Multimodal Spontaneous Expressive Speech Corpus for Hungarian

Márk Fék[1], Nicolas Audibert[2], János Szabó[1], Albert Rilliard[3], Géza Németh[1], Véronique Aubergé[2]

1: Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Magyar tudósok körútja 2., 1117 Budapest, HUNGARY

2: ICP/Gipsa-lab, UMR CNRS 5009 INPG-Université Stendhal 1180, avenue Centrale - BP 25, 38040 Grenoble, FRANCE

3: LIMSI-CNRS, B.P. 133, F-91403 ORSAY CEDEX, FRANCE

E-mail: fek@tmit.bme.hu, Nicolas.Audibert@gipsa-lab.inpg.fr, szabo.janos@eestec.hu, rilliard@limsi.fr, nemeth@tmit.bme.hu, Veronique.Auberge@icp.inpg.fr

Abstract

A Hungarian multimodal spontaneous expressive speech corpus was recorded following the methodology of a similar French corpus. The method relied on a Wizard of Oz scenario-based induction of varying affective states. The subjects were interacting with a supposedly voice-recognition driven computer application using simple command words. Audio and video signals were captured for the 7 recorded subjects. After the experiment, the subjects watched the video recording of their session and labelled the recorded corpus themselves, freely describing the evolution of their affective states. The obtained labels were later classified into one of the following broad emotional categories: satisfaction, dislike, stress, or other. A listening test was performed by 25 naïve listeners in order to validate the category labels originating from the self-labelling. For 52 of the 149 stimuli, listeners' judgements of the emotional content were in agreement with the labels. The result of the listening test was compared with an earlier test validating a part of the French corpus. While the French test had a higher success ratio, validating the labels of 79 tested stimuli, out of the 193, the stimuli validated by the two tests can form the basis of cross linguistic comparison experiments.

1. Introduction

Recent experiments on expressive speech synthesis and recognition for Hungarian (Zainkó et al., 2007; Tóth et al., 2007; Fék et al., 2005; Zainkó et al., 2006) used solely acted speech corpora. The use of acted speech allows great control over the type of emotions and the content of the utterances in the recorded speech corpus. Acted speech, on the other hand, can result in less authentic (e.g. exaggerated, overacted) expressions, unless the actor is using the technique of elicitation (Enos, 2006). Even in the latter case, acted speech may not rely on the same set of cues as spontaneous expressive speech. Therefore it is important to collect spontaneous data for a refined analysis of the cues conveying emotions and affect in speech.

Collecting spontaneous expressive speech data is difficult because of the lack of control over the quality of recordings, the emotions expressed, and the content of the utterances. To tackle these problems, a methodology for spontaneous expressive corpus collection was developed for the recording of a French corpus (Aubergé et al., 2004). The method relies on a Wizard of Oz scenario-based induction of varying affective states, designed and set up in lab using the dedicated software platform E-Wiz[1].

This paper presents a Hungarian multimodal spontaneous expressive speech corpus, recorded following the framework of the French E-Wiz corpus. As the recorded data is parallel to the French corpus, cross-cultural analyses can be performed in the future. In the following, we summarize the method of corpus collection for the French data, then, we describe the modifications of the setup for Hungarian recordings. A listening test was performed to assess the emotions expressed in the collected speech material. We evaluate its results and compare them with the results of an earlier French listening test.

2. Collection of the French Corpus

The method used for the recording of the French corpus relies on a Wizard of Oz scenario-based induction of varying affective states, designed and set up in lab using the dedicated software platform E-Wiz. Considering the expressions of affects in speech as shared between the 3 levels of (1) direct emotional expressions as a consequence of emotion-related physiological changes, (2) attitudinal expressions of social affects and (3) linguistic expressiveness (Aubergé, 2002), this corpus aimed at collecting more specifically expressions of direct emotions. This was achieved by allowing the recorded subjects to interact only with a computer application supposed to be voice-recognition driven, without the presence of anyone else in order to prevent subjects from producing social expressions of affects. Moreover the values of linguistic prosodic functions were also frozen by restraining the interaction in a major part of the corpus to a strict command language composed of French monosyllabic colour names ([bʁik], [ʒon], [ʁuʒ], [sabl] and [vɛʁ]), to which was added the command [paʒsʏivãt] (next page). Another goal of the use of such a command language was to make possible further perceptive experimentations using the standard content paradigm and acoustic analyzes on standardized utterances.

