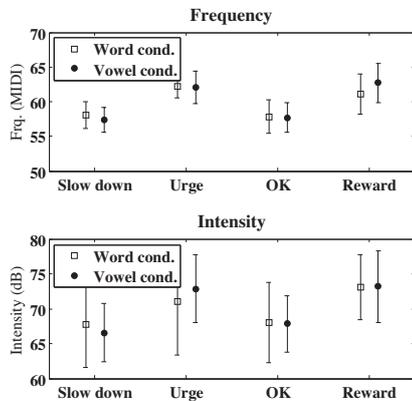


Figure 5: Mean F_0 and intensity across utterances and conditions.

Next the differences in the mean F_0 s across the conditions and functions were tested using two-way analysis of variance of condition (two levels: Word and Vowel) and function (four levels: Slow down, Urge, OK, and Reward). This analysis yielded a highly significant main effect across the function ($F[3,319]=143.9, p<0.001$) but no differences across the conditions ($F[1,319]=1.8, p=0.46$). When the same analysis was repeated with intensity, a similar pattern of results was obtained (see Figure 5). While the condition did not have an impact on these acoustic features, a similar analysis of other features revealed differences across the condition. This result was not surprising as the Word-condition was expected to provide some determinants over the vocalisation. The largest differences across the condition ($F[1,319]=55.1, p<0.001$) were found in the proportion of pauses. Differences across the conditions were also found in trend measures (F_0 and intensity) as well as in the length of the utterances. Still, despite these statistical parameters, many utterances from both conditions appeared surprisingly similar. This can clearly be observed from F_0 contours of utterances (see Figures 3 and 4), and it is also indicated by the ANOVA results of mean F_0 and intensity across the conditions.

The subsequent analysis of prosodic features for each function was carried out using one condition. We decided to focus on the Word condition as it was the preferred method for the participants (see 3.1.). A summary of comparison of acoustic features using ANOVA is given in Table 2. In addition to the means across the functions, Table 2 displays how many of the possible comparisons

Table 2: Means for acoustic features across the 4 functions.

FEATURE	Slow	Urge	OK	Rew.	Post-hoc
F_0 (M)	58.0	62.2	57.8	61.1	8/12 **
F_0 (SD)	1.9	1.6	2.4	2.8	6/12 **
F_0 trend	-0.31	0.04	-0.45	-0.28	6/12 **
Length	0.53	0.65	0.46	0.93	6/12 **
Pause prop.	0.56	0.48	0.24	0.05	12/12 **
VoInt (M)	67.7	70.9	67.9	73.1	10/12 **
VoInt (SD)	6.13	7.62	5.74	4.65	10/12 **
VoInt trend	-0.32	-0.11	-0.15	-0.20	4/12 *

ANOVA significant at * $p < 0.01$, ** $p < 0.001$.

between the functions ($4 \times 3 = 12$) contained significant differences in post-hoc (Scheffé) comparisons of the means. As can be seen, all the features are able to separate several communicative functions, although the most effective ones seem to be the Proportion of pauses and the voice intensity measures.

3.3. Classifying utterances according to acoustic features

To demonstrate the effectiveness of F_0 and intensity cues for separating the selected four communicative functions, a linear discriminant analysis (LDA) was used to classify individual utterances into the communicative functions. For this, two acoustic features were chosen, the F_0 (M) and the proportion of pauses (Pause prop.) from the previous analyses. The results of this analysis indicated that these two features were able to predict correctly 88% of the observations (see Figure 6) and thus highlighted how effective can two simple acoustic cues be in separating the functions from each other. In figure 6, the utterances can be clearly seen to cluster into distinct groups according to the proportion of pauses and mean F_0 .

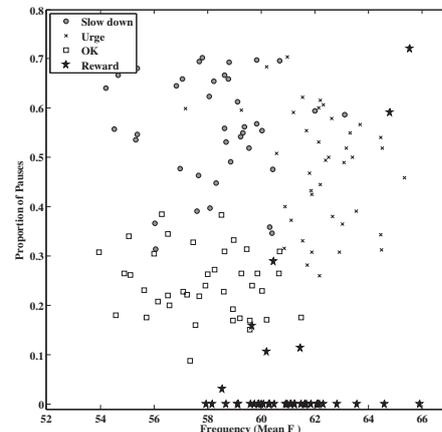


Figure 6: Scatterplot of the mean F_0 (X-axis) and Proportion of pauses (Y-axis) for each utterance representing the four communicative functions.

4. DISCUSSION

The universal, everyday usage of prosodic cues in human communication makes the prosody based information exceptionally potential source for common affective sound-meaning relations. In this study we examined whether four communicative functions of vocal utterances would produce distinct function-specific prosodic characteristics. The results demonstrated that the acoustic features of the utterances were highly successful in discriminating the functions from each other. This indicates that these vocalisations for four different communicative functions certainly have specific prosodic qualities (or invariant patterns in the Gibsonian sense), which can in turn be imitated in the design of user interface sounds for similar communicative purposes. The acoustic descriptors were fairly simple, which we interpret as an advantage, as these features of pitch and intensity are easy to manipulate and generate in applications. Moreover, the fact that even simple cues of monophonic pitch contour are effective in discriminating communicative functions (see 3.2. and Figure 6) affords the prosody based sound design even for devices that have limited sound generating capabilities.

While this study validates the assumed function-specific relations of prosodic cues, we admit that in a sense this is a halfway-result. More detailed analyses of the function-specific cues are needed in order to better understand their role in meaning-creation. In future studies we also need to perform recognition tests with listeners that will use synthesised sound examples of prosodic features in order to validate their communicative attributes. Still, even with the limited knowledge of stereotyped prosodic features, there are clear adaptation possibilities for sound design by imitating selected prosodic cues. The simplest form of adaptation would be more or less complete imitation of prosodic contours (pitch and/or intensity) that are found to represent characteristic qualities of a certain communicative function. To demonstrate this, we prepared special versions of audio examples that were portrayed in Figure 2. These sounds (see Table 3) are otherwise direct renditions of the original pitch contours except that they are transposed to a higher register and the contours are quantized to follow discrete pitches (in semitones). By listening to these modified versions, one is able to get an idea of how these intonations might work as typical, monophonic beeper sounds.

By using traditional terminology of auditory display research, the prosody based sound design may be seen as a relative to the design of *auditory icons* by Gaver [17] or *representational earcons* by Blattner et al. [18], which both share the same idea of imitating familiar aspects of our everyday environment. However, it is important to note that the prosodic encodings of sound engage primarily the listeners' empathetic and functional listening modes (i.e., levels of meaning-creation, see [12]), and they will not necessarily rule out the concurrent usage of, for instance, symbolic codes or other types of iconic resemblances. The utilisation of the prosodic features of speech in sound design can be seen as a *design paradigm* of its own. As such, the prosody based perspective emphasises affective and functional (pragmatic) viewpoints on meaning-creation. It can be applied to the design of many types of communicative sounds, and the sound designer should be able to utilise it in tandem with other design paradigms.

The methodology used for collecting the utterances representing various functions seemed to work in a way intended. The participants were able to produce utterances that fitted with each communicative function and were satisfied with their performance and

Table 3: Discrete pitch level renditions of frequency contours of the four utterances displayed in Figure 2. Sound examples can be triggered by clicking the corresponding file name.

Communicative function	Sound example
Slow down	(beeper-slow.wav)
Urge	(beeper-urge.wav)
OK	(beeper-ok.wav)
Reward	(beeper-reward.wav)

the experimental setup. Thus the method can be recommended for similar purposes of gathering function-specific vocalisations that matches the communicative functions of intended user interface sounds. As the condition (i.e., the method of vocalisation) did not seem to have too dramatic impact to the prosodic qualities of utterances, one might prefer to use the more natural verbal or pseudo-verbal form of expression. According to our observations, utterances in the Word condition produced somewhat more brisk and solid expressions. In fact, the choice of a vocalisation's verbal form can be considered as a way by which the sound designer can determine some aspects of the collected utterances. It should be noted, however, that the participant should be encouraged to communicate nonverbally in the experiment. Indeed, putting too much emphasis on the verbal side of an expression can also be misleading.

As a consideration for future research, cross-cultural studies would be beneficial for studying the possible cultural differences in encoding and decoding prosodic information beyond the already observed similarities [16, 6]. Another issue concerns the communicative functions: What kind of - and how many different (prosodically non-redundant) - functions of nonverbal vocal communication can be found that are compatible with human-computer interaction? Such taxonomical charting would provide the crucial framework for the future investigations of prosody based sound design.

5. ACKNOWLEDGEMENTS

This work is funded by Finnish Funding Agency for Technology and Innovation, and the following partners: Nokia Ltd., GE Healthcare Finland Ltd., Sunit Ltd., Suunto Ltd., and Tampere City Council.

6. REFERENCES

- [1] A. L. Blackler and J. Hurtienne, "Towards a unified view of intuitive interaction: definitions, models and tools across the world," *MMI-Interaktiv*, vol. 13, pp. 37-55, 2007.
- [2] B. Arons and E. Mynatt, "The future of speech and audio in the interface: a chi '94 workshop," *SIGCHI Bull.*, vol. 26, no. 4, pp. 44-48, 1994.
- [3] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychological Bulletin*, vol. 129, no. 5, pp. 770-814, 2003.
- [4] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, pp. 614-636, 1996.

- [5] A. Fernald, "Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective," in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, J. H. Barkow, L. Cosmides, and J. Tooby, Eds., pp. 391–428. Oxford University Press, 1992.
- [6] A. Fernald, "Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages," *Child Development*, vol. 64, no. 3, pp. 657–674, 1993.
- [7] K. R. Scherer, "Feelings integrate the central representation of appraisal-driven response organization in emotion," in *Feelings and Emotions: The Amsterdam Symposium*, A.S.R. Manstead, N.H. Frijda, and A.H. Fischer, Eds., pp. 136–157. Cambridge University Press, Cambridge, 2004.
- [8] K. R. Scherer and H. Ellgring, "Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal?," *Emotion*, vol. 7, no. 1, pp. 113–130, 2007.
- [9] D. Huron, "A six-component theory of auditory-evoked emotion," in *Proceedings of the 7th International Conference on Music Perception and Cognition, Sydney, 2002*, C. Stevens, D. Burnham, G. McPherson, E. Schubert, and J. Renwick, Eds., pp. 673–676. Causal Productions, 2002.
- [10] L. Carr, M. Iacoboni, M.-C. Dubeau, J. C. Mazziotta, and G. L. Lenzi, "Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas," *PNAS*, vol. 100, no. 9, pp. 5497–5502, April 2003.
- [11] M. Iacoboni, I. Molnar-Szakacs, V. Gallese, G. Buccino, J. C. Mazziotta, and G. Rizzolatti, "Grasping the intentions of others with one's own mirror neuron system," *PLoS Biology*, vol. 3, no. 3, pp. e79, 2005.
- [12] K. Tuuri, M.-S. Mustonen, and A. Pirhonen, "Same sound - different meanings: A novel scheme for modes of listening," in *Proceedings of Audio Mostly 2007, 2nd Conference on Interaction with Sound*, 2007, pp. 13–18, Fraunhofer IDMT.
- [13] S. Rosenthal, *Speculative pragmatism*, University of Massachusetts Press, Amherst, MA, US, 1986.
- [14] K. R. Scherer, R. Banse, and H. G. Wallbott, "Vocal affect expression: a review and a model for future research," *Psychological Bulletin*, vol. 99, no. 2, pp. 143–165, Mar 1986.
- [15] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glott International*, vol. 5, no. 9/10, pp. 341–345, 2001.
- [16] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Crosscultural Psychology*, vol. 32, pp. 76–92, 2001.
- [17] W. Gaver, "Auditory icons: Using sound in computer interfaces," *Human-Computer Interaction*, vol. 2, pp. 167–177, 1986.
- [18] M. Blattner, D. Sumikawa, and R. Greenberg, "Earcons and icons: Their structure and common design principles," *Human-Computer Interaction*, vol. 4, pp. 11–44, 1989.

PVI

**IDENTIFYING FUNCTION-SPECIFIC PROSODIC CUES FOR
NON-SPEECH USER INTERFACE SOUND DESIGN**

by

Kai Tuuri & Tuomas Eerola 2008

In J. Pakarinen, C. Erkut, H. Penttinen & V. Välimäki (Eds.), Proceedings of the
11th International Conference on Digital Audio Effects, Helsinki University of
Technology, Espoo, pp. 185–188

IDENTIFYING FUNCTION-SPECIFIC PROSODIC CUES FOR NON-SPEECH USER INTERFACE SOUND DESIGN

Kai Tuuri

Dept. of Computer Science and Information Systems
University of Jyväskylä
P.O. Box 35, FI-40014, Finland
krtuuri@jyu.fi

Tuomas Eerola

Department of Music
University of Jyväskylä
P.O. Box 35, FI-40014, Finland
tuomas.eerola@campus.jyu.fi

ABSTRACT

This study explores the potential of utilising certain prosodic qualities of function-specific vocal expressions in order to design effective non-speech user interface sounds. In an empirical setting, utterances with four context-situated communicative functions were produced by 20 participants. Time series of fundamental frequency (F_0) and intensity were extracted from the utterances and analysed statistically. The results show that individual communicative functions have distinct prosodic characteristics that can be statistically modelled. By using the model, certain function-specific prosodic cues can be identified and, in turn, imitated in the design of communicative interface sounds for the corresponding communicative functions in human-computer interaction.

1. INTRODUCTION

Finding ways to produce intuitively communicative non-speech sounds is a major challenge in the sound design for user interfaces. An interface sound can be seen intuitively communicative if the users' unconscious application of previous experience facilitates effective interaction [1]. One way to exploit our familiarity and facility in experiencing the everyday world is to mimic the ways in which we naturally use sound with social interactions. In addition to linguistic means of expression, the human vocal communication contains an important nonverbal channel. This affective content of speech is conveyed by various *prosodic* cues, which refer to certain characteristics in intonation, stress, timing and voice quality - or by acoustic terms - in dimensions such as pitch, intensity and spectrum. While many professional sound designers might tacitly mimic various prosodic cues in their work, there is a definitive lack of explicit knowledge of how certain prosodic characteristics are related with the human meaning-creation.

Vocal expressions are in many ways dependent on the situational context in which they take place and which they serve. Emotional and motivational states reflect the current situation and provide various effects to the determinants of the vocalisation. A wealth of evidence exists that emotional and intentional states are communicated non-verbally through vocal expressions [2]. The ability to catch the emotional and motivational state of mind of other people has been considered as crucial in forming and maintaining social relationships [3]. The unveiling of emotional and motivational states can also be utilised for manipulation and persuasion. The speaker also instrumentally uses the expression to convey information to the others and to influence the communicational process.

Communicative functions of vocalisations refer to the communicative intentions of the speaker as well as to the vocalisation's pragmatic meaning. We suggest that the evoked functional meaning¹ of nonverbal vocal patterns is indicated by the empathetic perception [5] of sound and its indexical relation to the situational context. The dependency to the situational conditions may vary. For example, an infant can perceive mother's vocal patterns as *prohibitive* in many different situations as long as the child is able to associate the utterance with her actions. On the other hand, the perception of certain functions may have more fine-tuned relationships between the vocalisation and its context. Communicative functions represent particular categories of vocal expression and also certain contexts of interactions. In this study we will use this term essentially to categorise certain context-specific communicative intentions for using sound.

It is pointed out by several authors [3, 2, 6] that the basis of encoding and decoding of prosodic features in vocal communication has a strong *phylogenetic* background. Such evolutionary perspective is supported e.g. by the evidence of cross-cultural prosodic similarities in infant-directed speech [7]. It is hardly the case that all codes related to nonverbal vocal expressions are "hard-wired" into the human species. One can assume that several parts of the coding consist of socio-culturally learned habits. But if the feature determinants and non-verbally evoked meanings of vocal patterns have even partial universality, these codes must be considered to be serving as a source of common sound-meaning relations.

In order to utilise function-specific prosodic cues in sound design, one must identify such stereotyped cues in certain function-related vocalisations. The goal of this study is to address this issue in the context of collaborated sound design case with *Suunto Ltd.*, a Finnish company designing and manufacturing sports instruments. The aim of the case is to design user interface sounds for training application in a wrist computer. One of the main functions of the sounds within that type of interaction is to persuade the user to control her running speed. Therefore the chosen communicative functions for this study were defined as "slow down" (decrease speed), "urge" (increase speed), "keep this / OK" (current speed is fine) and finally "reward" (positive cheer). Due to the typical limitations of wrist devices' sound output, the focus of prosodic features is on frequency and intensity instead of spectral qualities of the sounds. The research questions of this study are: In context-situated controlled setting of trainer-runner interaction, will participants encode function-specific (communicative functions mentioned above) vocal patterns in their utterances? And

¹See Tuuri et. al. [4] for a discussion about the perspectives of sonic meaning-creation.

can we identify such discriminating prosodic cues by analysing the patterns of fundamental frequency (F_0) and intensity?

2. METHOD

2.1. Participants

Vocalisations were gathered from a group of 20 Finnish-speaking students and personnel of University of Jyväskylä. Of the participants, 9 were male and 11 were female. The average age in the group was 24.8 years (with SD of 2.8 years). The participants were recruited from various departments of the university. Of these participants, 55% were IT-students, 25% were students of education and 15% were music students. One of the participants belonged to the university staff.

2.2. Experiment

The basic idea of the experiment was a controlled production task for gathering context-situated utterances from participants by recording them in a realistic setting. The prosodic content of those vocal expressions is the dependent variable of the study. The primary independent variable is the communicative function divided into four distinct functions ("slow down", "urge", "keep this/ok" and "reward").

To set different conditions for the usage of nonverbal means in the expression, we also chose to use an additional *moderator variable* which determines two different methods for vocalisations: *Word condition* is a verbal form of expression using specified words for each function². *Vowel condition* is a fully nonverbal form of expression (using "a"-vowel for all the functions). For a more detailed description of the experiment, see Tuuri and Eerola [8].

2.3. Pre-processing of material

In the experiment, each participant produced 10 takes under each (Word and Vowel) condition. A take here refers to recorded vocalisations that a participant produced under a single function-specific experimental trial. Due to the extra warm-up trials under "slow down" and "urge" categories, one trial from both of these categories under both conditions were rejected resulting in a total of 16 takes ($4 \text{ functions} \times 2 \text{ repetitions} \times 2 \text{ conditions}$) per participant. The selection of the most relevant utterance from each take was made by automatically marking out any undivided vocalisations in the material and then choosing and labelling the most prominent vocalisation of each take. For this, an automatic marking was successfully implemented by using the sound intensity based annotation feature in Praat 4.6 software [9].

The extraction of the prosodic features from audio was carried out using Praat software [9]. The fundamental frequency (F_0) and the voice intensity (energy in dBs) was obtained for each utterance using a 10 ms time-window. Even though the autocorrelation-based pitch extraction generally yielded reliable estimation of F_0 , some utterances contained minor inaccuracies, mostly unwanted jumps (octaves or fifths). These errors were corrected in Praat using its pitch editor and re-evaluated by playing back the synthesised pitch contours simultaneously with the original utterances.

²Finnish and pseudo-Finnish words that were used to express different communicative functions were "top" (for *slow down*), "hop" (for *urge*), "pidä tämä" (for *keep this / OK*) and "jee" (for *reward*).

For all utterances, F_0 s (in Hz) were converted into linear scale using the pitch numbering convention of MIDI standard ($C_4 = 60$). Note that this scaling does not alter the resolution of the F_0 as they were not reduced to the integers of the MIDI note standard. Next, the F_0 contours were centred to MIDI note 60 (261.6 Hz) within each participant to remove the obvious F_0 differences between the participants caused by gender, size, etc. For intensity, a similar operation was carried out (centred to 70 dB).

3. RESULTS

We first investigated whether there were differences between the repeated utterances each participant gave for each function and condition. One-way ANOVA yielded no statistically significant differences in the mean F_0 s ($F[1,158] = 1.22, p=n.s.$) or in mean intensities ($F[1,158] = 0.04, p=n.s.$) and hence both utterances are retained in the following analyses. This also suggests that prosodic information is robust in communicating these functions and minimally altered across repetitions in the experiment. Within the scope of this paper, the subsequent analysis of prosodic features for each function was carried out using solely the utterances of Word condition.

3.1. Acoustic predictors

The utterances were summarised by 15 descriptors related to frequency, intensity and length: means, standard deviations and slopes were calculated for frequency and intensity of the utterances. Also three periodicity measures of the frequency and intensity time-series were computed to characterise the possible oscillatory patterns of the utterances. For this, auto-correlation function was applied to the time-series, and the maximum amplitude and the period at the maximum as well as the entropy of the auto-correlated signal were used as descriptors of periodic patterns for frequency and intensity contours. Finally, the proportion of unvoiced frames within the utterances, the overall length of the utterances and mean voiced segment length within the utterances were computed. These predictors are listed in Table 1. This summary table also contains an index (the ANOVA column) of the predictors' ability to discriminate the four communicative functions using an analysis of variance and the subsequent posthoc test (Scheffé). The index refers to the proportion of comparisons that gave a positive result in this analysis (max. of 6 group comparisons). More advanced descriptors such as the attack slope, spectral centroid or formant variables could have been used as well, although we wanted to focus on frequency and intensity rather than on spectral measures, as these are easily manipulated in applications with limited audio generating capacities.

