Computing and Informatics, Vol. 33, 2014, 1312–1332

EXPRESSIVE SPEECH SYNTHESIS FOR CRITICAL SITUATIONS

Milan Rusko, Sakhia Darjaa, Marián Trnka
Róbert Sabo, Marian Ritomský

Institute of Informatics
Slovak Academy of Sciences
Dúbravská cesta 9
845 07 Bratislava, Slovakia
e-mail: `Milan.Rusko@savba.sk, Trnka@savba.sk`

Abstract. Presence of appropriate acoustic cues of affective features in the synthesized speech can be a prerequisite for the proper evaluation of the semantic content by the message recipient. In the recent work the authors have focused on the research of expressive speech synthesis capable of generating naturally sounding synthetic speech at various levels of arousal. Automatic information and warning systems can be used to inform, warn, instruct and navigate people in dangerous, critical situations, and increase the effectiveness of crisis management and rescue operations. One of the activities in the frame of the EU SF project CRISIS was called "Extremely expressive (hyper-expressive) speech synthesis for urgent warning messages generation". It was aimed at research and development of speech synthesizers with high naturalness and intelligibility capable of generating messages with various expressive loads. The synthesizers will be applicable to generate public alert and warning messages in case of fires, floods, state security threats, etc. Early warning in relation to the situations mentioned above can be made thanks to fire and flood spread forecasting; modeling thereof is covered by other activities of the CRISIS project. The most important part needed for the synthesizer building is the expressive speech database. An original method is proposed to create such a database. The current version of the expressive speech database is introduced and first experiments with expressive synthesizers developed with this database are presented and discussed.

Keywords: Crisis management, warning speech synthesis, soothing speech synthesis, acoustic cues of arousal, expressive speech database

Mathematics Subject Classification 2010: 68T10, 62-07

1 INTRODUCTION

Speech communication is the most common and most effective natural information transfer means used by humans. Thus, automatic information systems increasingly use interfaces allowing human speech communication between the user and the automatic system. However, for the time being artificial speech generating systems (synthesizers) are not able to generate speech with correct prosodic (supra-segmental, paralinguistic and extra-linguistic) features bearing additional information such as emotional content, urgency, warning or reassuring tone, etc. Information systems should be able to present this important information as well, for instance in emergency situations. It is known that investigators of some aircraft accidents believe that neutral and even reassuring mode of warning messages read by synthetic voice probably contributed to incorrect evaluation of the level of danger by the pilots.

Both psychologists and engineers categorize emotions expressed in speech in different manner and into different number of so-called basic emotions (most often 4 or 6). In our work, however, we do not want to model joy, sorrow, etc.; rather, we want to study and synthesize speech expressing warning and strong caution. We want to implement reading urgent messages or holding a reassuring dialogue.

Since almost all currently used speech synthesis methods apply large speech databases to acquire the data necessary for their development and activities (training and testing), design and subsequent creation (recording, annotation and processing) of specialized speech databases is necessary as the first step. Simultaneously, a new type of speech synthesizer is being developed, using special speech databases and modeling of features on the basis of Hidden Markov Models (HMMs) [1] with more accurate modeling of supra-segmental features.

2 EXPRESSIVE HMM-TTS FOR PUBLIC SAFETY

2.1 The Emergency Situations

In many emergency situations and collective crises of different scale the responsible management would need to make use of the information system equipped with expressive speech synthesizer capable of delivering automatic and updatable urgent messages to the needed places.

To mention just some of the critical situations, they range from everyday community emergencies, like traffic accidents, smaller fires, flooding, chemical, radioactive, or biological spills, medical emergencies, criminal activities, bomb threats, workplace violence, major power outages and many others, to disasters like large scale fires, catastrophic floods, ecological catastrophes, earthquakes, tsunami, technological catastrophes or even space catastrophes. Obviously, quantitatively different emergency situations require very different emergency management procedures.

2.2 The CRISIS Project

The European Structural Funds project "Research and Development of New Information Technologies for Prediction and Solution of Critical Situations of Inhabitants" – CRISIS (ITMS 26240220060) is aimed at predicting and solving critical situations when the inhabitants are endangered, and the environment has to be protected. The project is targeting:

- creating information systems that support grid technology to provide massive computational power to solve difficult tasks during the management of critical situations
- computer simulation of fires and its visualization
- mechatronic systems for critical situations
- nanotechnologies with special focus on sensors development
- development of safe communication and software platforms
- extremely expressive (hyper-expressive) speech synthesis for urgent warning messages generation.

2.3 The Role of Expressive TTS

Previous attempts at building synthesizers in Slovak were limited to emotionally neutral speech and used diphone concatenative, unit-selection [2, 3] and recently also statistical-parametrical [4, 5] speech synthesis. The present work is aimed at generating expressive synthetic speech. The synthesizer should be able to generate warning messages with different degrees of urgency, serious comments, read texts in neutral style and, last but not least, to generate soothing utterances and a very calm speech. From the point of view of affective phenomena, these speech styles reflect different levels of arousal.

The goal of the "Expressive speech synthesis" activity is to perform research and development of a system which would be capable of generating information system messages and dialogue system replies in naturally sounding speech with considerable content of paralinguistic and extra-linguistic information representing properties such as warning tone, urgency, but also soothing and reassuring speech tone. The application result is represented by a new type of expressive speech synthesizer using large speech databases and hidden Markov modeling.

The synthesizer is applicable for generating warning system messages in case of fire, flood, state security threats, etc. Early warning in relation to the situations mentioned above can be made thanks to fire and flood spread forecasting; modeling thereof is covered by other activities of the CRISIS project. One of the possible configurations of components of the crisis management system with application to fire emergency in buildings is shown in Figure 1.

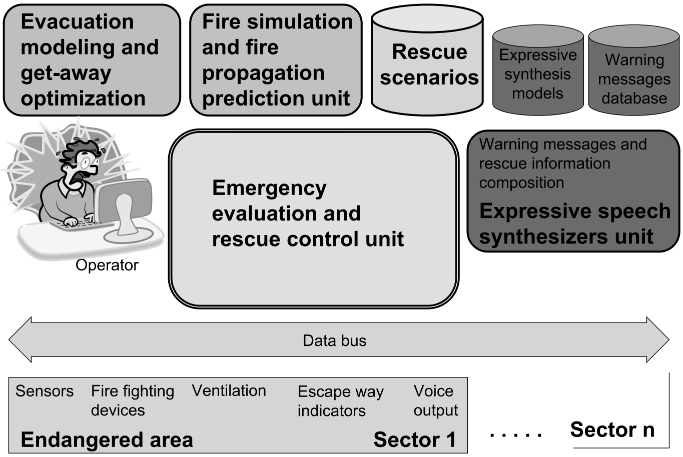

Evacuation modeling and get-away optimization
Fire simulation and fire propagation prediction unit
Rescue scenarios
Expressive synthesis models
Warning messages database
Operator
Emergency evaluation and rescue control unit
Warning messages and rescue information composition
Expressive speech synthesizers unit
Data bus
Sensors
Fire fighting devices
Ventilation
Escape way indicators
Voice output
Endangered area
Sector 1
.....
Sector n

Figure 1. One of the possible configurations of components of the crisis management system with application to fire emergency in buildings

3 WARNING AND SOOTHING SPEECH, AROUSAL, ACTIVITY

3.1 Warning and Soothing Speech Synthesis

A more precise definition of the task would be that the warning speech synthesizer should generate utterances intended to inform the listener about the emergency in acoustically appropriate way. The warning messages should make the listener prepared for solving the dangerous situation, to be ready for reaction, to be careful, watchful, etc.

