10.2478/v10048-010-0017-3

MEASUREMENT SCIENCE REVIEW, Volume 10, No. 3, 2010

An Experiment with Evaluation of Emotional Speech Conversion by Spectrograms

J. Přibil [1, 2] and A. Přibilová [3]

[1] Institute of Photonics and Electronics, Academy of Sciences CR, v.v.i., Chaberská 57, CZ-182 51 Prague 8, Czech Republic
[2] Institute of Measurement Science, SAS, Dúbravská cesta 9, SK-841 04 Bratislava, Slovakia, umerprib@savba.sk
[3] Slovak University of Technology, Faculty of Electrical Engineering & Information Technology, Department of Radio Electronics, Ilkovičova 3, SK-812 19 Bratislava, Slovakia

The spectrogram is a useful tool for visual quality comparison of different types of emotional synthetic speech. This paper is focused on application of this method to evaluation of sentences after spectral and prosodic modifications for emotional speech conversion. Performed experiments with spectrogram evaluation for several male and female speakers and four emotional states (joy, sadness, anger, neutral state) are also described.

Keywords: emotional speech, spectrogram, periodogram, evaluation of synthetic speech

1. INTRODUCTION

SEVERAL subjective methods can be used for the evaluation of synthetic speech quality [1] - [4]. As we focus on emotional speech transformation (conversion) method with cepstral speech description [5] and modification of speech spectral and prosodic parameters in the text-to-speech (TTS) system enabling expressive speech production [6], the users' opinion is very important for us. Therefore, listening tests must often be performed. However, a problem exists with their collective realization (more people together – for keeping the same test conditions) and also with repeating of the test in a short time interval. These personal problems motivated us to find another method for evaluation of emotional speech synthesis.

The spectrogram can be successfully used for visual quality comparison of different approaches to emotional speech conversion. This method works in the time / frequency domain and we can perform synthesis comparison of a short sentence or an isolated word by this approach. Disadvantage of subjectivity of this method can be eliminated by spectrogram classification with the help of statistical parameter analysis. This statistical approach is also often applied in other areas of biomedical research [7], [8].

The second approach is based only on the analysis of specific parts of the spectrograms (regions of interest – ROI). From these ROIs the mean periodogram by Welch's method [9], [10] can be determined. Finally, calculation of the RMS spectral distance between normal and converted emotional styles of these periodograms can be used for objective matching and comparison. These methods are used for comparison on the segmental or phoneme level of synthesized speech [11].

2. SUBJECT & METHODS

2.1 Spectrograms of transformed emotional speech

Emotional speech conversion method based on spectral modification of male and female voice using the cepstral and harmonic speech models was described in our previous work [12], [13]. Our approach to spectral modification consists of non-linear spectral envelope transformation with the effect of the first formant shift to the left and the higher ones to the right for pleasant emotions, and the first formant to the right and higher ones to the left for unpleasant emotions according to the knowledge of psychological and phonetic research [14]. The proposed spectral modification is combined with modification of F0 mean, F0 range, energy, and F0 linear trend superposition (at the end of the sentence, rising for "joy" and falling for "anger").

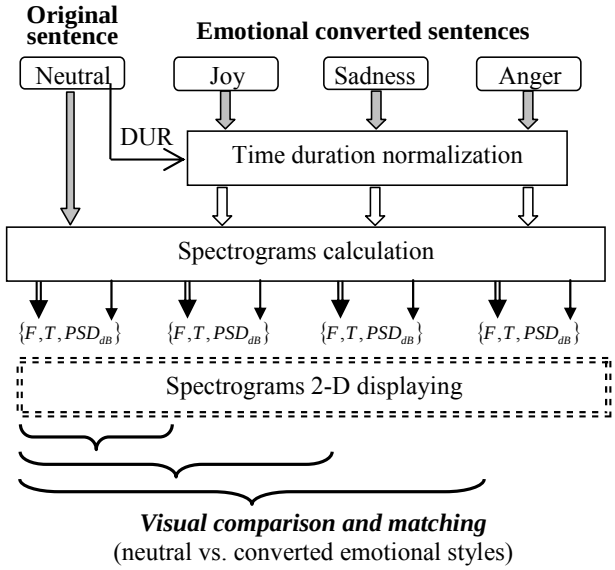

Fig.1. Block diagram of spectrogram calculation of original neutral and transformed emotional speech.