The software was presented as a futuristic phonetic-learning tool called Sound Teacher and based on the perception-action theory findings that would enable subjects to easily produce vowels of foreign languages

[1] http://www.icp.inpg.fr/EMOTION/EWiz/

using their brain plasticity and needed to be further tested prior to its commercialization, and was indeed manipulated by the experimenter. After a training phase in which subjects could get familiar with phonetic characters and with the layout of the interface, most of the tasks consisted of discriminating a pair of synthetic audio stimuli visually highlighted in the acoustic space by telling the corresponding colour name. In addition of that, production tasks in which the subject had to pronounce vowels were performed at the beginning of each phase of the scenario, and subjects were regularly requested to give free comments. In order to induce positive emotional states to the subjects, their performances were presented in the first, deliberately easy, training phase as better than previous mean performances, and in the second phase as within the best 3 performances ever realized, allowing the subject to skip directly to the last and more complex phase. Negative induction was then performed by attributing very bad scores to the subjects, telling them that their skills for the perception of French might have been damaged by the software and had to be checked again. In this latter phase the synthetic audio stimuli were tuned to force subjects to answer randomly and give credit to the claim that their skills for perceiving the vowels of their mother tongue could have been perturbed.

Audio, video and bio-physiological signals were captured and synchronized for the 17 recorded subjects, including 7 actors who were asked to reproduce the utterances of the command language using their actors' skills immediately after the recording. In addition of that the articulatory signal (EGG[2]) was recorded for 2 subjects. Each recording session lasted around 45 minutes, the collected speech material in the spontaneous part consisted of 195 command language stimuli per speaker in average, as well as free comments.

The corpus was labelled by the recorded subjects themselves shortly after the recording session, in order to take advantage of the autobiographic memory accuracy for the recollection of emotional episodes (Aubergé et al., 2006). This labelling task was performed using the video recording and a spreadsheet with pre-filled times, but without additional instructions given to the subjects in order to let them freely describe the evolution of their affective state. Analysis of labels collected from the French corpus revealed that subjects tended to label expressions rather than felt emotions when they were not congruent with each other. Though reactions to the same induction differed according to subjects' psychological profile, broad classes of reactions to the induction could be found, including satisfaction/amusement, stress/anxiety, irritation and boredom.

3. Collection of the Hungarian Corpus

Linguistic adaptation of the E-Wiz platform was necessary before collecting data for the Hungarian corpus. Most Hungarian colour names consist of more than one syllable, therefore we have chosen monosyllabic words as command words, whose meanings (e.g. grass, snow) could be easily associated with colors. The following words were used: [la:ng], [ho:], [fy:], [vɒj] and [ke:k], along with the [køvɛtkɛzø:oldɒl] (next page) command. Further, it was necessary to adapt the screen of the vowel recognition task to the Hungarian vowel system (Figure 1). An initial test showed that it was easy to forget the Hungarian command words, therefore we have explicitly indicated them on screen during each task.

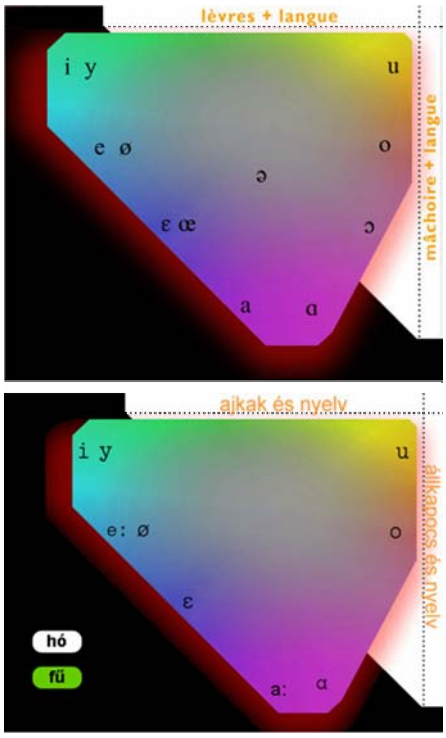

Figure 1. Screens showing the vowel recognition task in the French (top) and the Hungarian (bottom) version of E-Wiz.

The recording was performed in a soundproof room. The audio signal was captured by a condenser microphone mounted before the subject, and digitized at 44 kHz, using 16 bits per sample. The video signal was captured by an analogue Hi-8 camcorder and was digitized with a resolution of 384x288 pixels. In contrast to the French setup, the camera was hidden, a wider zoom angle was applied, and neither special lighting nor blue background was used (see Figure 2). The reason behind this decision was that the layout of the Hungarian soundproof room did not enable the filming of subjects under similar conditions.