The soothing speech synthesizer should generate synthetic utterances intended to calm the listener down in acoustically appropriate way, to make him/her tranquil, composed, easeful, peaceful, restful, or to soothe him/her.

It is of course possible to try to calm an upset, hysterical, or frightened person by a forceful command. We have however limited the present study to soothing via calm, easeful, peaceable voice.

3.2 Psychological Point of View

As pointed out in the previous paragraph, the affective state that we want to evoke in the listener by warning and soothing speech would fall in emotional and mood categories like careful, prepared, watchful, or calm, tranquil, composed, easeful, peaceful, restful, or soothed. This kind of categorization is language and culture specific and uneasy to translate. For better generalization we have decided not to explain the task by means of categorical, but rather by dimensional model of emotional space.

Dimensional models of emotion attempt to conceptualize human emotions by their position in two or three dimensional space. Practically all of them incorporate valence and arousal or intensity dimensions.

Wundt (1897) proposed three dimensions for emotion description, "pleasurable versus un-pleasurable", "arousing or subduing" and "strain or relaxation" [6]. Schlosberg (1954) named three dimensions of emotion: "pleasantness-unpleasantness", "attention-rejection" and "level of activation" [7]. Russell (1980) suggests that emotions are distributed in a two-dimensional space, containing arousal and valence dimensions [8].

In his work from 1989 Thayer updates his sooner concept of Activation-Deactivation Adjective Checklist (AD ACL). This was designed to assess two dimensions of subjective arousal. He comments that currently it is referred to as energetic arousal (EA) and tense arousal (TA). Energetic arousal is associated with readiness for vigorous and muscular-skeletal activation (energy versus tiredness and fatigue). Tense arousal represents a preparatory-emergency system, activated by some real or imagined danger that prepares the person for "fight or flight" and inhibiting ongoing activity to maintain readiness for reacting to threat [9].

We think that the notions of "tense arousal" and "emergency-preparatory activation" best describe the affective phenomenon which influences the speech we want to study.

4 SPEECH RESOURCES

Almost all up-to-date speech synthesis methods apply large speech databases to acquire the data necessary for their development, training, and testing. Specialized expressive speech databases have to be built for every particular language and have to cover the affective phenomena to be studied.

A bigger neutral speech database is needed to create a basic neutral voice in HMM synthesis. This can then be adapted to expressive voices with different levels of tense arousal using smaller expressive speech databases. We have chosen records of phonetically rich sentences from our VoiceDat-Sk database for the training of the neutral voice of one speaker (labeled as MR) [10]. For the other speakers we have similar, but smaller neutral databases available too.

The speech databases that were used in this project for phonetic-acoustic research on the speech with different levels of expressive load and for the development of expressive synthesizers are the following:

- neutral database for speech synthesis in Slovak – VoiceDat
- expressive speech database with semantically loaded texts – Crisis
- expressive speech database with semantically neutral texts – Euron.

4.1 Text Resources

All the texts are in Slovak. (In fact, one of our speakers has also recorded the databases in Serviko Romani, the language of Roma minority in Slovakia too [11], but this topic – bilinguality – is not addressed in this work.)

The texts of the "big" neutral databases consist of phonetically rich and semantically neutral sentences. (level 0 of tense arousal)

The texts of the CRISIS emotional databases contain certain emotional load. For higher tense arousal (levels 1, 2, 3) the texts of warning messages were used. The speech material consists of warning messages with lengths ranging from one word to four sentences. For lower tense arousal (levels -1, -2, -3) the sentences had soothing texts using adjectives from Activation-Deactivation Adjective Check List (AD ACL) [9].

The texts of the EURONOUNCE expressive speech databases were adopted from our former project [12] and they consist of semantically neutral phonetically rich sentences (same texts for levels 1, 2, 3 and levels -1, -2, -3.).

4.2 Database Recording

The databases were recorded in an acoustically treated recording studio using RODE K2 microphone and Emu Tracker Pre USB audio interface with 48 KHz sampling frequency and 16 bit resolution.

One of the biggest problems with recording acted expressive speech databases is that the actor is often unable to keep the level of portrayed emotion consistent for a longer time interval. After a while the expressive load in his/her speech changes. We therefore designed a three-step method of recording of the expressive database [13].

In this method the speaker does not try to maintain the same level of expressivity during the entire recording, but s/he rather varies the emotional load three times with every sentence. So s/he produces triples of lexically identical utterances trying to keep same steps in tense arousal levels. The speaker was instructed to utter the message once in a neutral manner (level 1 of tense arousal). This level is denoted "level 1" to express that it represents a reference for the triple of "high arousal" utterances. Then the sentence is uttered with higher imperativeness, like a serious command or directive (level 2), and finally like an extremely urgent command or statement being declared in a situation when human lives are directly endangered (level 3).

When recording the "lower tense arousal" triple of databases the speaker was instructed to utter the prompted message once in a neutral way, which was natural and comfortable for him/her. We assume that this level reflects the neutral state of the speaker at that particular recording session. The same sentence is then uttered in the second level of expressivity with lower activation (level -2). The speaker is instructed to imagine that s/he has to announce to a group of adult people that the emergency situation has passed, that the alarm was called off and they can calm down and stay at ease.

Then the same sentence is uttered with extremely low tense arousal (level −3). The speaker should imagine that s/he is speaking to scared small children, or to a seriously ill or wounded person. His/her speech should not be motherese, or whispered, but has to be very peaceful. After recording the message in the third level, the speaker relaxes for several seconds and then s/he starts with a new prompted sentence. A schematic representation of the emotional space and the position of our databases on the arousal axis is shown on Figure 2 (valence is not addressed in this study).

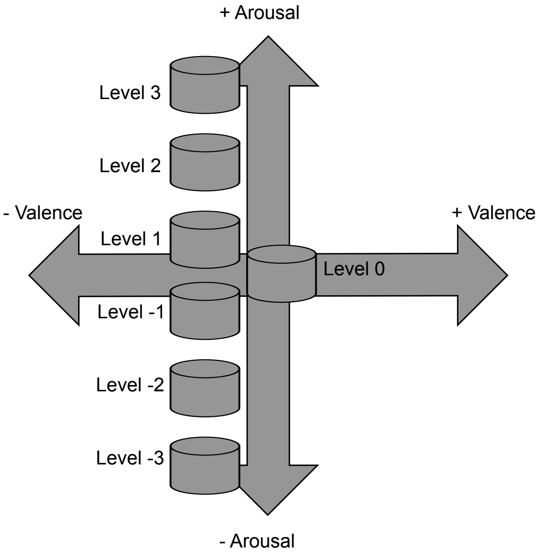


Figure 2. Schematic two-dimensional model of emotional space and expected position of our databases on the arousal axis

We assume this approach produces three emotionally consistent sets of utterances with three relatively consistent and distinguishable degrees of tense arousal in one recording session.

