Alert! … Calm Down, There Is Nothing to Worry About. Warning and Soothing Speech Synthesis

Milan Rusko, Sakhia Darjaa, Marián Trnka, Marian Ritomský, Róbert Sabo
Institute of Informatics of the Slovak Academy of Sciences
Dúbravská cesta 9, 845 07 Bratislava, Slovakia

E-mail: {Milan.Rusko, Sakhia.Darjaa, Marian.Trnka, Marian.Ritomsky, Robert.Sabo}@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. The synthesizer should be able to produce speech in Slovak in different styles from extremely urgent warnings, insisting messages, alerts, through comments, and neutral style speech to soothing messages and very calm speech. A three-step method was used for recording both - the high-activation and low-activation expressive speech databases. The acoustic properties of the obtained databases are discussed. Several synthesizers with different levels of arousal were designed using these databases and their outputs are compared to the original voice of the voice talent. A possible ambiguity of acoustic cues is pointed out and the relevance of the semantic meaning of the sentences both in the sentence set for the speech database recording and in the set for subjective synthesizer testing is discussed.

Keywords: warning speech synthesis, soothing speech synthesis, emotional speech database

1. Introduction

Previous attempts to build synthesizers in Slovak were limited to emotionally neutral speech and used diphone concatenative, unit-selection and recently also statistical-parametrical speech synthesis (Rusko, Trnka , Darjaa, 2006). 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. This paper describes the recording of the expressive databases in Slovak dedicated for basic phonetic research and synthesizer building which started recently (Rusko et al., 2012). It is no more focused only on warning and insisting speech, but addresses also the low levels of arousal.

2. Warning and Soothing Speech, Arousal, Activity, Tension

2.1 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 soothing speech synthesizer should generate synthetic utterances intended to calm the listener down in acoustically appropriate way.

2.2 Psychological point of view

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.
Thayer distinguishes two dimensions of subjective arousal; energetic arousal (EA) and tense arousal (TA). Energetic arousal is associated with readiness for vigorous and muscular-skeletal activation. Tense arousal represents a preparatory-emergency system, activated by some real or imagined danger. (Thayer 1989)
We think that the notions of "tense arousal" and "emergency-preparatory activation" best describe the affective phenomenon which influence on speech we want to study.

3. 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 a subset of phonetically rich sentences from our VoiceDat-Sk database for the training of the neutral voice of one speaker (Rusko et al., 2004).

4. Expressive Speech Databases Recording

4.1 Text resources

All the texts are in Slovak. The texts of the "big" neutral database, VoiceDat-Sk consist of 2500 phonetically rich and semantically neutral sentences. The level of arousal in this database will be indicated as level 0.

The texts of the CRISIS expressive databases contain certain emotional load. The texts of 150 warning messages were used for higher tense arousal databases (levels 1, 2, 3) and 150 sentences for lower tense arousal (levels -1, -2, -3) had soothing texts.

The texts of the EURONOUNCE expressive speech databases consist of 150 semantically neutral phonetically rich sentences. The texts of these sentences were adopted from the EURONOUNCE project (Jokisch et al., 2008). Same texts were used for databases with increasing arousal (levels 1, 2, 3) and for decreasing arousal (levels -1, -2, -3) in the EURONOUNCE speech database.

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 speech changes. We therefore designed a three step method of recording of the expressive database [1]. In this method the speaker does not try to maintain the same level of expressivity during the entire recording, but he rather varies the emotional load three times with every sentence. So he produces triples of lexically identical utterances trying to keep same steps in tense arousal levels. The authors think, that that coming back to the neutral, natural setting of the speaker's voice functions like a "reset" and gives the speaker a robust reference for further changes in his voice in the other two depicted levels of arousal.

The speaker was therefore instructed to utter the message once in a neutral manner (referred to as level 1 of tense arousal), then 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 in danger (level 3).