The data was collected from 7 (1 male, 6 female) subjects. All of them were students of a phonetics course.

[2] EGG: Electroglottograph, a non-invasive device of measuring vocal fold contact during voicing without affecting speech production.

Figure 2. Video snapshots from the French (top) and Hungarian (bottom) database.

Table 1 shows the mean number of utterances (per subject) collected in the French and Hungarian experiments. The Hungarian corpus contains less monosyllabic command words, because the reduced number of vowels (compared to the French vocalic system) allowed for fewer vowel discrimination tasks. The Hungarian subjects were less active in giving comments compared to the French subjects, which might be explained by their younger age and by cultural differences.

Similarly to the French corpus, the recorded subjects labelled themselves the recorded data. An initial analysis showed that the broad classes of reactions to the induction are similar to that found in the French corpus: satisfaction/amusement, stress/anxiety, irritation and boredom

4. Listening Test

We performed a perceptual test to assess the emotions expressed in the collected speech data and to create a subset in which emotions perceived by naïve listeners match the self-labelling.

The retained subset is meant to be used in a cross-linguistic evaluation together with French stimuli. Therefore the experimental design of the listening test for Hungarian data follows as much as possible the design of an earlier experiment evaluating the French data as a part of (Laukka et al., 2007).

The methodology of the two tests differed in some respects, however, which makes the direct statistic comparison of the results infeasible. We will report individual analyses and discuss the possible reasons explaining the differences in the outcome of the two experiments.

.

Per subject mean values	**French**	**Hungarian**
# of 'next page' utterances	54	54
#of monosyllabic command words	141	98
# of vowels	35	34
# of comments	20	7

Table 1. Mean number of utterances per subject in the French and Hungarian databases

The French test used the productions of 6 (3 male and 3 female) of the 17 recorded speakers. All 6 speakers were actors who were asked, immediately after having participated in the Wizard of Oz experiment, to portray expressions of the same affects they had just been feeling. In addition, the actors were asked to portray expressions of six basic emotions still using the same utterances (French colour names and "next page"). Both the acted and the spontaneous stimuli were included in the listening test, but we will only report here the results obtained for the non-acted stimuli.

The self-reports of the 6 subjects were classified by the authors into four broad categories: anger, fear, joy, and other. These categories were intended to be suitable for both spontaneous and acted stimuli, of which the acted stimuli were supposed to express stronger emotions. A subsequent expert listening was performed by the authors where a set of 246 stimuli was selected consisting of 146 spontaneous and 100 acted stimuli. The goal of this selection was to retain a number of stimuli that can be considered as good exemplars of each emotion class, with as much variation as possible in activation level and prosodic morphology.

The selected 246 stimuli were evaluated by 15 French speaking naïve listeners (8 male and 7 female, mean age = 29.8 years), who were allowed to listen to the stimuli as many times as they wanted and had to select one of 3 emotion classes: anger, fear, or joy, or an "other emotion" label. The listeners received a detailed explanation on how they should interpret the broad classes of anger, fear and joy. Examples of emotions that should be considered to belong to the broad classes were also given. For example, they were told that "fear" could designate any phenomenon ranging from anxiety or stress to terror or panic. Moreover, the listeners were asked to rate the perceived emotion intensity of each stimulus using a scale ranging from 0 (very weak intensity) to 10 (very strong intensity). Stimuli were presented in a different

random order to each listener. The presentation and the recording of the answers were automated using a graphical interface.

The Hungarian listening test included recordings from all the 7 subjects (1 male, 6 female) participating in the experiment. The self-reports given by the subjects were not as easily interpretable as the French ones. The reports of three subjects were partially incomplete, and one of the subjects misinterpreted the task and mostly reported his thoughts instead of his feelings. Moreover, some of the reports included ambiguous descriptions of emotional states.

The quality difference of the French and Hungarian self-reports may be explained by the younger age of the Hungarian participants compared to the French ones, and by the supposed expert knowledge of the French actors regarding their own emotions. The accuracy difference between self-reports given by actors vs. non-actors was not systematically evaluated, and the number of recorded speakers does not allow us to draw strong conclusions. However, the majority of actors recorded in the French corpus seem to have indeed provided self-labels easier to interpret than non-actors.

