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 nonverbal 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 re-enactment. 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 meaning. Hence we suggest that the evoked functional meaning[1] (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 in 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 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 functions[2], 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?

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

[2] For 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 function [3]. 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.

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 HHB Portadisc audio recorder, an HP laptop computer (for running the animation), Olympus earphones, and a Samsung 17” LCD display.

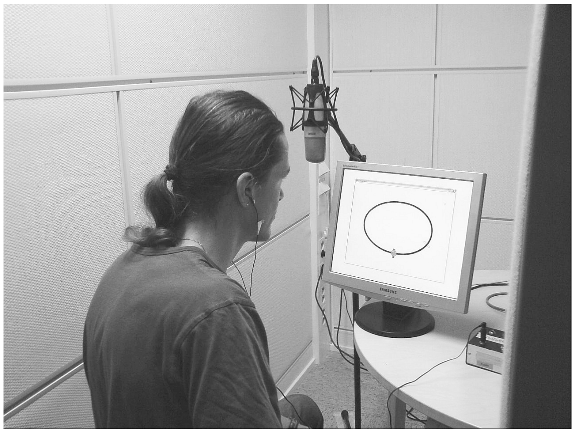


Figure 1: The experimental setting showing the computer display with the animation and a participant.

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,

[3] Finnish 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*).

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.

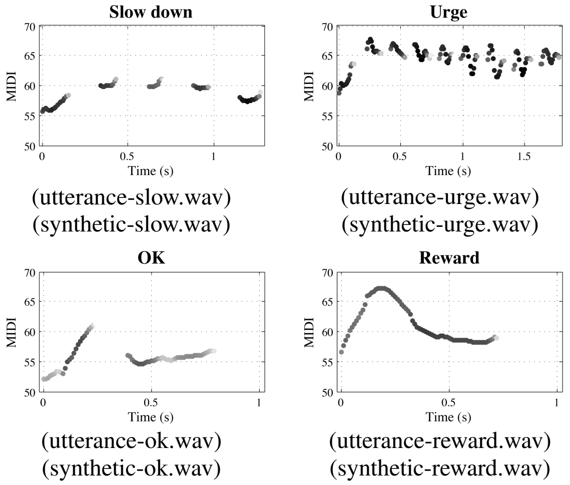
2.5. Acoustic analysis

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_0s (in Hz) were converted into linear scale by

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

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).



(utterance-slow.wav)
(synthetic-slow.wav)

(utterance-urge.wav)
(synthetic-urge.wav)

(utterance-ok.wav)
(synthetic-ok.wav)

(utterance-reward.wav)
(synthetic-reward.wav)

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.

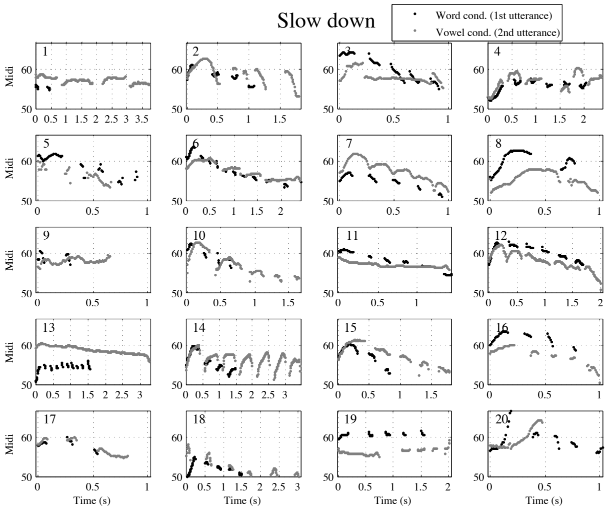


Figure 3: The F_0 contours of two utterances by all the participants for the *Slow down* communicative function.

The utterances were then summarised by 8 simple descriptors: mean frequency, F_0 *(M)*, frequency variation, F_0 *(SD)*, voice intensity, *VoInt (M)*, intensity variation, *VoInt (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 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.