Applied emotional speech conversion method includes also time duration changes:

– time duration lengthening for "sad" emotional style,

– time duration shortening for "joyous" and "angry" styles.
– Therefore, normalization in the time domain must be performed before the spectrogram calculation and comparison.

For time duration normalization we can use linear or non-linear time scale mapping function, or dynamic time warping (DTW) algorithm [15] – see the block diagram of spectrogram calculation in Fig.1 and results of the applied method – the set of four spectrograms in Fig. 2. Every spectrogram set consists of the original neutral sentence and three transformed sentences in emotional styles – the calculated power spectral density (PSD) values in [dB] are used for next comparison and evaluation.

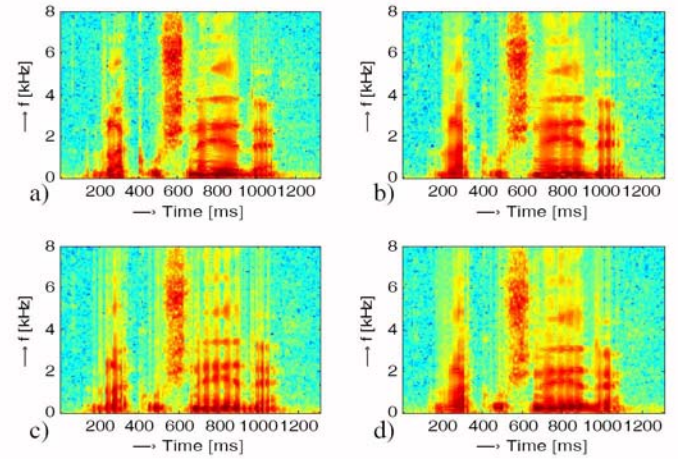

Fig.2. Spectrograms of the sentence "*Vlak už nejede*" (male voice, F0 ≈ 110 Hz, f_s = 16 kHz): original in neutral emotional style (a), resynthesis with applied normalization in the time domain – in "joyous" style (b), in "sad" style (c), and in "angry" style (d).

2.2 Subjective and objective comparison of spectrograms

Our first evaluation alternative to the listening test was visual comparison and matching of displayed spectrograms (see block diagram in Fig.1). However, results of visual comparison of the whole spectrogram depend strongly on the person that makes this matching. For objective comparison and matching of the whole spectrogram, the statistical approach based on analysis of variances (ANOVA) and hypothesis tests can be applied [8].

To obtain more precise matching results it is necessary to select typical or interesting parts – the ROI areas. From the chosen ROI area, the mean periodogram calculated by the Welch method can be determined – see demonstration Fig.3. The resulting Welch's periodogram in [dB] can be used for subsequent comparison. For exact numerical comparison (objective matching method) it is possible to calculate the spectral distance D_{RMS} (by the RMS method) between different periodograms (between the basic sentence in neutral style and other three sentences in transformed emotional styles). This method, enabling objective comparison of the same ROI area ($\Delta T_N = \Delta T_J = \Delta T_S = \Delta T_A$) of four spectrograms after speech signal time normalization based on the neutral style, is illustrated by the block diagram in Fig.4.

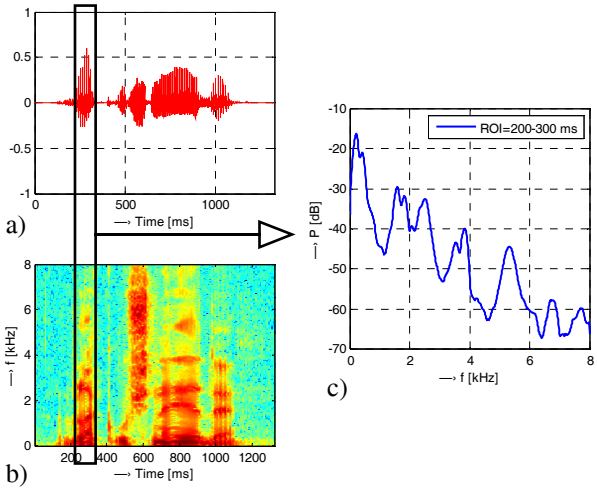

Fig.3. Example of the processed sentence "*Vlak už nejede*" in neutral emotional style: input signal in the time domain (a), corresponding spectrogram (b), Welch's periodogram in [dB] of selected ROI in 100-ms time interval (from 200 ms to 300 ms) (c).

