SPEMOTICONS: TEXT-TO-SPEECH BASED EMOTIONAL AUDITORY CUES

Géza Németh, Gábor Olaszy, Tamás Gábor Csapó

Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics,
H-1117 Budapest, Magyar tudósok körútja 2., Hungary
{nemeth, olaszy, csapot}@tmit.bme.hu

ABSTRACT

There are various methods of providing auditory cues in human-computer user interfaces. The two basic traditional methods are the application of real-life sounds (auditory icons) and artificially generated audio signals (earcons). Recently in-between solutions have been developed based on text-to-speech (TTS) technology. Spearcons are speeded-up versions of TTS output of a particular text-template while spindex cues are generated as auditory index items from the first letter of menu list elements. Auditory emoticons are the non-verbal human sound based audible equivalents of emoticons. Auditory emoticons are the non-verbal human sound based audible equivalents of emoticons. However we are not aware of any attempt for generating auditory emotional and intentional state representation (comparable to emoticon characters) based on a TTS solution. We denote these meaningless cues as *spemoticons*. The interactive development environment of our TTS system is applied as a modification tool for generating spemoticons. The intensity, duration and pitch structure of the generated speech is manipulated. An experimental sound inventory of 44 elements was compiled and tested by 54 adult subjects for the selection of spemoticons.

1. INTRODUCTION

The scientific literature on acoustic units used in human-computer interfaces has at least 25 years of history [1]. The initial spread of personal computers raised the idea that it would be useful if the user would not only see a visual feedback, but also could be informed of certain events by audio feedback. Since the appearance of the first acoustic units with this aim, several types of sounds have been applied recently in auditory menus, of which we will refer to five (auditory icon, earcon, spearcon, spindex and auditory emoticon) and extend it with a novel prototype.

Auditory icons are typically simplified versions of sounds occurring in our natural environment [2]. In the first experiments it has been found that sounds which occur in the real world are more acceptable to users than artificial sounds. In the computer environment such an acoustic icon can be bound to an event, which provides information about a certain event, data or action. An example for this could be the arrival of an email: if we receive a short email, we hear an “easy” sound; in another case with an email containing a huge attachment the sound will be “heavy”. An analogy of this to real life is the sound of throwing a postal mail or package to the mailbox. This way, the auditory icon for an incoming email could be the sound of the thump of an object into the mailbox. Icons of this type typically require little training and are easily learned. However, it is sometimes difficult to find an auditory icon for a certain function (e.g. save or load a file). There have been a lot of applications using auditory icons, including the help to blind users [3]; collaboration between several people working on the same task in an office [4]; a navigation system [5] and an information system of a mobile provider [6].

Earcons are usually composed of musical motives, which are rhythmic sequences of pitches with variable intensity, timbre and register [7]. The word earcon was created as an analogy of the word icon; to express the meaning of an audible icon. The goal of these non-verbal representations is to provide information to the users about some objects, operations or interactions. In the comparative study of [6] it was shown that auditory icons are much more suitable for mobile phone users than earcons. The problem with earcons is that they use an arbitrary mapping between a sound and an object; which means they are slower to learn and quicker to forget. The conclusion was that they are less useful than auditory icons. The advantage of this arbitrary mapping is that earcons can represent any concept, so the previous example of saving or loading a file can be expressed easily with earcons.

Spearcons are basically spoken phrases, which are speeded up, even to the point of being unintelligible [7]. They can be produced with a TTS system which has a capability of accelerating the output speech in a high degree. Spearcons are also referred to as “fingerprints” of speech, because they are originating from real speech samples. Generating spearcons is an easy task which makes them highly useful in menus which change often. According to some experiments, they are quick to learn by the users, as spearcons derive from original speech.