When recording the "lower tense arousal" triplet of databases the speaker was instructed to utter the prompted message once in a natural way, comfortable for him. We again assume that this level reflects the neutral state of the speaker at that particular recording session. This first, neutral, reference level of tense arousal is denoted "level -1" to distinguish it from the "level 0" related to the "big" neutral database and "level 1" which is the label of the neutral database from the triplet with increasing arousal. The same sentence is then uttered in the second (decreased) level of expressivity with lower activation (level -2). The speaker is instructed to imagine that 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.

And then is the same sentence uttered with extremely low tense arousal (level -3). The speaker should imagine that he is speaking to scared small children, or to a seriously ill or wounded person. His 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 he starts with a new prompted sentence.

The authors assume that this approach produces three emotionally consistent sets of utterances with three relatively consistent and distinguishable degrees of tense arousal for every the two triples of the databases (with increasing, and with decreasing activation). Even for amateur speaker it should be easy to maintain three levels of arousal relatively well separated throughout the recording. Recording of 450 sentences (three levels by 150) take about one hour, which is a time interval long enough for the untrained voice to become tired. That is why the number of sentences was limited to 150.

A schematic representation of the emotional space and the expected position of our databases in the emotional space (or rather on the arousal axis) are shown on Figure 1. Valence is not addressed in this study.

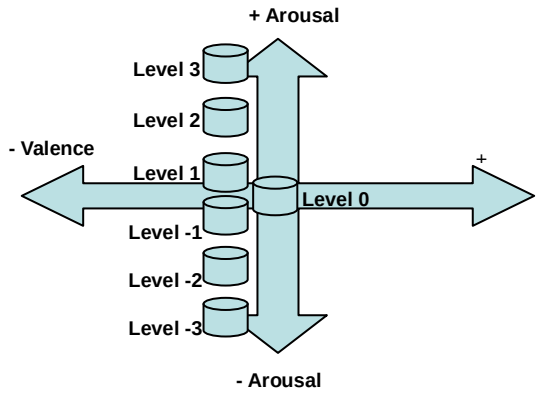

Figure 1: Schematic two-dimensional model of emotional space and expected position of our databases on the arousal axis

5. Acoustical Analyses

To get an idea what are the acoustic cues of the arousal we investigated several basic acoustic characteristics of the recorded speech databases of one speaker. These were F0 features, formants of vowels, intensity features and time-domain characteristics. We also investigated whether the functional relationship between the variables and the level of arousal is monotonic.

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 with mean μ and standard deviation σ.

For the normal distribution two standard deviations from the mean account for about 95.449% of the values, therefore we will use the interval $\mu \pm 2\sigma$ 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). (Hazewinkel, 2001; Rodgers, Nicewander, 1988).

5.2 Analyses of F0

From the statistical characteristics of the fundamental frequency of vocal cords vibration (F0) we have chosen to measure the mean F0 of the all voiced parts of the signal and the frequency range of F0. In Figure 2 and Figure 3 we present histograms of measured F0 on databases with increasing and decreasing tense arousal respectively.

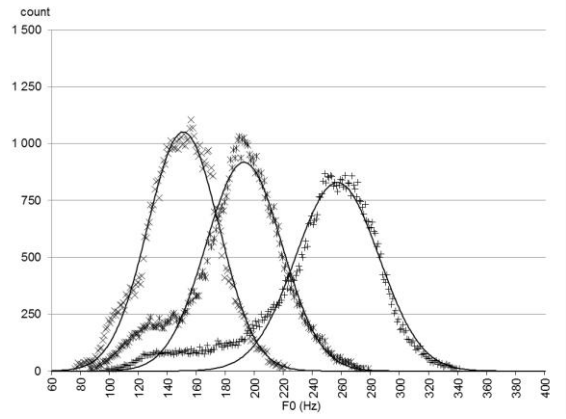

Figure 2: Histograms of F0 obtained from the measurements on the triplet of the CRISIS databases with increasing arousal
(from left: level 1, level 2, level 3)

With increasing tense arousal (Fig. 2.) 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 (-1 and -2) are clearly observable. The distributions of F0 for levels -2 and -3 are very similar in shape, mean and standard deviation. Our assumption is that mean F0 generally decreases with decreasing emergency-preparatory activity. With further decreasing the tense arousal level the speaker probably reaches his physiological lower limit of glottal fundamental frequency.