In order to make a first selection of a reasonable number of stimuli to be included in the Hungarian listening test, a pre-test was performed by 4 listeners who volunteered to evaluate all the 962 stimuli consisting of monosyllabic words and 'next page' utterances, independently of self-reports. The stimuli were presented in a different random order to each listener. The listeners had to select one of the following labels best describing the perceived emotion for each stimulus: satisfaction, dislike, stress, feeling low, boredom, determination, uncertainty, or "other emotion". During the subsequent evaluation "boredom" answers were incorporated into the "dislike" category, while "uncertainty" answers were merged into the "stress" category. Next, stimuli for which the answers of at least 3 listeners were in agreement were selected, and they were labelled according to the selected category. Finally, 149 stimuli labelled as "satisfaction", "dislike" or "stress" were retained for the subsequent test.

The 149 pre-selected stimuli were evaluated by 25 Hungarian native speakers (8 male and 17 female, mean age = 25.8 years) in a second listening test. The test was performed using the (translated) graphical interface applied in the French 15-listener test. The broad category labels used in the French test were changed here to "satisfaction", "dislike" and "stress", according to the labels used in the pre-selection. These labels were considered to better represent weak or mild expressions, that were expected to be found in the spontaneous stimuli, than the stronger labels used in the French test. It allowed us to give less complicated explanations to the listeners, decreasing the possibility of confusion. The listeners, similarly to the French test, were instructed to consider the labels as broad categories representing a range of emotions. The 10-point intensity scale was replaced by a 100-point (practically continuous) scale, with labels at 15, 50, and 85 indicating "mild", "medium", and "strong" respectively.

The self-reports corresponding to the 149 selected stimuli were carefully evaluated by the authors and classified into one of the following 5 categories: "satisfaction", "dislike", "stress", "other", or "no data". 103 of the 149 stimuli were labelled as satisfaction, dislike, or stress. Only these 103 stimuli are analyzed in the Results section.

5. Results

The purpose of the listening test was to retain a set of stimuli for which the self-evaluation provided by the subject matched the judgement of naïve listeners regarding the emotional content of the stimuli. The judgement of naïve listeners was determined as the emotional category having the maximum recognition ratio for the given stimulus. As a result, only stimuli recognized above chance level were retained in the final selections.

French	**joy**	**anger**	**fear**
pre-selected	46	53	47
validated	28	26	25
Hungarian	**satisf.**	**dislike**	**stress**
pre-selected	33	42	28
validated	27	13	12

Table 2. Distribution of self-evaluation labels among the broad emotion categories for the pre-selected and validated stimuli sets in case of both languages.

The validation of the self-evaluation was successful for 79 of the 146 tested spontaneous stimuli in the French experiment, while 52 of 103 stimuli were validated in the Hungarian experiment.

Table 2 shows the distribution of self-evaluation labels among the broad emotion categories for the pre-selected and validated stimuli sets in case of both languages.

	joy	**anger**	**fear**	**other**
joy	51,16%	18,12%	14,78%	15,94%
anger	14,84%	40,38%	22,89%	21,89%
fear	12,32%	18,12%	47,97%	21,59%

Table 3. Overall confusion among the identification of 146 French spontaneous stimuli.

First we report two confusion matrices (one per language) representing the confusion among the self-evaluation and the emotions recognized by the listeners. The matrices were calculated for all the validated and non validated stimuli. Next we report two additional confusion matrices calculated over the validated stimuli only. Finally we compare the intensity histograms of the retained stimuli.

Table 3 shows the overall confusion matrix obtained for the French experiment. In all confusion matrices, the rows indicate the labels obtained from self-evaluation and the columns indicate the emotion perceived by the listeners. The analysis was performed on all the 146

spontaneous stimuli. All three categories were recognized with moderate confusions. The "joy" and "fear" categories are the best separated with the minimum confusion between the two.

Table 4 shows the overall confusion matrix obtained for the Hungarian experiment. The analysis was performed on the 103 stimuli for which the self-evaluation was unambiguously interpretable as related to satisfaction, stress or dislike. The "satisfaction" and "stress" categories are the best separated, which is in accord with the French results where "joy" and "fear" were the best separated categories. There is an important confusion, however, for stimuli reported as "dislike". The confusion is equally important with both the "satisfaction" and "stress" categories. Further, stimuli reported as "stress" were recognized as "dislike" in a number of cases.

	satisf.	**dislike**	**stress**	**other**
satisf.	63.27%	17.82%	10.79%	8.12%
dislike	30.67%	29.43%	25.90%	14.00%
stress	11.00%	30.86%	38.86%	19.29%

Table 4. Overall confusion among the identification of 103 Hungarian stimuli.