4.3 Current State of the Databases

The current state of the recordings in the expressive database is shown in Table 1.

5 ACOUSTICAL ANALYSES

To get an idea what are the acoustic cues of the tense arousal level changes in speech we investigated several basic acoustic characteristics of the speech databases

Spk.	Sex/Language	Neutral Level 0 (senten.)	CRISIS Level 1	CRIS. L. 2	CRIS. L. 3	CRIS. L. -1	CRIS. L. -2	CRIS. L. -3	EURON Level 1	EUR. L. 2	EUR. L. 3	EUR. L. -1	EUR. L. -2	EUR. L. -3
MR	male Sk	3 500	150	150	150	150	150	150	150	150	150	150	150	150
JK	male Sk	450	150	150	150	150	150	150	450	450	450	450	450	450
MD	male Sk	450	150	150	150	150	150	150	450	450	450	450	450	450
VR	female Sk	450	150	150	150	150	150	150	450	450	450	450	450	450
SC	male Sk	1 500	150	150	150	–	–	–	–	–	–	–	–	–
SC	male Roma	1 500	150	150	150	–	–	–	–	–	–	–	–	–

Table 1. Current state of the databases

recorded in the way described above. These were F0 features, formants of vowels, intensity features and time-domain characteristics.

5.1 Statistical Processing of the Measured Acoustic Characteristics

The objects of investigation were the acoustic characteristics of the speech signal contained in the databases such as Intensity (I), fundamental frequency (F0), their means and ranges. To represent the characteristics in an illustrative and easily comparable manner we decided to approximate the distribution of measured values over each particular database with k-multiplied normal distribution in the form:

$$f(x) = k\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

where μ is the mean of the distribution, σ is its standard deviation and k is a constant. For the normal distribution two standard deviations from the mean account for about 95.449 % of the values, therefore we will use the interval $\mu-2\sigma \leq x \leq \mu+2\sigma$ to define the range. This characteristic is sometimes denoted as SD95 %. Pearson product-moment correlation coefficient is used to express the quality of approximation of the measured data with the chosen model (which is in our case the k-multiplied normal distribution) [14, 15].

5.2 Acoustical Analyses – F0

In Figure 3 we present typical histograms of the fundamental frequency F0 (dB) on the databases with decreasing and increasing tense arousal.

With increasing tense arousal the distributions of F0 are well differentiated for the three levels (1, 2 and 3) of emergency-preparatory activity. Conversely, in an effort to decrease the level of tense arousal only the differences between the two first levels are clearly observable, the differences between the lowest two levels are very small. Our assumption which is in accordance with our observations so far is that mean F0 generally decreases with decreasing emergency-preparatory activity. In an effort to further decrease the tense arousal at the very low level the speaker probably reaches his/her physiological lower limit of glottal fundamental frequency.

To identify which of the phenomena are speaker dependent it is of course inevitable to have the data of more people available. The database sets with low tense arousal and high tense arousal were recorded by four speakers – three males and one female. We present the results of the analyses of F0 of our four speakers in Figure 4.

5.3 Acoustical Analyses – Intensity

In Figure 5 we present typical histograms of the Intensity I (dB) on the databases with decreasing and increasing tense arousal.

Similarly to F0 with increasing tense arousal the distributions of intensity are well differentiated for the three levels (1, 2 and 3) of emergency-preparatory activity. When decreasing the level of tense arousal the differences between the two first levels are bigger than between the second (−2) and third (−3) levels. Intensity decreases with decreasing emergency-preparatory activity; and again, the difference between the lowest two levels is significantly smaller than between others. Intensity seems to be closely correlated to F0. We present the results of the analyses of F0 of these speakers in Figure 6.

6 ACOUSTICAL ANALYSES – FORMANT POSITIONS

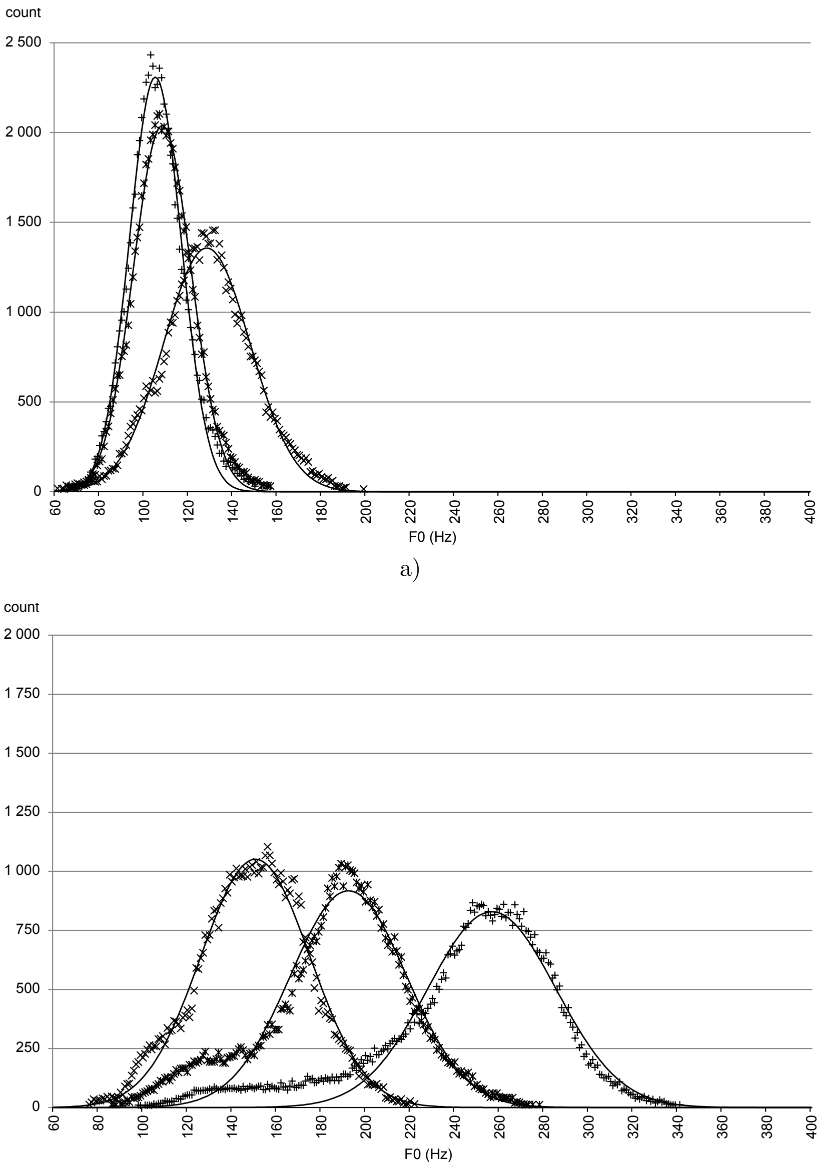
One of the easily measurable characteristics representing the changes in the voice quality is the position of the formants. Figure 7 shows the changes in formant central frequencies of vowels with increasing level of arousal.