Spindex is an auditory index based on speech sounds [7]. The main idea here is to use short speech syllables starting with the same phoneme as the word or phrase where the user is currently navigating in a menu. Before listening to the whole prompt using TTS, a short spindex utterance is played first with a small pause after that. If the spindex gives enough information for the user about the navigation (e.g. in a contact list which is sorted by the alphabet), (s)he can navigate further before listening to the whole phrase. The study of [7] shows that spindex-enhanced navigation has a higher performance in menu navigation compared to TTS-only auditory menus. Spindex is particularly useful in long menus – e.g. in the playlist of an MP3 player or the contact list of a cell phone. It is

easy to learn for the users (it is expected that no training is necessary) and quick to generate with a TTS engine. However, several users reported it to be annoying when listening to them through a long usage.

Auditory emoticons are the non-verbal audible representations of smileys [8]. Wersényi conducted an extensive evaluation and comparison of the above-mentioned sound effects (auditory icons, earcons and spearcons) and extended it with a new design method for emotionally driven auditory events, as a parallel to the manner smileys (emoticons) express emotions in an easy but limited way. These auditory emoticons contain mostly non-verbal human sounds, which are language independent (e.g. laughter, chucking and kiss) and can reflect the emotional mood of the speaker. Experiments with blind and sighted users showed that auditory emoticons are received well, and users especially welcomed such female voices.

We have introduced five types of sounds used in auditory interaction; of which the auditory icons have the longest history and spearcon plus spindex seem to be the easiest to use by mobile phone users. From the above, only auditory emoticons have expressive content, using non-speech human sounds. In this paper a new sound type –spemoticon- (speech-based emoticon) is introduced for expressing emotional and intentional states.

2. TOOL SET

2.1. Motivation

When communicating with machines we suppose that human speech based but nonsense short sound sequences may be used successfully for expressing emotions, intentions related to certain situations and activities. The novelty of the experiment is the use of synthetic speech [9] as a basis on the one hand and our interactive parameter modification tool to generate the final sound sample (if selected then spemoticon) on the other hand.

Thus we get synthetic acoustic events that are close to human speech, meaningless, and do not occur in real life, but may express emotional content. Three speech parameters: time structure, intensity and pitch can be modified by the tool on sound and sound part level. Inserting breaks of different length, the rhythm of the sound event can also be diversified. The person, who generates the sound samples is free to decide how to adjust the three parameters. Thus theoretically the number of generated sound samples is only the question of fantasy.

2.2. Hypothesis

We suppose that speech-like, but nonsense samples based on TTS technology can express certain situations, emotions for humans. The expressive samples can be selected by perceptual tests from a sound inventory, where several such artificially generated sounds are stored (referred to as sound inventory). The selected samples are the *spemoticons* (stored in an inventory of spemoticons). The sound inventory is always larger than the inventory of spemoticons.

2.3. Creating the sound inventory

To build up the experimental sound inventory a Hungarian TTS system [9] was used together with the sound modification tool. The procedure has two steps: generate the basic sound from text input and freely modify its acoustic structure until a characteristic sound is achieved (this is not yet a spemoticon). This part of the experiment is fully open.

The generation of a sound inventory for selecting spemoticons depends on the person, who uses the TTS and the modification tool (editor). Why use TTS? The initial data for sound generation are objective. The change of parameters and the sound effect can be controlled. There is no need for a human speaker. The modification of sound samples can be planned exactly. The re-synthesis of a sound always gives the same result. Sound samples generated for this experiment can be tested at: http://hungarianspeech.tmit.bme.hu/spemoticon.

2.4. The TTS system

The Hungarian version of the Profivox TTS sytem (male voice) was used for generating the basic sound for modification [9]. The synthesizer uses diphone (CV, VC, VV, CC) and triphone (CVC) waveform sequences for speech generation (C=consonant; V=vowel). The input of the system is a simple text-like character sequence. At this point no intonation is used.

The prosodic modification (the final character of the sound) can be implemented by the interactive TTS modification tool. This tool is tightly coupled with the Profivox software (client-server). Several parameters of the synthesized items are displayed on the screen: the data matrix, the waveform (time-amplitude function), the intensity structure or the pitch curve (Figure 1.). By changing the parameters of the data matrix (upper layer of Figure 1), a speech based acoustic sample can be generated.