Table 5 shows the confusion matrix obtained for the 79 stimuli validated in the French experiment. The recognition ratios are improved over the previous results, due to the selection criterion applied. The "joy" and "fear" categories remained the best separated ones.

	joy	**anger**	**fear**	**other**
joy	72.38%	8.33%	5.71%	13.57%
anger	10.00%	65.13%	7.95%	16.92%
fear	6.67%	5.87%	74.13%	13.33%

Table 5. Confusion among the identification of the 79 validated French stimuli.

Table 6 shows the confusion matrix obtained for the 52 stimuli validated in the Hungarian experiment. The identification ratios also improved much compared to the results for all the 103 stimuli, but an important confusion remained among the dislike and stress categories.

	satisf.	**dislike**	**stress**	**other**
satisf.	74.81%	9.78%	7.85%	7.56%
dislike	6.77%	56.00%	22.77%	14.46%
stress	5.00%	19.00%	56.67%	19.33%

Table 6. Confusion among the identification of the 52 validated Hungarian stimuli.

We have also compared how the listeners rated the emotional intensity in the French and Hungarian experiments. Figure 3 shows the intensity histograms obtained for all the stimuli in both experiments. The across-listener mean intensity scores were quantized to 10 equidistant values in both cases. The French results were multiplied by 10 to equate the two scales. The histograms show the number of stimuli in each quantization bin. The overall mean intensity values (French: 34, Hungarian: 45) calculated for all stimuli differ in the two experiments. The mean intensity of non-acted stimuli in the French experiment was probably lowered because the listeners rated the spontaneous stimuli along with more intense acted stimuli. Apart from the difference in the mean value, the profiles of the two histograms are remarkably similar.

The mean perceived intensities for each emotion category were also calculated. Figure 4 compares the obtained values for both languages. Stimuli judged as satisfaction were rated as relatively more intense in the Hungarian experiment when compared to the French experiment. This is probably due to a speaker showing more confidence in the Hungarian experiment, whose utterances expressing satisfaction were rated as more intense than those of other speakers (a mean value of 58.0 for this speaker vs. 41.6 for the others).

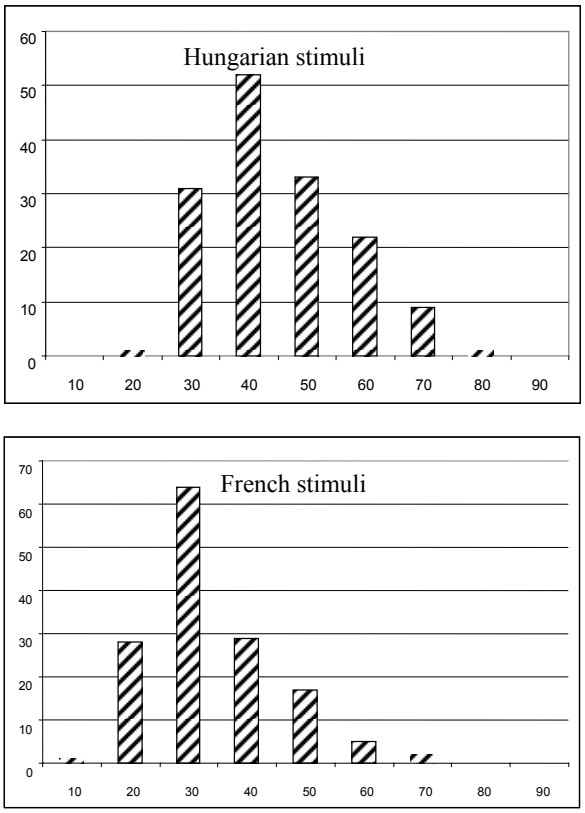

Figure 3. Intensity histograms obtained for all the Hungarian (top) and French (bottom) stimuli presented during the listening tests. The vertical axis shows the number of stimuli, the horizontal axis shows the quantized (across-listener) mean intensity value. The intensity values of the French data were multiplied by 10 to match the scale of the Hungarian data.

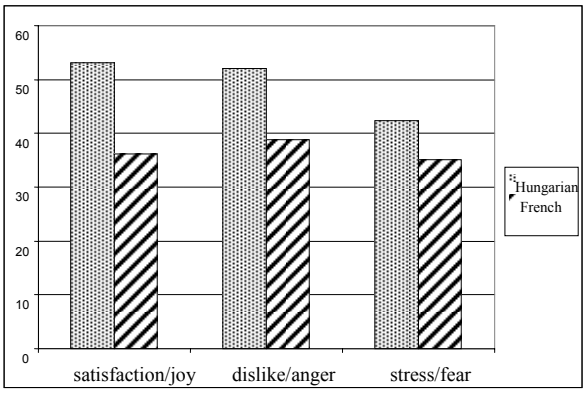

Figure 4. Mean perceived intensity per emotion category. The means were calculated over the stimuli validated by the experiments and they were labelled according to the self-evaluation. The intensity values of the French data were multiplied by 10 to match the scale of the Hungarian data.