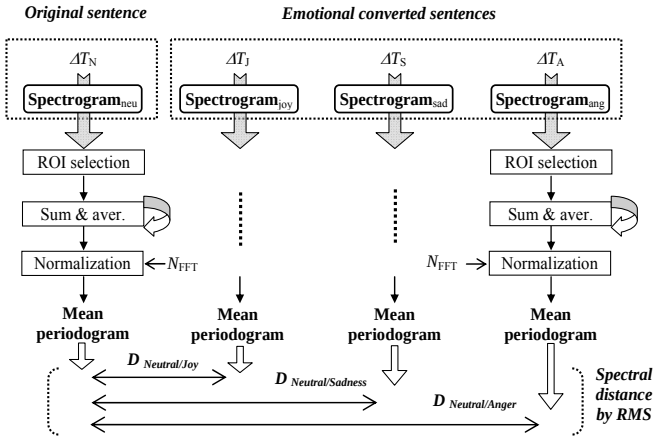

Fig.4 Block diagram of the averaged Welch's periodograms calculation and the spectral distance determination from the selected ROI areas between neutral and converted emotional styles.

3. Material, Experiments and Results

The speech material for this experiment was obtained from CDs of stories in Czech and Slovak languages uttered by professional actors. It was divided into two databases (separately from male voices – 77 sentences, and female voices – 51 sentences, 7+3 speakers altogether) consisting of short sentences in neutral emotional style with duration from 0.5 to 2.5 seconds, resampled at 16 kHz.

Spectral properties modification together with prosodic changes of F0, time duration, and energy were applied to the source sentences in neutral speaking style. Resulting sentences were obtained with corresponding settings of emotional styles conversion (joy, sadness, and anger). In the case of male speakers, changes described in [12] were carried out, the sentences of female speakers were modified by the transformation described in [13]. Three series of experiments were performed:

1) Spectrogram calculation of all sentences collected in two databases (male / female voices): calculation of ANOVA of PSD values, multiple comparison tests of groups.
2) Processing of sentences used in performed listening tests [12], [13]: evaluation by the hypothesis test – comparison with the obtained results of the listening test.
3) Selecting the ROI areas (voiced sounds) of spectrogram: calculation of mean periodograms by Welch's method, numerical matching of results from the calculated spectral distances between corresponding periodograms by the RMS method.

3.1 Experiment with processing of all database sentences

Since the speech material collected in both databases (male / female voice) originates from speakers (S_m / S_f) with different mean F0 value (see Table 1), different parameter settings for spectrogram calculation (window length L_W and window overlapping L_O) [10] must be applied. Therefore three classes of spectrogram parameter settings were realized (C1-3$_m$ and C1-3$_f$.). In every speaker class the set of four spectrograms was calculated consisting of the original and three modified sentences in emotional speaking styles – 308 sentences of male speakers and 204 sentences of female speakers were analyzed in total.

Results of the ANOVA analysis of the set of spectrograms (sums of squares, degrees of freedom, mean squares) in graph form together with visualization of differences between group means are presented in Fig.5, Fig.6 and Fig.7 (for male voice), and in Fig.8, Fig.9, and Fig.10 (for female voice). Every group represents one type of speaking style: neutral, joy, sadness and anger.

*)	$S1_m$	$S2_m$	$S3_m$	$S4_m$	$S5_m$	$S6_m$	$S7_m$	$S1_f$	$S2_f$	$S3_f$
$F0_{mean}$ [Hz]	133	127	98	132	99	101	88	228	177	207

*) $C1_m$={$S1_m$, $S2_m$, $S4_m$}, $C2_m$={$S3_m$, $S5_m$, $S6_m$}, $C3_m$={$S7_m$}, $C1_f$ = $S1_f$, $C2_f$ = $S2_f$, and $C3_f$ = $S3_f$

Table 1: Speaker mean F0 values – male and female voice.

*)	$C1_m$	$C2_m$	$C3_m$	$C1_f$	$C2_f$	$C3_f$
L_W **[samples]**	256	328	164	128	180	164

*) $L_O = L_W/2$, $N_{FFT} = 1024$, $f_s = 16$ kHz

Table 2: Input parameters for spectrogram calculation.