It can be seen that for "u" and "a" the change of F1 in the lower range of arousal is smaller than for the higher levels, but for "i", "e" and "o" it is the opposite. Much more detailed research on statistically representative data will have to be done on these phenomena in order to better understand them.

6.1 Acoustical Analyses – Segmental Lengths

Many investigators observed that vowel (plus dipthongs, syllabic consonants and some voiced consonants) durations increase, while consonant durations become shorter in loud speech (see e.g. [16]).

From the analyses of the segmental durations in the Crisis and EURONOUNCE expressive databases not all the data confirm this trend. We will illustrate this with



a)

b)

Figure 3. a) Histograms of F0 from the three databases with decreasing emergency-preparatory activity (from the left: level −3, level −2, level −1). b) Histograms of F0 obtained from the measurements on the three databases with increasing emergency-preparatory activity (from the left: level 1, level 2, level 3). Database: CRISIS; speaker: MR.

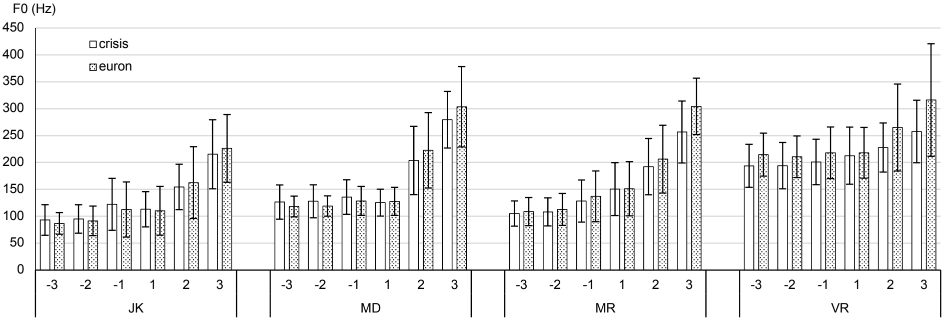


Figure 4. Means and ranges (SD95 %) of F0 for all levels of tense arousal. (Databases: CRISIS, EURONOUNCE; speakers JK, MD and MR – male, and VR – female)

the results of measurement of mean durations of short vowels (a, e, i, o, u). While their lengths are monotonously increasing with higher tense arousal for the higher three levels in practically all speakers and in all vowels, for low levels the results are not consistent.

Recording the databases with lower and higher levels of arousal represent two independent tasks that were realized in different sessions. This inter-session and inter-task difference is easily recognizable in our data. The mean segmental lengths of the short vowels are presented in Figures 8 to 12.

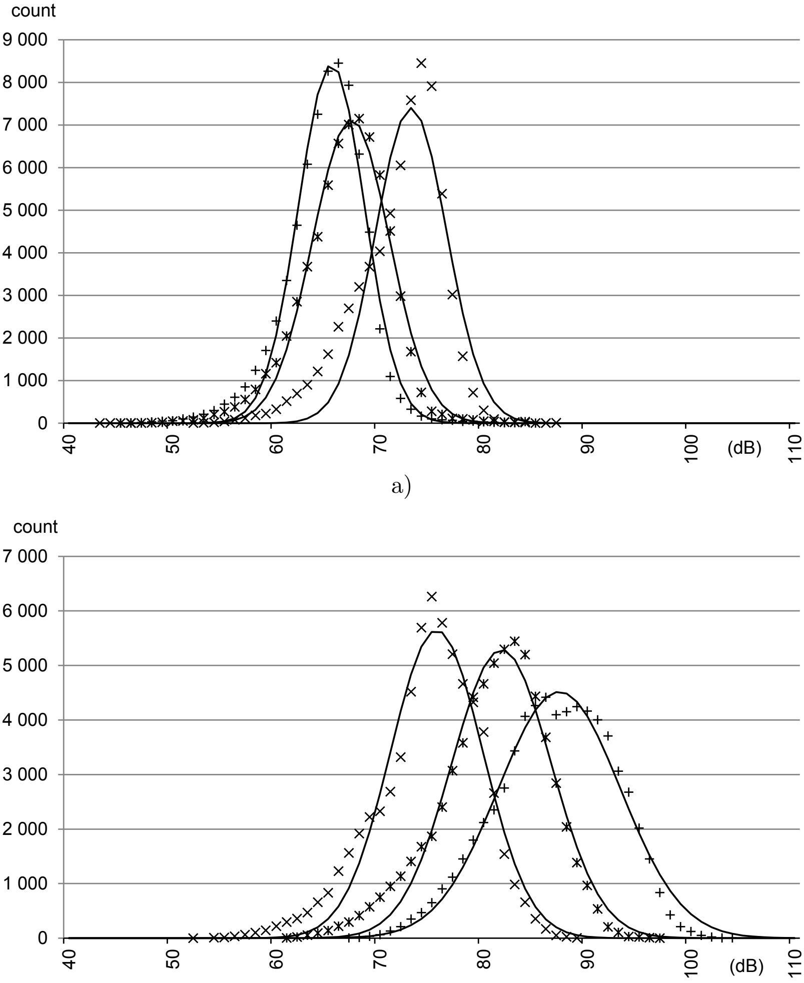
We can only conclude that in the present state of the research we are not able to explain adequately the behavior of the segmental lengths of vowels and their dependence on tense arousal. One of the possible ways to find a characteristic derived of segmental lengths that would change monotonously in the whole range of tense arousal could probably be a ratio of combination of means of selected voiced phonemes to the combination of means of the segmental lengths of selected unvoiced phonemes.

7 EXPERIMENTS WITH EXPRESSIVE SPEECH SYNTHESIS

The HTS system [1] was used for experiments in speech synthesis. The HMM-TTS neutral voice was trained from the Neutral database. This voice was then adapted using Constrained Structural Maximum A-Posteriori Linear Regression (CSMAPLR) technique [17].

According to our informal listening tests the synthesized speech keeps the voice quality, rhythmical and pitch features from the source recordings very well.