3.2. Classification using Regression Tree Analysis

As already shown by the ANOVA column of the Table 1, most predictors demonstrate differences across the functions and few can be observed to show differences between most of the group comparison (Prop. of unvoiced frames, intensity measures). However, in order to better understand which *combination* of the available acoustic features contributes the most to the separation of the four function categories, a classification approach was adopted. To classify properly the utterances into four function-specific groups, we chose to apply Regression Tree Analysis (RTA) [10]. RTA constructs rules by recursively partitioning the observations into

Table 1: Summary of predictors.

Nro	Predictor	ANOVA
1.	Frequency (Mean)	4/6
2.	Frequency (Standard deviation)	3/6
3.	Frequency (Slope)	2/6
4.	Frequency Periodicity (Max. ampl.)	3/6
5.	Frequency Periodicity (Max. period)	3/6
6.	Frequency Periodicity (Entropy)	0/6
7.	Intensity (Mean)	5/6
8.	Intensity (Standard deviation)	5/6
9.	Intensity (Slope)	2/6
10.	Intensity Periodicity (Max. ampl.)	4/6
11.	Intensity Periodicity (Max. period)	3/6
12.	Intensity Periodicity (Entropy)	2/6
13.	Proportion of unvoiced frames	6/6
14.	Total length	3/6
15.	Mean voiced segment length	3/6

ANOVA column displays the number of functions that are statistically different at $p < 0.05$ level in Scheffé posthoc comparisons.

smaller groups based on a single variable at a time. These splits are created to maximise the between groups sum of squares. The resulting tree diagram initially has a large number of tree nodes (logical *if-then* conditions) which are pruned by cross-validation to reduce the overfitting. This approach provides several advantages over discriminant analysis (DA) or classical regression techniques: it is able to uncover structures in observations which are hierarchical, it is nonparametric, and allows interactions and nonlinearities between the predictors [11]. The rules that describe the splitting into groups are also easy to interpret and provide insights into the process of classification.

For the analysis, all predictors were checked for normality, and those violating the normality assumption were transformed into normal distribution using a Box-cox power transformation. However, 3 predictors (Prop. of unvoiced frames, Period maximum amplitude and period measures in frequency) could not be successfully transformed. It should be noted, however, that this does not pose problems for the RTA analysis. All predictors were converted into z-scores and entered into the RTA analysis, which yielded classification accuracy of 88.75% with 10-fold cross-validation in which excessive tree nodes were trimmed. This final model had 3 nodes and 2 predictors, as displayed in Figure 1. Thus the Proportion of unvoiced frames of the utterances was the most discriminative feature³ of the function-specific categories as it separates *Reward* category from the other categories and further distinguishes *OK* utterances from *Slow down* utterances. Moreover, *Urge* utterances are also clearly separated by the higher mean frequency from the other categories. This simple RTA model and the actual observations are visualised in Figure 2). The smaller markers denote the predictions by the model (the four areas marked by the RTA decision tree) and the larger markers represent the 160 observations (20×2 utterances \times 4 categories). Note that the Proportion of unvoiced frames clearly has a non-normal distribution (a large amount zeroes), which would have been problematic for classical classification analyses. In figure 2, the utterances can be clearly be

³One should interpret the Proportion of unvoiced frames in the classification as an index of segmented utterances which could consist of, for example, series of short bursts of sound.

seen to cluster into distinct groups according to the Proportion of unvoiced frames and mean frequency.

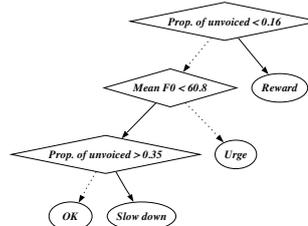


Figure 1: Decision tree based on the final, pruned and cross-validated RTA model. Solid lines indicate the path taken when a rule is filled.

We also compared the results to those obtained with the linear discriminant analysis using all 15 variables with a stepwise option to trim the amount of variables according Rao's V [12]. This analysis resulted in a 2 variable solution that correctly classified 88.8% of the observations (without a cross-validation procedure). These two variables were again the Proportion of unvoiced frames and mean frequency. Hence similar results were demonstrated using a more traditional technique.

Additionally, by using the RTA classification model we explored a set of the categorically best-ranking utterances (ranked by the distance from the group centroids). These utterances should be, according to the model, statistically the best representatives of a given communicative function. As an example, one high-ranking utterance of each communicative function (given from a single participant) is shown in Figure 3, where the frequency and intensity contours of utterances are visualised.

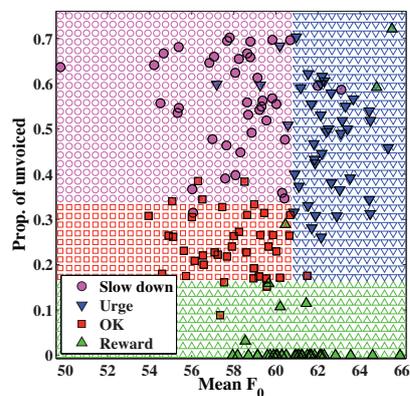


Figure 2: Scatterplot of the two predictors (*Mean frequency & Proportion of unvoiced frames*) that were able to classify most utterances into the four function-specific categories. Note that the original predictor values (not the z-scores) are displayed.

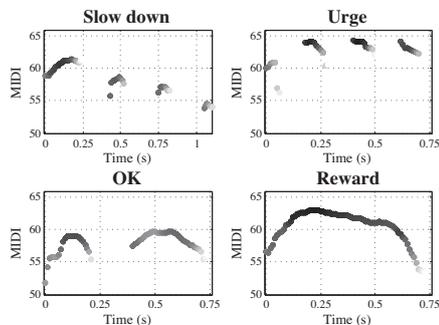


Figure 3: Examples of the frequency and intensity contours for each four functions from a single participant (Word condition). Darker colour indicates higher dB (intensity) value.

4. DISCUSSION

The universal, everyday usage of prosodic cues in human communication makes the prosodic information exceptionally potential source for common affective sound-meaning relations. In this study we examined whether four different communicative functions of vocal utterances would produce distinct function-specific prosodic characteristics. The results demonstrate that the acoustic features of the utterances were highly successful in discriminating the functions from each other. This indicates that these vocalisations for four different communicative functions certainly have specific prosodic qualities, which can, in turn, be imitated in the design of user interface sounds of similar communicative purposes. The acoustic descriptors were fairly simple, which we interpret as an advantage, as these features of pitch and intensity are easy to manipulate and generate in applications.

While this study sheds light on the characteristics of function-specific prosodic cues, we admit that this is a halfway-result. Further studies of the function-specific cues are needed in order to better understand their role in meaning creation. For example, recognition tests with listeners that will use synthesised sound examples of prosodic features should be carried out in order to validate their communicative attributes. Still, even with the limited knowledge of stereotyped prosodic features, clear possibilities exist for utilising prosodic information as a basis of user interface sound design.

The prosody based sound design may be seen as a relative to the design of *auditory icons* by Gaver [13], as both share the same idea of imitating familiar aspects of our everyday environment. Iconic references to the original vocalisations should be considered in two levels: imitation of prosodic features and imitation of communicative function. Hence, for the sake of functional matching and as a natural part of interaction design, it is crucial to define the communicative functions (i.e., purposes) for every sound occurring in the interaction. It is also important to note that the prosodic encodings of sound engage primarily the listeners' empathetic and functional listening modes (i.e., levels of meaning-creation, see [4]), and they will not necessary rule out the concurrent usage of, for instance, symbolic codes or other types of iconic resemblances. Prosody based sound design can thus be applied to the design of many types of communicative sounds, and the sound

designer should be able to utilise it in tandem with other design paradigms.

5. ACKNOWLEDGEMENTS

This work is funded by Finnish Funding Agency for Technology and Innovation, and the following partners: Nokia Ltd., GE Healthcare Finland Ltd., Sunit Ltd., Suunto Ltd., and Tampere City Council.

6. REFERENCES

- [1] A. L. Blackler and J. Hurtienne, "Towards a unified view of intuitive interaction: definitions, models and tools across the world," *MMI-Interaktiv*, vol. 13, pp. 37–55, 2007.
- [2] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, pp. 614–636, 1996.
- [3] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychological Bulletin*, vol. 129, no. 5, pp. 770–814, 2003.
- [4] K. Tuuri, M.-S. Mustonen, and A. Pirhonen, "Same sound - different meanings: A novel scheme for modes of listening," in *Proceedings of Audio Mostly 2007, 2nd Conference on Interaction with Sound*, 2007, pp. 13–18, Fraunhofer IDMT.
- [5] M. Iacoboni, *Understanding Others: Imitation, Language, and Empathy*, pp. 77–100, Perspectives on Imitation From Neuroscience to Social Science - Volume 1: Mechanisms of Imitation and Imitation in Animals. MIT Press, 2005.
- [6] A. Fernald, "Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective," in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, J. H. Barkow, L. Cosmides, and J. Tooby, Eds., pp. 391–428. Oxford University Press, 1992.
- [7] A. Fernald, "Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages," *Child Development*, vol. 64, no. 3, pp. 657–674, 1993.
- [8] K. Tuuri and T. Eerola, "Could function-specific prosodic cues be used as a basis for non-speech user interface sound design?," in *Proceedings of the 14th International Conference on Auditory Display, Paris, France, 2008* (in press).
- [9] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glott International*, vol. 5, no. 9/10, pp. 341–345, 2001.
- [10] L. Breiman, J. Freidman, R. Olshen, and C. Stone, *Classification and regression trees*, Wadsworth, Belmont, CA, USA, 1984.
- [11] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.
- [12] G. J. McLachlan, *Discriminant analysis and statistical pattern recognition*, Wiley-Interscience, New York, NY, USA, 2004.
- [13] W. Gaver, "Auditory icons: Using sound in computer interfaces," *Human-Computer Interaction*, vol. 2, pp. 167–177, 1986.

PVII

**DESIGN AND EVALUATION OF PROSODY BASED
NON-SPEECH AUDIO FEEDBACK FOR PHYSICAL TRAINING
APPLICATION**

by

Kai Tuuri, Tuomas Eerola & Antti Pirhonen

International Journal of Human-Computer Studies, Accepted pending revision

Reproduced with kind permission of Elsevier.

Design and Evaluation of Prosody-Based Non-Speech Audio Feedback for Physical Training Application[☆]

Kai Tuuri^a, Tuomas Eerola^b, Antti Pirhonen^a

^a*Department of Computer Science and Information Systems, University of Jyväskylä,
P.O. Box 35, FI-40014, Finland*

^b*Department of Music, University of Jyväskylä, P.O. Box 35, FI-40014, Finland*

Abstract

Methodological support for the design of non-speech user interface sounds for human-computer interaction is still fairly scarce. To meet this challenge, this paper presents a sound design case which, as a practical design solution for a wrist computer physical training application, outlines a prosody-based method for designing non-speech user interface sounds. The principles used in the design are based on nonverbal communicative functions of prosody in speech acts. These principles are founded on an interpersonal extension to the prevailing paradigm of sonic interaction design. The stages of the design process are justified with a theoretical analysis and three empirical sub-studies, which comprise production and recognition tasks involving four communicative functions. The final evaluation study indicates that the resulting sounds of the design process successfully served these functions. In all, this study suggests that prosody-based sound design provides widely applicable means to attribute meaningful, interaction-derived qualities to non-speech sounds for interactive applications.

Key words: sound design, interaction design, non-speech sounds, design process, prosody, embodied cognition, intentionality

[☆]This work is funded by Finnish Funding Agency for Technology and Innovation, and the following partners: GE Healthcare Finland Ltd., Suunto Ltd., Sandvik Mining and Construction Ltd. and Bronto Skylift Ltd.

Email addresses: kai.tuuri@jyu.fi (Kai Tuuri), tuomas.eerola@jyu.fi (Tuomas Eerola), antti.pirhonen@jyu.fi (Antti Pirhonen)

Preprint submitted to International Journal of Human-Computer Studies January 17, 2011

1. Introduction

Even though hearing is one of our primary senses for interacting with the world, using sounds as user interface (UI) elements is still far from being a matter-of-course. Development of user interfaces has strongly focused on textual and graphical forms of presentation. Traditionally, sounds – mainly as simple beeps – have been used in warnings and to grab attention. Using sound for these functions is fairly natural considering that the sense of hearing has had an important primordial role as "watchman" to warn about dangers looming in the environment. Hence, sonic interaction between individual and the environment provides a pivotal framework in understanding the roles of sound in a UI. But apart from interaction with the environment, sonic interaction has also played an equally important role in maintaining social relationships and in the development of culture. It is this communicative role of sound and its ability to help us understand each other that deserves more attention in the development of user interfaces.

There is a clear need to learn about the interaction between a user and a technical device when there is no keyboard, large display or mouse available. The role of sound becomes more important in applications with, e.g., small or ubiquitous devices where the visual attention of the user cannot be taken for granted. Also at a more general level, sounds can offer a uniquely rapid and affective means for communication. In the pursuit for more effective, intuitive and enjoyable user interfaces, sounds can increasingly take on common and natural roles in the communication with machines.

To facilitate intuitive interaction, a user interface sound needs to involve something familiar. Our answer to this requirement for familiarity is to exploit the everyday usage of sound in social encounters. Maybe the most straightforward approach is to mimic the ways in which we use our voice for nonverbal communication. In addition to linguistic means of expression, human vocal communication includes an important nonverbal channel. This "affective component" of speech consists of various *prosodic cues*, which refer to certain characteristics in intonation, stress, timing and voice quality or – in acoustic terms – in dimensions such as pitch, intensity and spectrum. In this paper, we focus on couplings between a sound structure and nonverbal meanings being communicated through vocal expressions. Through descriptions of a case study and the related sub-studies, the aim is to exemplify that this perspective has considerable potential to contribute to human-computer interaction (HCI) and to the design of usable products.

1.1. Paradigms of Non-Speech UI Sound Design

In a sense it would be logical to consider speech as the foremost auditory approach for a UI. Speech indeed is a promising mode of human-computer interaction, but while having many advantages, it is not suitable for all purposes (see, e.g., Ramloll et al., 2001). Non-speech sounds, on which this study is focused, have been proposed as a qualitatively different information presentation mode (Brewster, 2003).

The field of non-speech sound research has formulated two main paradigms of sound design. One focuses on the *symbolic* relation between sound and meaning (Blattner et al., 1989; Brewster et al., 1995). By using an arbitrary coding, it is seen that sounds themselves do not have to resemble what they represent. Conversely, the other design paradigm focuses on resemblances that sound may evoke. This *ecological* paradigm is based on accounts of *everyday listening* (Gaver, 1989) and an ecological view of perception (Gaver, 1993b), and it emphasises sound design that utilises familiar acoustic aspects of the everyday environment by mimicking them (Gaver, 1993a; Rocchesso and Fontana, 2003). According to Gibson’s theory of ecological perception (1979), our interaction with the world is full of meanings that we can perceive rapidly without much effort. Drawing upon the action-oriented bias of the human organism, these meanings are based on understanding of action-relevant invariant properties of the environment that relate to our sensory-motor experiencing of the world (i.e., affordances). In other words, understanding is embodied and inseparably coupled with the environment due to the experiential history of using our bodies to interact with it (see Varela et al., 1991).

The ecological paradigm has been adopted into the recent trend of sonic interaction design, which particularly recognises the strong coupling between meaning and interaction (Rocchesso et al., 2008). However, the concept of sonic interaction has been applied mostly in a way that concerns interactions within a material environment, i.e., recognising the material and inertial properties of interacting objects (Gaver, 1993a; Rocchesso and Fontana, 2003). Since we are not alone in this world, interactions with material objects and the environment surely provide only a partial or restricted framework for ecologically focused sound design. We thus intend to expand the account of interaction-based sound design to also emphasise the couplings between human individuals – counting especially the aspects of embodied *interpersonal attunement* and *communication of intentionality*. Human-caused sounds, which are recurrently used and interpreted as communication, can also be

seen in terms of everyday listening. If such recurrent experiences correlate with stereotypical acoustic patterns, those patterns are "directly" meaningful invariants to be utilised in sound design. The usage of prosodic cues in vocal communication is a universal everyday phenomenon, which makes it a very promising source for such invariants.

1.2. Goals of the Study

With a theoretical review, we firstly intend to justify the interpersonal dimension within the ecological paradigm of UI sound design – working our way into the design principles that are based on an embodied perspective on human cognition. Secondly, we will apply this interpersonal approach to a UI design case for a wrist computer physical training application. Within the design process, our goal is to justify and evaluate the utilisation of acoustic properties of vocal expressions used in interpersonal interaction analogous to the interactional context of the application.

Sound design inevitably consists of analytical and creative processes. Indeed, it is essentially a creative and expressive task, rather than just consisting of engineering of acoustic features. Despite the widely accepted advantages of UI sounds and extensive research in the area¹, there is not sufficient methodological support for sound design. The tacit craftsmanship of a professional sound designer, in many ways comparable to work of an artist, has proven difficult to formalise into guidelines. It has been recognised (Malouin and Landry, 1983) that guidelines for such complex processes should have heuristic characteristics (as opposed to algorithmic), hence their potential should not be hindered by attempting to make them exhaustive.

By describing the complete sound design case, this study outlines a method, which is analytical and based on a solid theoretical and empirical basis, while still being holistic by building on the context of use. Our design method proposes to bind creativity and the use of analytical methods together in an effective design process. The method can be utilised as such, when appropriate, or the general ideas of the approach can be modified to other contexts as needed.

¹For example, see the work of International Community for Auditory Display <http://www.icad.org/>.

2. Interpersonal Approach to Sonic Interaction Design

One of the most fundamental aspects of the human mind is its ability to have different mental states. Intentionality, the central concept of this study, refers to directed mental states (e.g., emotions and intentions) that are *about* something (Searle, 1983). For us, as social beings, it is important to be able to "catch" and understand the intentionality of other people.

2.1. Embodied Basis of the Communication of Intentionality

For the phenomenology of Merleau-Ponty (Merleau-Ponty, 1962), the basis of interpersonal understanding lies in the bodily existence which intrinsically mediates the intentional relationship between corporeal subject and the world. This perspective strongly emphasises one's embodied experience – how the surrounding world exists for us *in* and *through* our bodies. A prominent aspect of this perspective lies in the concept of a *corporeal intentionality* (or motor intentionality); a mode of knowledge and expression which acknowledges the "flesh" as intrinsically purposive in relation to its worldly needs (Dillon, 1997). Since the physical constitution of the human body is universal, the bodily existence of another subject manifests intentionality which can be perceived in terms of the perceiver's own embodied ontology of intentional states. In line with such an idea of bodily projection of intentionality, many authors have suggested that humans possess a primordial capacity for interpersonal understanding (see a review in Wachsmuth et al., 2008). Such capacity is arguably based on sensory-motoric *mirroring mechanisms* which enable embodied attunement to the actions of others (Rizzolatti and Craighero, 2004; Iacoboni et al., 2005; Gallese et al., 2007).

"Mirrored", ideomotoric experiences of action are thus understood in terms of our own bodily existence, action repertoire and experiential background, i.e., as if they were actions of our own. Decades ago, motor theories of speech perception (Lieberman and Mattingly, 1985) already suggested that speech understanding could be mediated by ideomotor understanding of vocal action mimicking the motor movements (phonetic gestures) by which sounds are produced in the phonatory apparatus. Interestingly, it was later found that, due to the mirror-mechanism, speech-related motor areas of the brain are indeed activated in the course of speech perception (Rizzolatti and Arbib, 1998). Therefore in a sound-producing action, such as a vocal expression, the resulting sound derives physical cues of motor movements involved *in* the action (i.e., gestural signatures). These cues, in turn, should be able

to evoke observer-dependent kinaesthetic attributions (Tuuri, 2010; Godøy, 2010).

There is a bias to perceive sound as being intentional, especially if it suggests biologically relevant movement patterns (Leman, 2008). The suggested capacity for motor-related action understanding thus essentially allows us to primordially attune to another person's corporeal intentionality being reflected in acoustic patterns. It is essential to keep in mind that the attribution of such a *derived* intentionality is observer-dependent (Searle, 1983) and its understanding is bound up with the situational context of the perceived action. The perception of intentionality is thus ultimately completed by empathetical, contextual inference about the motivations of the imagined/observed individual being engaged in the interaction (Leman, 2008; Tuuri et al., 2007).

It is interesting to notice that motor-based attributions of intentionality can be extended to listening experiences of basically any kind of sounds, even artificial ones, as long as these sounds are able to imply motor movements. For instance, music can be understood as an intentional object (Leman, 2008), even when it does not directly denote the presence of another person. Sounds should thus be able to convey interpersonal affective qualities, without needing to render direct implications of another person, or to be excessively realistic in their anthropomorphism. Therefore, the utilisation of human attributes in sounds does not necessarily restrict the design, nor should it necessarily lead to problems that have been associated with the use of anthropomorphism in UIs (Shneiderman, 1997).

Whether or not the aim is to use any kind of anthropomorphism in design, the communication of intentionality is still a matter of relevance to UI sound design, as interaction with UI elements inherently carries with it intentional connotations (Dourish, 2001). Therefore, in the context of an HCI application, sound events in a UI inevitably suggest themselves as being purposive. For this reason, sound design should ensure that *every* UI sound instance appropriately illustrates its premeditated (communicative) purpose, that is, an intention to be understood in a certain, context-relevant way. Being put into the context of an interpersonal approach, the sound designer should thus basically conceive her role as an intentional person who is "speaking" to the user.

2.2. Prosody-Based Sound Design

This study specifically aims at utilising nonverbal prosodic characteristics of vocal expressions in sound design. The idea of prosody-based design is based on an assumption that there exists a coupling between prosodic structures and the intentional stance of a vocaliser within the interactional context. Within a theoretical framework of embodied communication of intentionality, this premise essentially defines the ontology of meanings upon which prosody-based sound design operates.

