Design and evaluation of prosody-based non-speech audio feedback for physical training application☆

Kai Tuuri³,*, Tuomas Eerola⁵, Antti Pirhonen⁴

³Department of Computer Science and Information Systems, University of Jyväskylä, P.O. Box 35, FI-40014, Finland
⁴Department of Music, University of Jyväskylä, P.O. Box 35, FI-40014, Finland

Received 18 December 2009; received in revised form 17 January 2011; accepted 27 June 2011
Communicated by P. Baudisch
Available online 2 July 2011

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, exemplifying an interpersonal approach to 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.

© 2011 Elsevier Ltd. All rights reserved.

Keywords: Sound design; Interaction design; Non-speech sounds; Design process; Prosody; Embodied cognition; Intentionality

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 (Blackler and Huri-enne, 2007). 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 (1979) theory of ecological perception, 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 area, there is not yet sufficient methodological support for identifying and utilising the potential of sound in UI design (e.g., Ronkainen, 2010). Moreover, 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/.

Author's personal copy
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 (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 (Liberman 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; Justlin 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 a 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 (1979) argues 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 interactive 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 projects, 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.

GEAR 2 and 3 (Grammar of Earcons), funded by the Finnish Funding Agency for Technology and Innovation.
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 Fig. 1). Such applications are based on extensive research on how to optimise physical training to enhance physical performance. The 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 programme, 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, 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

<table>
<thead>
<tr>
<th>Function</th>
<th>Intended content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow down</td>
<td>Slow down your pace (your pace is too fast)</td>
</tr>
<tr>
<td>Urge</td>
<td>Quicken your pace (your pace is too slow)</td>
</tr>
<tr>
<td>Ok</td>
<td>Keep this pace (your pace is fine)</td>
</tr>
<tr>
<td>Reward</td>
<td>Well done! (you reached a training goal)</td>
</tr>
</tbody>
</table>
(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 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 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, nine 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 novel) each participant produced two experimental trials for each of the four functions (20 participants × 2 conditions × 4 functions × 2 trials).

3.3.2. Results of the production task

The extraction of the prosodic features from recorded utterances was carried out using the 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

\[
P = 69 + 12 \times \log_2 \left( \frac{F_0}{440} \right),
\]

where \( P \) represents the pitch numbering convention used in MIDI standard (\( C_4 = 60 \)). 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 and voice quality. Voice intensity information (energy in dBs) was also extracted from utterances. However, intensity measures could not be used in this...
Similar overall patterns are indicated also within the utterances of vowel condition.

In Fig. 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 observable: 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 pitch contour is mostly descending. Ok utterances are also low in pitch and relatively short in overall length. They seem to consist of two segments 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

---

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

---

*Author's personal copy*
(rise-fall pattern) to the pitch contours of the Approval category in the Fernald’s (1992b) study. 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 the 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 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 ($\eta^2 = 0.07–0.57$) 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 (first utterance $M=0.71$ s, 95%CI=0.63–0.80 s and second 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 five out of six 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, $p < 0.001$) and four out of six comparisons for the mean $F_0$ and variance of $F_0$ ($F=125.7$ and 13, both $p < 0.001$). 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).

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

<table>
<thead>
<tr>
<th>Source</th>
<th>Feature</th>
<th>df</th>
<th>$F_0$ (M)</th>
<th>$F_0$ (SD)</th>
<th>Length</th>
<th>Unvoiced%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition (C)</td>
<td></td>
<td>1</td>
<td>0.5</td>
<td>0.2</td>
<td>36.7**</td>
<td>56.4**</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>$\eta^2$</td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Function (F)</td>
<td></td>
<td>3</td>
<td>144.8**</td>
<td>10.6**</td>
<td>9.3**</td>
<td>67.8**</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>2</td>
<td>0.57</td>
<td>0.09</td>
<td>0.07</td>
<td>0.32</td>
</tr>
<tr>
<td>$\eta^2$</td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Repetition (R)</td>
<td></td>
<td>1</td>
<td>5.1</td>
<td>0.08</td>
<td>16.0**</td>
<td>2.3</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>2</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>$\eta^2$</td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C $\times$ F</td>
<td></td>
<td>3</td>
<td>5.7**</td>
<td>3.19*</td>
<td>3.2*</td>
<td>21.4</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>2</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>$\eta^2$</td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>C $\times$ R</td>
<td></td>
<td>1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.7</td>
<td>4.6*</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>$\eta^2$</td>
<td></td>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*p < 0.05, ** p < 0.001 (Greenhouse–Geisser corrected).
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 (Fig. 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 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 three 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 Fig. 3 by the boxed areas and the actual observations are denoted by four types of individuals 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 two 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, centred 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 ($\alpha = 0.94$, 0.97, 0.94, and 0.87 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 Fig. 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 a two-dimensional representation by convolution with a two-dimensional 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 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 three were the same in the slow down category, eight in the urgent category, three in the ok category and five 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 Fig. 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 recorded. 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 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 programme 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 Fig. 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 Fig. 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 Fig. 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, 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 interactive 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

---

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 Fig. 2). Example number 1 of each function (bolded) was chosen as the basis for sound design.

<table>
<thead>
<tr>
<th>Communicative functions</th>
<th>Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ex. 1</td>
</tr>
<tr>
<td>Slow down</td>
<td>61</td>
</tr>
<tr>
<td>Urge</td>
<td>121</td>
</tr>
<tr>
<td>Ok</td>
<td>201</td>
</tr>
<tr>
<td>Reward</td>
<td>61</td>
</tr>
</tbody>
</table>

---

5Note that the recorded tones had an approximately 3-4% higher pitch level in relation to the MIDI note they represent.
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, eight female and four male. All of the participants were students at our university, and aged 23–29. They were from different departments, representing seven 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 × 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 min 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 computer and video camera.

![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.](image-url)
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.

Fig. 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, interpreted 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 s.

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 minimize 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 a 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 ($\alpha = 0.77$). Two participants were identified as outliers ($> 2$ SDs from the mean intersubject correlation) and removed, which resulted in an acceptable inter-rater reliability ($\alpha = 0.90$). 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 $p < 0.001$ except for OK, $p < 0.01$). 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 $\%25, \chi^2 = 21.6, p < 0.001$). 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 horizontal dimension. A successful expression is thus indicated by a high percentage in the diagonal of the table. $\chi^2$ 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 eight participants used an expression that directly referred to the intended meaning, and only two had conflicting descriptions.

- **Urge**: Seven participants used an expression referring to acceleration. Words referring to warning or alarm were used by four 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 four 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, eight 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”. Three participants out of 12 gave clearly conflicting expressions.

- **reward**: Five 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, seven 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 two, who found the sound “shripping” 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 realised 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 Section 3.2, should be heuristically applicable to various cases of design. Carefully implemented modeling 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 Section 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-centred 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, 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 Rinnott, 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