The statistical parametric synthesis has a tendency to give greater weight to average values and to limit the extremes. This causes several phenomena observable on the measured data. For instance, the difference between means of F0 in the three higher levels of arousal is slightly smaller for synthesized sentences than for the original voice, the difference between means of Intensity for the lowest two arousal



a)

b)

Figure 5. a) Histograms of Intensity from the three databases with decreasing emergency-preparatory activity (from the left: level −3, level −2, level −1). b) Histograms of F0 obtained from the measurements on the three databases with increasing emergency-preparatory activity (from the left: level 1, level 2, level 3). Database: CRISIS; speaker: MR.

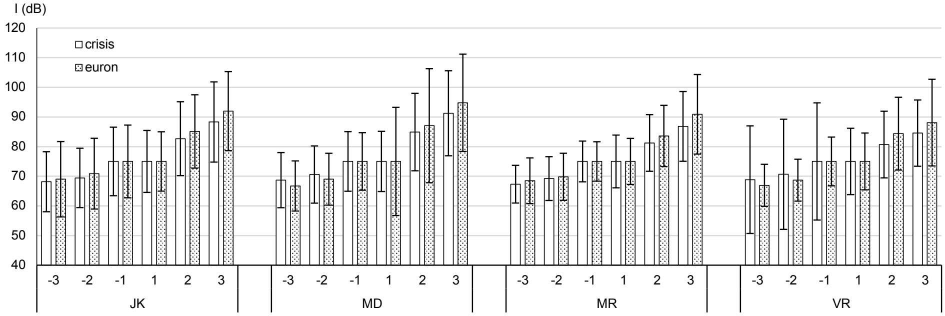


Figure 6. Means and ranges (SD95 %) of Intensity for all levels of tense arousal. (Databases: CRISIS, EURONOUNCE; speakers JK, MD and MR – male, and VR – female)

levels is smaller for synthesized sentences than for the original voice, etc. Therefore the expressive load in the synthesized voices will be lower than in the original speech. There are also some outlying local maxima clearly observable in the histograms that show that the probability distribution of values of F0 in the synthesized sentences is even less Normal (Gaussian), than in the natural speech (i.e. the diversity and randomness of values are smaller).

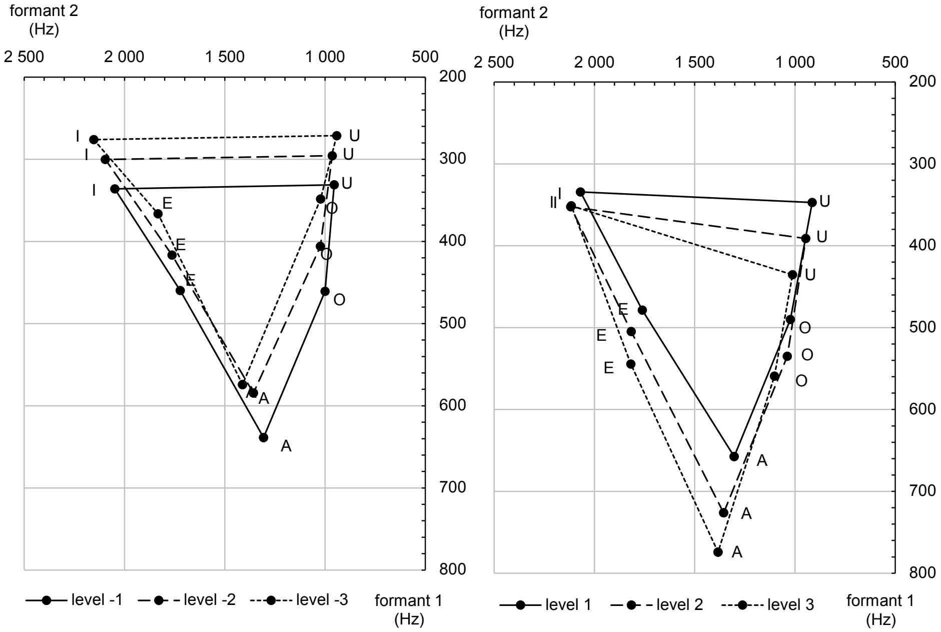


Figure 7. Changes of the formant central frequencies with increasing tense arousal (Database: CRISIS; speaker MR)

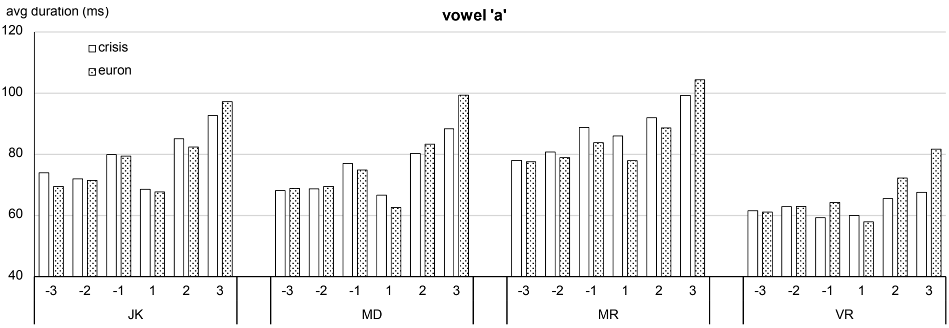


Figure 8. Mean segmental lengths of the vowel "a" (Databases: CRISIS, EURONOUNCE; speakers: JK, MD, MR, VR)

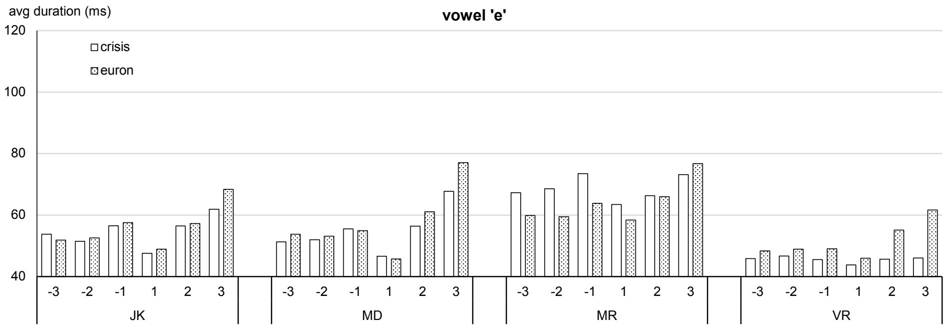


Figure 9. Mean segmental lengths of the vowel "e" (Databases: CRISIS, EURONOUNCE; speakers: JK, MD, MR, VR)

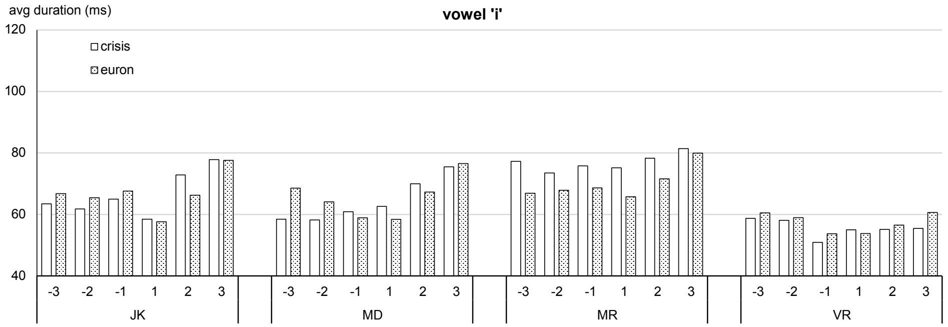


Figure 10. Mean segmental lengths of the vowel "i" (Databases: CRISIS, EURONOUNCE; speakers: JK, MD, MR, VR)