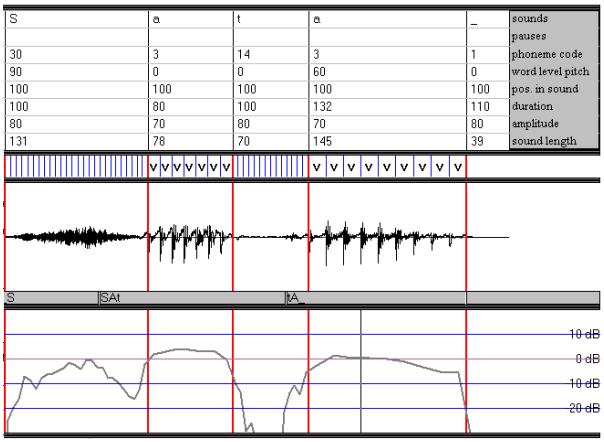

Figure 1. The screen of the modification tool: parameters for modification (upper), waveform (middle) and the intensity (lower) of a two syllable CVCV synthesized meaningless text.

On the upper right corner of Figure 1. the definition of the parameters can be seen: pauses (insertion of pauses); phoneme code (by changing the number, another sound can

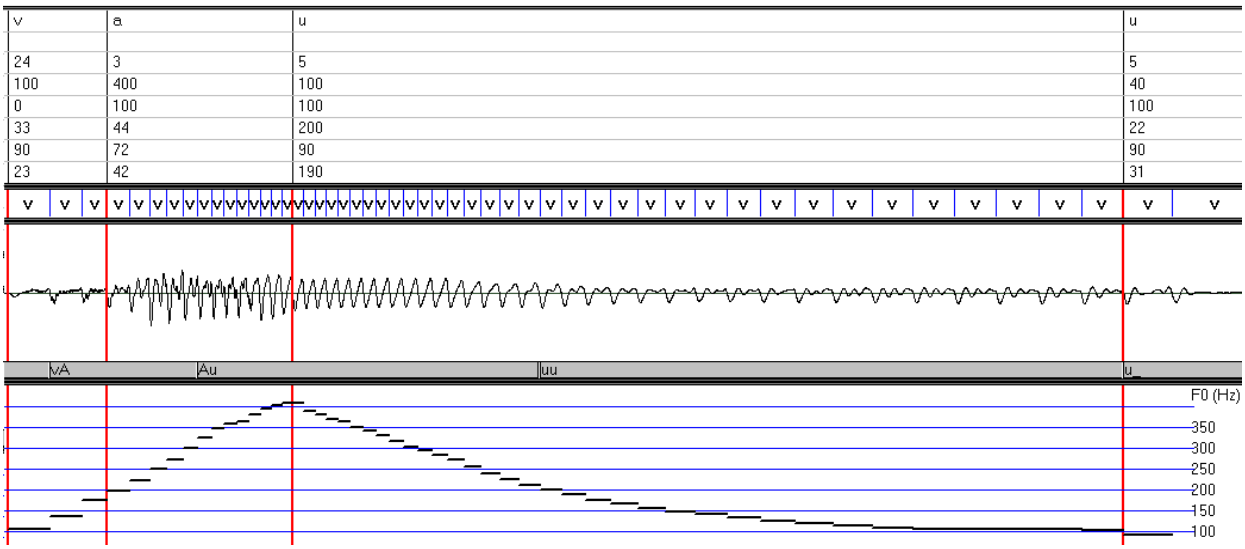

Figure 2. Sample sound based on the *vau* text input and modified to reach a characteristic sound

be synthesized); word level pitch (the Fo change is defined in relative scale); position in sound (defines the beginning point of the Fo change inside the sound); duration (defines the length of the sound in %); amplitude (the sound intensity is defined in %), the row of sound length gives the physical duration of the sound in ms. The absolute values of duration can be seen on the waveform and the absolute pitch values can be determined from the default Fo value that is defined in the settings of the TTS system.