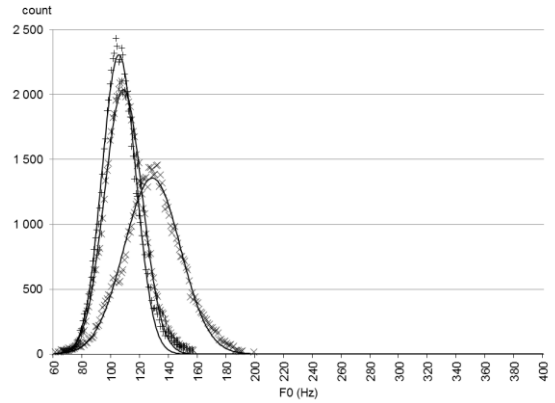

Figure 3: Histograms of F0 obtained from the measurements on the triplet of the CRISIS databases with decreasing arousal
(from left: level -3, level -2, level -1)

Comparison of F0 means and ranges for the databases with semantically loaded texts (CRISIS) and semantically neutral texts (EURONOUNCE) of the same speaker is presented in Figure 4.

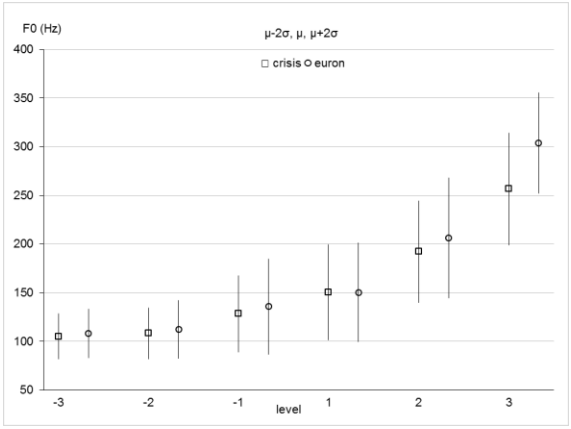

Figure 4: Comparison of F0 means and ranges for the databases with semantically loaded texts (CRISIS) and semantically neutral texts (EURONOUNCE)

This comparison shows that the changes of F0 are bigger in the databases with the semantically neutral texts than in the set with semantically loaded texts. The databases in which the speaker had to utter neutral texts in expressive way seem to be "over-acted". The speaker does not know which of the words should be accented and therefore he accents too many of them.

5.3 Formant positions

One of the measurable characteristics representing the changes in the voice quality is the position of the formants. For better understanding of changes in phonetic quality of the vowels with changing tense arousal we have therefore measured the positions of the first three formants of all the vowels in the databases and computed means of formant centre frequencies per vowel in each of the databases. We present the results in Figure 5.

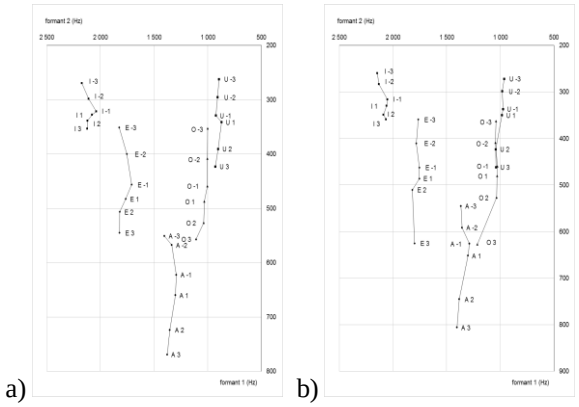

Figure 5: Comparison of vowels formant positions for the databases with
a) semantically loaded texts (CRISIS) and
b) semantically neutral texts (EURONOUNCE)

The formant positions and trajectories are very similar and the differences in the quality of the corresponding vowels seem not to be dramatic in the two sets of databases. More data from more speakers will be needed to study the behaviour of vowel formant frequencies with changing arousal in general, but it is obvious that the frequency of the first formant increases with the increased value of the tense arousal in our data.