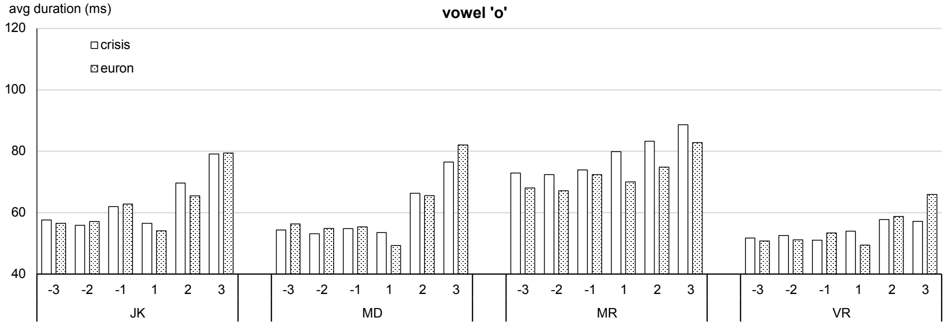


Figure 11. Mean segmental lengths of the vowel "o" (Databases: CRISIS, EURONOUNCE; speakers: JK, MD, MR, VR)

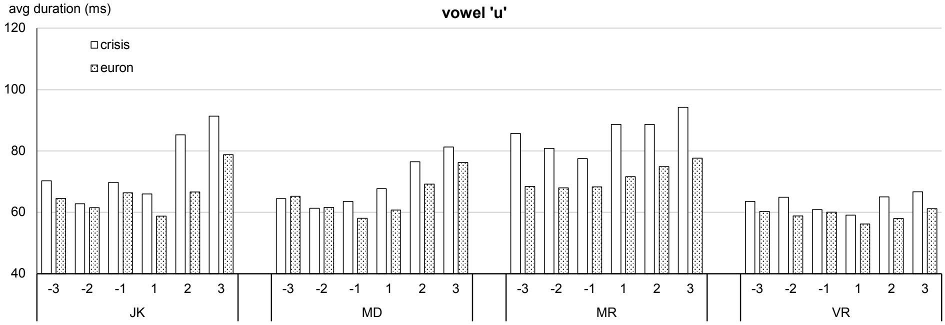


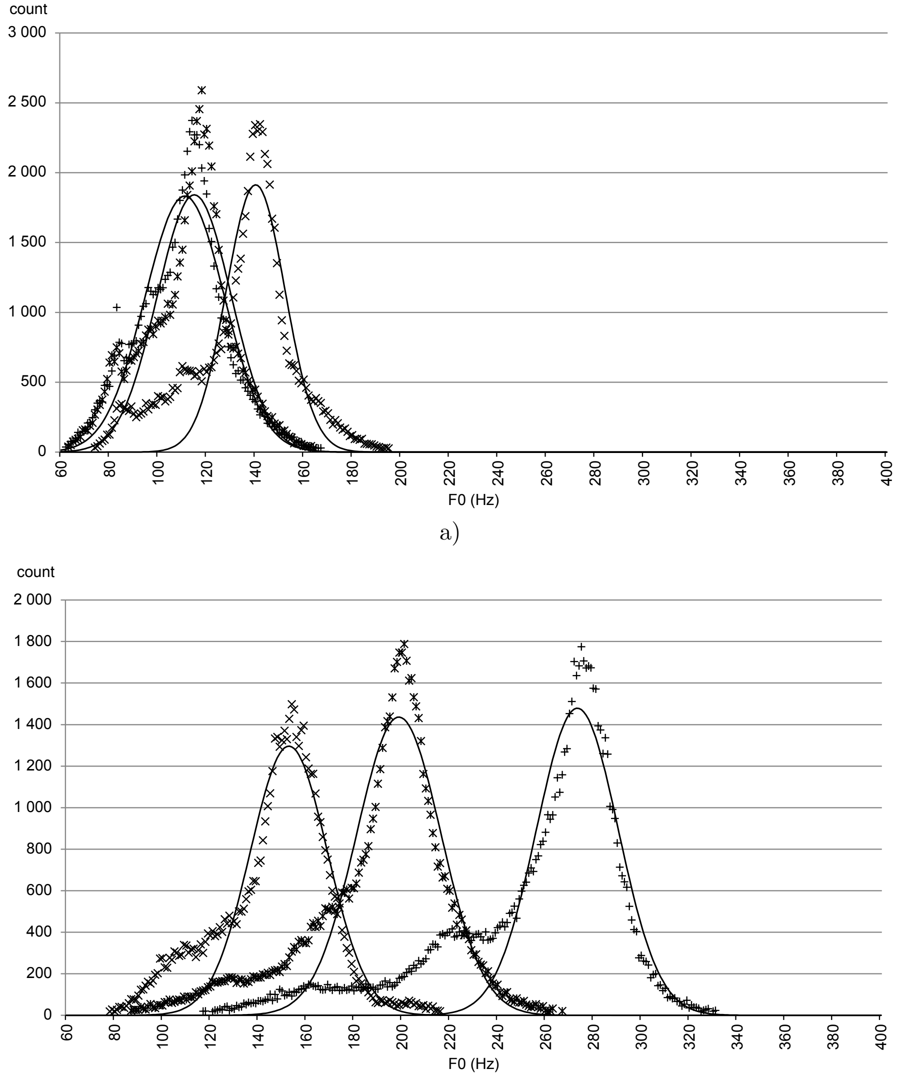
Figure 12. Mean segmental lengths of the vowel "u" (Databases: CRISIS, EURONOUNCE; speakers: JK, MD, MR, VR)

8 UTILIZATION OF THE CRISIS EXPRESSIVE SYNTHESIS SYSTEM

To apply the results of the research in praxis the authors have designed an expressive synthesis system which contains both Unit-selection and Statistical parametric synthesizers and give a wide opportunity to fine-tune their characteristics.