A wealth of evidence exists that emotional and intentional states are communicated nonverbally through vocal expressions (Banse and Scherer, 1996; Fernald, 1989). It has been proposed (Fernald, 1992a; Banse and Scherer, 1996; Juslin and Laukka, 2003) that the encoding and decoding of prosodic features in vocal communication have a strong evolutionary background. Such a claim is supported, for example, by the evidence of cross-cultural prosodic similarities in infant-directed speech (Fernald, 1992b). There is also convincing evidence that musical and vocal expressions of emotions have acoustic similarities, and thus communicative attributes of musical expressions could be based on principles that are derived from nonverbal vocal communication (Juslin and Laukka, 2003).

In the studies of animal vocalisations, some basic cross-species regularities has been found between structure (i.e., form) and motivation (i.e., function) of a vocal expression (Owings and Morton, 1998). It has been suggested that similar regularities are demonstrated in certain human uses of prosody which are similar cross-linguistically and cross-culturally (Ohala, 1984). Animals, such as squirrel monkeys, also produce vocalisations acoustically specific to certain purposes or affective situations, e.g., for warnings, threats and social calls (Ploog, 1992). Similarly, humans also rely on intent-specific forms of speech melody (intonation), especially in communication with pre-linguistic infants (Fernald, 1989). It has been suggested that these *communicative functions* of prosody (e.g., for prohibiting, approving or calming) represent the first semantic correspondences for the infant (Fernald, 1992a). In all, there is reason to believe that at least part of the prosodic communication lies on an evolutionarily developed pre-linguistic foundation, thus suggesting an intriguing potential for cross-cultural applicability.

Function of a sound arguably refers to a distinct mode of human meaning creation (Tuuri et al., 2007). We are using the concept of communicative functions to categorise vocal expressions according to their intended and

perceived purposes. Functional meaning thus simultaneously refers to embodied experience of both expressing and perceiving a communicative intent. As already suggested in this paper, communicative function, referring to a premeditated purpose of a sound, is also one of the most elementary design principles that the sound designer must consider for every UI sound.

But how many different communicative functions of expressions are there? According to Searle (1979) we can distinguish a part of the speech act that constitutes its illocutionary force F (the speaker's intention in producing the utterance) from the part that constitutes its propositional content (p). A speech act, as a whole, is therefore seen as $F(p)$. The most prominent subset of illocutionary force is called *illocutionary point*. Searle argues (1979) that there are only five fundamentally different types:

- *Assertive* point is to present a proposition (p) as representing a state of affairs in the world (e.g., *describing, explaining, claiming*).
- *Directive* point is to try to get the hearer to undertake the course of action matching the (p) (e.g., *asking, ordering, persuading*).
- *Commissive* point is to commit to undertake the course of action proposed in (p) (e.g., *promising, guaranteeing, vowing*).
- *Expressive* point is to express the sincerity condition about the state of affairs specified in (p) (e.g., *praising, welcoming, expressing arousal*).
- *Declarative* point is to create a state of affairs just by representing it by (p) as created (e.g., *pronouncing* a couple as a husband and a wife).

Although prosodically relevant communicative functions do not necessarily relate to any verbal content of speech, it seems appropriate to categorise them with the taxonomy presented above.

In this study, we focus on building sound design on the prosodic features which correlate with communicative functions. Our hypothesis is that the prosodic features specific to a certain communicative function in its interactional context would function similarly in communication which is mediated with UI sounds within an analogous HCI context. Maintaining the analogy in the type of communicative use of sound is important, since prosody-based UI sound design should essentially consider the coupling between acoustic forms and meanings in terms of the situationally and ecologically valid use of prosody in the interaction.

We acknowledge that prosodic attributes have been used, for instance, in speech synthesis to provide artificial speech with natural and emotional qualities, but any research about using prosody in HCI independently from the verbal context is rare. It is, however, very likely that prosodic characteristics already form an intuitive design basis for many professional sound designers. The best known examples of the use of prosodic features in film sound relate to nonverbal use of voice in conversational expressions; for example in the robot-voice of R2D2 (in the movie 'Star Wars') or in animated cartoons which omit the use of any verbal content. But the use of prosodic characteristics in film sound design is not founded on explicit knowledge of prosodic correlates of communication but on the implicit/tacit craftsmanship of a designer or a voice-actor. For doing systematic prosody-based design, we do not usually have enough explicit knowledge of how certain prosodic characteristics are related to communication.

Our intent is to provide a non-speech UI-sound design method which is based on systematic utilisation of prosody. The method consists of 1) producing relevant interaction-derived utterances, 2) performing explicit analysis of their prosodic characteristics and 3) utilising them systematically as design parameters. In this study, we also perform an evaluation of the communicative attributes of the design result.

3. Sound Design Case: Physical Training Application

Thanks to the industrial partners of our research projects², we have access to some real-world challenges of interaction design in product development. This section describes the design process of a case-study in collaboration with *Suunto Ltd.*, a Finnish company designing and manufacturing sports instruments.

3.1. Case Background

The assignment was to design non-speech user interface sounds to support interaction between a user and a training application designed for a wrist computer (see Figure 1). Such applications are based on extensive research on how to optimise physical training to enhance physical performance. The

²GEAR 2 and 3 (Grammar of Earcons), funded by the Finnish Funding Agency for Technology and Innovation.



Figure 1: Wrist computer (Suunto T4)

user of the device enters certain parameters, such as age, weight, height, activity class (ranging from "no regular physical activity" to "training daily"), maximum heart rate and gender. The application then generates a personal exercise program, for a five-day period at a time. Training is monitored with a heart beat sensor, and speed can be measured with an optional Global Positioning System (GPS) device. The aim is to reach a certain training effect (TE) in each training session, a quantity which is calculated on the basis of personal information and the sensory data input of the ongoing exercise session (duration, heart rate, etc.).

The design of interaction between a wrist computer and its user has challenges, which are familiar in the design of most mobile applications. Above all, how to enable interaction when the user's gaze is engaged with other things than monitoring the device's small visual display? In a training application, non-visual interaction would be extremely beneficial: while concentrating on an extensive physical performance, like running, staring at your wrist computer is often distracting. Moreover, the role of a UI presentation in such a training application is related to the need to provide feedback to the user about the progress of training, and at the same time to persuade the user to control his or her performance accordingly. Sound, as a gaze-independent presentation mode, is well suited for this kind of interaction.

The primary need in the sound design assignment concerned the intuitivity of sonic interaction: i.e., sounds should be able to communicate their messages as effortlessly as possible to facilitate an effective training. This

essentially means that the user should not be left to just learn what the purpose of each different kind of beep is. Rather, the sounds should possess some communicative attributes in themselves which would evoke a suitable affective resonance in the user and thus help the user in creating a proper understanding with a minimum requirement of learning. Such an aim for intuitive understanding outlines a major challenge for the design – to which the use of prosodic attributes in sound design seeks to offer an answer.

Technical restrictions also presented challenges concerning the sound producing capabilities of a wrist computer. These capabilities are usually limited to "beeper-sounds", i.e., the possibility to produce only simple tones in discrete pitch-levels without any control of sound intensity.

3.2. Defining Communicative Functions and the Design Process

We first familiarised ourselves with both the functionality and the intended usage of the training application in its context. By carefully exploring the interaction model of the application (with use scenarios), the need for certain types of sounds in interaction were pinpointed. As a result, four different communicative functions were determined. Three of the functions *Slow down* (decrease speed), *Urge* (increase speed) and *Ok* (keep the current speed) relate to the regulation of running speed. The fourth function, *Reward* (positive cheer), is essentially a praising expression of "well done" while also indicating that a certain goal of the training session has been reached.

In order to gather prosodic information that relates to the chosen communicative functions, an embedded sub-study was needed. In a production task participants, relying essentially on nonverbal communication, were asked to perform context-relevant vocalisations for each function. As all the functions were analogous to a typical trainer-runner interaction, they were easily suited to such a task. Table 1 shows a summary of the functions and their intended content.

By analysing the prosodic characteristics of the gathered vocalisations, we were able to find out if there exist any function-specific prosodic invariants. These acoustic analyses also enable the utilisation of prosodic information as design parameters. Due to the technical restrictions of the sound producing capabilities of the target device, only pitch-related information could be used in the design. This was, however, not necessarily a weakness as the intonation (i.e., melody) of speech has proved its ability to serve communicative functions (Fernald, 1992b). Moreover, we argue that pitch-related features are likely to be effective in terms of their greater scalability and resistance

Function:	Intended content:
Slow down	Slow down your pace (your pace is too fast)
Urge	Quicken your pace (your pace is too slow)
Ok	Keep this pace (your pace is fine)
Reward	Well done! (you reached a training goal)

Table 1: Summary of communicative functions of the required feedback sounds. For each function the intended nonverbal content of the utterances to-be-performed is shown. The basic function class of Slow down, Urge and Ok is primarily *directive* (persuades the runner) and secondarily *assertive* (indicates the pace expediency). For the Reward function the basic class is primarily *expressive* (praises the runner) and secondarily *declarative* (declares a training achievement).

to noise and room acoustics than timbre, and have been dominating emotion communication research for years (see Scherer, 2003). We also see that the need to reduce the prosodic features to very simple beeper-sound "melodies" will truly put the scalability of prosody-based communication to the test.

The sound design process, as a whole, had the following phases:

1. Modelling phase: pinpointing the communicative functions for UI sound elements in the course of conceptual application design.
2. Production phase: production of context-situated vocal utterances for specific communicative functions.
3. Analysis phase: extraction of acoustic features, analysis of acoustic features and evaluation of utterances.
4. Application phase: choosing best suited utterances and using their acoustic features as parameterised projections of prosody in the sound design and implementation.
5. Evaluation phase: assessing the effectiveness of extracted and implemented prosodic features in communicating intended meanings as non-speech UI sounds.

The detailed descriptions of design phases 2-5 are presented in the subsequent sections. Phases 2 and 3 relate to the sub-studies of prosodic material (Studies 1 and 2) and phase 5 relates to the evaluation of the implemented design (Study 3).

3.3. Study 1: Gathering and Analysis of Prosodic Information

A production study was carried out to discover if the vocalisations of certain kind of communicative functions would produce prosodic invariants,

i.e., function-specific prosodic characteristics. Such characteristics could then be used in sound design to serve similar communicative functions within the context of HCI as physical cues of intentionality.

3.3.1. Production Task Experiment

The basic idea of the experiment was to gather a sufficient amount of context-situated utterances for the four chosen communicative functions mentioned above. These utterances were produced by 20 Finnish speaking students of the University of Jyväskylä. Of the participants, 9 were male and 11 were female. The average age in the group was 24.8 years. Two separate types of production tasks were conducted (uttering with words and with a vowel), but in this sound design case we ultimately took into account only the task where predefined words were used as a method of vocalisation. The chosen wordings were brief but appropriate for the communicative use with which they were associated. Preferred utterances were short vocal gestures rather than spoken sentences. The participants were encouraged to communicate nonverbally, use the given words freely, and stress them as they wished. Additional details about the experimental design and the procedure of the production task can be found in an earlier report (Tuuri and Eerola, 2008a).

Vocal expressions are in many ways dependent on the situational context in which they take place and which they serve. Emotional and intentional states reflect the current situation and they provide various effects (both voluntary and involuntary) that determine the acoustic characteristics of an utterance (Scherer and Bänziger, 2004). This interaction between mental states and physiological determinants of vocal sound can be considered from the perspective of bodily projected intentionality, which we introduced earlier. One of the main priorities in the design of the experiment was to create a natural and immersive interactional context, in which the utterances could be produced spontaneously. To make the experiment as natural as possible for the participants, the contextual scenario was analogous to normal trainer-runner interaction by not being application specific to any extent. Despite this the intended communicative functions had to remain adaptable for application use.

The context of trainer-runner interaction was brought into the experimental setting in two ways: 1) by a short written scenario, which provided the background for a participant's role as a trainer in an imaginary setting and 2) by a simplified computer animation, which guided the procedure. The

animation visualised the running process with a dot (runner) moving along a circle (running track). Towards the end of each lap, animation informed whether the runner’s lap time was too fast, too slow, fine, or if the participant need to reward the runner with a cheer. These conditions implied which one of the corresponding ”messages” (Slow down, Urge, Ok or Reward) should be communicated to the imaginary runner. The participant then had a few seconds to respond vocally to the ”passing runner”.

In total, 320 utterances were gathered in the production task. In two conditions (*Word* and *Vowel*) each participant produced two experimental trials for each of the four functions (20 participants \times 2 conditions \times 4 functions \times 2 trials).

3.3.2. Results of the Production Task

The extraction of the prosodic features from recorded utterances was carried out using Praat software (Boersma and Weenink, 2001). The fundamental frequency (F_0) was extracted by means of Praat’s autocorrelation analysis using a 10 ms time-window and then converted into a linear scale by

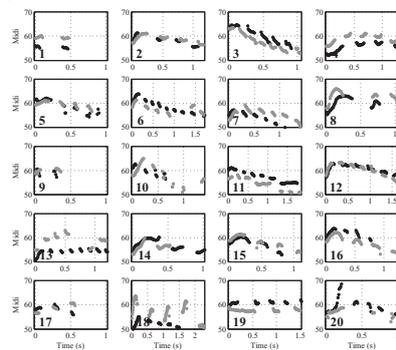
$$P = 69 + 12 \times \log_2 \left(\frac{F_0}{440} \right), \quad (1)$$

where P represents the pitch numbering convention used in MIDI standard ($C_4 = 60$). Next, the F_0 contours were centered to MIDI note 60 (261.6 Hz) within each participant to remove the obvious F_0 differences between the participants caused by gender, size and voice quality. Voice intensity information (energy in dBs) was also extracted from utterances. However, intensity measures could not be used in this design case, since the target device does not support the control of intensity.

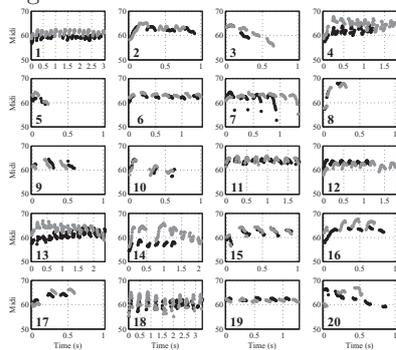
In Figure 2 the visualised pitch contours (i.e., intonations) of all 40 Word condition utterances for each communicative function are displayed to demonstrate the similarities within each function and trial. From the figure the function-specific overall patterns are clearly observable³: Urge utterances seem to have a high overall pitch and a rapidly segmented contour with an ascending or level trend. Slow down utterances are lower in pitch but segmented as well, although the segments are longer within the utterances and less variable in pitch compared to the Urge segments. Also, the

³Similar overall patterns are indicated also within the utterances of Vowel condition

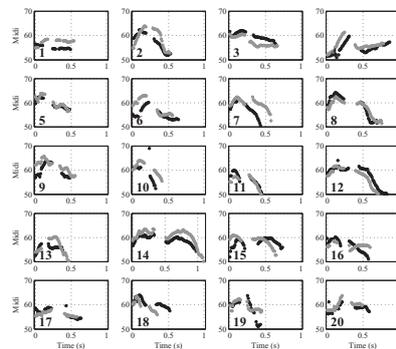
Slow down



Urge



Ok



Reward

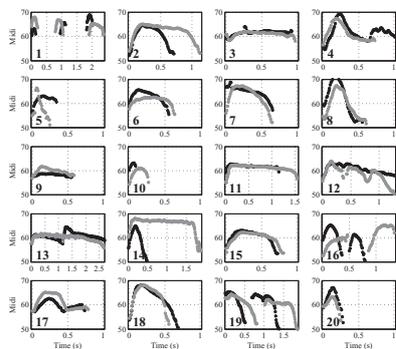


Figure 2: Visualisation of Word condition utterances for each communicative function from all 20 participants. Black contours indicate the utterance of the first trial and grey contours indicate the utterance of the second trial for each participant.

pitch contour is mostly descending. Ok utterances are also low in pitch and relatively short in overall length. They seem to consist of two segments⁴ and have a descending or level overall trend in pitch contour. Reward utterances are high in pitch and mostly have no segmentation at all. The pitch contour often has a solid arc, which is first ascending and then descending strongly. It should be noted that pitch contours of the Reward category have clear similarities (rise-fall pattern) to the pitch contours of the Approval category in the Fernald’s study (1992b). This finding is consistent with the fact that utterances in both categories have similar communicative intent to praise the hearer.

In order to perform statistical analysis, the utterances were summarised by four simple acoustic features: mean fundamental frequency, F_0 (M), frequency variation, F_0 (SD), the length of the utterances, $Length$, and proportion of unvoiced segments within utterances, $Unvoiced \%$. More sophisticated descriptors such as the attack slope, brightness or formant measures could be viable additions but there is ample evidence that relatively simple measures such as the ones outlined above are able to account for most differences in, for example, vocal expressions of emotions (Juslin and Laukka, 2003; Banse and Scherer, 1996). Also, we needed to focus on F_0 rather than spectral measures or intensity, as F_0 due to the limited audio synthesis capacities of the target device. Finally, the normality of each of features was investigated and when violated, converted to normality using box-cox power transform.

Next, we ran a three-way repeated ANOVA ($2 \times Condition \ 4 \times Function \ 2 \times Repetition$) analysis for each of the four features. A summary of the results is given in Table 2.

Main effects of Condition (Word and Vowel) were found for both utterance Lengths and Unvoiced proportions of the segments. The utterances in Word condition were significantly shorter (Word $M = 0.64$ s, $95\%CI = 0.57-0.71$ s, Vowel $M = 1.05$ s, $95\%CI = 0.93-1.18$ s) and contained more unvoiced segments (Word $M = 33 \%$, $95\%CI = 30-37 \%$, Vowel $M = 20 \%$, $95\%CI = 17-23 \%$). This observation is consistent with the speech research literature, where the consonantal structures markedly differ in voice production principles from those of the non-consonantal (such as those produced with vowels) consisting of fast changes of spectra and amplitude that result in voiceless

⁴This segmentation is likely due to the verbal content (two words instead of one) used in expressions of the Ok category performed in Word condition.

Table 2: Summary of acoustic features across the 3 treatment conditions (ANOVA).

Source	df	Feature			
		F_0 (M)	F_0 (SD)	Length	Unvoiced %
Condition (C)	1				
F		0.5	0.2	36.7**	56.4**
η^2		0	.00	.09	.09
Function (F)	3				
F		144.8**	10.6**	9.3**	67.8**
η^2		.57	.09	.07	.32
Repetition (R)	1				
F		5.1	0.08	16.0**	2.3
η^2		.01	.00	.04	.00
C \times F	3				
F		5.7**	3.19*	3.2*	21.4
η^2		.02	.02	.00	.10
C \times R	1				
F		0.1	0.0	.7	4.6*
η^2		.00	.00	.00	.01
F \times R	3				
F		0.6	0.1	1.42	1.8
η^2		.00	.00	.01	.00

* $p < .05$, ** $p < .001$ (Greenhouse-Geisser corrected)

moments in the signal in comparison to non-consonantal segments (Stevens and Blumstein, 1981).

The two fundamental frequency measures (Mean and SD of F_0) were not significantly different across the two Conditions. However, significant main effects of Function were found from all four features, exhibiting the largest effect sizes ($\eta^2 = .07-.57$) of all analysis conditions. Repetition of the sequences was significant only for Length of the utterances, probably due to latter sequences often being somewhat longer (1st utterance M = 0.71 s, 95%CI = 0.63-0.80 s and 2nd utterance M = 0.98 s 95%CI = 0.86-1.10 s). The effect size of the Repetition, however, was small. There were few interactions between the conditions, mainly between Condition and Function, implying that the utterance production type (Condition) modulated the effects of the Functions, mainly by enhancing the discrimination of the function in the Word condition in terms of the utterance lengths and the unvoiced proportions. In sum, Function provided the most differences between the utterances.

Post hoc analyses were carried out for four different *Functions* across the four features. This revealed statistically significant differences between 5 out of 6 possible paired comparisons between the four communicative functions using Holm-Sidak multiple t-test adjusted values for Length and Proportion of Unvoiced (F = 17.7 and 66.9, respectively, $p < .001$) and 4 out of 6 comparisons for the mean F_0 and variance of F_0 (F = 125.7 and 13, both $p < .001$). The functions that were *not* discriminated in these analyses were Urge-Slow down for Length and Unvoiced %, Urge-Reward and Slow down-Ok for F_0 (M), and Slow down-Urge and Urge-Ok for F_0 (SD).

In sum, all of the chosen four features were able to discriminate the communicative functions, thus indicating significant differences in the acoustic features between the functions. It should also be noted that no significant differences were found between repeated utterances each participant gave for each function except for the length of the utterance. These observations can be seen from the visualised pitch contours (Figure 2), suggesting that prosodic information is robust in communicating these functions with minimally altered repetitions in separate experimental trials.

The Condition mainly influenced the Proportion of Unvoiced and the Length of the utterances rather than the frequency-based features. Consonantal passages contain fast changes of spectra and amplitude in comparison to non-consonantal segments and, interestingly, the former are considered as regions of high information in speech research due to richer information available in the temporal dynamics of the signal (Stevens and Blumstein,

1981). Furthermore, in the subsequent analyses the Word condition yielded better classification results (classification accuracy of 86 %) than the Vowel condition (59 %), as measured by Regression-Tree Analysis (RTA), described in the next analysis phase. For both theoretical and data-analytic reasons, we will focus on Word condition in the subsequent analyses of the production material although fairly similar overall results would be obtained with Vowel condition as well. This decision is also consistent with the findings from infant-directed speech, in which the melodic characteristics are mainly influenced by expressive intentions, emotional state, and personal style rather than the actual words used (Fernald, 1992b; Bergeson and Trehub, 2007).