5.4 Analyses of Intensity

In Figure 6 and Figure 7 we present typical histograms of the Intensity I (dB) measured on the databases with increasing and decreasing tense arousal respectively.

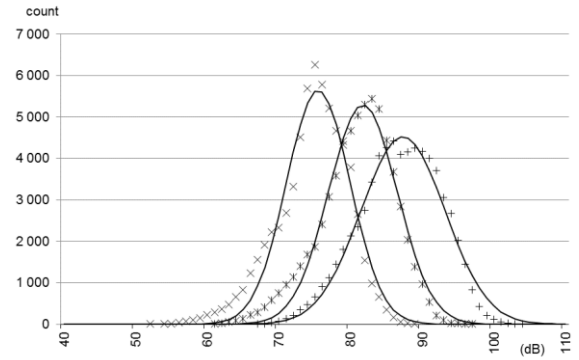

Figure 6: Histograms of Intensity obtained from the measurements on the triplet of the CRISIS databases with increasing emergency-preparatory activity (from left: level 1, level 2, level 3)

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, then 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.
Comparison of Intensity means and ranges for the databases with semantically loaded texts (CRISIS) and semantically neutral texts (EURONOUNCE) of the same speaker is presented in Figure 8.

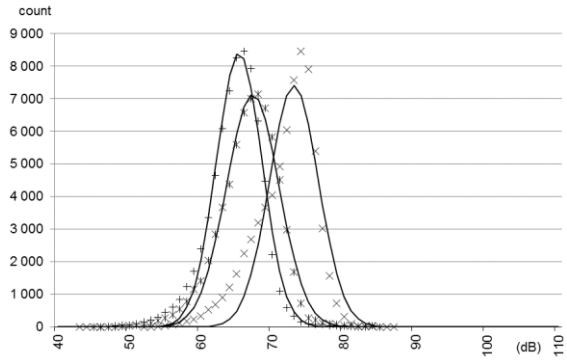

Figure 7: Histograms of F0 obtained from the measurements on the triplet of the CRISIS databases with decreasing emergency-preparatory activity (from left: level -3, level -2, level -1)

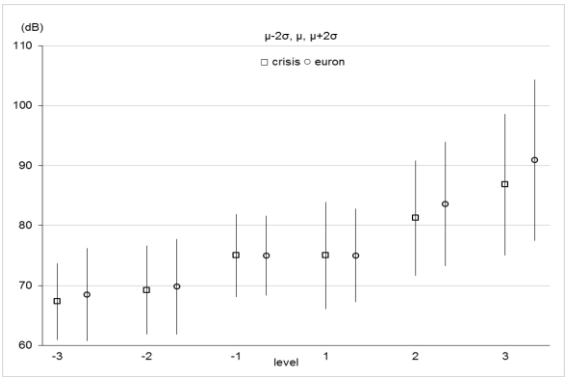

Figure 8: Comparison of Intensity means and ranges for the databases with semantically loaded texts (CRISIS) and semantically neutral texts (EURONOUNCE)

Similarly to F0 this comparison shows that the changes of Intensity are bigger in the expressive databases with the semantically neutral texts than in the set with semantically loaded texts. The databases in with neutral texts way seem to be "over-acted" in comparison to those with emotionally loaded texts.

5.5 Acoustic characteristics in the time domain

Many investigators observed that vowel durations (including diphthongs and syllabic consonants) increase, while consonant durations become shorter in loud speech (see e.g. Geumann, 2002). Our data confirm this observation only partly. While in the databases with increasing activity the length of vowels increases and consonants are shortened with arousal level, in the databases with decreasing activity this tendency does not apply. We have found the same phenomenon also in the expressive speech databases of other three speakers.
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 its

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 lengths of selected voiced phonemes to the combination of means of selected unvoiced phonemes.