3. GENERATING SOUND SAMPLES

Using the TTS and the modification tool, anyone can generate speech-like sounds. The final form of the sound can be defined by changing the numbers in the data matrix. The result can be heard and seen immediately. Thus interactive feedback assists in forming the desired sound character. The process is flexible, the desired sound character can be formed quickly. This solution allows immediate comparison among samples and the change of the parameters if needed (reversibility). Easy fine-tuning of the sound is also ensured. The visual feedback shows the result of the change immediately. Let us follow the generation procedure of a simple sound [vau:] step by step according to the example in Figure 2:

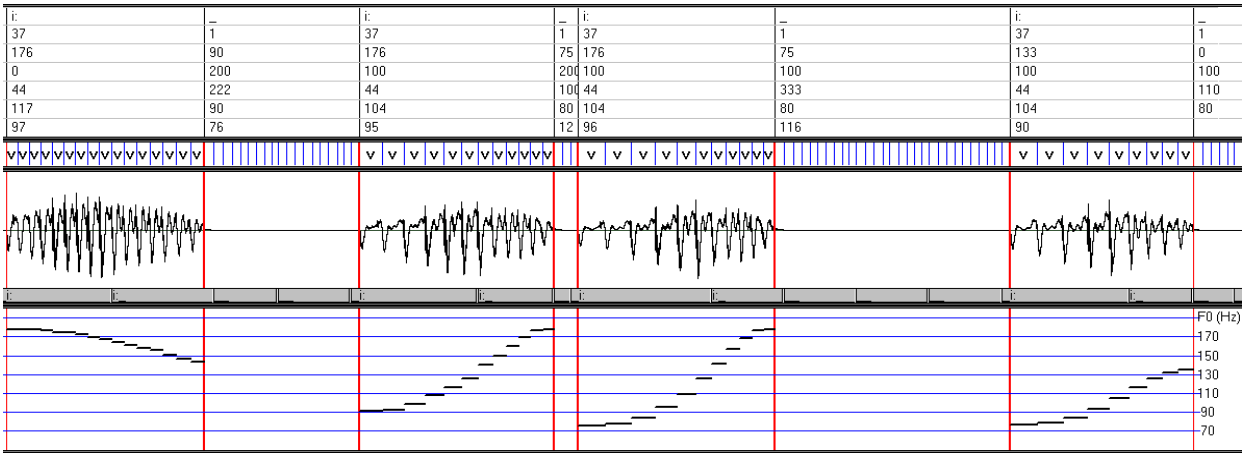

Figure 3. A sound sequence having four [i:] vowels (97, 95, 96, 90 ms), and three pauses in different lengths (76, 12 and 116 ms) and a dynamic Fo change structure