8.1 Project

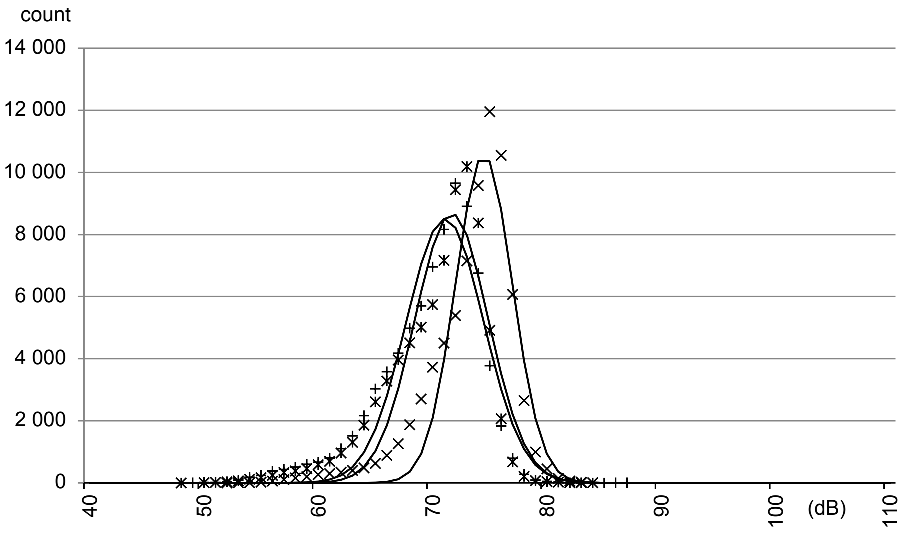
The user needs first to prepare a Project, which is a text document with the texts of all the messages that are to be uttered. S/he assigns voice names to all of the messages. The names are user-defined (e.g. EMERGENCY 1: Attention, there is imminent danger of gas leak!).



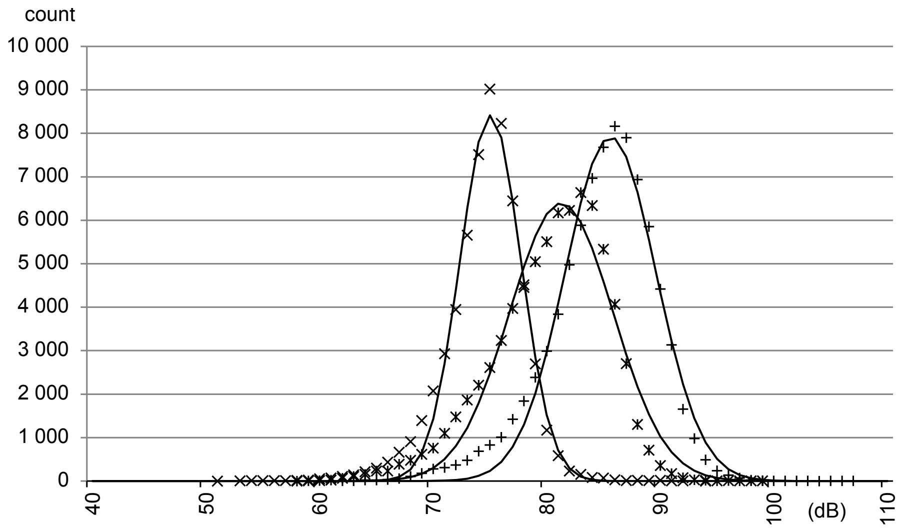
a)

b)

Figure 13. Histograms of F0 from the synthesized speech adapted to six levels of tense arousal (from the left: a) level −3, level −2, level −1, b) level 1, level 2, level 3)



a)

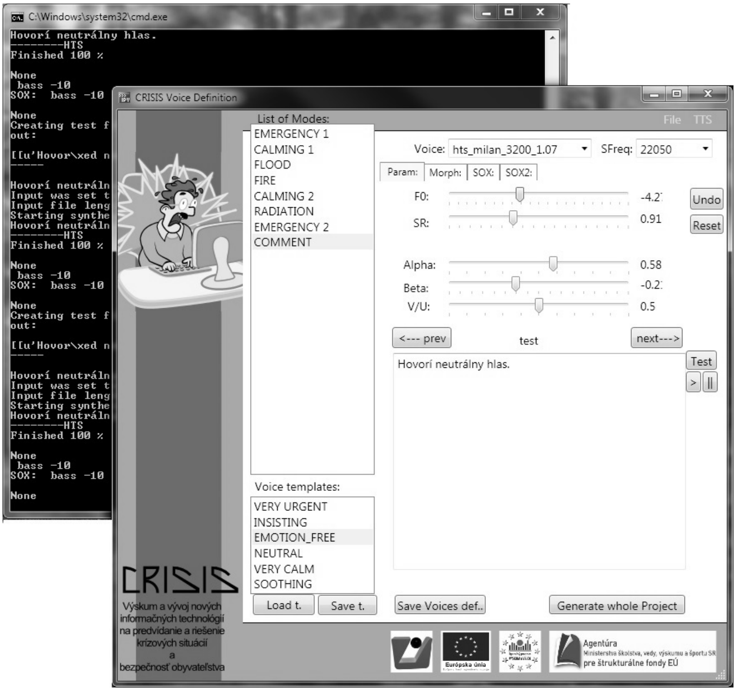


b)

Figure 14. Histograms of Intensity from the synthesized speech adapted to six levels of tense arousal (from the left: a) level −3, level −2, level −1, b) level 1, level 2, level 3)

8.2 Graphical User Interface

Graphical user interface shows the list of the names of the desired voices (called modes) and offers a set of voice templates which can be assigned to these voices and further fine-tuned. Morphing – interpolation between two voices – is also possible in the interface. After the characteristics of all the user voices are defined, the user presses the "Generate whole Project" button and the system generates all the required utterances in corresponding voices. Typical appearance of the CRISIS expressive synthesis system graphical interface can be seen in Figure 15.



```
Hovorí neutrálny hlas.
--------HTS
Finished 100 %

None
 bass -10
SOX:  bass -10

None
Creating test f
out:

[[u'Hovor\xed n
-----

Hovorí neutráln
Input was set t
Input file leng
Starting synthe
Hovorí neutráln
--------HTS
Finished 100 %

None
 bass -10
SOX:  bass -10

None
Creating test f
out:

[[u'Hovor\xed n
-----

Hovorí neutráln
Input was set t
Input file leng
Starting synthe
Hovorí neutráln
--------HTS
Finished 100 %

None
 bass -10
SOX:  bass -10

None
```

Voice: hts_milan_3200_1.07 SFreq: 22050

F0: -4.2
SR: 0.91
Alpha: 0.58
Beta: -0.2
V/U: 0.5

Hovorí neutrálny hlas.

Figure 15. Typical appearance of the CRISIS graphiclal user interface

9 CONCLUSION

We introduced an original method of recording expressive speech database for the collection of speech resources that can be used for the development of hyper-expressive speech synthesis in Slovak, aimed at the use in public safety domain. The experiments with HMM speech synthesis confirmed that the proposed method of speech database development is suitable for the design of expressive speech synthesizers for emergency situations. The first objective tests (see Section 7) confirm the ability of parametric statistical method to model the statistical distribution of the F0 and intensity and its change with the change of tense arousal.

The informal listening tests done by the authors and the audience at the presentations of CRISIS system show that levels of arousal are easily distinguishable and warning voices are much more suitable for the synthesis of urgent messages.