6. Discussion

The shortcomings of the Hungarian self-evaluations can in part explain why the French experiment produced more validated stimuli. We examined the non-validated stimuli in an informal post-test listening to seek for further explanations. We have found, that although utterances labelled as ”dislike” by a particular speaker seemed to be easily recognizable when compared to the expressions of satisfaction by the same speaker, those utterances were largely perceived as dislike when presented alone. We have also noticed that some of the Hungarian stimuli which were labelled as “dislike” (or “stress”) based on the self-reports but were recognized as “stress” (or “dislike”) could be considered as acceptable expressions for both emotion categories. This might be an explanation to the weaker perceptual separation between these two categories, and therefore to the lesser number of stimuli expressing stress or dislike validated in the Hungarian data.

7. Conclusion

This paper presented a Hungarian multimodal spontaneous expressive speech corpus containing speech and video data. The corpus was recorded following the methodology of a similar French corpus, using a Wizard-of-Oz scenario based induction of affective states.

The collected speech material contains vowels, monosyllabic words, “next page” utterances, and longer comments from 7 subjects in several versions. The subjects labelled the data themselves describing the evolution of their affective states during the recording. Most of the expressed emotions fell into the following broad classes: satisfaction/amusement, stress/anxiety, irritation and boredom.

A 25-person listening test was performed to validate a subset of the recorded material. Listener judgements validated the self-evaluation for a set of 52 stimuli. The validated set contains expressions falling into the broad classes of satisfaction, dislike or stress. Moreover, this stimuli set is suitable for cross-linguistic comparisons with a similarly established set of 79 validated French stimuli.

8. Acknowledgements

We express our gratitude to all the subjects and listeners who participated in the experiments.

The research was partially supported by the Franco-Hungarian bilateral cooperation programme Balaton.

9. References

Aubergé, V. (2002). A Gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP, 1st Speech Prosody, Aix-en-Provence, France, pp. 151-154.

Aubergé, V., Audibert, N., Rilliard, A. (2004). E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. 4th LREC, Lisbon, Portugal, pp. 179-182.

Aubergé, V., Audibert, N. & Rilliard, A. (2006). Auto-annotation: an alternative method to label expressive corpora. LREC Workshop “Corpora for research on emotion and affect”, Genova, Italy, pp. 45-46.

Enos, F. & Hirschberg, J. (2006). A Framework for Eliciting Emotional Speech: Capitalizing on the Actor’s Process. LREC Workshop “Corpora for research on emotion and affect”, Genova, Italy, pp. 6-10.

Fék, M., Szabó, J., Olaszy, G., Németh, G., Gordos, G. (2005). Érzelem kifejezése gépi beszéddel (Expressing Emotions by Machine Generated Speech). Beszédkutatás 2005, pp.134-143. (In Hungarian)

Laukka, P., Audibert, N., Aubergé, V. (2007). Graded structure in vocal expression of emotion: What is meant by "prototypical expressions"? In Proceedings of Paralinguistic Speech - between models and data. Saarbrücken, Germany, 6-10 August 2007. pp. 1-4.

Tóth, Sz. L., Sztahó, D., Vicsi, K. (2007). Emotion perception by human and machine. COST2102 International Conference on Nonverbal Features of Human-Human and Human-Machine Interaction, Patras, Greece 29-31 October 2007. pp. 223-236.

Zainkó, Cs., Fék M. (2006). Beszédadatbázis prozódiájának szerepe a gépi beszéd hangzásában és érzelmi tartalmak kifejezésében – Vidám avagy szomorú a beszédszintetizátor? (Influence of the speech database prosody on synthetic speech – is the TTS happy or sad?). Beszédkutatás 2006, pp.208-217. (In Hungarian)

Zainkó, Cs., Fék M., Németh, G. (2007). Expressive Speech Synthesis Using Emotion-Specific Speech Inventories. COST2102 International Conference on Nonverbal Features of Human-Human and Human-Machine Interaction, Patras, Greece 29-31 October 2007. pp. 237-246.