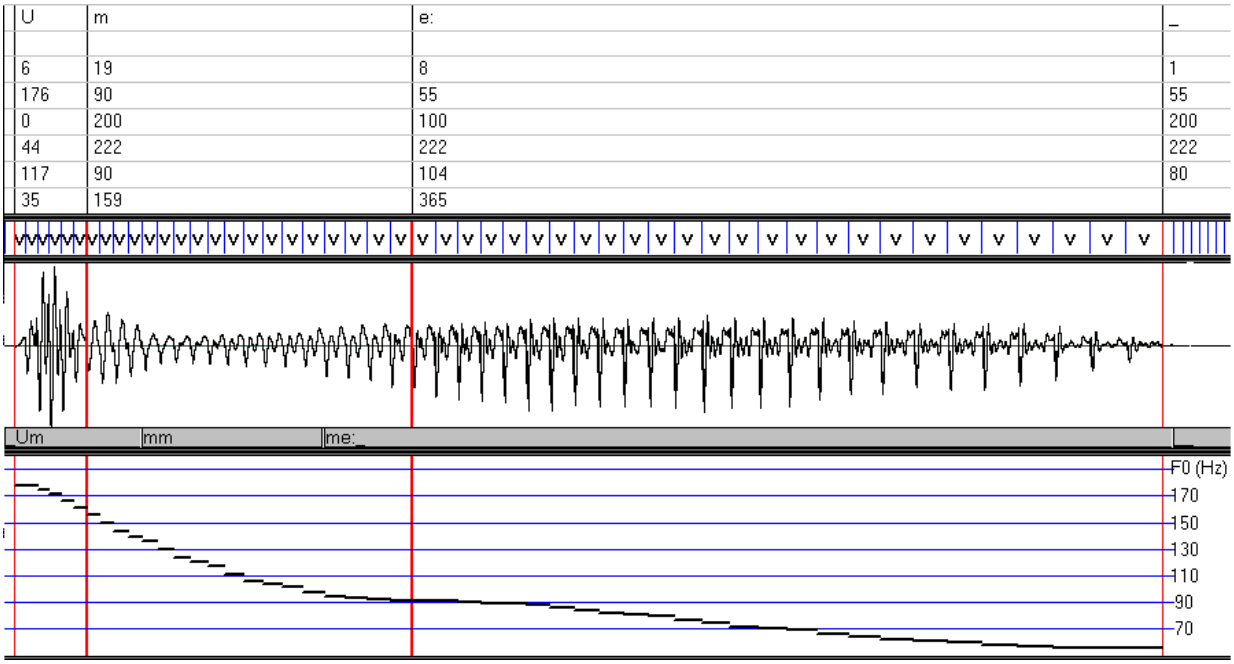

Figure 4: Use of a nasal consonant in intervocalic position. The sound sequence is [ym:e:]

(1) some text input is defined, let it be a CVV sequence *vau*. The synthesis results appear on the screen: the data matrix; the waveform and the Fo change.

(2) the Fo form is defined, having a dynamically increasing part (starting from 100 Hz) at the beginning of the sequence, and a slowly decreasing character from the beginning of [u] until the end. The peak value at the end of the rising part will be high (400 Hz).

(3) The definition of the sound durations will finalize the sound sequence.

Thus the following features will form the final sound: the character of the vowels, the speed of pitch change (quick and slow), the range of pitch change (300 Hz), and the length of the pitch change sequences, ie. the sound durations. If we use pauses, a rhythm structure can be added to the planned sound sequence. A sample sound for this is shown in Figure 3.

The variation of sounds can be escalated by using consonants as well. Nasal sounds have unique properties.

Using fricatives may make the sound more rich. An example is shown in Figure 4 for the use of a nasal consonant in intervocalic position.

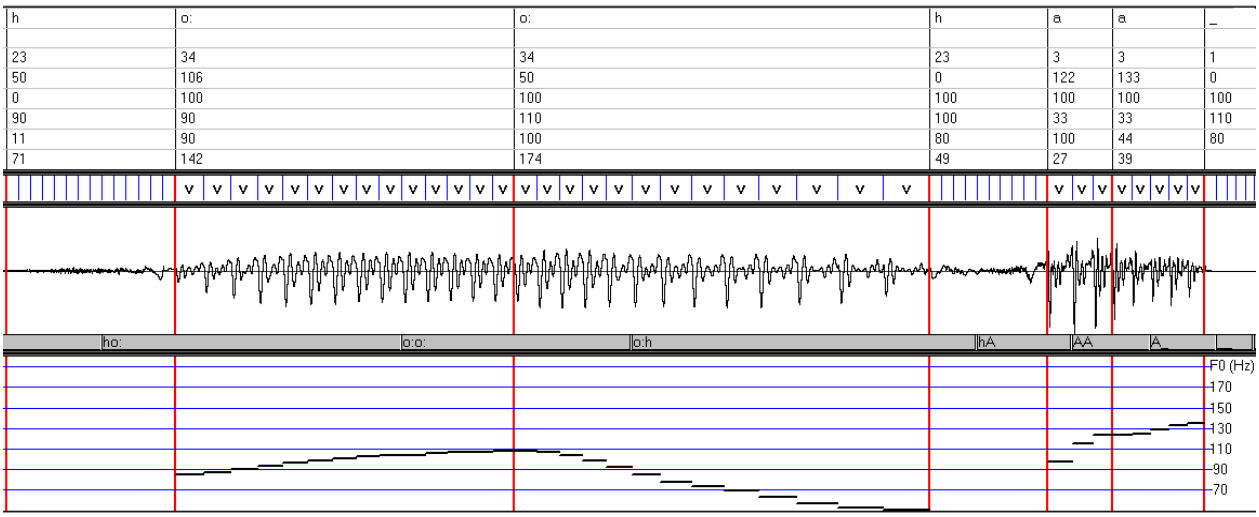

Figure 5: Use of a fricative in a CVCV sequence. The sounds are [ho:ha]