6. Synthesizers Trained on the Databases

The voices are based on Statistic-parametric speech synthesis. We created the neutral voice in the HTS system (Zen et al., 2007) and adapted it to six new voices using smaller expressive databases using Constrained Structural Maximum A-Posteriori Linear Regression (CSMAPLR) technique (Nakano et al., 2006).

We present the results of the synthetic voices adapted with CRISIS database. As the first tests for objective evaluation of the new voices we have done the same acoustical analyses of F0 and Intensity (see Figure 9 to 12) as we have done on the databases. The set of synthesized recordings used the same set of sentences that was used in the CRISIS expressive speech database. The results from synthesized speech can now be compared with those from the original speech.

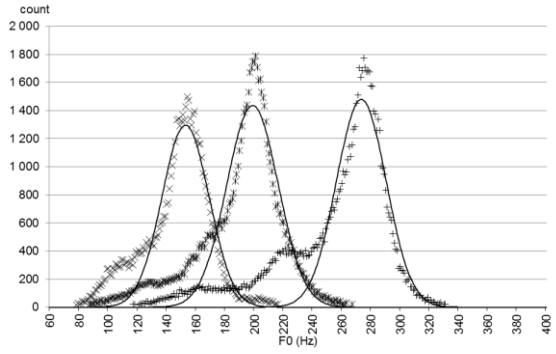

Figure 9: Histograms of F0 of the three sets of synthesized utterances with increasing arousal (from left: level 1, level 2, level 3)

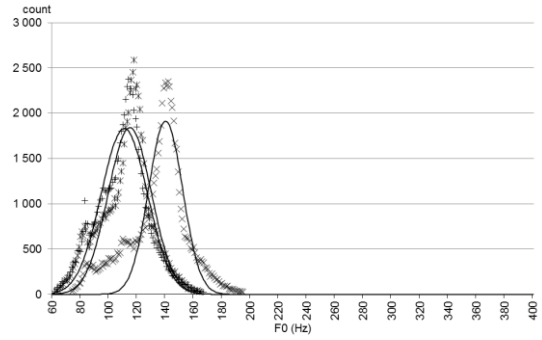

Figure 10: Histograms of F0 of the three sets of synthesized utterances with decreasing emergency-preparatory activity (from left: level -3, level -2, level -1).

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 levels is smaller for synthesized sentences than for the original voice, etc.

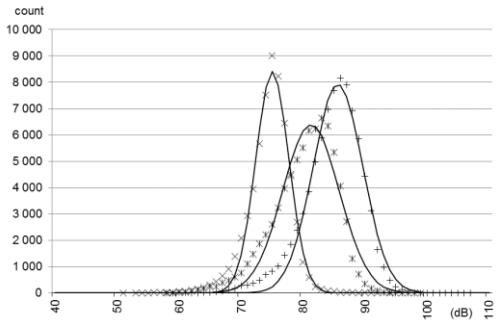

Figure 11: Histograms of Intensity of the three sets of synthesized utterances with increasing arousal (from left: level 1, level 2, level 3)

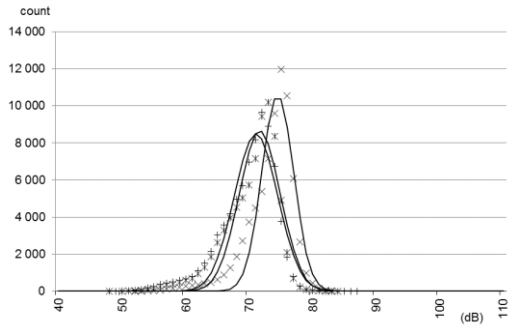

Figure 12: Histograms of F0 of the three sets of synthesized utterances with decreasing arousal (from left: level -3, level -2, level -1)

Therefore the expressive load in the synthesized voices will be lower then 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 is smaller).