3.3.3. Classification Using Regression Tree Analysis

As already shown by the ANOVA column of Table 2, most features demonstrate differences across the functions. However, in order to better understand which *combinations* of the available acoustic features contribute the most to the separation of the four function categories, a classification approach was adopted. To classify the utterances properly into four function-specific groups, we chose to apply Regression Tree Analysis (RTA) (Breiman et al., 1984). RTA constructs rules by recursively partitioning the observations into smaller groups based on a single variable at a time. These splits are created to maximise the between groups sum of squares. The resulting tree diagram initially has a large number of tree nodes (logical *if-then* conditions) which are pruned by cross-validation to reduce the overfitting. This approach provides several advantages over discriminant analysis (DA) or classical regression techniques: it is able to uncover structures in observations which are hierarchical, it is nonparametric, and allows interactions and nonlinearities between the predictors (Ripley, 1996). The rules that describe the splitting into groups are also easy to interpret and provide insights into the process of classification.

Using the four features, a successful model correctly classified 94.4 % of the utterances. However, an optimal combination in terms of prediction accuracy and model simplicity was obtained by means of cross-validation and pruning of the regression tree. This resulted in a solution with two features (Mean F_0 and Unvoiced %), which had a classification accuracy of 85.6 % with 10-fold cross-validation. This model contains only 3 *if-then* rules for deciding the classification into four functions: (1) If the utterance is continuous (Unvoiced % is lower than 16 %), it is probably Reward. (2) If not and the Mean Frequency is higher than 60.8, it is highly likely Urge. (3) If the

utterance is non-continuous and low in mean frequency and the Proportion of Unvoiced is higher than 35 %, it is probably Slow down. If these criteria are not satisfied, the utterance is likely to be Ok. This simple RTA model is visualised in Figure 3 by the boxed areas and the actual observations are denoted by four types of individual markers. It is worth noticing that even when a larger set of features were used in analysis, including also intensity-related descriptors, the optimal RTA model still resulted in the same two-feature solution (Tuuri and Eerola, 2008b).

Classification of Vowel condition data reveals the same underlying pattern with 2 optimal features (Mean F_0 and Unvoiced %), but the classification rate is significantly poorer (with four features, 88.1 % and for optimal, cross-validated, 58.7 %) and more complex (30-40 % more branches and nodes are needed for the classification).

3.4. Study 2: Evaluations of the Utterances

To assess the goodness of each utterance, an evaluation of the utterances in terms of their ability to convey any one of the four functions was carried out. For this purpose, only the fundamental frequency information was considered. The F_0 contours of the Word condition utterances were synthesised using sine waves, centered to 440 Hz within each source utterer. Five participants, consisting mostly of academic colleagues, rated the amount of each of the four functions that were represented by each of the 160 utterances. The ratings were carried out in a random order and using a Likert scale (1-5). The results of these assessments showed a surprisingly high inter-rater agreement ($\alpha = .94, .97, .94,$ and $.87$ for Slow, Urge, Ok, and Reward, respectively) considering the length of the experimental session (approx. 1 h). The differences in inter-rater reliability are indicative of the difficulty of choosing the appropriate functional category.

The dark areas in Figure 3 denote high ratings for the function in question. For each utterance, the mean value concerning each function was calculated and then linearly interpolated to two-dimensional representation by convolution with a 2D-gaussian kernel. The result is functional appropriateness across the two-dimensional space, showing darker areas in locations where the best examples of the function category could be found.

3.5. Utilising Prosodic Information

The results of the production study were extremely encouraging in demonstrating that the utterances of the four communicative functions indeed have

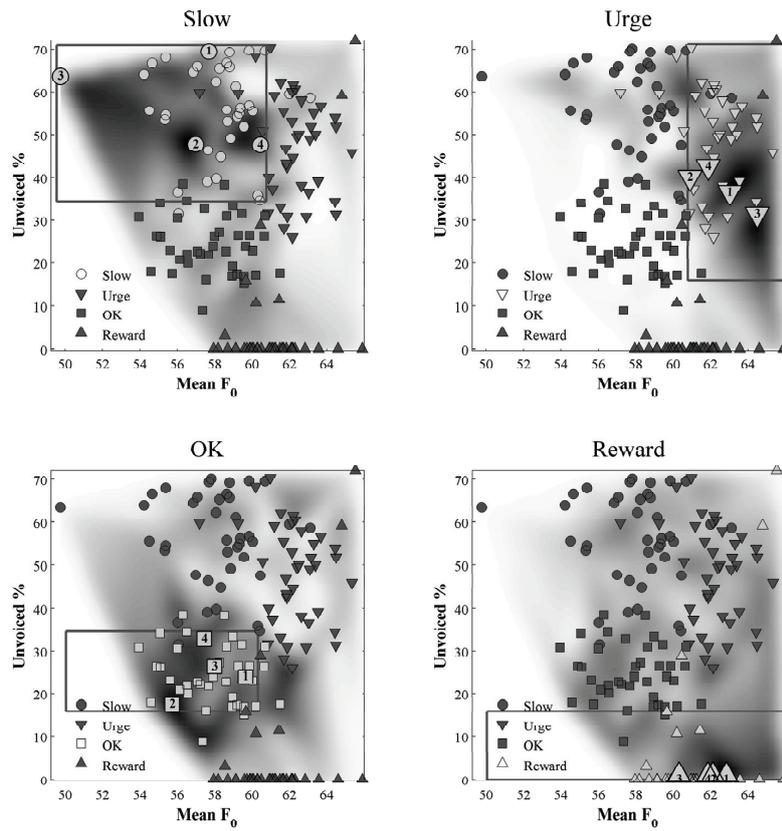


Figure 3: Visualisation of 160 utterances for each communicative function from all 20 participants across the two acoustic features (Mean F_0 & Unvoiced %). Marker types indicate functions, and the utterances of the highlighted function are shown in light colour. The dark areas indicate the interpolated ratings of goodness-of-fit for each function category across the utterances. The examples chosen to best suit sound design have been labelled within each category with numbers 1-4. The boxed regions indicate the areas identified by the optimal RTA model (see text for details).

function-specific prosodic characteristics. But how could these acoustic cues of the utterances be used as meaningful attributes in sound design? We suggest two different approaches. If we can analytically demonstrate that certain kinds of prosodic features are derived from the utterances of a certain communicative intention, those features could be adapted as separate acoustic cues into the design and sound composition. This approach would require a more comprehensive study of the function-specific prosodic cues in order to understand their relation in subjective meaning creation. However, the second approach is more straightforward. If we can find certain utterances to be representatives of a certain communicative function, why not simply imitate, for example, the original pitch contour of these utterances? In that way, we can also ensure that the sound design will include such prosodic qualities that might not have been captured by the limited analytic descriptors.

In this sound design case, direct imitation of pitch contour is a fitting method because the target device only allows the control of pitch in a discrete manner. Utterances are also conveniently short and thus an appropriate basis for UI audio feedback. After choosing the suitable source utterances, the remaining procedure for implementing sounds would roughly consist of making "device-reduced" renditions (described in more detail later) of chosen pitch contours and, if necessary, carrying out final refinements depending on functional and aesthetic requirements. As parameterised projections of the original vocal gestures, the pitch contours of the chosen utterances form a design basis for synthesising final UI sounds. The design results therefore can be seen as parameterised auditory icons (Gaver, 1993a), as they do not focus on the recognition of sound source but rather focus on the attribution of ecologically meaningful parameters of the source event.

3.5.1. Ranking Utterances Within Functions

Next, we had to find the most suitable utterances for sound design. This task can also be approached from two directions; either from the direction of the statistical information about the characteristics of communicative functions or from the direction of subjective evaluation. Both approaches were tested in our effort to pick the "best" source utterances for each function category (out of 40 Word utterances).

Based on the RTA classification model (Study 1) we formulated a list of best-ranking utterances for each function (ranked by the distance from the group centroids of the RTA model). According to the model, the utterances close to the group centroids should be statistically the best representatives

of a given communicative function. However, we also ranked utterances on the basis of subjective evaluation test of the utterances (Study 2).

There was not a high level of correspondence between the highest ranking utterances of the two different ranking methods (computational and subjective). Within the top 15 utterances of both ranking lists, only 3 were the same in the Slow down category, 8 in the Urge category, 3 in the Ok category and 5 in the Reward category. This result arguably demonstrates differences between the more complex nature of subjective ranking compared to the use of relatively simple acoustic descriptors as classifiers. Differences between the basis of these rankings is also evident when comparing the darker areas with the boxed RTA regions in Figure 3. So, which ranking should be trusted? Of course, sound designers must also trust their own intuitive judgement, and in this case it favoured the top utterances of the subjective ranking list. Despite the good results in the statistical classification approach, it seems that such a mechanistic ranking method might still be inadequate as a design principle.

3.5.2. Choosing and Implementing Final Sounds

The top 15 utterances of each function category in the subjective ranking were considered as candidates for the UI sound basis. Next we synthesised device-reduced renditions of the pitch contours of the candidate utterances in order to evaluate their adaptability to the restrictions of the device’s capabilities. We approximated the specifications of general ”beeper-sounds” as being able to control discrete pitches in semitones at the 50 ms time window. We also determined the available pitches to range from MIDI note number 83 (987.8 Hz) to MIDI note number 107 (3951.1 Hz). The intensity of tones could not be controlled, so it remained constant for all tones. With the aim to produce realistic sounds, some sample tones from Suunto’s T4 device were recorded⁵. By using these samples, a MIDI controllable instrument was implemented with the ability to produce relatively authentic wrist-computer renditions from the sequences of MIDI note information.

Pitch contours of all the candidate utterances were first shifted into the high register and normalised for each participant. This was done by shifting the participant’s pitch values up until the highest pitch measure matched

⁵Note that the recorded tones had an approximately 3-4% higher pitch level in relation to the MIDI note they represent.

Communicative functions	Utterances			
	Ex. 1	Ex. 2	Ex. 3	Ex. 4
Slow down	6 ₁	11 ₁	18 ₁	12 ₁
Urge	12 ₁	13 ₁	4 ₂	12 ₂
Ok	20 ₂	4 ₂	6 ₂	1 ₂
Reward	6 ₁	18 ₁	2 ₁	7 ₂

Table 3: List of four example utterances chosen to best suit each communicative function. The first number refers to the number of participant and following index refers to the first or second trial (for visualisations, see Figure 2). Example number 1 of each function (bolded) was chosen as the basis for sound design.

the specified ceiling frequency (3951.1 Hz). In this manner, most utterances were accommodated nicely within the restricted pitch range. Next, the pitch contours of all candidate utterances were re-sampled across time (at a 50 ms interval) and pitch (in semitones) in order to conform with a reduced resolution of the target device. The processed data was then converted into MIDI information. This allows the data to be easily edited with a MIDI sequencer program and to be rendered into realistic beeper sounds by using the custom-made wrist-computer instrument.

By evaluating the device-reduced renditions of the candidate utterances, the design team was able to trim down the number of candidate sounds from 15 to 4 for each function category, which are listed in Table 3. These chosen examples are also highlighted in the scatterplot visualisations for each function in Figure 3. After the second iteration of informal evaluation, the design team finally chose one utterance for each category to become the basis for the UI sound. These chosen examples are labelled with a number 1 both in Table 3 and in Figure 3.

To finish the sound implementation, sounds were checked for any inconveniently sounding re-sampling artefacts. As a result, a minor alteration was made to one of the reduced pitch contours (Reward). In this stage of sound design, more extensive stylising could be performed in order to make aesthetic refinements. Visualisations of the finalised sounds, together with the original utterances, are presented in Figure 4.

3.6. Study 3: Evaluation by Contextual Recognition Test

The evaluation of mobile applications has been a topical issue in the HCI community for a long time. The traditional usability evaluation methods,

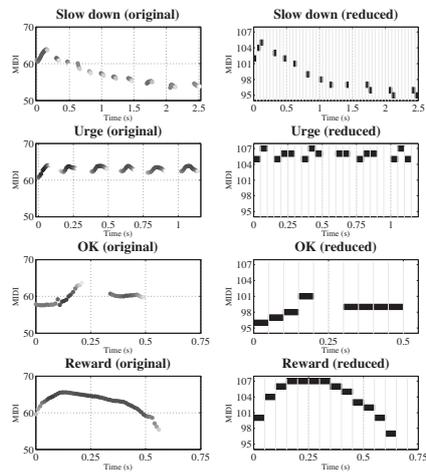


Figure 4: Visualisations of chosen original utterances and their device-reduced final contours for each communicative function. In the contours of the original utterances, a darker colour indicates a higher intensity (dB) value. In reduced contours, vertical grey lines represent 50 ms time intervals.

which were applied in desktop settings, were suddenly of not much use. The debate on whether to conduct experiments in the field or in a laboratory is never-ending. For assessing the intuitiveness of interpreting prosody-based sound feedback, considerations about the interactional context are essential, as the functions of prosody are bound up with the situation which it is intended to serve. Moreover, as we have already pointed out in the beginning of this paper, the coupling of sound and meaning is never merely a property of a sound object, but rather an intentional connection between user and sound in the course of interaction.

3.6.1. Experimental Design

For our recognition test of the four finalised UI sounds, we decided to resort to a laboratory setting in which the context of a mobile application was simulated to a certain degree. As it was a question of a sports application, we used simple exercise equipment, a mini-stepper, to introduce physical performance and the idea of exercise into the experiment. We used a similar setting in our previous study (Pirhonen et al., 2002) and found it an appropriate simulation of a mobile setting, still enabling precise, laboratory-like control. Thus, it should be noted that our experimental setting is a compromise solution: while carefully providing some contextual elements of exercise, training application and physical activity of a participant, the test is still essentially conducted in a laboratory setup context.

We had 12 participants in the experiment, 8 female and 4 male. All of the participants were students at our university, and aged 23-29. They were from different departments, representing 7 different major subjects.

In the recognition test, each of the four sounds was played four times. In all, there were 16 sound stimuli (4 examples \times 4 repetitions). No training or familiarising stimuli was presented, although participants were told to expect "beeper-quality" sounds.

Each subject participated in a test of approximately 10 minutes' duration, and was rewarded with a movie ticket. The test started with instructions, which were read aloud from a paper by the researcher, to make sure that all the participants got the instructions in exactly the same format. This was important since small changes in the nuances might have changed the orientation of a participant. For instance, the participants were strongly encouraged to rely on intuition rather than logic or reasoning. The researcher who read the instructions was keeping track of the choices during the actual test. In addition, there was another researcher in the laboratory, using the



Figure 5: The experimental setting: the stepper, the table of alternatives, the audio speakers and the illustration of a participant.

computer and video camera.

To make it easier to imagine a run in a forest, there was also a background soundscape providing a constant background sound stream of surrounding forest-ambience with a light steady rain and distant sounds of birds. It was carefully implemented so that it would not disturb the listening to sound stimuli in any way. The sound was processed to avoid any masking or perceptual disorientation. Moreover, the loudness difference between sound stimuli and ambient sound in the room was very clear (approximately 75 dB in RMS power and 20 dB at peak levels). The function of the soundscape was solely to introduce a contextually suitable mood into the laboratory setting – from the reading of instructions to the start of an interview.

Figure 5 illustrates the experiment in practice. After having heard the instructions, a participant was asked to start stepping with the stepper at a convenient frequency. She or he was then supposed to follow the sounds which were played through the speakers.

For the recognition of sounds we used a forced-choice paradigm. Whenever a sound sign was played, the participant was supposed to point with his or her finger at the table in front, to indicate what he or she thought the sound meant. The alternatives were in four categories: "Urge on", "Slow down", "Keep on going like this" and "Well done!". The participants were also asked to react to the perceived message; if someone, for instance, in-

terpreted a sound to mean "Urge on", that participant was supposed to accelerate the stepping speed.

The sequence of the sounds was fixed, but the timing was flexible: Using a computer the researcher launched each sound when the participant had made a choice and reacted accordingly – e.g., changed the stepping speed according to the perceived message. In practice, this meant that the gap between individual sounds was approximately 15 seconds.

According to our previous experience (Pirhonen, 2005), even in a short experiment a certain level of learning of the meaning of sounds can occur. In the current evaluation, we encouraged participants to rely on their first impressions and thus aimed at figuring out the intuitive responses rather than learned meanings. In addition, to minimise the learning effect, the number of each type of sound was low (4) in the evaluation session. We thus assume that the data reflects intuition more than reasoning or learned meanings. This is an important distinction, because the aim of the study is – rather than to explore human ability to learn abstract sounds – to learn to design audio cues which would communicate intended meanings effectively.

After the stepping session, the participant was asked to describe the experience. In addition, each sound was played again, one at time, and the participant was asked to describe what kind of meanings each sound evoked.

3.6.2. Results

First, inter-rater reliabilities of the target choices were calculated ($\alpha = .77$). Two participants were identified as outliers (>2 SDs from the mean intersubject correlation) and removed, which resulted in an acceptable inter-rater reliability ($\alpha = .90$). An ANOVA was run to compare the correct identification rates across the four function categories ($F = 254.10, 8.45, 4.15,$ and 7.03 for Slow down, Urge, OK, and Reward, respectively, all $p < .001$ except for OK, $p < .01$). This analysis yielded significant results for each target category in a post-hoc (Scheffé) analysis with one exception that concerned the Ok category. This target category could not be reliably discriminated from the Reward and Slow down categories, although it was nevertheless predicted with above chance level (in 45 % of the cases compared to the null hypothesis of %25, $\chi^2 = 21.6, p < .001$). The recognition accuracy across the functions together with the associated pattern of confusions and statistical test values are shown in Table 4.

The recognition accuracy is presented in Table 4. The vertical dimension denotes the intended meanings, with the perceived meanings given in the

Intended meaning	Perceived meaning				χ^2	p value
	Slow down	Urge	Ok	Reward		
Slow down	92.5	5.0	2.5	0	97.4	<.001
Urge	20.0	62.5	10.0	7.5	31.4	<.001
Ok	15.0	0	45.0	40.0	21.6	<.001
Reward	17.5	12.5	10.0	60.0	26.6	<.001

Table 4: Recognition accuracy in percentages for each communicative function.

horizontal dimension. A successful expression is thus indicated by a high percentage in the diagonal of the table. χ^2 and the associated p value display the results of goodness-of-fit test between the chance level and the obtained pattern within each function. As can be seen, the recognition of the intended meaning was fairly robust across the functions, despite the large differences between the functions (from 45 to 92.5 %). It is worth pointing out that the prevalent confusions between the functions occur between the Ok and Reward functions (40 % of misattributions of the Ok function).

In the interviews, the participants were allowed to describe the evoked meanings in their own words for each sound. Since the interview was held after the stepping test, it is obvious that many of the participants had expressions in mind, which were similar to those used in the test. The results of the interviews, taking all 12 participants into account, can be summarised as follows:

- **Slow down.** Out of 12 participants 10 described the sound as slowing down, descending or reducing. It is worth noticing that 8 participants used an expression that directly referred to the intended meaning, and only 2 had conflicting descriptions.
- **Urge.** 7 participants used an expression referring to acceleration. Words referring to warning or alarm were used by 4 participants. This was not a surprise because the utterances of Urge, in general, resulted in acoustic patterns (high pitched rapid bursts) associated with perceived urgency and known to function effectively in alarm signals (Edworthy et al., 1991). Two participants said that they interpreted the sound as a warning about going "too fast". This interpretation also shows in the test results where 20 % of Urge sounds were recognised as Slow down. If this test was conducted in real exercise conditions, such an affordance for slowing down would have been unlikely, due to the "slow

pace” condition required for the sound feedback in the first place. However, as may be observed in the recognition accuracy results (Table 4), the message was quite successfully communicated, but not as well as in the Slow down function.

- **Ok.** The expressions which the participants used for these sounds were more heterogeneous than for Slow down and Urge sounds. Out of 12 participants 4 used words which undoubtedly referred to the directive intended meaning (“Go on like this”, etc.). This result is in accordance with Table 4: under a half of the choices were as intended. But, in all, 8 participants out of 12 used expressions that can be counted as appropriate for Ok category. They were either directive (like “Go on!” , or “Keep going”) or assertive (like “You are doing fine”, or just “Fine”). Some of the participants also called the sound “Neutral”. 3 participants out of 12 gave clearly conflicting expressions.
- **Reward.** 5 participants used words which refer to a reward function (“Great!”, “...rewarding”). Other expressions, which could be classified as appropriate, were “Happy” and “New level begins”. In all, 7 participants out of 12 regarded the sound either as positive, praising expression or as declaration of a milestone. Two of the participants misinterpreted the sound as an Ok function (“Keep going”). A clearly contradictory meaning was perceived by 2, who found the sound “Shrilling” and thus alarming.
- **General observations.** Among the comments, there were statements which show that the device-specific restrictions had an important role in the connotations. “Old electronic game”, “Sounds like a computer game”, “Switching off a mobile phone”, for instance, were comments which show that the central human strategy of adapting to the new is to find familiar elements in it.

Right after the stepper session, the participants were encouraged to comment freely on the experiment. Practically all found the use of sounds effective and pleasant. Some stated spontaneously that they would apply that kind of sound-based interaction if it were available. Interestingly, one participant said that he felt that the sounds were speech acts, even though he did not know anything about the design background. The same participant recognised the intended meanings almost perfectly.