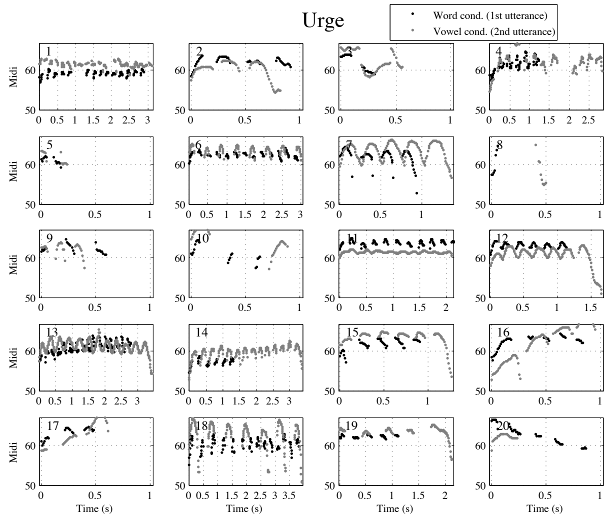


Figure 4: The F_0 contours of two utterances by all the participants for the *Urge* communicative function.

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_0s (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.

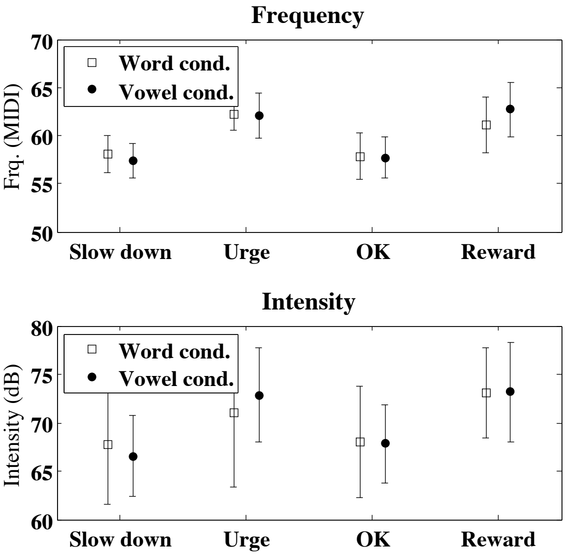


Figure 5: Mean F_0 and intensity across utterances and conditions.

Next the differences in the mean F_0s 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 between the functions (4 × 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.

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.

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.

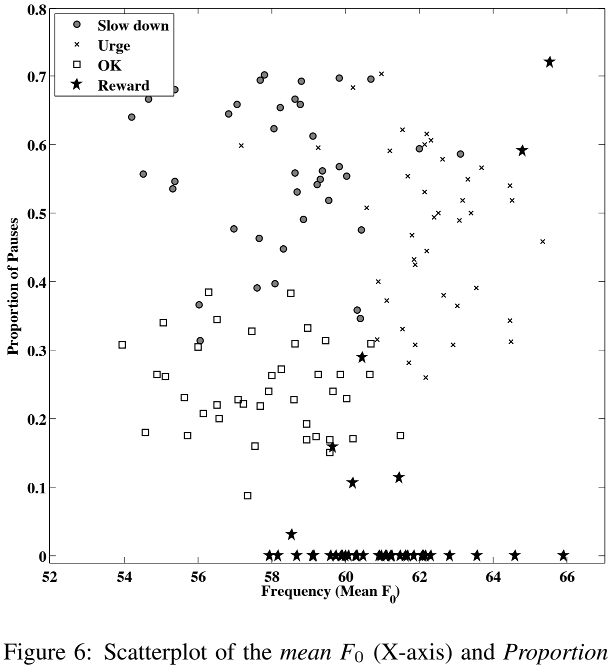


Figure 6: Scatterplot of the *mean* F_0 (X-axis) and *Proportion of pauses* (Y-axis) for each utterance representing the four communicative functions.

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 necessary 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 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.

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)

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 Conferencce 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,” *Glot 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.