According to informal listening tests the synthesized speech keeps the voice-quality, rhythm, pitch features, and the resulting expressive load from the source recordings very well. The rising urgency is reliably distinguishable across levels.

7. Utilization of Expressive Synthesis

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.

The user needs first to prepare a Project, which is a text document where he lists his user-defined voice names

and includes messages for these voices. (e.g. EMERGENCY 1: Be careful, there is imminent danger of gas leak!). The interface automatically derives the list of voices from the Project.
Graphical user interface shows the list of user's voices and offers a set of voice templates which can be assigned to user 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 system is ready to generate all the required utterances in corresponding voices.
Typical appearance of the CRISIS expressive synthesis system graphical interface can be seen on Figure 13.

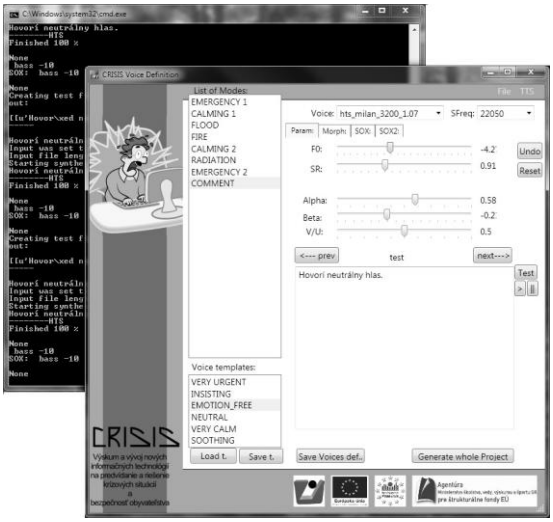

Figure 13: Typical appearance of the graphical user-interface of the expressive speech synthesizer

8. 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 expressive speech synthesis. The method makes it possible to create expressive databases at five consistent and distinguishable levels of arousal.
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 capable to generate insisting, warning, neutral, calming and soothing speech.

9. Acknowledgements

This publication is the result of the project implementation: RPKOM, ITMS 26240220064 supported by the Research & Development Operational Programme funded by the ERDF.

10. References

Hazewinkel, M., ed. (2001). "Normal distribution", *Encyclopedia of Mathematics*, Springer Verlag.

Jokisch O. et al. (2008). The EURONOUNCE Project – an intelligent language tutoring system with multimodal feedback functions, roadmap and specification. *Proc. ESSV 2008*, Frankfurt/M., pp. 116-123.

Rodgers, J. L., Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1), pp. 59–66.

Rusko, M., et al. (2004). Slovak Speech Database for Experiments and Application Building in Unit-Selection Speech Synthesis. In: P. Sojka, I. Kopecek, and K. Pala, editors, *Text Speech Dialogue, Lecture Notes in Computer Science*, Springer Verlag, Volume 3206, 2004, pp 457-464.

Rusko, M., Trnka M., Darjaa S. (2006). Three Generations of Speech Synthesis systems in Slovakia, *Proceedings of the XI. International Conference SPECOM 2006*, St. Petersburg, Russia, pp. 449 – 454.

Rusko, M., et al. (2012). Expressive speech synthesis database for emergent messages and warnings generation in critical situations. In *LREC 2012 Proceedings*, Istanbul, Turkey, pp. 50-53.

Thayer, R. E. (1989). *The Biopsychology of Mood and Arousal*. New York: Oxford University Press, Appendix 1.

Zen, H. et al. (2007), The HMM-based speech synthesis system version 2.0, *Proceedings of ISCA SSW6*, Bonn, Germany.

Nakano, Y., et al. (2006), Constrained Structural Maximum A Posteriori Linear Regression for Average-Voice-Based Speech Synthesis, *Proceedings of ICSLP'06.*