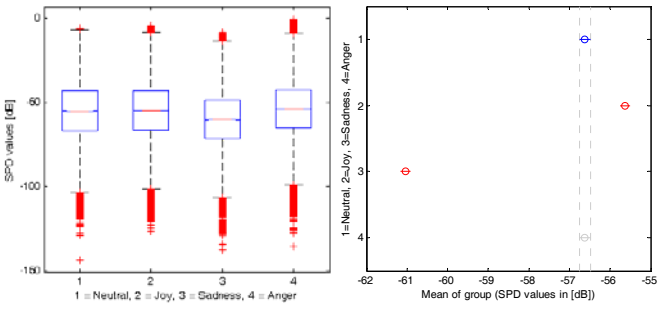

Fig.5. Visualization of the class $C1_m$ results: ANOVA parameters, (left), the difference between group means (right).

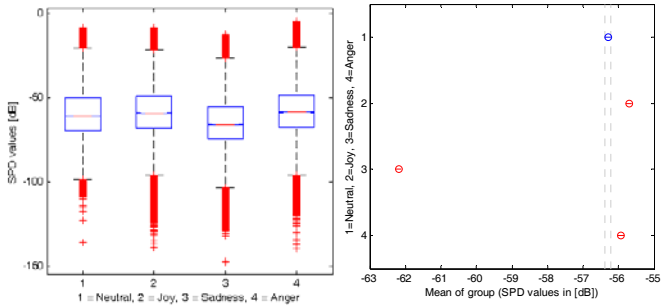

Fig.6 Visualization of the class $C2_m$ results: ANOVA parameters (left), the difference between group means (right).

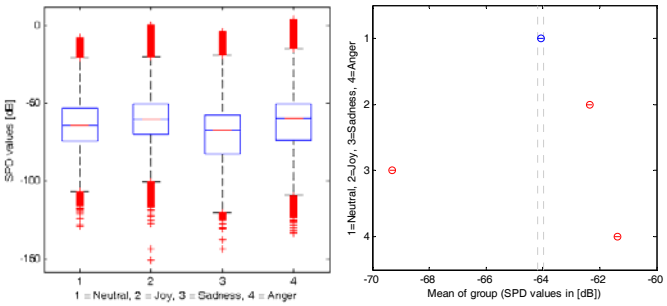

Fig.7. Visualization of the class $C3_m$ results: ANOVA parameters (left), the difference between group means (right).

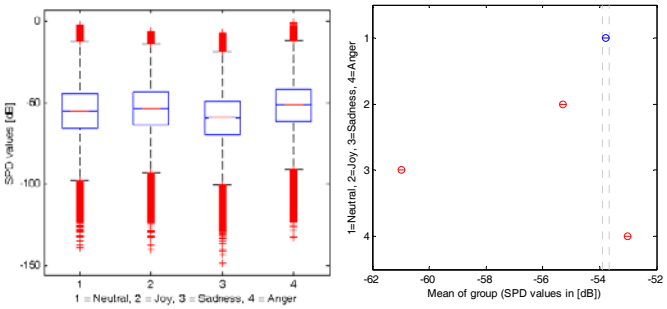

Fig.8. Visualization of the class $C1_f$ results: ANOVA parameters (left), the difference between group means (right).

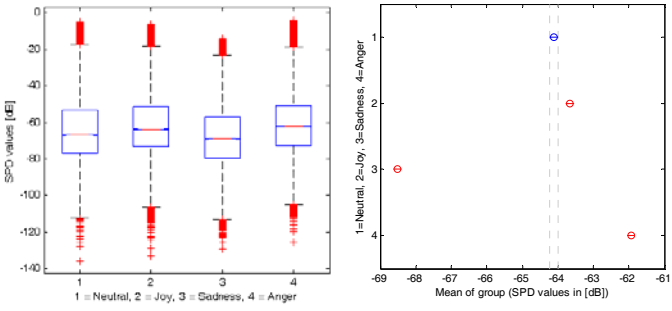

Fig.9. Visualization of the class $C2_f$ results: ANOVA parameters (left), the difference between group means (right).

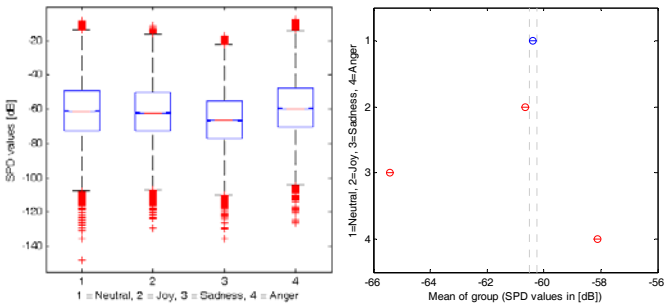

Fig.10. Visualization of the class $C3_f$ results: ANOVA parameters (left), the difference between group means (right).