Acknowledgements

This publication is the result of the project implementation: CRISIS, ITMS 2624022-0060 supported by the Research & Development Operational Programme funded by the ERDF and the acoustic analyses were funded by VEGA grant, number 2/02-02/11, Automatic speech recognition in the communication of more persons with the application for a transcript of the court hearings.

REFERENCES

[1] Zen, H.—Nose, T.—Yamagishi, J.—Sako, S.—Masuko, T.—Black, A. W.—Tokuda, K.: The HMM-Based Speech Synthesis System Version 2.0. Proc. of ISCA SSW6, Bonn, Germany, 2007.

[2] Conkie, A. D.: Robust Unit Selection System for Speech Synthesis. Joint Meeting of ASA, EAA, and DAGA, paper 1PSCB_10, Berlin, Germany, March 15–19, 1999.

[3] Rusko, M.—Trnka, M.—Darjaa S.: Three Generations of Speech Synthesis Systems in Slovakia. Proceedings of the XI. International Conference SPECOM 2006, St. Petersburg, Russia 2006, pp. 449–454, ISBN 5-7452-0074-x.

[4] Black, A. W.—Zen, H.—Tokuda, K.: Statistical Parametric Speech Synthesis. Proceedings of ICASSP, April 2007, pp. 1229–1232.

[5] Darjaa, S.—Trnka, M.—Cerňak, M.—Rusko, M.—Sabo, R.—Hluchý, L.: HMM Speech Synthesizer in Slovak. Proceedings of the 7th International Workshop on Grid Computing for Complex Problems (GCCP 2011), Bratislava 2011, pp. 212–221.

[6] Wundt, W. M.: Outlines of Psychology, 1897. Classics in the History of Psychology, An Internet Resource Developed by Christopher D. Green, York University, Toronto 2010. Available on: `http://psychclassics.asu.edu/index.htm`.

[7] Schlosberg, H.: Three Dimensions of Emotion. Psychological Review, Vol. 61, 1954, No. 2, pp. 81–88.

[8] Russell, J. A.: A Circumplex Model of Affect. Journal of Personality and Social Psychology, Vol. 39, 1980, No. 6, pp. 1161–1178.

[9] Thayer, R. E.: The Activation-Deactivation Adjective Check List (AD ACL). APPENDIX I, The Biopsychology of Mood and Arousal, Oxford University Press, New York 1989.

[10] Rusko, M.—Trnka, M.—Darjaa, S.—Cerňak, M.: Slovak Speech Database for Experiments and Application Building in Unit Selection Speech Synthesis. In: Sojka, P., Kopecek, I., Pala, K. (Eds.): Text Speech Dialogue, Lecture Notes in Computer Science Vol. 3206, 2004, pp. 457–464.

[11] Rusko, M.—Trnka, M.—Darjaa, S.—Ritomský, M.: Expressive Speech Synthesis for Urgent Warning Messages Generation in Romani and Slovak. In: Habernal, I., Matoušek, V. (Eds.): Text, Speech and Dialogue, Proceedings of the 16th International Conference TSD 2013, Springer-Verlag, Berlin 2013, pp. 257–264. ISSN 0302-9743.

[12] Jokisch, O.—Jäckel, R.—Rusko, M.—Demenko, G.—Cylwik, N.—Ronzhin, A.—Hirschfeld, D.—Koloska, U.—Hanisch, L.—Hoffmann, R.: The EURONOUNCE Project – An Intelligent Language Tutoring System With Multimodal Feedback Functions, Roadmap and Specification. Proc. ESSV 2008, Frankfurt/M., September 2008, pp. 116–123.

[13] Rusko, M.—Darjaa, S.—Trnka, M.—Cerňak, M.: Expressive Speech Synthesis Database for Emergent Messages and Warnings Generation in Critical Situations. LREC 2012 Proceedings, International Convention and Exhibition Centre, Istanbul 2012, pp. 50–53.

[14] Hazewinkel, M. (Ed.): Normal Distribution, Encyclopedia of Mathematics. Springer 2001. ISBN 978-1-55608-010-4.

[15] Rodgers, J. L.—Nicewander, W. A.: Thirteen Ways to Look at the Correlation Coefficient. The American Statistician, Vol. 42, 1988, No. 1, pp. 59-66.

[16] Geumann, A.: Segmental Durations in Loud Speech. Proceedings of the ITRW on Temporal Integration in Perception of Speech, Aix-en-Provence, France, 2002.

[17] Nakano, Y.—Makoto, T.—Yamagishi, J.—Kobayashi, T.: Constrained Structural Maximum a Posteriori Linear Regression for Average-Voice-Based Speech Synthesis. Proc. of ICSLP '06, 2006.

Milan Rusko graduated from the Slovak Technical University, Bratislava in 1994, Ph. D. degree at the Slovak Technical University, Košice in 2013. Since 1993 he has been the head of the Department of Speech Analysis and Synthesis at the Institute of informatics of the Slovak Academy of Sciences. His research interests include speech acoustics, speech corpora, digital speech and audio processing, speech recognition and speech synthesis. Responsible leader of several national and international projects. Head of the section "Physiological, psychological and musical acoustics" in Slovak Acoustical Society. Management implementation and application of scientific results into practice.

Sakhia Darjaa received his Eng. degree from the Slovak Technical University, Košice in 1979, and Ph. D. degree in 2013. Design of algorithms and their implementation. Programming, design of innovative solutions, robustness and reliability assessment, application development. Member of the Department of Speech Analysis and Synthesis with extensive research and publication activities. Expert in speech synthesis in English and other languages. He is the author of the first software speech synthesizer (Institute of Informatics of the Slovak Academy of Sciences, 1992), and a primary co-author of its next two generations (unit selection and HMM).

Marián Trnka received his Eng. degree in technical cybernetics in 1994 from the Slovak Technical University in Bratislava. He is expert analyses and solution proposals, software and application development. Deputy head of the Department of Speech Analysis and Synthesis. Rich scientific and publication activities. Expert on modeling using Hidden Markov Models, statistical parametric speech synthesis, voice conversion, speaker recognition and speaker verification, automatic signal processing and speech recognition; analyst and programmer.

Róbert Sabo graduated from Comenius University, Bratislava in 1994, Ph. D. degree at the Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences in 2013. Linguist, phonetician, teacher. His research interests include linguistics, phonetics, natural language processing, building language models, speech prosody. At the Department of Speech Analysis and Synthesis he is responsible for the designing of specialized speech databases and their creation. Recording speech databases, management and training of the team of annotators of databases.

Marian Ritomský received his RNDr. degree in physical electronics in 1982 from the Faculty of Mathematics and Physics, Comenius University in Bratislava. He is a programmer, analyst and database expert. He has many years of experience with programming and database design. He gains expertise on statistical processing of acoustic and prosodic speech parameters, in development and evaluation of the quality of expressive speech synthesizers also in optimization of parameters of speech synthesizer to achieve desired performance.