The results of the interviews revealed more heterogeneity and ambiguity in the interpretations of Ok and Reward sounds, when compared to Slow down and Urge. The confusions found between Ok and Reward functions in the recognition test (see Table 4) indicate that there may exist some overlap between these functions. According to the results of the interviews, however, only one participant took the Ok sound for Reward and only two participants misinterpreted the Reward sound as being Ok. One reason for the weaker recognition of Ok sounds could be related to their less directive characteristics than what was expected on the basis of the test instructions (the instruction for Ok was "Keep on going like this"). In the production experiment (Study 1), the articulated prosody of the Ok function seemed to emphasise the assertive intent with the propositional content of "this is fine" rather than giving the direction to keep up the pace (see also Table 1). The interview results for the Ok function back this assumption. Such a positive and approving assertion for the Ok function could be easily interpreted as a rewarding praise, thus explaining why Ok sounds were more often misinterpreted as Reward (40%) than the other way around (10%) in the recognition test (see Table 4).

The results of the evaluation demonstrated that the participants intuitively attributed meanings that were fairly well in line with the intended meanings of the four communicative functions. The results also indicate that sounds indeed conveyed prosodic cues of intentionality, even though the prosodic information was in a heavily reduced form.

4. Conclusions

The design goal of the case was to produce non-speech audio feedback elements for four different communicative functions (Slow down, Urge, Ok and Reward) in a training application. The focus of this design case was to provide familiar communicative attributes in sounds that would communicate specific meanings to a user by non-linguistic means. The described design process, as a practical solution, outlines a method of prosody-based sound design. At a more general level, it illustrates an interpersonal approach to the design paradigm, which is based on everyday listening and sonic interactions.

Prosodic information, in social situations, represents acoustically realised derived intentionality, which we are able to understand in terms of empathetic involvement. However, the way we suggest using prosody in interaction design does not require the presence of any explicit agent, whose actions

the user could empathise with. Rather, in our method, the empathetical involvement of the user is outlined as his or her natural tendency to conceive the sound event of a UI as intentional, when experienced in the context of interaction.

The method draws upon interaction, emphasising the embodied nature of a situational experience as an inseparable factor in subjective couplings of sound and meaning. The method is analytical, and within it, the described design process is based on principles founded on a solid theoretical framework as well as on the empirical findings of three separate sub-studies.

In Study 1, utterances were gathered for each function within an interactional context. The study demonstrated the existence of function-specific characteristics in prosodic information, thus strengthening the justification for using prosody as a design principle. Analysis also showed that function-specificity taps into such properties of an expression that are relatively simple and robust to detect even by combination of two basic F_0 descriptors. In Study 2, sine wave renditions of utterances were subjectively evaluated, in order to find out which are best for each communicative function. Following this, one utterance for each of the four communicative roles was chosen as a basis for implementing the final set of sounds for the application. Study 3, the contextual evaluation of the heavily reduced final sounds (beeper implementations), assessed the intuitive and spontaneous recognition of the four intended meanings. On the basis of the results, it could be argued that the final set of sounds did have communicative attributes (derived from the intentional stance of a source utterance) that facilitated the recognition of their intended purpose. Therefore, in the actual use of the application, the process of getting accustomed to these sounds should be effortless.

Considering the technically coarse nature of the implemented ("beeper-style") sounds, the communicative effectiveness that mere pitch-related prosody demonstrated is very encouraging. In future studies, it would be interesting to examine the extent to which the prosodic information can be further reduced without losing its communicative value. It would also be important to examine the possible cross-cultural differences in encoding and decoding the prosodic information of speech acts. Assuming that there is an evolutionary continuity in couplings between the function and the form of a vocalisation (see discussion in 2.2), at least some pre-linguistic universality in prosodic communication can be expected. However, it remains to be seen to what extent the pitch contours of this study rely on conventionalised uses of prosody (such as emblems or linguistic dependencies) that are specific, for example, to

Finnish or Western culture. Another idea for further studies could take the form of analysis of existing sounds in video games, user interfaces, cartoons or movies utilising similar functions of non-speech communication. Possible regularities, demonstrated in sounds used for similar functions, could be compared to prosodic structures gathered for these same functions in order to find out whether they share any characteristics.

The main phases of the suggested design method, which were listed in 3.2, should be heuristically applicable to various cases of design. Carefully implemented modelling and production phases of design should ensure that design is sufficiently tailored to a given application context, and thus builds upon interaction-derived ecological validity in the use of prosody for the pinpointed functions. Basically this process is scalable to non-speech sound design tasks for any communicative functions which can be conceived in terms of nonverbal vocal interaction (either directly or metaphorically). Searle's taxonomy, presented in 2.2, provides a general-level outline, which can be used in conceptualising a spectrum of different communicative functions and their possible combinations. Due to the interaction-centered approach, it should be noted that the results of prosody-based design would not automatically work outside the intended context. For example, the agitative nature of the Urge sound of the presented design case is well suited for physical training situations but might not be fully appropriate for all directive functions referring to acceleration.

Due to the explorative and highly systematic nature of the described case study, it might not be regarded as a typical representation of a sound design case in practice. Therefore it is important to note that prosody-based design can also be approached through less analytic endeavours. For example, the designer's intuition could have a more important role in choosing the most functional source utterances – or even in producing those utterances by oneself. Hence, sound designer can exploit the intuitivity of vocal expressions in generating, refining and communicating design ideas. In this sense, prosody-based design can be seen in the light of vocal sketching (see Ekman and Rinott, 2010).

The results of this study demonstrated that prosody served distinct communicative functions. This study also demonstrated that compatible communicative functions of sound can be found between the contexts of interpersonal interaction and HCI. Within such communicative functions of sound, function-specific prosodic characteristics provide useful correlates between intended communicative functions and acoustic descriptions of sound. We

thus argue that the principles presented in this paper would contribute to sound design of interactive products.

References

- Banse, R., Scherer, K., 1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology* 70, 614–636.
- Bergeson, T. R., Trehub, S., 2007. Signature tunes in mothers’ speech to infants. *Infant Behavior & Development* 30 (4), 648–654.
- Blattner, M., Sumikawa, D., Greenberg, R., 1989. Earcons and icons: Their structure and common design principles. *Human-Computer Interaction* 4 (1), 11–44.
- Boersma, P., Weenink, D., 2001. Praat, a system for doing phonetics by computer. *Glott international* 5 (9/10), 341–345.
- Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Wadsworth, Belmont, CA.
- Brewster, S., Wright, P., Edwards, A., 1995. Experimentally derived guidelines for the creation of earcons. In: *Adjunct Proceedings of HCI*. Vol. 95. pp. 155–159.
- Brewster, S. A., 2003. Non-speech auditory output. In: Jacko, J., Sears, A. (Eds.), *The Human Computer Interaction Handbook*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 220–239.
- Dillon, M., 1997. Merleau-Ponty’s ontology. Northwestern University Press.
- Dourish, P., 2001. *Where the Action Is: The Foundations of Embodied Interaction*. MIT Press, Cambridge, MA.
- Edworthy, J., Loxley, S., Dennis, I., 1991. Improving auditory warning design: relationship between warning sound parameters and perceived urgency. *Human factors* 33 (2), 205.
- Ekman, I., Rinott, M., 2010. Using vocal sketching for designing sonic interactions. In: *Proceedings of the 8th ACM Conference on Designing Interactive Systems*. ACM, pp. 123–131.

- Fernald, A., 1989. Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child development*, 1497–1510.
- Fernald, A., 1992a. Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In: Barkow, J. H., Cosmides, L., Tooby, J. (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, pp. 391–428.
- Fernald, A., 1992b. Meaningful melodies in mothers' speech to infants. In: Papoušek, H., Jürgens, U. (Eds.), *Nonverbal vocal communication: Comparative and developmental approaches*. Cambridge University Press, pp. 262–282.
- Gallese, V., Eagle, M., Migone, P., 2007. Intentional attunement: Mirror neurons and the neural underpinnings of interpersonal relations. *Journal of the American Psychoanalytic Association* 55 (1), 131–175.
- Gaver, W., 1989. The SonicFinder: An interface that uses auditory icons. *Human-Computer Interaction* 4 (1), 67–94.
- Gaver, W., 1993a. Synthesizing auditory icons. In: *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. ACM, New York, NY, pp. 228–235.
- Gaver, W., 1993b. What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology* 5 (1), 1–30.
- Gibson, J., 1979. *The ecological approach to visual perception*. Houghton Mifflin Boston.
- Godøy, R., 2010. Gestural Affordances of Musical Sound. In: Godøy, R., Leman, M. (Eds.), *Musical gestures: sound, movement, and meaning*. Routledge, pp. 103–125.
- Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J., Rizzolatti, G., 2005. Grasping the intentions of others with one's own mirror neuron system. *PLoS Biol* 3 (3), e79.
- Juslin, P., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological Bulletin* 129 (5), 770–814.

- Leman, M., 2008. *Embodied Music Cognition and Mediation Technology*. The MIT Press, Cambridge, MA.
- Lieberman, A., Mattingly, I., 1985. The motor theory of speech perception revised. *Cognition* 21 (1), 1–36.
- Malouin, J., Landry, M., 1983. The mirage of universal methods in systems design. *Journal of Applied Systems Analysis* 10, 47–62.
- Merleau-Ponty, M., 1962. *Phenomenology of Perception* (1945). Trans. Colin Smith. London: Routledge.
- Ohala, J., 1984. An Ethological Perspective on Common Cross-Language Utilization of F_0 of Voice. *Phonetica* 41 (1), 1–16.
- Owings, D., Morton, E., 1998. *Animal Vocal Communication: A New Approach*. Cambridge University Press.
- Pirhonen, A., 2005. Supporting a user facing a novel application: learnability in oobe. *Personal and Ubiquitous Computing* 9 (4), 218–226.
- Pirhonen, A., Brewster, S., Holguin, C., 2002. Gestural and audio metaphors as a means of control for mobile devices. In: *Proceedings of the CHI 2002 conference on Human factors in computing systems: Changing our world, changing ourselves*. ACM, pp. 291–298.
- Ploog, D., 1992. The evolution of vocal communication. In: Papoušek, H., Jürgens, U. (Eds.), *Nonverbal vocal communication: Comparative and developmental approaches*. Cambridge University Press, pp. 6–30.
- Ramloll, R., Yu, W., Riedel, B., Brewster, S., 2001. Using non-speech sounds to improve access to 2d tabular numerical information for visually impaired users. *Proceedings of BCS IHM-HCI 2001, Lille, France*, 515–530.
- Ripley, B. D., 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, MA.
- Rizzolatti, G., Arbib, M., 1998. Language within our grasp. *Trends in neurosciences* 21 (5), 188–194.
- Rizzolatti, G., Craighero, L., 2004. The Mirror-neuron System. *Annu. Rev. Neurosci* 27, 169–92.

- Rocchesso, D., Fontana, F., 2003. *The Sounding Object*. Edizioni Mondo Estremo.
- Rocchesso, D., Serafin, S., Behrendt, F., Bernardini, N., Bresin, R., Eckel, G., Franinovic, K., Hermann, T., Pauletto, S., Susini, P., et al., 2008. Sonic interaction design: sound, information and experience. In: *CHI '08 extended abstracts on Human factors in computing systems*. ACM, New York, NY, pp. 3969–3972.
- Scherer, K., 2003. Vocal communication of emotion: A review of research paradigms. *Speech communication* 40 (1-2), 227–256.
- Scherer, K., Bänziger, T., 2004. Emotional expression in prosody: a review and an agenda for future research. In: *Speech Prosody 2004, International Conference*.
- Searle, J., 1979. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.
- Searle, J., 1983. *Intentionality, an essay in the philosophy of mind*. Cambridge University Press.
- Shneiderman, B., 1997. *Designing the user interface: strategies for effective human-computer interaction*. Addison-Wesley, Boston, MA.
- Stevens, K., Blumstein, S., 1981. The search for invariant acoustic correlates of phonetic features. In: Eimas, P., Miller, J. (Eds.), *Perspectives on the study of speech*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 1–38.
- Tuuri, K., 2010. Gestural attributions as semantics in user interface sound design. In: Kopp, S., Wachsmuth, I. (Eds.), *Gesture in Embodied Communication and Human-Computer Interaction*. No. 5934 in LNAI. Springer-Verlag, pp. 257–268.
- Tuuri, K., Eerola, T., 2008a. Could Function-Specific Prosodic Cues Be Used as a Basis for Non-Speech User Interface Sound Design. In: *Proceedings of International Conference on Auditory Display, IRCAM, Paris*.
- Tuuri, K., Eerola, T., 2008b. Identifying function-specific prosodic cues for non-speech user interface sound design. *Proceedings of the 11th International Conference on Digital Audio Effects*, 185–188.

- Tuuri, K., Mustonen, M., Pirhonen, A., 2007. Same sound–different meanings: A Novel Scheme for Modes of Listening. In: Proceedings of Audio Mostly: 2nd Conference on Interaction with Sound. pp. 13–18.
- Varela, F., Thompson, E., Rosch, E., 1991. The Embodied Mind: Cognitive Science and Human Experience. MIT Press, Cambridge, MA.
- Wachsmuth, I., Lenzen, M., Knoblich, G., 2008. Introduction to embodied communication: Why communication needs the body. In: Wachsmuth, I., Lenzen, M., Knoblich, G. (Eds.), Embodied communication in Humans and Machines. Oxford University Press, pp. 1–28.

PVIII

LEAPING ACROSS MODALITIES: SPEED REGULATION MESSAGES IN AUDIO AND TACTILE DOMAINS

by

Kai Tuuri, Tuomas Eerola & Antti Pirhonen 2010

In Nordahl, R., Serafin, S., Fontana, F. & Brewster, S. (Eds.), Haptic and Audio
Interaction Design (LNCS 6306), Proceedings of 5th International Workshop
HAID 2010, Copenhagen, Springer-Verlag, Heidelberg, pp. 10–19

Reproduced with kind permission of Springer-Verlag.

Leaping Across Modalities: Speed Regulation Messages in Audio and Tactile Domains

Kai Tuuri¹, Tuomas Eerola², and Antti Pirhonen¹

¹ Department of Computer Science and Information Systems

² Department of Music

FI-40014 University of Jyväskylä, Finland

{krtuuri,ptee,pianta}@jyu.fi

Abstract. This study examines three design bases for speed regulation messages by testing their ability to function across modalities. Two of the design bases utilise a method originally intended for sound design and the third uses a method meant for tactile feedback. According to the experimental results, all designs communicate the intended meanings similarly in audio and tactile domains. It was also found that melodic (frequency changes) and rhythmic (segmentation) features of stimuli function differently for each type of message.

Key words: audio, tactile, crossmodal interactions, crossmodal design

1 Introduction

When designing applications for mobile or ubiquitous contexts, the means to present information redundantly in audio and tactile domains can provide valuable flexibility. Since the actual context of use is hard to anticipate, it is important that there are options for interacting with the application. Users can also have different preferences about either hearing feedback in audio or feeling it on their skin in a more intimate manner. This is the usual rationale for *crossmodal design approach* [1].

Physical training applications, used in a varying contexts, would benefit of crossmodal design. They use the information from, e.g., heartbeat or Global Positioning System (GPS) sensors, and provide feedback for the user for controlling his or her training performance accordingly. This study focuses on messages that relate to the regulation of running speed. Sounds and vibrations, as gaze-independent presentation modes, are well suited for such an interaction.

As a starting point, this study picks up two existing methods for designing messages for the regulation of running speed: one for sounds [2], and one for vibrotactile feedback [3]. Both methods aim at intuitive interaction, so that sounds and vibrations would communicate their messages as effortlessly as possible – with a minimum requirement for learning. This study examines empirically how these design methods can be applied for crossmodal design, i.e., utilising them in creating both non-speech sounds and vibrotactile stimuli.

1.1 The Quest for Amodality

The concept of crossmodal design is based on an assumed existence of *amodal* content, which can be presented more or less interchangeably in different sensory domains. Communication of such content would then mean the articulation of certain amodal attributes in terms of the chosen presentation modality.

Traditional cognitivist view [4] has seen amodality as information processing on an abstract level, operating independently from modalities, well above sensory domains. But in the light of recent neurostudies, it seems that the thing we call amodality actually refers to close and early interconnections between widely integrated sensory-motor aspects of perception and the very roots of conceptual thinking [5]. According to the embodied perspective to human cognition [6, 7], understanding and thinking indeed are modality dependent, essentially bound with the human body and all of its modalities of interaction. Even the concepts of language are thus strongly dependent on senses and bodily experiencing of the physical world, and on how these experiences are schematised [8, 6, 5].

We must stress that when amodal attributes are referred to in this study, we are talking about certain mental imagery relating naturally to multiple modalities rather than being modality independent (see also [9]). For example, when thinking of an amodal concept of "roughness", one can easily express the mental image with hand gestures, or describe physical attributes that relate to different sensory domains: seeing, hearing and feeling a rough surface as an action-related perception.

According to *image schema theory* [8, 6], "image-like" schematic structures multimodally capture the contours of recurrent sensory-motor experiences of interacting with the environment. They simultaneously act as pre-conceptual, directly meaningful *gestalt* structures of perception, thinking and acting. Despite being called "image", such gestalts have a kinaesthetic character which integrates sensory-motor information. Therefore, in principle, they should be key points of crossmodal design, when trying to find certain sounds and vibrations that evoke similar associations as gestalt completions. Basic image schemas refer to experiences of, e.g., spatial motions or relations, forces and object interactions. In thinking, image schemas are often projected metaphorically.

1.2 Communicative Functions and Design Methods

Within the context of physical training, speed regulation feedback has three main communicative functions: to tell the runner 1) to decrease the pace (*Slow*), 2) to increase the pace (*Urge*) or 3) to keep the current pace (*Ok*). In terms of communication, Slow and Urge functions are directive, as they try to get the runner to undertake a change in speed. The Ok function, in contrast, primarily approves the current state of speed.

For these functions, a *prosody-based* (PB) method for designing non-speech sounds has been proposed [2]. This method aims to utilise "speech melodies" (i.e., intonation) of short vocal expressions which are spontaneously produced for each communicative function. It has previously been found that independently of

verbal content, humans rely on intention-specific forms of intonation in communication, especially when interacting with infants [10]. Studies on prosody-based sound design [2, 11] have found intonation patterns specific to Slow, Urge and Ok functions (along with an additional Reward function), which can be utilised as "musical" parameters in design.

The other design approach is a method which simply utilises the *direct analogy* (DA) between changes in frequency and the corresponding messages of "decelerate", "accelerate" and "keep it constant" [3]. It has been utilised in the design of vibrotactile stimuli, in which the vibration frequency decelerates for the Slow function, accelerates for the Urge function and keeps unchanged for the Ok function.

At the physical level of implementation, both methods basically concern simple frequency-related features along the temporal dimension. In this study, we test their crossmodal functionality by using the physical vibrations as stimuli for different sensory domains.

1.3 Research Questions

1. *How effectively do the different designs serve the intended communicative functions?* Both design methods (PB and DA) have already proven their usefulness in their original domains [12, 3]. Although not being the main focus of this study, it is interesting to see how the designs based on vocal gestures of human expressions compare to the "mechanically" straightforward direct analogy designs.

2. *How does communication vary across the audio and tactile domains?* Both design methods are based on principles which suppose the attribution of certain amodal meanings to physical cues presented in a temporal continuum. In the PB method, such meanings refer to kinaesthetic imagery of projected gestural forms which reflect "bodily affect" in vocalisation [13]. In the DA method, amodal meanings also refer to kinaesthetic imagery such as "decelerating" or "falling". We hypothesise that the designs would work across the domains; i.e., stimuli would still crossmodally resonate with similar "embodied gestalts" and evoke the corresponding spatio-motor mental imagery. We are also interested to see how crossmodal attributions vary across design bases and functions.

3. *What is the role of melodic and rhythmic factors (i.e., frequency changes and segmentation) in communicating the intended functions?* We want to explore how important are the roles these features serve in communication, and whether these roles are weighed differently across functions, domains and different design bases. We especially want to see what kind of effect the melodic factors have within tactile domain, which is not commonly thought of as being in compliance with "melody".

2 Method

2.1 Apparatus

Two Engineering Acoustic C-2 vibrotactile actuators (<http://www.eaiinfo.com/>) were used for tactile presentation and high-quality active speakers were used for



Fig. 1. Experimental setting with a participant performing the tactile task.

audio. As actuators are driven by audio signal (usually in sine waves), it was possible to use the same sound file as a source for both the audio and tactile stimulus. Optimal signal levels were set separately for both domains. To enhance tactile sensation, two actuators were used concurrently, both attached under a wristband in the backside of a left wrist (see Figure 1).

2.2 Stimuli

Three different design bases were prepared, two of them utilising the PB method. The first design base (*PB1*) consists of the same intonation contours for Slow, Urge and Ok functions that were used in the previous evaluation study [12]. These contours were "designerly" chosen from the bulk of 40 source utterances for each function. In contrast, the second design base (*PB2*) uses intonation contours chosen by a statistical classification model based on function-specific prosodic characteristics [11]. The third design base of this experiment represents the DA principle. From the previous study [3], we chose the stimuli designs of 1750 ms duration as they were highly rated and are also better in accordance with the stimuli durations of other design bases.