3.2 Comparison experiment with the listening test

In this second experiment, the Ansari-Bradley hypothesis tests of the PSD values of spectrograms were carried out [16]. In the case of male voices, the sentences of speakers $S1_m$ and $S4_m$ were processed, therefore the spectrogram input settings were the same as for the class $C1_m$ (see Table 2). For processing of the female voice sentences, the setting was the same as for the class $C1_f$, because the speaker $S1_f$ uttered all sentences.

Comparisons were performed between original listening test[1] results obtained in correspondence with [12] (see Table 3), and [13] (see Table 4), and current obtained results from spectrograms. Resulting null hypothesis/probability values for 5% significance level of the Ansari-Bradley test are presented in Table 5 (male voice), and Table 6 (female voice).

	Joy	Sadness	Anger	Other
Joy	**67.86 %**	0.36 %	6.07 %	25.71 %
Sadness	1.07 %	**86.79 %**	2.14 %	10.00 %
Anger	5.71 %	2.14 %	**61.08 %**	31.07 %

Table 3: Original confusion matrix of the listening test – male voice.

	Joy	Sadness	Anger	Other
Joy	**59.0 %**	16.0 %	0.5 %	24.5 %
Sadness	0.5 %	**90.0 %**	0.5 %	9.0 %
Anger	2.5 %	2.0 %	**73.5 %**	22.0 %

Table 4: Original confusion matrix of the listening test – female voice.

	Neutral	Joy	Sadness	Anger
Neutral	0/1	$1/1.66\ 10^{-5}$	$1/3.73\ 10^{-30}$	**0/0866**
Joy		0/1	$1/9.64\ 10^{-46}$	$1/1.51\ 10^{-75}$
Sadness			0/1	$1/1.14\ 10^{-89}$
Anger				0/1

Table 5: Results of the Ansari-Bradley hypothesis tests of spectrogram PSD values - male voice.

1 For each sentence there was a choice from four possibilities: "joy", "sadness", "anger", or "not classified" (values in the column "other" represent choice "not classified"), the original sentences in neutral state were not evaluated.

	Neutral	Joy	Sadness	Anger
Neutral	0/1	**0/0.309**	$1/3.73\ 10^{-30}$	$1/8.66\ 10^{-55}$
Joy		0/1	$1/9.64\ 10^{-46}$	$1/3.14\ 10^{-25}$
Sadness			0/1	$1/1.05\ 10^{-169}$
Anger				0/1

Table 6: Results of the Ansari-Bradley hypothesis tests of spectrogram PSD values - female voice.

3.3 Analysis of voiced sounds by Welch's periodograms

From the main database of sentences, the derived one consisting of manually selected basic vowels "a", "e", "i", "o", "u" and voiced consonants "m", "n" and "l" was created. Spectral analysis with the help of Welch's periodograms was performed in two steps:

1) visual comparison of calculated Welch's periodograms (for selected ROI from the database of vowels and voiced consonants),
2) numerical matching of results from the calculated spectral distances between corresponding periodograms by the RMS method.

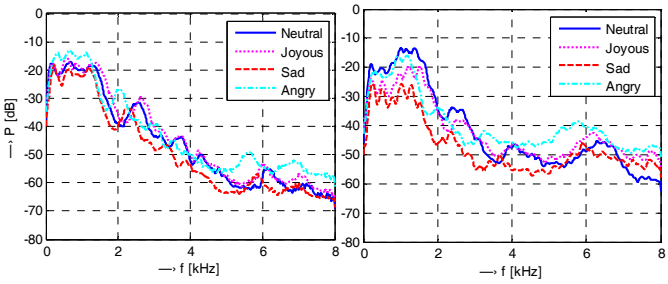

Fig.11 Mean periodograms of analyzed voiced speech parts corresponding to sound "a": male voice (left), female voice (right).