The generated sound sequence is [ym:e:]. Here the first vowel is short and the pitch is rising, the nasal sound is much longer than in real speech (159 ms) and the pitch is decreasing continuously. The final vowel is extremely long (365 ms), having a further decreasing Fo.

This time structure and the Fo change results in an expressive acoustic icon. The sample sound [ho:ha] is shown as an example of using fricatives (Figure 5.). Using all parameter possibilities described above the authors formed 44 sound samples, which represent the sound inventory for the selection of tpossible spemoticons. Further samples can be generated anytime.

4. PERCEPTUAL TEST FOR SELECTING SPEMOTICONS

To perform the test emotional and situation categories were defined. The test subjects had to classify a sound to one of the following seven spemoticon categories. Depending on the application other categories may also be defined.

1. Continue, I like it, Please repeat it.
   (positive emotion)
2. I enjoy it, I agree, this is good.
   (positive emotion)
3. This is not good for me, do not do it, I do not like it. This bothers me!
   (negative action)
4. I am angry, I hate you. This bothers me.
   (conflict, negative situation)
5. I am sad! I am not in good mood.
   (bad mood and its consequence)
6. Attention, Look!
   (warning, anxiety)
7. Congratulations, This is success!
   (positive evaluation, commendation)

4.1. The perceptual test

A web based listening test was conducted to get information about the perceived meaning of the sound samples of the sound inventory. 54 native speakers of Hungarian with normal hearing participated in the test. The listening test took 15 minutes to complete, on average. The task of the subjects was to assign a spemoticon category from the seven options listed above to each sound. The test was self-paced. The listeners had the option to replay a stimulus as many times as they wished, but they were not allowed to go back to a preceding stimulus, once they rated it. The analysis of the number of positive answers for a given spemoticon category it was determined that a sound sample may be regarded a spemoticon or not.

4.2. Results

The number of answers for the 7 emotional categories was collected automatically. In those cases, where many sound samples were assigned to the same category the result showed that this emotional category cannot be characterized by the sound inventory of this test (it should be noted that new, different sound samples may be created). Spemoticon test category no. 3 fell in this class. The corresponding diagram is shown in Figure 6. In those cases where only some of the sound samples were assigned to a spemoticon category, it is supposed that the sounds expressed that situation. This was the case for spemoticon test category no.1 (Figure 7.), no. 7. (Figure 8.) and no. 5. (Figure 9.).

Alltogether 9 sound samples were found to be spemoticons from the 44 element sound inventory. For negative categories no. 4. and 5. four different spemoticons were assigned. For positive categories no. 1., 2. and 7. five spemoticons were defined.

5. DISCUSSION

The results of the perceptual test clearly demonstrate that the proposed method is suitable for generating acoustic cues that are easy to identify even by untrained subjects in a relatively large space of forced choice (seven options). The distribution of the answers show that some elements of the 44 unit sound inventory may be used to identify some of the seven spemoticon categories.

The current study aimed at verifying the concept. In the future a larger sound inventory is planned. It will be based on the analysis of the parameters of the selected spemoticons. This analysis may also reveal correspondence between objective parameters and perceived emotional and contextual settings. It may also provide targets for generating new spemoticons.

The current study was conducted within the framework of the BelAmi project (http://www.belami-project.hu/) in a planned ambient assisted living scenario. The work is also related to the ETOCOM project in a robotic and in a mobile phone (http://cis.coginfo.sztaki.hu/etocom/project_aims.html) emotional interface context. The concept is to be tested in intelligent transport applications as well. The generated spemoticon inventory has been provided for the above mentioned scenarios for application integration and further testing.