The stimuli of the DA base were already optimised for C-2 actuators [3], but all source contours of the PB bases were preprocessed to conform with the tactile presentation technology. Intonation contours were first centered to 250 Hz within each source utterer to remove the pitch differences caused, e.g., by the utterer's gender. This center frequency is the recommended operating frequency of C-2. A pilot testing revealed that the original pitch ranges of contours were too ample for the actuator's optimal range. Also, in terms of temporal sensitivity of touch, the contours felt too quick. Therefore the pitch ranges were scaled down by a factor of 0.75 and the contour lengths were scaled up by a factor of 1.25. Excess fluctuations in pitch were finally smoothed out. All modifications were subtle enough to retain the original characteristics of the contours for audio domain.

For each function within each design base, three different versions of stimuli were prepared: one with frequency changes and segmentation (*FC+Seg*) and other ones without any frequency changes (*NoFC*) or segmentation (*NoSeg*). All *FC+Seg* versions are illustrated in Figure 2. The segmentation in the PB bases is derived from the original utterances. As the DA pitch contours originally had no segmentation, it was implemented by inserting short gaps of silence to

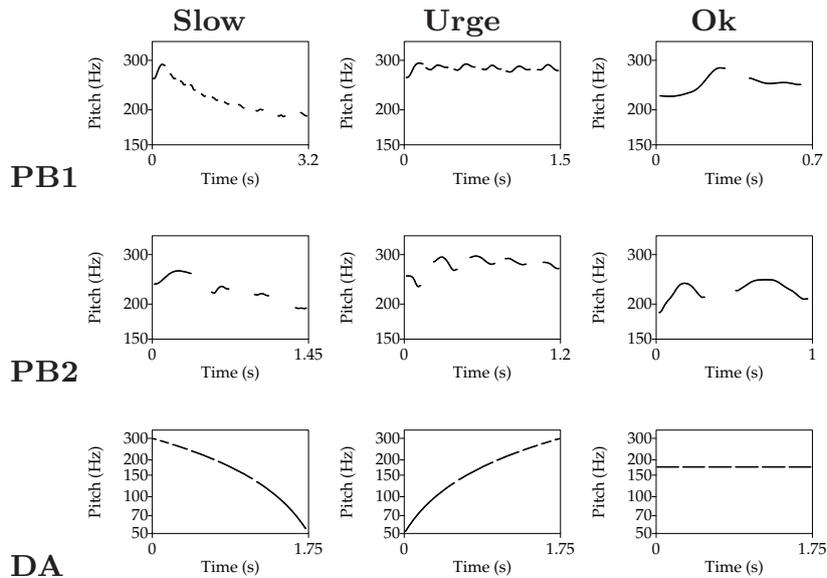


Fig. 2. Visualisations of stimuli containing frequency changes and segmentation.

the FC+Seg and NoFC contours in accordance with the DA principle: for Slow function the onset time intervals of consecutive segments decelerate, for Urge they accelerate and for Ok they remain even. All NoFC versions retain the segmentation but their pitch is flattened to the mean frequency (in Hz) of the contour. Within the DA base, the NoFC versions have the same pitch level (175 Hz) for all functions, but in other cases flattened pitch levels varied across functions. For prosody based NoSeg versions, gaps in the contour were filled by using interpolation. Short rising and falling ramps of intensity envelopes were added to all segment onsets and offsets respectively, to prevent audible "pops".

All 27 stimuli (3 design bases \times 3 functions \times 3 versions) were finally synthesised as sine waves. Intensity levels were the same for all. Despite the optimisations for tactile domain, they all are also in a comfortable hearing range. The preparation of stimuli was made with Praat software (<http://praat.org/>).

2.3 Participants and procedure

Twenty-two students of our university took part in the experiment. They were from different departments, representing many different major subjects. Of the participants, 8 were male and 14 were female. The average age in the group was 25.5 years. All participants reportedly had normal hearing and sense of touch.

Each participant performed rating tasks for both the audio domain and the tactile domain. Half of the participants did the tasks in reverse order, in order to counterbalance the learning effect. In both tasks, all stimuli were presented twice. The participants rated the amount of each of the three functions that were represented by the 54 stimuli (2×27). The ratings were carried out in a random order and using a five-level Likert scale (0-4). In order to block the audible sounds produced by the actuators during the tactile task, the participants wore closed earphones through which white noise was played at a comfortable level.

Before the first task, instructions were read aloud to each participant. Participants were encouraged to rely on their intuition and give ratings relatively quickly. Three novel training stimuli were presented before each task to help the participants to adapt themselves to the type of stimuli. After both tasks, they were asked to freely comment on the experiment.

3 Results

3.1 Effectiveness of Design Bases

For each rating scale (Slow, Urge, and Ok), a separate two-way within-subjects ANOVA was conducted with the within-subjects factors being function (3 levels, Slow, Urge, Ok) and design (3 design bases: PB1, PB2, DA) and the dependent variable being the ratings across the two domains, repetitions and pitch and segmentation variants. The means for all three ratings across the two variables are displayed in Figure 3. It is noticeable that within the ratings for each function, the mean ratings for correct target function were clearly the highest ones.

For the ratings of Slow, ANOVA yielded a significant effect of function, $F(2,42) = 83.05$, $p < .001$ and design, $F(2,42) = 19.44$, $p < .001$, as well as a significant interaction between the two, $F(4,84) = 14.03$, $p < .001$. The ratings of Slow for the correct function (Slow) were clearly separated from the other two functions. The design bases worked significantly differently from each other, PB1 producing the highest overall ratings, followed by the DA base.

The ratings of Urge showed a significant effect of function, $F(2,42) = 118.9$, $p < .001$, but not design, $F(2,42) = 1.13$, $p = .33$. However, a small but significant interaction between the two exists, $F(4,84) = 3.86$, $p < .01$. In other words, overall, the design bases produced equal results, but within the correct communicative function the DA base conveyed the intended meanings more effectively.

Finally, an analysis of the ratings of Ok, ANOVA indicated a significant effect of function, $F(2,42) = 81.70$, $p < .001$, and design, $F(2,42) = 12.47$, $p < .001$, as well as a significant interaction between the two, $F(4,84) = 6.99$, $p < .001$. The ratings for the correct function were distinct from the other target functions (Slow and Urge), and the direct analogy (DA) produced the highest overall ratings. The prosody-based designs were not statistically significantly different from each other.

3.2 Domain-Related Differences

To explore the effect of domain on the ratings, separate three-way ANOVAs were conducted for each rating scale. This time the within-subjects factors consisted of domain (2: audio and tactile), design (3 design bases: PB1, PB2, DA), and function (3 communicative functions). ANOVAs yielded a non-significant effect of domain, $F(1,21) = 0.07, 2.38, 0.05$, respectively for the Slow, Urge and Ok ratings (all $p > 0.20$). However, there were few interactions between the domain and the other two factors, especially in the ratings of Slow. The interactions between

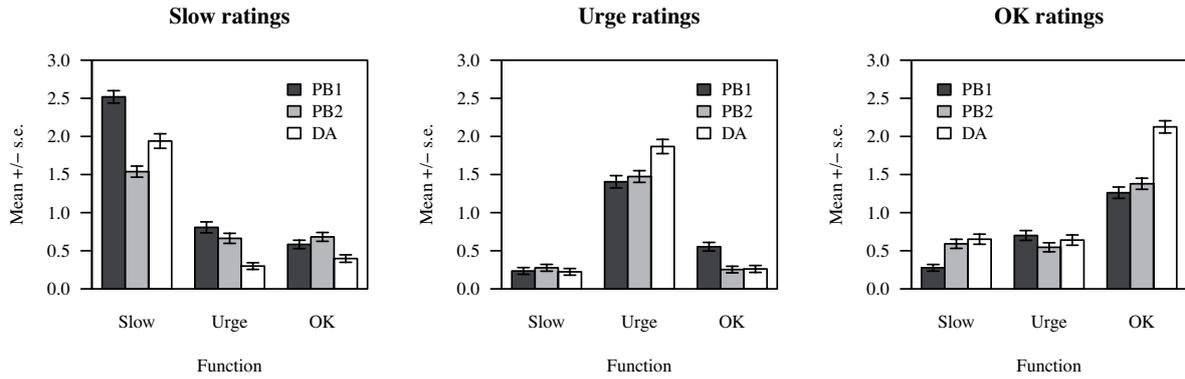


Fig. 3. Mean ratings of Slow, Urge and Ok across design bases and communicative functions.

Table 1. Recognition rates across domains, functions and design bases.

Domain	<i>Audio</i>			<i>Tactile</i>		
Design	PB1	PB2	DA	PB1	PB2	DA
Slow	0.82	0.64	0.75	0.82	0.64	0.55
Urge	0.47	0.50	0.68	0.46	0.52	0.55
Ok	0.64	0.61	0.81	0.47	0.60	0.80

domain and function were significant, $F(2,42) = 6.82$, $p < .01$, and the interactions between domain and design were also significant, $F(2,42) = 10.98$, $p < .001$. In Urge ratings, the domain and design interaction was significant, $F(2,42) = 4.39$, $p < .05$. The sources of these interactions seem to relate to better functionality of the direct analogy (DA) for Slow and Urge within the audio domain (see Table 1). In all, differences due to the domain were surprisingly small.

To illustrate the differences between the domains and other factors, the ratings of the three communicative functions were converted into recognition rates. In this, the highest rating across the three rating scales was compared with the corrected intended function for each example. If the highest rating and the target function matched, the item received a value of 1 (correct) and mismatching items received a value 0 (incorrect). This individual classification was aggregated across participants, domains, functions and design bases, and the mean recognition accuracy is shown in Table 1. These numbers illustrate the sparsity of the domain effect in recognising the functions. The overall recognition rate was somewhat better with the audio (66%) than with the tactile stimulation (60%), and this difference was statistically significant with Kruskal-Wallis test, $\chi^2 = 8.83$, $p < .01$. The fact that the DA design base, in particular, seemed to work better in the audio domain was surprising.

3.3 Effects of Melodic and Rhythmic Manipulations

The roles of frequency changes and segmentation were investigated with a series of three-way ANOVAs. Frequency change factor (two levels: FC and NoFC), segmentation factor (two levels: Seg and NoSeg) and communicative function

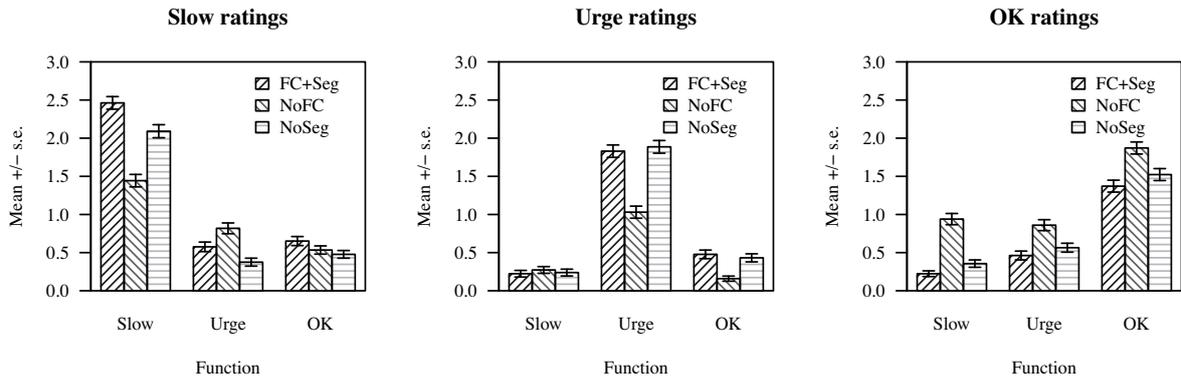


Fig. 4. Mean ratings of Slow, Urge and Ok across pitch and rhythm manipulations and communicative functions.

were the within-subject factors for the three ratings given by the participants. To constrain the number of interactions, only frequency change \times function and segmentation \times function were tested in this analysis since the previous analyses contain most of the other possible interactions.

In Slow, both the frequency change and segmentation factors achieved statistical significance, $F(1,21) = 4.55$, and 14.44 $p < .05$ and $.001$, respectively. Also the interactions between the function and the frequency change and segmentation factors were statistically highly significant, $F(2,42) = 74.2$ and 47.9 , both $p < .001$. These interactions are interestingly demonstrated in Figure 4: non-segmented versions seem to convey the intended meaning of the message more effectively than the segmented version with no frequency changes. Therefore, leaving out the pitch information would harm the communicative function more than leaving out the segmentation. The best rated stimuli for Slow function, in all design bases and in both domains, were the versions that contained both features (FC+Seg).

The Urge ratings were different across frequency changes, $F(1,21) = 28.8$, $p < .001$, but not across segmentation despite the interactions between function and frequency changes, $F(2,42) = 21.7$, $p < .001$, and function and segmentation, $F(4,84) = 97.4$, $p < .001$. As can be observed in Figure 4, the frequency changes seems to be the most effective feature for the Urge function, and leaving out the segmentation does not seem to harm the communication. For the Urge function, the best rated stimuli in the DA and PB1 bases indeed were the ones with no segmentation (NoSeg), while in PB2 it contained both features (FC+Seg).

In the Ok ratings, only a significant main effect of frequency changes was observed, $F(1,21) = 21.9$, $p < .001$, although an interaction effect between function and segmentation was also observed, $F(4,84) = 65.1$, $p < .001$. The interpretation, evident from the Figure 4 as well, points out that the apparently best communication of this function was without frequency changes (NoFC). The best rated stimuli for the Ok function also lacked frequency changes within all designs and in both domains. It must be noted that the DA principle for the Ok function did not permit frequency changes (FC+Seg and NoFC were identical). The lack of this feature might partly explain the superior success of the DA base for the

Ok function, illustrated in Figure 3 and Table 1. Figure 4 also shows that NoFC stimuli for other functions were given relatively high ratings in the Ok scale.

With two-way ANOVAs, we finally tested if the domain factor had any interactions with either the frequency change factor or the segmentation factor. The only statistically significant interaction was found between domain and frequency changes in the ratings of Urge, $F(3,63) = 17.2$, $p < .001$. In other words, the melodic and rhythmic features generally functioned similarly for each function, regardless of domain. The tactile domain thus did not have any apparent handicap with respect to the usage of melodic features.

4 Conclusions and Discussion

All design bases performed well in terms of communicating the intended meaning, bearing in mind that the ratings were given on the basis of intuitive associations rather than any learnt or accustomed coding. Due to its straightforward nature, the DA design base generally seemed to function best. However, both PB design bases functioned effectively as well, especially PB1 which scored the best ratings in communicating the Slow function. The affect-based character of PB designs was evident in the spontaneous expressions of some of the participants, stating that certain stimuli "...just felt like someone were telling you to slow down", for example.

When compared with the previous studies concerning DA and PB1 designs [12, 3], the new results accords with some earlier findings. For example, the NoSeg DA version for Urge (88% recognition) and the FC+Seg PB1 version for Slow (89% recognition) performed especially well. In the previous evaluation of the PB1 design base [12], some participants interpreted the "agitating" imagery associated with Urge samples as warning against going too fast. A similar recognition ambiguity was found in this experiment as well, weakening the ratings for the Urge function. One participant pondered this issue spontaneously: "...it felt like rushing, but it was similar to the warnings in heart-rate meters".

Although there are similarities in the function-specific features between the DA and PB designs, they also differ in many aspects. This indicates that the coupling between the features and the related attributions is not exclusionary. Thus, it should be possible to combine the features relating to the same function. For example, the ascending pitch could be applied to PB Urge designs to potentially reduce the ambiguity in interpretation. Similarly, DA designs could benefit from affect-related features of PB designs.

The most important finding of this study is that domains indeed seem to function in an interchangeable manner, thus supporting the hypothesis. This finding suggests that, regardless of the original usage of any design principle or presentation feature, it might be worth exploring their applicability across modality domains. Many of the participants expressed that "...understanding was easy to 'catch' in both domains", and that "...both domains felt comprehensive" or "...in tactile domain, I played the rhythm in my mind". The audio domain, however, was preferred by the majority of the participants.

In the experiment, the same stimuli were used directly in both domains. This might not be the optimal usage for real-life designs. Of course, we would recommend better utilisation of the domain-related strengths and restrictions: for instance, using the most suitable pitch register and timbre for audio. When audio and tactile stimuli are presented concurrently, the "fused" perception (i.e., *synchresis* [14]) can be something different from the sum of its "parts". Therefore we also recommend creative uses of crossmodal attributes, which would not only be justified as a modality option but also as a multimodal enrichment in supporting the contextually appropriate perception.

Acknowledgments. This work is funded by Finnish Funding Agency for Technology and Innovation, and the following partners: GE Healthcare Finland Ltd., Suunto Ltd., Sandvik Mining and Construction Ltd. and Bronto Skylift Ltd.

References

1. Hoggan, E. & Brewster, S.: Designing audio and tactile crossmodal icons for mobile devices. In Proc. of the 9th International Conference on Multimodal Interfaces. NY: ACM, 162–169 (2007)
2. Tuuri, K. & Eerola, T.: Could function-specific prosodic cues be used as a basis for non-speech user interface sound design? In Proc. of ICAD 2008, Paris: IRCAM (2008)
3. Lylykangas, J., Surakka, V., Rantala, J., Raisamo, J., Raisamo, R. & Tuulari, E.: Vibrotactile Information for Intuitive Speed Regulation. In Proc. of HCI 2009. 112–119 (2009)
4. Fodor, J. A.: The language of thought. Cambridge, MA: Harvard University Press (1975)
5. Gallese, V. & Lakoff, G.: The brain's concepts: The role of the sensory-motor system in reason and language. *Cognitive Neuropsychology*, 22, 455–479 (2005)
6. Johnson, M. & Rohrer, T.: We are live creatures: Embodiment, American pragmatism and the cognitive organism. In: J. Zlatev, T. Ziemke, R. Frank, & R. Dirven (Eds.), *Body, language, and mind*, vol. 1. Berlin: Mouton de Gruyter, 17–54 (2007)
7. Lemay, M.: *Embodied Music Cognition and Mediation Technology*. Cambridge, MA: MIT Press (2008)
8. Johnson, M.: *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago, IL: University of Chicago (1987)
9. Pirhonen, A. & Tuuri, K.: In Search for an Integrated Design Basis for Audio and Haptics. In Proc. of HAID 2008, LNCS 5270. Springer-Verlag, 81–90 (2008)
10. Fernald, A.: Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child development*, 1497–1510 (1989)
11. Tuuri, K. & Eerola, T.: Identifying function-specific prosodic cues for non-speech user interface sound design. In Proc. of the 11th International Conference on Digital Audio Effects, 185–188 (2008)
12. Tuuri, K., Eerola, T. & Pirhonen, A.: Design and Evaluation of Prosody Based Non-Speech Audio Feedback for Physical Training Application. (journal submission)
13. Tuuri, K.: Gestural attributions as semantics in user interface sound design. In: Kopp, S., Wachsmuth, I. (Eds.), *Gesture in Embodied Communication and Human-Computer Interaction*, LNAI 5934. Springer-Verlag, 257–268 (2010)
14. Chion, M.: *Audio-vision: Sound on screen*. NY: Columbia University Press (1990)

