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

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

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

ABSTRACT

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

1. INTRODUCTION

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

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

Communicative functions of vocalisations refer to the communicative intentions of the speaker as well as to the vocalisation's pragmatic meaning. We suggest that the evoked functional meaning[1] of nonverbal vocal patterns is indicated by the empathetic perception [5] of sound and its indexical relation to the situational context. The dependency to the situational conditions may vary. For example, an infant can perceive mother's vocal patterns as *prohibitive* in many different situations as long as the child is able to associate the utterance with her actions. On the other hand, the perception of certain functions may have more fine-tuned relationships between the vocalisation and its context. Communicative functions represent particular categories of vocal expression and also certain contexts of interactions. In this study we will use this term essentially to categorise certain context-specific communicative intentions for using sound.

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

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

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

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

2. METHOD

2.1. Participants

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

2.2. Experiment

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

To set different conditions for the usage of nonverbal means in the expression, we also chose to use an additional *moderator variable* which determines two different methods for vocalisations: *Word condition* is a verbal form of expression using specified words for each function [2]. *Vowel condition* is a fully nonverbal form of expression (using "a"-vowel for all the functions). For a more detailed description of the experiment, see Tuuri and Eerola [8].

2.3. Pre-processing of material

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

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

[2]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*).

For all utterances, F_0s (in Hz) were converted into linear scale using the pitch numbering convention of MIDI standard (C_4 = 60). Note that this scaling does not alter the resolution of the F_0 as they were not reduced to the integers of the MIDI note standard. Next, the F_0 contours were centred to MIDI note 60 (261.6 Hz) within each participant to remove the obvious F_0 differences between the participants caused by gender, size, etc. For intensity, a similar operation was carried out (centred to 70 dB).

3. RESULTS

We first investigated whether there were differences between the repeated utterances each participant gave for each function and condition. One-way ANOVA yielded no statistically significant differences in the mean F_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. Within the scope of this paper, the subsequent analysis of prosodic features for each function was carried out using solely the utterances of Word condition.

3.1. Acoustic predictors

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

3.2. Classification using Regression Tree Analysis

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

Table 1: Summary of predictors.

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

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

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

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

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

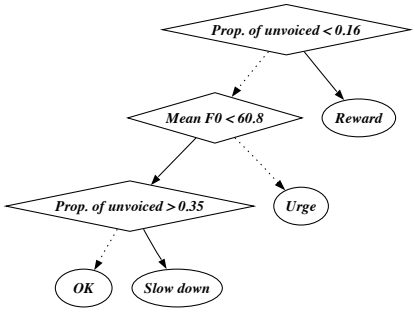

Figure 1: Decision tree based on the final, pruned and cross-validated RTA model. Solid lines indicate the path taken when a rule is filled.

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

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

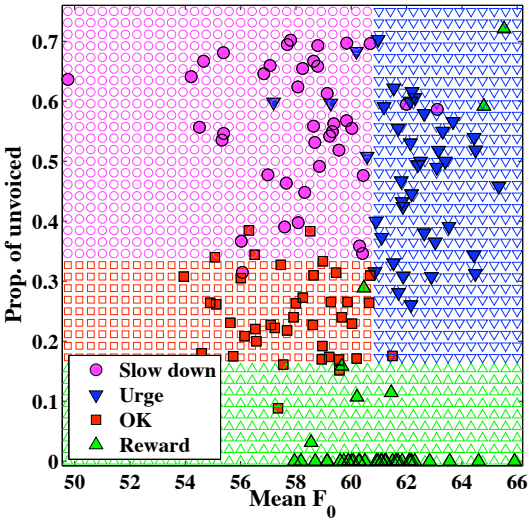

Figure 2: Scatterplot of the two predictors (*Mean frequency* & *Proportion of unvoiced frames*) that were able to classify most utterances into the four function-specific categories. Note that the original predictor values (not the z-scores) are displayed.

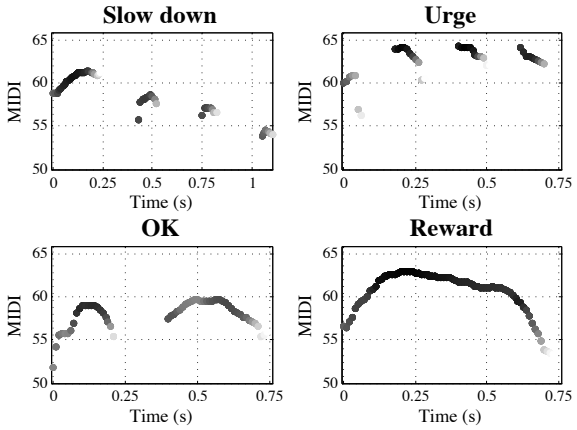

Figure 3: Examples of the frequency and intensity contours for each four functions from a single participant (Word condition). Darker colour indicates higher dB (intensity) value.

4. DISCUSSION

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

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

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

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

[1] A. L. Blackler and J. Hurtienne, "Towards a unified view of intuitive interaction: definitions, models and tools across the world," *MMI-Interaktiv*, vol. 13, pp. 37–55, 2007.

[2] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, pp. 614–636, 1996.

[3] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychological Bulletin*, vol. 129, no. 5, pp. 770–814, 2003.

[4] K. Tuuri, M.-S. Mustonen, and A. Pirhonen, "Same sound - different meanings: A novel scheme for modes of listening," in *Proceedings of Audio Mostly 2007, 2nd Conference on Interaction with Sound*. 2007, pp. 13–18, Fraunhofer IDMT.

[5] M. Iacoboni, *Understanding Others: Imitation, Language, and Empathy*, pp. 77–100, Perspectives on Imitation From Neuroscience to Social Science - Volume 1: Mechanisms of Imitation and Imitation in Animals. MIT Press, 2005.

[6] A. Fernald, "Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective," in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, J. H. Barkow, L. Cosmides, and J. Tooby, Eds., pp. 391–428. Oxford University Press, 1992.

[7] A. Fernald, "Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages," *Child Development*, vol. 64, no. 3, pp. 657–674, 1993.

[8] K. Tuuri and T. Eerola, "Could function-specific prosodic cues be used as a basis for non-speech user interface sound design?," in *Proceedings of the 14th International Conference on Auditory Display, Paris, France*, 2008 (in press).

[9] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[10] L. Breiman, J. Freidman, R. Olshen, and C. Stone, *Classification and regression trees*, Wadsworth, Belmont, CA, USA, 1984.

[11] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.

[12] G. J. McLachlan, *Discriminant analysis and statistical pattern recognition*, Wiley-Interscience, New York, NY, USA, 2004.

[13] W. Gaver, "Auditory icons: Using sound in computer interfaces," *Human-Computer Interaction*, vol. 2, pp. 167–177, 1986.