6. SUMMARY

In this paper a new solution for generating emotional and intentional auditory cues was presented. It is based on TTS technology and allows strict control over the creation and modification of TTS-based speech emoticons, spemoticons. In order to evaluate the concept a sound inventory of 44 sound samples was generated.

A subjective perception test was completed in order to evaluate the sound samples related to the representation of seven emotional and situation categories. Those samples that had been allocated to different categories were discarded. The remaining 9 sounds that could be well associated with one of the seven categories are regarded as *spemoticons*. The readers are welcome to complete the open perceptual test at:

http://hungarianspeech.tmit.bme.hu/spemoticon/test .

Both the spemoticons themselves and the emotional and situation categories may be application and language dependent. In the future we intend to extend our studies in both

directions and we look forward to co-operation with other researchers of the field.

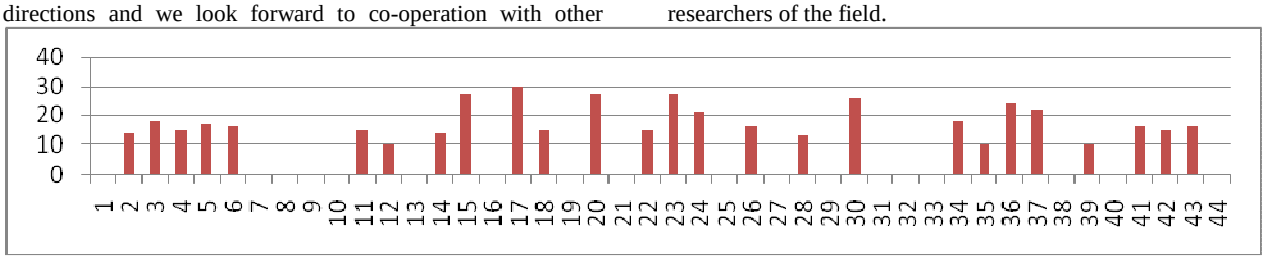

Figure 6. The distribution of the answers for the spemoticon category no 3. (This is not good for me, do not do it, I do not like it. This bothers me!)

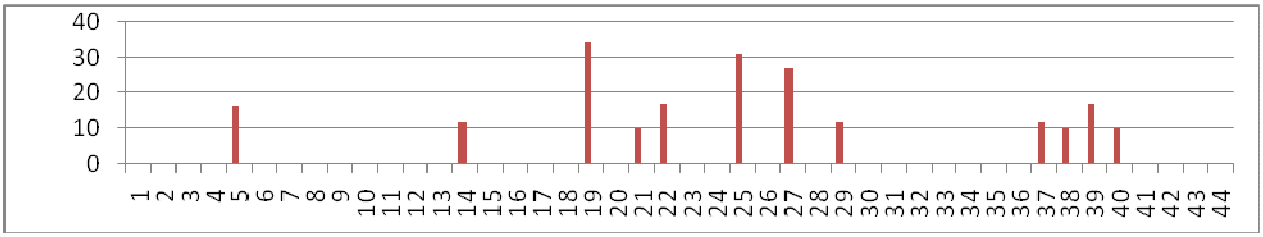

Figure 7. Answer distribution for the positive category no. 1. (Continue, I like it, Please repeat it). The spemoticons are samples no. 19, 25 and 27

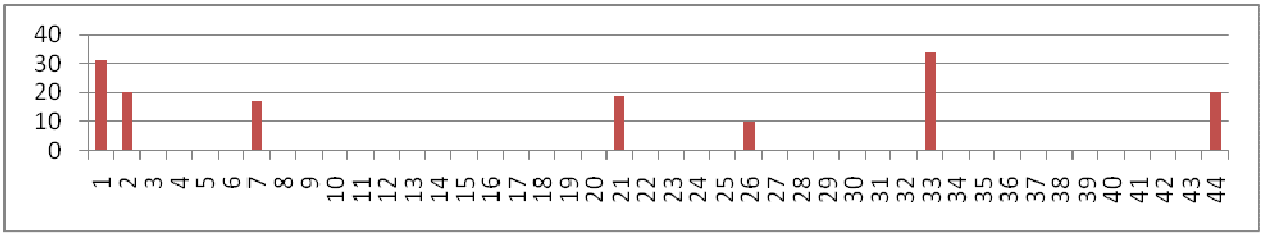

Figure 8. Answer distribution for the positive category no. 7. (Congratulations, This is success!). The spemoticons are samples no. 1 and 33