The lengths of selected ROI areas were chosen as multiples of basic L_W length setting (see Table 2). Detailed mean periodograms of analyzed sound "a" selected from the speaker classes $C1_m/C1_f$ are shown in Fig.11. The spectral distances calculated between mean periodograms in original "neutral" and transformed emotional styles are summarized in Table 7 (male voice) and Table 8 (female voice).

Neutral – to:	**joyous**	**sad**	**angry**
D_{RMS} of "a" [dB]	2.517	4.516	5.845
D_{RMS} of "e" [dB]	2.608	3.551	4.862
D_{RMS} of "i" [dB]	2.427	3.769	5.279
D_{RMS} of "o" [dB]	2.710	3.639	6.110
D_{RMS} of "u" [dB]	3.509	4.596	5.846
D_{RMS} of "m"[dB]	2.205	4.809	6.595
D_{RMS} of "n"[dB]	2.160	3.852	4.615
D_{RMS} of "l" [dB]	2.839	3.408	6.063

Table 7: Summary results of spectral distances of analyzed sounds (D_{RMS} are calculated between periodograms of "neutral" and transformed emotional styles) – male voice.

Neutral – to:	joyous	sad	angry
D_{RMS} of "a" [dB]	4.598	7.223	8.382
D_{RMS} of "e" [dB]	5.708	6.599	13.012
D_{RMS} of "i" [dB]	5.794	7.236	8.060
D_{RMS} of "o" [dB]	3.841	5.866	6.370
D_{RMS} of "u" [dB]	2.692	4.973	6.298
D_{RMS} of "m"[dB]	4.186	4.724	4.774
D_{RMS} of "n"[dB]	3.985	6.597	8.909
D_{RMS} of "l" [dB]	3.207	6.929	8.064

Table 8: Summary results of spectral distances of analyzed sounds (D_{RMS} are calculated between periodograms of "neutral" and transformed emotional styles) – female voice.

6. CONCLUSION

Results of evaluation experiments with full spectrograms confirm our premise that spectrogram can also be successfully used for visual comparison of different types of emotional synthetic speech. This alternative additional approach to the listening tests enables to make the objective statistical comparison by ANOVA, and hypothesis tests. This approach is more simple than recent speech recognition methods using evaluation by hidden Markov models [17] or in real-time emotion recognition systems [18].

In our first experiment, spectrogram calculation of all sentences collected in the main databases (from male / female voices, processed per class in dependence on the speakers' mean F0) and evaluation by one-way ANOVA were carried out. Using this statistical approach we test, whether the PSD values from several groups have a common mean. Next, series of *t*-tests, and visualization of differences between group means were performed. From the obtained results, it is evident that PSD data of the speaker class $C1_m$ have small differences between means of the group "neutral" and the group "angry" (see Fig.5) – it means, that emotional transformation of this type has not fully succeeded. The situation was better in the case of the speaker class $C2_m$ – where the interval of means of the group "neutral" and the group "angry" were close together, but without the coincidence. The best result was obtained for the class $C3_m$. In the case of the female speaker classes $C2_f$ and $C3_f$, the position between the groups "neutral" and "joy" (small differences of means, but groups do not overlap) show, that the used transformation parameters were probably not optimal. The greatest differences between groups were obtained for the class $C1_f$.

In a classical listening test, only a limited number of sentences from databases (male and female voice) can be used for comparison. Therefore, in our second comparison experiment, only the spectrograms of sentences originally used in the listening tests were calculated. ANOVA gives also *F* statistic and results of the hypothesis test including probability values. However, a different type of hypothesis test was chosen in our comparison. Unlike the ANOVA *F* statistic, the Ansari-Bradley test compares whether two independent samples come from the same distribution against the alternative that they come from distributions having the same median and shape but different variances. With chosen 5% significance level, the null hypothesis of identical distributions cannot be rejected in the case of tested PSD values of sentences with transformed "angry" style in the case of a male speaker. This result corresponds with obtained values from the originally performed listening test – result of evaluation of "angry" style has a low score and many confused values (see Table 3). For a female speaker, the null hypothesis cannot be rejected in the case of the transformation to the "joyous" style, which is also in agreement with the obtained result of the listening test (see Table 4).