- 1 KOSTIAINEN, EMMA, Viestintä ammattiosaamisen ulottuvuutena. - Communication as a dimension of vocational competence. 305 p. Summary 4 p. 2003.
- 2 SEPPÄLÄ, ANTTI, Todellisuutta kuvaamassa – todellisuutta tuottamassa. Työ ja koti television ja vähän radionkin uutisissa. - Describing reality – producing reality. Discourses of work and home in television and on a small scale in radio news. 211 p. Summary 3 p. 2003.
- 3 GERLANDER, MAIJA, Jännitteet lääkärin ja potilaan välisessä viestintäsuhteessa. - Tensions in the doctor-patient communication and relationship. 228 p. Summary 6 p. 2003.
- 4 LEHIKONEN, TAISTO, Religious media theory - Understanding mediated faith and christian applications of modern media. - Uskonnollinen mediateoria: Modernin median kristilliset sovellukset. 341 p. Summary 5 p. 2003.
- 5 JARVA, VESA, Venäläisperäisyys ja ekspressiivisyys suomen murteiden sanastossa. - Russian influence and expressivity in the lexicon of Finnish dialects. 215 p. 6 p. 2003.
- 6 USKALI, TURO, "Älä kirjoita itseäsi ulos" Suomalaisen Moskovan-kirjeenvaihtajuuden alkutaival 1957–1975. - "Do not write yourself out" The beginning of the Finnish Moscow-correspondency in 1957–1975. 484 p. Summary 4 p. 2003.
- 7 VALKONEN, TARJA, Puheviestintätaitojen arviointi. Näkökulmia lukioikäisten esiintymis- ja ryhmätaitoihin. - Assessing speech communication skills. Perspectives on presentation and group communication skills among upper secondary school students. 310 p. Summary 7 p. 2003.
- 8 TAMPERE, KAJA, Public relations in a transition society 1989–2002. Using a stakeholder approach in organisational communications and relation analyses. 137 p. 2003.
- 9 EEROLA, TUOMAS, The dynamics of musical expectancy. Cross-cultural and statistical approaches to melodic expectations. - Musiikillisten odotusten tarkastelu kulttuurien välisen vertailujen ja tilastollisten mallien avulla. 84 p. (277 p.) Yhteenveto 2 p. 2003.
- 10 PAAANANEN, PIIRKKO, Monta polkua musiikkiin. Tonaalisen musiikin perusrakenteiden kehittyminen musiikin tuottamis- ja improvisaatio-tehtävissä ikävuosina 6–11. - Many paths to music. The development of basic structures of tonal music in music production and improvisation at the age of 6–11 years. 235 p. Summary 4 p. 2003.
- 11 LAAKSAMO, JOUKO, Musiikillisten karakterien metamorfoosi. Transformaatio- ja metamorfoosiprosessit Usko Meriläisen tuotannossa vuosina 1963–86. - "Metamorphosis of musical characters". Transformation and metamorphosis processes in the works of Usko Meriläinen during 1963–86. 307 p. Summary 3 p. 2004.
- 12 RAUTIO, RIITTA, *Fortspinnungstypus* Revisited. Schemata and prototypical features in J. S. Bach's Minor-Key Cantata Aria Introductions. - Uusi katsaus kehitysmuotoon. Skeemat ja prototyyppiset piirteet J. S. Bachin kantaattien molliarioiden alkusoitoissa. 238 p. Yhteenveto 3 p. 2004.
- 13 MÄNTYLÄ, KATJA, Idioms and language users: the effect of the characteristics of idioms on their recognition and interpretation by native and non-native speakers of English. - Idiomien ominaisuuksien vaikutus englannin idiomien ymmärtämiseen ja tulkintaan syntyperäisten ja suomea äidinkielenään puhuvien näkökulmasta. 239 p. Yhteenveto 3 p. 2004.
- 14 MIKKONEN, YRJÖ, On conceptualization of music. Applying systemic approach to musicological concepts, with practical examples of music theory and analysis. - Musiikin käsitteellistämisestä. Systemisen tarkastelutavan soveltaminen musikologisiin käsitteisiin sekä käytännön esimerkkejä musiikin teoriasta ja analyysistä. 294 p. Yhteenveto 10 p. 2004.
- 15 HOLM, JAN-MARKUS, Virtual violin in the digital domain. Physical modeling and model-based sound synthesis of violin and its interactive application in virtual environment. - Virtuaalinen viulu digitaalisella alueella. Viulun fysikaalinen mallintaminen ja mallipohjainen äänisynteesi sekä sen vuorovaikutteinen soveltaminen virtuaalitodellisuus ympäristössä. 74 p. (123 p.) Yhteenveto 1 p. 2004.
- 16 KEMP, CHRIS, Towards the holistic interpretation of musical genre classification. - Kohti musiikin genreluokituksen kokonaisvaltaista tulkintaa. 302 p. Yhteenveto 1 p. 2004.
- 17 LEINONEN, KARI, Finlandssvenskt sje-, tje- och s-ljud i kontrastiv belysning. 274 p. Yhteenveto 4 p. 2004.
- 18 MÄKINEN, EEVA, Pianisti cembalistina. Cembalotekniikka cembalonsoittoa aloittavan pianistin ongelmana. - The Pianist as cembalist. Adapting to harpsichord technique as a problem for pianists beginning on the harpsichord. 189 p. Summary 4 p. 2004.
- 19 KINNUNEN, MAURI, Herätysliike kahden kulttuurin rajalla. Lestadiolaisuus Karjalassa 1870–1939. - The Conviction on the boundary of two cultures. Laestadianism in Karelia in 1870–1939. 591 p. Summary 9 p. 2004.
- 20 Лилия Сибберг, "БЕЛЫЕ ЛИЛИИ". ГЕНЕЗИС ФИНСКОГО МИФА В БОЛГАРИИ. РОЛЬ РУССКОГО ФЕННОИЛЬСТВА. ФИНСКО-БОЛГАРСКИЕ КОНТАКТЫ И ПОСРЕДНИКИ С КОНЦА XIX ДО КОНЦА XX ВЕКА. 284 с. - "Belye lilii". Genezis finskogo mifa v Bolgarii. Rol' russkogo fennoil'stva. Finsko-bolgarskie kontakty i posredniki s konca XIX do konca XX veka. 284 p. Yhteenveto 2 p. 2004.

JYVÄSKYLÄ STUDIES IN HUMANITIES

- 21 FUCHS, BERTOLD, *Phonetische Aspekte einer Didaktik der Finnischen Gebärdensprache als Fremdsprache*. - Suomalainen viittomakieli vieraana kielenä. Didaktinen fonetiikka. 476 p. Yhteenveto 14 p. 2004.
- 22 JÄÄSKELÄINEN, PETRI, *Instrumentatiivisuus ja nykysuomen verbinjohto. Semanttinen tutkimus*. - Instrumentality and verb derivation in Finnish. A semantic study. 504 p. Summary 5 p. 2004.
- 23 MERTANEN TOMI, *Kahdentoista markan kapi- na? Vuoden 1956 yleislakko Suomessa*. - A Rebellion for Twelve Marks? - The General Strike of 1956 in Finland. 399 p. Summary 10 p. 2004.
- 24 MALHERBE, JEAN-YVES, *L'œuvre de fiction en prose de Marcel Thiry : une lecture d'inaboutissements*. 353 p. Yhteenveto 1 p. 2004.
- 25 KUHNA, MATTI, *Kahden maailman välissä. Marko Tapion Arktinen hysteria Väinö Linnan haastajana*. - Between two worlds. Marko Tapio's Arktinen hysteria as a challenger to Väinö Linna. 307p. Summary 2 p. 2004.
- 26 VALTONEN, HELI, *Minäkuvat, arvot ja menta- liteetit. Tutkimus 1900-luvun alussa syntynei- den toimihenkilönaisten omaelämäkertoista*. - Self-images, values and mentalities. An autobiographical study of white collar women in twentieth century Finland. 272 p. Summary 6 p. 2004.
- 27 PUSZTAL, BERTALAN, *Religious tourists. Constructing authentic experiences in late modern hungarian catholicism*. - Uskontotu- ristit. Autenttisen elämyksen rakentaminen myöhäismodernissa unkarilaisessa katoli- suudessa. 256 p. Yhteenveto 9 p. Summary in Hungarian 9 p. 2004.
- 28 PÄÄJOKI, TARJA, *Taide kulttuurisena kohtaa- mispaikkana taidekavatuksessa*. - The arts as a place of cultural encounters in arts education. 125 p. Summary 3 p. 2004.
- 29 JUPPI, PIRITA, *"Keitä me olemme? Mitä me haluamme?" Eläinoikeusliike määrittely- kamppailun, marginalisoinnin ja moraalisen paniikin kohteena suomalaisessa sanomaleh- distössä*. - "Who are we? What do we want?" The animal rights movement as an object of discursive struggle, marginalization and moral panic in Finnish newspapers. 315 p. Summary 6 p. 2004.
- 30 HOLMBERG, JUKKA, *Etusivun politiikkaa. Yhteiskunnallisten toimijoiden representointi suomalaisissa sanomalehti uutisissa 1987-2003*. - Front page politics. Representation of societal actors in Finnish newspapers' news articles in 1987-2003. 291 p. Summary 2 p. 2004.
- 31 LAGERBLOM, KIMMO, *Kaukana Kainuussa, valtavyylän varrella. Etnologinen tutkimus Kontiomäen rautatieläisyhteisön elinkaaresta 1950 - 1972*. - Far, far away, nearby a main passage. An ethnological study of the life spans of Kontiomäki railtown 1950 - 1972. 407 p. Summary 2 p. 2004.
- 32 HAKAMÄKI, LEENA, *Scaffolded assistance provided by an EFL teacher during whole- class interaction*. - Vieraan kielen opettajan antama oikea-aikainen tuki luokkahuoneessa. 331 p. Yhteenveto 7 p. 2005.
- 33 VIERGUTZ, GUDRUN, *Beiträge zur Geschichte des Musikunterrichts an den Gelehrtschulen der östlichen Ostseeregion im 16. und 17. Jahrhundert*. - Latinankoulujen musiikinopetuksen historiasta itäisen Itämeren rannikkokaupungeissa 1500- ja 1600-luvuilla. 211 p. Yhteenveto 9 p. 2005.
- 34 NIKULA, KAISU, *Zur Umsetzung deutscher Lyrik in finnische Musik am Beispiel Rainer Maria Rilke und Einjuhani Rautavaara*. - Saksalainen runous suomalaisessa musiikis- sa, esimerkkinä Rainer Maria Rilke ja Einju- hani Rautavaara. 304 p. Yhteenveto 6 p. 2005.
- 35 SYVÄNEN, KARI, *Vastatunteiden dynamiikka musiikkiterapiassa*. - Counter emotions dynamics in music therapy. 186 p. Summary 4 p. 2005.
- 36 ELORANTA, JARI & OJALA, JARI (eds), *East-West trade and the cold war*. 235 p. 2005.
- 37 HILTUNEN, KAISA, *Images of time, thought and emotions: Narration and the spectator's experience in Krzysztof Kieslowski 's late fiction films*. - Ajan, ajattelun ja tunteiden kuvia. Kerronta ja katsojan kokemus Krzysztof Kieslowskin myöhäisfiktiossa. 203 p. Yhteenveto 5 p. 2005.
- 38 AHONEN, KALEVI, *From sugar triangle to cotton triangle. Trade and shipping between America and Baltic Russia, 1783-1860*. 572 p. Yhteenveto 9 p. 2005.
- 39 UTRIAINEN, JAANA, *A gestalt music analysis. Philosophical theory, method, and analysis of Iégor Reznikoff's compositions*. - Hahmope- rustainen musiikkianalyysi. Hahmofilosofi- nen teoria, metodi ja musiikkianalyysi Iégor Reznikoffin sävellyksistä. 222 p. Yhteenveto 3 p. 2005.
- 40 MURTORINNE, ANNAMARI, *Tuskan hauskaa! Tavoitteena tiedostava kirjoittaminen. Kirjoittamisprosessi peruskoulun yhdek- sännellä luokalla*. - Painfully fun! Towards reflective writing process. 338 p. 2005.
- 41 TUNTURI, ANNA-RIITTA, *Der Pikareske Roman als Katalysator in Geschichtlichen Abläufen. Erzählerische Kommunikationsmodelle in Das Leben des Lazarillo von Tormes, bei Thomas Mann und in Einigen Finnischen Romanen*. 183 p. 2005.
- 42 LUOMA-AHO, VILMA, *Faith-holders as Social Capital of Finnish Public Organisations*. - Luottojoukot - Suomalaisten julkisten organisaatioiden sosiaalista pääomaa. 368 p. Yhteenveto 8 p. 2005.

- 43 PENTTINEN, ESA MARTTI, Kielioppi virheiden varjossa. Kielitiedon merkitys lukion saksan kieliopin opetuksessa. - Grammar in the shadow of mistakes. The role of linguistic knowledge in general upper secondary school German grammar instruction. 153 p. Summary 2 p. Zusammenfassung 3 p. 2005.
- 44 KAIVAPALU, ANNEKATRIN, Lähdekieli kielenoppimisen apuna. - Contribution of L1 to foreign language acquisition. 348 p. Summary 7 p. 2005.
- 45 SALAVUO, MIikka, Verkkoavusteinen opiskelu yliopiston musiikkikasvatuksen opiskelukulttuurissa - Network-assisted learning in the learning culture of university music education. 317 p. Summary 5 p. 2005.
- 46 MAIJALA, JUHA, Maaseutuyhteisön kriisi-1930-luvun pula ja pakkojuutokaupat paikallisena ilmiönä Kalajokilaaksossa. - Agricultural society in crisis - the depression of the 1930s and compulsory sales as a local phenomenon in the basin of the Kalajoki-river. 242 p. Summary 4 p. 2005.
- 47 JOUHKI, JUUKA, Imagining the Other. Orientalism and occidentalism in Tamil-European relations in South India. -Tulkintoja Toiseudesta. Orientalismi ja oksidentalismi tamileiden ja eurooppalaisten välisissä suhteissa Etelä-Intiassa. 233 p. Yhteenveto 2 p. 2006.
- 48 LEHTO, KEIJO, Aatteista arkeen. Suomalaisten seitsenpäiväisten sanomalehtien linjapaperien synty ja muutos 1971-2005. - From ideologies to everyday life. Editorial principles of Finnish newspapers, 1971-2005. 499 p. Summary 3 p. 2006.
- 49 VALTONEN, HANNU, Tavallisesta kuriositeetiksi. Kahden Keski-Suomen Ilmailumuseon Messerschmitt Bf 109 -lentokoneen museoarvo. - From Commonplace to curiosity - The Museum value of two Messerschmitt Bf 109 -aircraft at the Central Finland Aviation Museum. 104 p. 2006.
- 50 KALLINEN, KARI, Towards a comprehensive theory of musical emotions. A multi-dimensional research approach and some empirical findings. - Kohti kokonaisvaltaista teoriaa musiikillisista emootioista. Moniulotteinen tutkimuslähestymistapa ja empiirisiä havain-toja. 71 p. (200 p.) Yhteenveto 2 p. 2006.
- 51 ISKANUIS, SANNA, Venäjänkielisten maahanmuuttajaopiskelijoiden kieli-identiteetti. - Language and identity of Russian-speaking students in Finland. 264 p. Summary 5 p. Пефепат 6 c. 2006.
- 52 HEINÄNEN, SEIJA, Käsityö - taide - teollisuus. Näkemyksiä käsityöstä taideiteollisuuteen 1900-luvun alun ammatti- ja aikakausleh-dissä. - Craft - Art - Industry: From craft to industrial art in the views of magazines and trade publications of the early 20th Century. 403 p. Summary 7 p. 2006.
- 53 KAIVAPALU, ANNEKATRIN & PRUULLI, KÜLVI (eds), Lähivertailuja 17. - Close comparisons. 254 p. 2006.
- 54 ALATALO, PIRJO, Directive functions in intra-corporate cross-border email interaction. - Direktiiviset funktiot monikansallisen yrityksen englanninkielisessä sisäisessä sähköpostiviestinnässä. 471 p. Yhteenveto 3 p. 2006.
- 55 KISANTAL, TAMÁS, „...egy tömegmészárlásról mi értelmes dolgot lehetne elmondani?” Az ábrázolásmód mint történelemkonceptió a holokauszt-irodalomban. - „...there is nothing intelligent to say about a massacre”. The representational method as a conception of history in the holocaust-literature. 203 p. Summary 4 p. 2006.
- 56 MATIKAINEN, SATU, Great Britain, British Jews, and the international protection of Romanian Jews, 1900-1914: A study of Jewish diplomacy and minority rights. - Britannia, Britannian juutalaiset ja Romanian juutalaisten kansain-väläinen suojele, 1900-1914: Tutkimus juuta-laisesta diplomatiasta ja vähemmistöoikeuk-sista. 237 p. Yhteenveto 7 p. 2006.
- 57 HÄNNINEN, KIRSI, Visiosta toimintaan. Museoi-den ympäristökasvatus sosiokulttuurisena jatkumona, säätelymekanismina ja innovatiivisena viestintänä. - From vision to action. Environmental education in museums as a socio-cultural continuum, regulating mechanism, and as innovative communication 278 p. Summary 6 p. 2006.
- 58 JOENSUU, SANNA, Kaksi kuvaa työntekijästä. Sisäisen viestinnän opit ja postmoderni näkö-kulma. - Two images of an employee; internal communication doctrines from a postmodern perspective. 225 p. Summary 9 p. 2006.
- 59 KOSKIMÄKI, JOUNI, Happiness is... a good transcription - Reconsidering the Beatles sheet music publications. - Onni on... hyvä transkriptio - Beatles-nuottijulkaisut uudelleen arvioituna. 55 p. (320 p. + CD). Yhteenveto 2 p. 2006.
- 60 HIETAHARJU, MIKKO, Valokuvan voi repiä. Valokuvan rakenne-elementit, käyttöym-päristöt sekä valokuvatulkinnan syntyminen. - Tearing a photograph. Compositional elements, contexts and the birth of the interpretation. 255 p. Summary 5 p. 2006.
- 61 JÄMSÄNEN, AULI, Matrikkeliteilijaksi valikoituminen. Suomen Kuvaamataiteilijat -hakuteoksen (1943) kriteerit. - Prerequisites for being listed in a biographical encyclopedia criteria for the Finnish Artists Encyclopedia of 1943. 285 p. Summary 4 p. 2006.
- 62 HOKKANEN, MARKKU, Quests for Health in Colonial Society. Scottish missionaries and medical culture in the Northern Malawi region, 1875-1930. 519 p. Yhteenveto 9 p. 2006.

- 63 RUUSKANEN, ESA, Viholliskuviin ja viranomaisiin vetoamalla vaiennetut työväentalot. Kuinka Pohjois-Savon Lapuan liike sai nimismiehet ja maaherran sulkemaan 59 kommunistista työväentaloa Pohjois-Savossa vuosina 1930-1932. - The workers' halls closed by scare-mongering and the use of special powers by the authorities. 248 p. Summary 5 p. 2006.
- 64 VARDJA, MERIKE, Tegelaskategoriat ja tegelase kujutamise vahendid Väinö Linna romaanis "Tundmatu sõdur". - Character categories and the means of character representation in Väinö Linna's Novel *The Unknown Soldier*. 208 p. Summary 3 p. 2006.
- 65 TAKÁTS, JÓZSEF, Módszertani berek. Írások az irodalomtörténet-írásról. - The Grove of Methodology. Writings on Literary Historiography. 164 p. Summary 3 p. 2006.
- 66 MIKKOLA, LEENA, Tuen merkitykset potilaan ja hoitajan vuorovaikutuksessa. - Meanings of social support in patient-nurse interaction. 260 p. Summary 3 p. 2006.
- 67 SAARIKALLIO, SUVI, Music as mood regulation in adolescence. - Musiikki nuorten tunteiden säätelyä. 46 p. (119 p.) Yhteenveto 2 p. 2007.
- 68 HUJANEN, ERKKI, Lukijakunnan rajamailla. Sanomalehden muuttuvat merkitykset arjessa. - On the fringes of readership. The changing meanings of newspaper in everyday life. 296 p. Summary 4 p. 2007.
- 69 TUOKKO, Eeva, Mille tasolle perusopetuksen englannin opiskelussa päästään? Perusopetuksen päättövaiheen kansallisen arvioinnin 1999 eurooppalaisen viitekehyksen taitotasoihin linkitetyt tulokset. - What level do pupils reach in English at the end of the comprehensive school? National assessment results linked to the common European framework. 338 p. Summary 7 p. Sammanfattning 1 p. Tiivistelmä 1 p. 2007.
- 70 TUUKKA, TIMO, "Kekkonen konstit". Urho Kekkonen historia- ja politiikkakäsitykset teoriasta käytäntöön 1933-1981. - "Kekkonen's way". Urho Kekkonen's conceptions of history and politics from theory to practice, 1933-1981. 413 p. Summary 3 p. 2007.
- 71 Humanistista kirjoja. 145 s. 2007.
- 72 NIEMINEN, LEA, A complex case: a morphosyntactic approach to complexity in early child language. 296 p. Tiivistelmä 7 p. 2007.
- 73 TORVELAINEN, PÄIVI, Kaksivuotiaiden lasten fonologisen kehityksen variaatio. Puheen ymmärrettävyyden sekä sananmuotojen tavoittelun ja tuottamisen tarkastelu. - Variation in phonological development of two-year-old Finnish children. A study of speech intelligibility and attempting and production of words. 220 p. Summary 10 p. 2007.
- 74 SIITONEN, MARKO, Social interaction in online multiplayer communities. - Vuorovaikutus verkkopeliyhteisöissä. 235 p. Yhteenveto 5 p. 2007.
- 75 STJERNVALL-JÄRVI, BIRGITTA, Kartanoarkkitehtuuri osana Tandefelt-suvun elämäntapaa. - Manor house architecture as part of the Tandefelt family's lifestyle. 231 p. 2007.
- 76 SULKUNEN, SARI, Text authenticity in international reading literacy assessment. Focusing on PISA 2000. - Tekstien autenttisuus kansainvälisissä lukutaidon arviointitutkimuksissa: PISA 2000. 227 p. Tiivistelmä 6 p. 2007.
- 77 KÖSZEGHY, PÉTER, Magyar Alkibiadés. Balassi Bálint élete. - The Hungarian Alcibiades. The life of Bálint Balass. 270 p. Summary 6 p. 2007.
- 78 MIKKONEN, SIMO, State composers and the red courtiers - Music, ideology, and politics in the Soviet 1930s - Valtion säveltäjiä ja punaisia hoviherroja. Musiikki, ideologia ja politiikka 1930-luvun Neuvostoliitossa. 336 p. Yhteenveto 4 p. 2007.
- 79 SIVUNEN, ANU, Vuorovaikutus, viestintä-tekniologia ja identifiointuminen hajautetuissa tiimeissä. - Social interaction, communication technology and identification in virtual teams. 251 p. Summary 6 p. 2007.
- 80 LAPPI, TIINA-RIITTA, Neuvottelu tilan tulkinnoista. Etnologinen tutkimus sosiaalisen ja materiaalisesta ympäristön vuorovaikutuksesta jyvaskyläläisissä kaupunkipuhunnoissa. - Negotiating urban spatiality. An ethnological study on the interplay of social and material environment in urban narrations on Jyväskylä. 231 p. Summary 4 p. 2007.
- 81 HUHTAMÄKI, ÜLLA, "Heittäydä vapauteen". Avantgarde ja Kauko Lehtisen taiteen murros 1961-1965. - "Fling yourself into freedom!" The Avant-Garde and the artistic transition of Kauko Lehtinen over the period 1961-1965. 287 p. Summary 4 p. 2007.
- 82 KELA, MARIA, Jumalan kasvot suomeksi. Metaforisaatio ja erään uskonnollisen ilmauksen synty. - God's face in Finnish. Metaphorisation and the emergence of a religious expression. 275 p. Summary 5 p. 2007.
- 83 SAAVINEN, TAINA, Quality on the move. Discursive construction of higher education policy from the perspective of quality. - Laatu liikkeessä. Korkeakoulupolitiikan diskursiivinen rakentuminen laadun näkökulmasta. 90 p. (176 p.) Yhteenveto 4 p. 2007.
- 84 MÄKILÄ, KIMMO, Tuhoa, tehoa ja tuhlausta. Helsingin Sanomien ja New York Timesin ydinaseutisoinnin tarkastelua diskurssi-analyttisestä näkökulmasta 1945-1998.