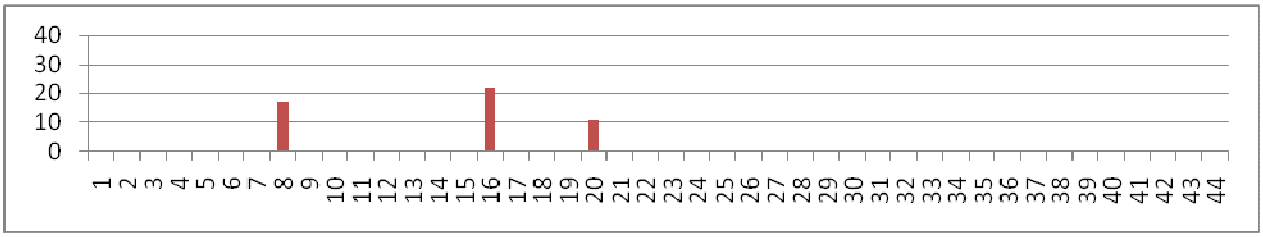

Figure 9. Answer distribution for the negative category no. 5. (I am sad! I am not in good mood). The spemoticon is sample no. 16

7. ACKNOWLEDGEMENT

The authors acknowledge the contribution of Mátyás Bartalis in formatting the text and figures. This research was supported by the BelAmi: ALAP2-00004/2005, the ETOCOM: TÁMOP-4.2.2-08/1/KMR-2008-0007 and the TÁMOP-4.2.1/B-09/1/KMR-2010-0002 projects.

8. REFERENCES

[1] Gaver, W. W. (1986). Auditory icons: Using sound in computer interfaces. Human-Computer Interaction. 2, 167 - 177.

[2] Gaver, W. W. (1988). Everyday listening and auditory icons. Doctoral Dissertation, University of California, San Diego.

[3] Mynatt, E. D. (1994), Designing with Auditory Icons, Proc. ICAD94, pp. 109-119.

[4] Gaver, W.W. (1991). Sound support for collaboration. Proceedings of the Second European Conference on Computer-Supported Collaborative Work.

[5] Skantze, D., Dahlbäck, N. (2003) Auditory Icon support for navigation in speech-only interfaces for room-based design metaphors, Proceedings of ICAD 2003, the International Conference on Auditory Display, 2003 pp. 140-143.

[6] Garzonis, S., Jones, S., Jay, T., O'Neill, E., (2009) Auditory icon and earcon mobile service notifications: intuitiveness, learnability, memorability and preference, Proceedings of the 27th international conference on Human factors in computing systems, Pages: 1513-1522

[7] Jeon, M., & Walker, B. N. (2009). "Spindex": Accelerated Initial Speech Sounds Improve Navigation Performance in Auditory Menus. Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society (HFES2009), San Antonio, TX (19-23 October).

[8] Wersényi, Gy., Auditory Representations of a Graphical User Interface for a Better Human-Computer Interaction. in S. Ystad et al. (Eds.): Auditory Display. CMMR/ICAD 2009 post proceedings edition, Lecture Notes in Computer Science (LNCS) 5954, Springer Verlag, Berlin, 2010. pp. 80-102.

[9] Olaszy Gábor – Németh Géza – Olaszi Péter – Kiss Géza – Zainkó Csaba – Gordos Géza 2000. Profivox - a Hungarian TTS system for telecommunications applications. International Journal of Speech Technology. Vol 3-4. Kluwer Academic Publishers. 2000. 201–215.