From visual comparison of the whole spectrogram and from principles of applied speech conversion method results, that emotional speech brings about the most significant spectral changes for voiced speech. Therefore the extended analysis of sounds based on Welch's periodograms was subsequently performed. The comparison of calculated spectral distances between "neutral" and transformed emotional styles of voiced sounds shows that the spectral changes (formant position and bandwidth) are the greatest for "angry" and the smallest for "joyous" style. These results are in correspondence with the applied emotional transformation method, which means this approach is fully usable for detailed spectral analysis of voiced parts of speech. But a weak point of this method is the manual selection of ROIs. Speech recognition can be used here (e.g. in the form of a simple phoneme alignment procedure) to get these ROIs automatically.

ACKNOWLEDGMENT

The work was done in the framework of the COST 2102 Action. It was also supported by the Grant Agency of the Czech Republic (GA102/09/0989), the Grant Agency of the Slovak Academy of Sciences (VEGA 2/0142/08) and the Ministry of Education of the Slovak Republic (VEGA 1/0693/08).

REFERENCES

[1] Möller, S., Jekosch, U., Mersdorf, J., Kraft, V. (2001). Auditory assessment of synthetized speech in application scenarios: two case studies. *Speech Communication*, 34, 229-246.

[2] Audibert, N., Vincent, D., Aubergé, V., Rosec, O. (2006). Evaluation of expressive speech resynthesis. In *LREC 2006 : Workshop on Emotional Corpora*. Genova, Italy, 37-40.

[3] Gobl, C., Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40, 189-212.

[4] Tučková, J., Holub, J., Duběda, T. (2009). Technical and phonetic aspects of speech quality assessment: the case of prosody synthesis. In Esposito, A., Vích, R. (eds.) *Cross-Modal Analysis of Speech, Gesture, Gaze and Facial Expressions*. Lecture Notes in Artificial Intelligence 5641. Berlin Heidelberg: Springer-Verlag, 106-115.

[5] Přibilová, A., Přibil, J. (2006). Non-linear frequency scale mapping for voice conversion in text-to-speech system with cepstral description. *Speech Communication*, 48, 1691-1703.

[6] Přibil, J., Přibilová, A. (2008). Application of expressive speech in TTS system with cepstral description. In Esposito, A. et. al (eds.) *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction 2007*. Lecture Notes in Artificial Intelligence 5042. Berlin Heidelberg: Springer-Verlag, 201-213.

[7] Volaufová, J. (2005). Statistical methods in biomedical research and measurement science. *Measurement Science Review*, 5 (1), 1-10.

[8] Hartung, J., Makambi, H.K, Arcac, D. (2001). An extended ANOVA F-test with applications to the heterogeneity problem in meta-analysis. *Biometrical Journal*, 43 (2), 135-146.

[9] Welch, P.D. (1967). The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, AU-15, 70-73.

[10] Stoica, P., Moses, R.L. (1997). *Introduction to Spectral Analysis.* Prentice-Hall, 52-54

[11] Slifka, J., Anderson, T.R. (1995). Speaker modification with LPC pole analysis. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing.* Detroit, 644-647.

[12] Přibilová, A., Přibil, J. (2009). Spectrum modification for emotional speech synthesis. In Esposito, A. et. al. (eds.) *Multimodal Signals: Cognitive and Algorithmic Issues.* Lecture Notes in Artificial Intelligence 5398. Berlin Heidelberg: Springer-Verlag, 232-241.

[13] Přibilová, A., Přibil, J. (2009). Harmonic model for female voice emotional synthesis. In Fierrez, J. et al. (eds.) *Biometric ID Management and Multimodal Communication.* Lecture Notes in Artificial Intelligence 5707. Berlin Heidelberg: Springer-Verlag, 49-56.

[14] Scherer, K.R. (2003). Vocal communication of emotions: review of research paradigms. *Speech Communication*, 40, 227-256.

[15] Benesty, J., Sondhi, M.M., Huang, Y. (eds.) (2008). *Springer Handbook of Speech Processing*. Berlin Heidelberg: Springer-Verlag.

[16] Khuri, A.I., Mathew, T., Sinha, B.K. (1998). *Statistical Tests for Mixed Models*. New York: John Wiley & Sons.

[17] Srinivasan, S., DeLiang, W. (2010). Robust speech recognition by integrating speech separation and hypothesis testing. *Speech Communication,* 52, 72-81.

[18] Atassi, H., Smékal, Z. (2008). Real-time model for automatic vocal emotion recognition. In *Proceedings of the 31st International Conference on Telecommunications and Signal Processing*. Parádfürdö, Hungary, 21-25.

Received February 15, 2010.
Accepted June 7, 2010.