- "Powerful, Useful and Wasteful". Discourses of Nuclear Weapons in the New York Times and Helsingin Sanomat 1945-1998. 337 p. Summary 7 p. 2007.
- 85 KANTANEN, HELENA, Stakeholder dialogue and regional engagement in the context of higher education. - Yliopistojen sidosryhmävuoropuhelu ja alueellinen sitoutuminen. 209 p. Yhteenveto 8 p. 2007.
- 86 ALMONKARI, MERJA, Jännittäminen opiskelun puheviestintätilanteissa. - Social anxiety in study-related communication situations. 204 p. Summary 4 p. 2007.
- 87 VALENTINI, CHIARA, Promoting the European Union. Comparative analysis of EU communication strategies in Finland and in Italy. 159 p. (282 p.) 2008.
- 88 PULKKINEN, HANNU, Uutisten arkkitehtuuri - Sanomalehden ulkoasun rakenteiden järjestys ja jousto. - The Architecture of news. Order and flexibility of newspaper design structures. 280 p. Yhteenveto 5 p. 2008.
- 89 MERILÄINEN, MERJA, Monenlaiset oppijat englanninkielisessä kielikylpyopetuksessa - rakennusaineita opetusjärjestelyjen tueksi. - Diverse Children in English Immersion: Tools for Supporting Teaching Arrangements. 197 p. 2008.
- 90 VARES, MARI, The question of Western Hungary/Burgenland, 1918-1923. A territorial question in the context of national and international policy. - Länsi-Unkarin/Burgenlandin kysymys 1918-1923. Aluekysymys kansallisen ja kansainvälisen politiikan kontekstissa. 328 p. Yhteenveto 8 p. 2008.
- 91 ALA-RUONA, ESA, Alkuarviointi kliinisenä käytäntönä psyykkisesti oireilevien asiakkaiden musiikkiterapiassa - strategioita, menetelmiä ja apukeinoja. - Initial assessment as a clinical procedure in music therapy of clients with mental health problems - strategies, methods and tools. 155 p. 2008.
- 92 ORAVALA, JUHA, Kohti elokuvallista ajattelua. Virtuaalisen todellisen ontologia Gilles Deleuzen ja Jean-Luc Godardin elokuvakäsityksissä. - Towards cinematic thinking. The ontology of the virtually real in Gilles Deleuze's and Jean-Luc Godard's conceptions of cinema. 184 p. Summary 6 p. 2008.
- 93 KECSKEMÉTI, ISTVÁN, Papyrusesta megabitteihin. Arkisto- ja valokuvakokoelmien konservoinnin prosessin hallinta. - From papyrus to megabytes: Conservation management of archival and photographic collections. 277 p. 2008.
- 94 SUNI, MINNA, Toista kieltä vuorovaikutuksessa. Kielellisten resurssien jakaminen toisen kielen omaksumisen alkuvaiheessa. - Second language in interaction: sharing linguistic resources in the early stage of second language acquisition. 251 p. Summary 9 p. 2008.
- 95 N. PÁL, JÓZSEF, Modernség, progresszió, Ady Endre és az Ady-Rákosi vita. Egy konfliktusos eszmetörténeti pozíció természetete és következményei. 203 p. Summary 3 p. 2008.
- 96 BARTIS, IMRE, „Az igazság ismérve az, hogy igaz”. Etika és nemzeti identitás Sütő András Anyám könnyű álmot ígér című művében és annak recepciójában. 173 p. Summary 4 p. 2008.
- 97 RANTA-MEYER, TUIRE, Nulla dies sine linea. Avauksia Erkki Melartinin vaikutteisiin, verkostoihin ja vastaanottoon henkilö- ja reseptiohistoriallisena tutkimuksena. - *Nulla dies sine linea*: A biographical and reception-historical approach to Finnish composer Erkki Melartin. 68 p. Summary 6 p. 2008.
- 98 KOIVISTO, KEIJO, Itsenäisen Suomen kanta-aliupseeriston synty, koulutus, rekrytointitausta ja palvelusehdot. - The rise, education, the background of recruitment and conditions of service of the non-commissioned officers in independent Finland. 300 p. Summary 7 p. 2008.
- 99 KISS, MIKLÓS, Between narrative and cognitive approaches. Film theory of non-linearity applied to Hungarian movies. 198 p. 2008.
- 100 RUUSUNEN, AIMO, Todeksi uskottua. Kansandemokraattinen Neuvostoliitto-journalismi rajapinnan tulkkina vuosina 1964-1973. - Believed to be true. Reporting on the USSR as interpretation of a boundary surface in pro-communist partisan journalism 1964-1973. 311 p. Summary 4 p. 2008.
- 101 HÄRMÄLÄ, MARITA, Riittääkö *Ett ögonblick* näytöksi merkonomilta edellytetystä kielitaidosta? Kielitaidon arviointi aikuisten näytötutkinnoissa. - Is *Ett ögonblick* a sufficient demonstration of the language skills required in the qualification of business and administration? Language assessment in competence-based qualifications for adults. 318 p. Summary 4 p. 2008.
- 102 COELHO, JACQUES, The vision of the cyclops. From painting to video ways of seeing in the 20th century and through the eyes of Man Ray. 538 p. 2008.
- 103 BREWIS, KIELO, Stress in the multi-ethnic customer contacts of the Finnish civil servants: Developing critical pragmatic intercultural professionals. - Stressin kokemus suomalaisen viranomaisten monietnisisissä asiakaskontaktteissa: kriittis-pragmaattisen kulttuurien välisen ammattitaidon kehittäminen. 299 p. Yhteenveto 4 p. 2008.
- 104 BELIK, ZHANNA, The Peshekhonovs' Workshop: The Heritage in Icon Painting. 239 p. [Russian]. Summary 7 p. 2008.
- 105 MOILANEN, LAURA-KRISTINA, Talonpoikaisuus, säädyllisyys ja suomalaisuus 1800- ja 1900-lukujen vaihteen suomenkielisen proosan kertomana. - Peasant values, estate society and the Finnish in late nineteenth- and early

- and early twentieth-century narrative literature. 208 p. Summary 3 p. 2008.
- 106 PÄÄRNILÄ, OSSI, Hengen hehkusta tietostrategioihin. Jyväskylän yliopiston humanistisen tiedekunnan viisi vuosikymmentä. 110 p. 2008.
- 107 KANGASNIEMI, JUKKA, Yksinäisyyden kokemuksen avainkomponentit Yleisradion tekstitelevisiion Nuorten palstan kirjoituksissa. - The key components of the experience of loneliness on the Finnish Broadcasting Company's (YLE) teletext forum for adolescents. 388 p. 2008.
- 108 GAJDÓ, TAMÁS, Színház történeti metszetek a 19. század végétől a 20. század közepéig. - Segments of theatre history from the end of the 19th century to the middle of the 20th century. 246 p. Summary 2 p. 2008.
- 109 CATANI, JOHANNA, Yritystapahtuma kontekstina ja kulttuurisena kokemuksena. - Corporate event as context and cultural experience. 140 p. Summary 3 p. 2008.
- 110 MAHLAMÄKI-KAISTINEN, RIIKKA, Mätänevän velhon taidejulistus. Intertekstuaalisen ja -figuraalisen aineiston asema Apollinairen L'Enchanteur pourrissant teoksen tematikassa ja symboliikassa. - Pamphlet of the rotten sorcerer. The themes and symbols that intertextuality and interfigurality raise in Apollinaire's prose work L'Enchanteur pourrissant. 235 p. Résumé 4 p. 2008.
- 111 PIETILÄ, JYRKI, Kirjoitus, juttu, tekstilementti. Suomalainen sanomalehtijournalismi juttutyypin kehityksen valossa printtimedian vuosina 1771-2000. - Written Item, Story, Text Element. Finnish print journalism in the light of the development of journalistic genres during the period 1771-2000. 779 p. Summary 2 p. 2008.
- 112 SAUKKO, PÄIVI, Musiikkiterapian tavoitteet lapsen kuntoutusprosessissa. - The goals of music therapy in the child's rehabilitation process. 215 p. Summary 2 p. 2008.
- 113 LASSILA-MERISALO, MARIA, Faktan ja fiktion rajamailla. Kaunokirjallisen journalismin poetiikka suomalaisissa aikakauslehdissä. - On the borderline of fact and fiction. The poetics of literary journalism in Finnish magazines. 238 p. Summary 3 p. 2009.
- 114 KNUUTINEN, ULLA, Kulttuurihistoriallisten materiaalien menneisyys ja tulevaisuus. Konservoinnin materiaalitutkimuksen heritologiset funktiot. - The heritological functions of materials research of conservation. 157 p. (208 p.) 2009.
- 115 NIIRANEN, SUSANNA, «Miroir de mérite». Valeurs sociales, rôles et image de la femme dans les textes médiévaux des *troubairitz*. - "Arvokkuuden peili". Sosiaaliset arvot, roolit ja naiskuva keskiaikaisissa *troubairitz*-teksteissä. 267 p. Yhteenveto 4 p. 2009.
- 116 ARO, MARI, Speakers and doers. Polyphony and agency in children's beliefs about language learning. - Puhujat ja tekijät. Polyfonia ja agenttiivisuus lasten kielenoppimiskäsityksissä. 184 p. Yhteenveto 5 p. 2009.
- 117 JANTUNEN, TOMMI, Tavu ja lause. Tutkimuksia kahden sekventiaalisen perusyksikön olemuksesta suomalaisessa viittomakielessä. - Syllable and sentence. Studies on the nature of two sequential basic units in Finnish Sign Language. 64 p. 2009.
- 118 SÄRKKÄ, TIMO, Hobson's Imperialism. A Study in Late-Victorian political thought. - J. A. Hobsonin imperialismi. 211 p. Yhteenveto 11 p. 2009.
- 119 LAIHONEN, PETTERI, Language ideologies in the Romanian Banat. Analysis of interviews and academic writings among the Hungarians and Germans. 51 p. (180 p) Yhteenveto 3 p. 2009.
- 120 MÁTYÁS, EMESE, Sprachlernspiele im DaF-Unterricht. Einblick in die Spielpraxis des finnischen und ungarischen Deutsch-als-Fremdsprache-Unterrichts in der gymnasialen Oberstufe sowie in die subjektiven Theorien der Lehrenden über den Einsatz von Sprachlernspielen. 399 p. 2009.
- 121 PARACZKY, ÁGNES, Näkeekö taitava muusikko sen minkä kuulee? Melodiadiktaatin ongelmat suomalaisessa ja unkarilaisessa taidemuusiikin ammattikoulutuksessa. - Do accomplished musicians see what they hear? 164 p. Magyar nyelvű összefoglaló 15 p. Summary 4 p. 2009.
- 122 ELOMAA, EEVA, Oppikirja eläköön! Teoreettisia ja käytännön näkökohtia kielten oppimateriaalien uudistamiseen. - Cheers to the textbook! Theoretical and practical considerations on enhancing foreign language textbook design. 307 p. Zusammenfassung 1 p. 2009.
- 123 HELLE, ANNA, Jäljet sanoissa. Jälkistrukturalistisen kirjallisuuskäsityksen tulo 1980-luvun Suomeen. - Traces in the words. The advent of the poststructuralist conception of literature to Finland in the 1980s. 272 p. Summary 2 p. 2009.
- 124 PIMIÄ, TENHO ILARI, Tähtäin idässä. Suomalainen sukukansojen tutkimus toisessa maailmansodassa. - Setting sights on East Karelia: Finnish ethnology during the Second World War. 275 p. Summary 2 p. 2009.
- 125 VUORIO, KAIJA, Sanoma, lähettäjä, kulttuuri. Lehdistöhistorian tutkimustraditiot Suomessa ja median rakennemuutos. - Message, sender, culture. Traditions of research into the history of the press in Finland and structural change in the media. 107 p. 2009.
- 126 BENE, ADRIÁN, Egyén és közösség. Jean-Paul Sartre *Critique de la raison dialectique* című műve a magyar recepció tükrében. - Individual and community. Jean-Paul Sartre's

- Critique of dialectical reason* in the mirror of the Hungarian reception. 230 p. Summary 5 p. 2009.
- 127 DRAKE, MERJA, Terveysviestinnän kipupisteitä. Terveystiedon tuottajat ja hankkijat Internetissä. - At the interstices of health communication. Producers and seekers of health information on the Internet. 206 p. Summary 9 p. 2009.
- 128 ROUHIAINEN-NEUNHÄUSERER, MAIJASTIINA, Johtajan vuorovaikutusosaaminen ja sen kehittyminen. Johtamisen viestintähaasteet tietoperustaisessa organisaatiossa. - The interpersonal communication competence of leaders and its development. Leadership communication challenges in a knowledge-based organization. 215 p. Summary 9 p. 2009.
- 129 VAARALA, HEIDI, Oudosta omaksi. Miten suomenoppijat keskustelevat nykynovelista? - From strange to familiar: how do learners of Finnish discuss the modern short story? 317 p. Summary 10 p. 2009.
- 130 MARJANEN, KAARINA, The Belly-Button Chord. Connections of pre-and postnatal music education with early mother-child interaction. - Napasointu. Pre- ja postnataalin musiikkikasvatuksen ja varhaisen äiti-vauva-vuorovaikutuksen yhteydet. 189 p. Yhteenveto 4 p. 2009.
- 131 BÖHM, GÁBOR, Önéletírás, emlékezet, elbeszélés. Az emlékező próza hermeneutikai aspektusai az önéletírás-kutatás újabb eredményei tükrében. - Autobiography, remembrance, narrative. The hermeneutical aspects of the literature of remembrance in the mirror of recent research on autobiography. 171 p. Summary 5 p. 2009.
- 132 LEPPÄNEN, SIRPA, PITKÄNEN-HUHTA, ANNE, NIKULA, TARJA, KYTÖLÄ, SAMU, TÖRMÄKANGAS, TIMO, NISSINEN, KARI, KÄÄNTÄ, LEILA, VIRKKULA, TIINA, LAITINEN, MIKKO, PAHTA, PÄIVI, KOSKELA, HEIDI, LÄHDESMÄKI, SALLA & JOUSMÄKI, HENNA, Kansallinen kyselytutkimus englannin kielestä Suomessa: Käyttö, merkitys ja asenteet. - National survey on the English language in Finland: Uses, meanings and attitudes. 365 p. 2009.
- 133 HEIKKINEN, OLLI, Äänitemoodi. Äänite musiikillisessa kommunikaatiossa. - Recording Mode. Recordings in Musical Communication. 149 p. 2010.
- 134 LÄHDESMÄKI, TUULI (ED.), Gender, Nation, Narration. Critical Readings of Cultural Phenomena. 105 p. 2010.
- 135 MIKKONEN, INKA, "Olen sitä mieltä, että". Lukiolaisten yleisönasastotekstien rakenne ja argumentointi. - "In my opinion..." Structure and argumentation of letters to the editor written by upper secondary school students. 242 p. Summary 7 p. 2010.
- 136 NIEMINEN, TOMMI, Lajien synty. Tekstilaji kielitieteen semioottisessa metateoriassa. - Origin of genres: Genre in the semiotic metatheory of linguistics. 303 p. Summary 6 p. 2010.
- 137 KÄÄNTÄ, LEILA, Teacher turn allocation and repair practices in classroom interaction. A multisemiotic perspective. - Opettajan vuoronanto- ja korjauskäytännöt luokkahuonevuorovaikutuksessa: multisemioottinen näkökulma. 295 p. Yhteenveto 4 p. 2010. HUOM: vain verkkoversiona.
- 138 SAARIMÄKI, PASI, Naimisen normit, käytännöt ja konfliktit. Esiaviollinen ja aviollinen seksuaalisuus 1800-luvun lopun keskisuomalaisella maaseudulla. - The norms, practices and conflicts of sex and marriage. Premarital and marital sexual activity in rural Central Finland in the late nineteenth century. 275 p. Summary 12 p. 2010.
- 139 KUUVA, SARI, Symbol, Munch and creativity: Metabolism of visual symbols. - Symboli, Munch ja luovuus - Visuaalisten symbolien metabolismi. 296 p. Yhteenveto 4 p. 2010.
- 140 SKANIAKOS, TERHI, Discoursing Finnish rock. Articulations of identities in the Saimaa-Ilmiö rock documentary. - Suomi-rockin diskursseja. Identiteettien artikulaatioita Saimaa-Ilmiö rockdokumenttielokuvassa. 229 p. 2010.
- 141 KAUPPINEN, MERJA, Lukemisen linjaukset - lukutaito ja sen opetus perusopetuksen äidinkielen ja kirjallisuuden opetussuunnitelmassa. - Literacy delineated - reading literacy and its instruction in the curricula for the mother tongue in basic education. 338 p. Summary 8 p. 2010.
- 142 PEKKOLA, MIKA, Prophet of radicalism. Erich Fromm and the figurative constitution of the crisis of modernity. - Radikalismen profeetta. Erich Fromm ja modernisaation kriisin figuratiivinen rakentuminen. 271 p. Yhteenveto 2 p. 2010.
- 143 KOKKONEN, LOTTI, Pakolaisten vuorovaikutussuhteet. Keski-Suomeen muuttaneiden pakolaisten kokemuksia vuorovaikutussuhteistaan ja kiinnittymisestäään uuteen sosiaaliseen ympäristöön. - Interpersonal relationships of refugees in Central Finland: perceptions of relationship development and attachment to a new social environment. 260 p. Summary 8 p. 2010.
- 144 KANANEN, HELI KAARINA, Kontrolloitu sopeutuminen. Ortodoksinen siirtoväki sotien jälkeisessä Ylä-Savossa (1946-1959). - Controlled integration: Displaced orthodox Finns in postwar upper Savo (1946-1959). 318 p. Summary 4 p. 2010.

J Y V Ä S K Y L Ä S T U D I E S I N H U M A N I T I E S

- 145 NISSI, RIIKKA, Totuuden jäljillä. Tekstin tulkin-
ta nuorten aikuisten raamattupiirikeskuste-
luissa. – In search of the truth. Text interpre-
tation in young adults' Bible study conversa-
tions. 351 p. Summary 5 p. 2010.
- 146 LILJA, NIINA, Ongelmista oppimiseen. Toisen
aloittamat korjausjaksot kakkoskielisessä kes-
kustelussa. – Other-initiated repair sequences
in Finnish second language interactions.
336 p. Summary 8 p. 2010.
- 147 VÁRADI, ILDIKÓ, A parasztpolgárosodás
„finn útja”. Kodolányi János finnországi
tevékenysége és finn útirajzai. – The “Finn-
ish Way” of Peasant-Bourgeoisization. János
Kodolányi's Activity in Finland and His
Travelogues on Finland. 182 p. Summary 3 p.
2010.
- 148 HANKALA, MARI, Sanomalehdellä aktiiviseksi
kansalaiseksi? Näkökulmia nuorten sanoma-
lehtien lukijuuteen ja koulun sanomalehti-
tiopetukseen. – Active citizenship through
newspapers? Perspectives on young people's
newspaper readership and on the use of
newspapers in education. 222 p. Summary 5
p. 2011.
- 149 SALMINEN, ELINA, Monta kuvaa menneisyy-
destä. Etnologinen tutkimus museoko-koelm-
ien yksityisyydestä ja julkisuudesta. – Images
of the Past. An ethnological study of the
privacy and publicity of museum collections.
226 p. Summary 5 p. 2011. HUOM: vain verk-
koversiona.
- 150 JÄRVI, ULLA, Media terveyden lähteillä. Miten
sairaus ja terveys rakentuvat 2000-luvun
mediassa. – Media forces and health sources.
Study of sickness and health in the media.
209 p. Summary 3 p. 2011.
- 151 ULLAKONOJA, RIIKKA, Da. Eto vopros! Prosodic
development of Finnish students' read-aloud
Russian during study in Russia. – Suoma-
laisten opiskelijoiden lukupuhunnan prosod-
inen kehittyminen vaihto-opiskelujakson
aikana Venäjällä. 159 p. (208 p.)
Summary 5 p. 2011.
- 152 MARITA VOS, RAGNHILD LUND, ZVI REICH AND
HALLIKI HARRO-LOIT (EDS), Developing a Crisis
Communication Scorecard. Outcomes of
an International Research Project 2008-2011
(Ref.). 340 p. 2011.
- 153 PUNKANEN, MARKO, Improvisational music
therapy and perception of emotions in music
by people with depression. 60 p. (94 p.)
Yhteenveto 1 p. 2011.
- 154 DI ROSARIO, GIOVANNA, Electronic poetry.
Understanding poetry in the digital environ-
ment. – Elektroninen runous. Miten runous
ymmärretään digitaalisessa ympäristössä?
327 p. Tiivistelmä 1 p. 2011.
- 155 TUURI, KAI, Hearing Gestures: Vocalisations
as embodied projections of intentionality in
designing non-speech sounds for communi-
cative functions. – Puheakteissa kehollisesti
välittyvä intentionaalisuus apuna ei-
kielellisesti viestivien käyttöliittymä-äänien
suunnittelussa. 50 p. (200 p.) Yhteenveto 2 p.
2011.