9

Spectral Properties and Prosodic Parameters of Emotional Speech in Czech and Slovak

Jiří Přibil[1] and Anna Přibilová[2]
[1]*Institute of Measurement Science, SAS,*
[2]*Faculty of Electrical Engineering & Information Technology, SUT,*
Slovakia

1. Introduction

The methods of analysis of human voice are based on the knowledge of speaker individuality. One of basic studies on speaker acoustic characteristics can be found in (Kuwabara & Sagisaka 1995). According to them the voice individuality is affected by the voice source (the average pitch frequency, the pitch contour, the pitch frequency fluctuation, the glottal wave shape) and the vocal tract (the shape of spectral envelope and spectral tilt, the absolute values of formant frequencies, the formant trajectories, the long-term average speech spectrum, the formant bandwidth). The most important factors on individuality are the pitch frequency and the resonance characteristics of the vocal tract, though the order of the two factors differs in different research studies.

According to (Scherer 2003) larynx and pharynx expansion, vocal tract walls relaxation, and mouth corners retraction upward lead to falling first formant and rising higher formants during pleasant emotions. On the other hand, larynx and pharynx constriction, vocal tract walls tension, and mouth corners retraction downward lead to rising first formant and falling higher formants for unpleasant emotions. Thus, the first formant and the higher formants of emotional speech shift in opposite directions in the frequency ranges divided by a frequency between the first and the second formant. In practice, the formant frequencies differ to some extent for different languages and their ranges are overlapped. According to (Stevens 1997) the frequency of vibration of the vocal folds during normal speech production is usually in the range 80 ÷ 160 Hz for adult males, 170 ÷ 340 Hz for adult females, and 250 ÷ 500 Hz for younger children. It means that female pitch frequencies are about twice the male pitch frequencies, pitch frequencies of younger children are about 1.5-times higher than those of females and about 3-times higher than those of males. As regards the formant frequencies, females have them on average 20 % higher than males, but the relation between male and female formant frequencies is nonuniform and deviates from a simple scale factor (Fant 2004).

Emotional state of a speaker is accompanied by physiological changes affecting respiration, phonation, and articulation. These acoustic changes are transmitted to the ears of the listener and perceived via the auditory perceptual system (Scherer 2003). From literature and our experiments follows that different types of emotions are manifested not only in prosodic patterns (F0, energy, duration) and several voice quality features (e.g. jitter, shimmer, glottal-to-noise excitation ratio, Hammarberg index) (Li et al. 2007) but also by significant

changes in spectral domain (Nwe et al. 2003). Several spectral features (spectral centroid, spectral flatness measure, Renyi entropy, etc.) quantify speaker-dependent as well as emotion-dependent characteristics of a speech signal (Hosseinzadeh & Krishnan 2008). It means these features provide information which complements the vocal tract characteristics. This paper describes analysis and comparison of basic spectral properties (values and ranges of cepstral coefficients, positions of formants), complementary spectral features (spectral flatness measure), and prosodic parameters (F0 and energy, microintonation, and jitter) of male and female acted emotional speech in Czech and Slovak languages. We perform statistical analysis for four emotional states: joy, sadness, anger, and a neutral state.

2. Subject and methods

Our experiments are aimed at statistical analysis and comparison of the spectral and prosodic features in emotional and neutral speech. It comprises comparison of basic statistical parameters (minimum, maximum, mean values, and standard deviation) and calculated histograms of distribution. Extended statistical parameters (skewness, kurtosis) are subsequently calculated from these histograms and/or the histogram can be evaluated by the analysis of variances (ANOVA) approach. Hypothesis tests are used for objective classification of neutral and different emotional styles.

2.1 Evaluation of results based on statistical analysis

The resulting parameters obtained from our analysis experiment in the form of histograms of distribution can be applied to visual classification (determination) of speech in different emotional states (typical shapes of particular histograms), or extended statistical parameters can be subsequently calculated from these histograms for objective matching. The skewness y and kurtosis k of a distribution is defined as

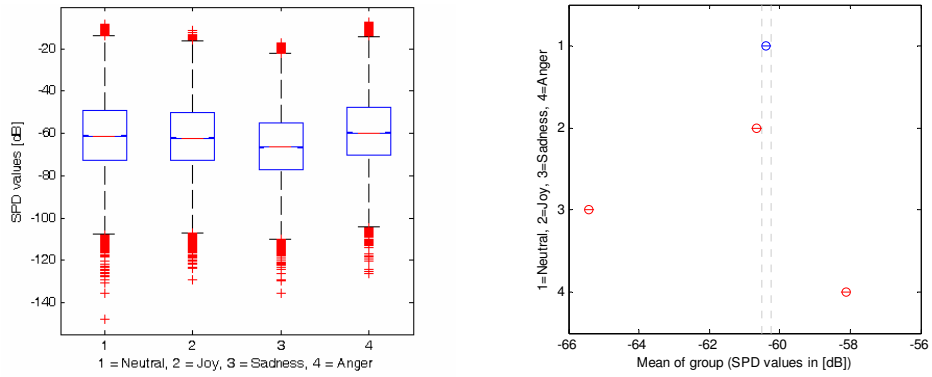
$$y = \frac{E(x-\mu)^3}{\sigma^3}, \qquad k = \frac{E(x-\mu)^4}{\sigma^4}, \tag{1}$$

where μ is the mean and σ is the standard deviation of the random variable X, and $E(t)$ represents the expected value of the quantity t. Skewness is a measure of asymmetry of the data around the sample mean. If skewness is negative, the data are spread out more to the left of the mean than to the right. If skewness is positive, the data are spread out more to the right. Kurtosis is a measure of how outlier-prone a distribution is. The kurtosis of the normal distribution is 3. Distributions that are more outlier-prone than the normal distribution have kurtosis greater than 3; distributions that are less outlier-prone have kurtosis less than 3. On the other hand, some definitions of kurtosis subtract 3 from the computed value, so that the normal distribution has kurtosis of 0 (Suhov & Kelbert 2005). We use this approach for calculation of kurtosis values in this study.

For objective comparison and matching the evaluation by ANOVA with multiple comparison of groups can be applied (Everitt 2006). This approach is more simple than recent speech recognition methods using evaluation by hidden Markov models (Srinivasan & DeLiang 2010) and is often used in other areas of biomedical research (Volaufova 2005), (Hartung et al. 2001). ANOVA gives also F statistics and results of the hypothesis test including probability values. Unlike the ANOVA F statistics, the Ansari-Bradley test (Suhov & Kelbert 2005) 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. The result is h = 0, if the null hypothesis of identical distributions cannot be rejected at the 5% significance level, or h = 1, if the null hypothesis can be rejected at the 5% level. The hypothesis test also returns the probability of observing the given result. Small values of this probability cast doubt on the validity of the null hypothesis.

Application of described evaluation approach is demonstrated on example of the Spectral Power Density (SPD) values in [dB] of spectrograms of the sentence "*Vlak už nejede*" (Czech male speaker) uttered in neutral and three emotional styles. Fig. 1a) contains the box plot of basic statistical parameters, Fig. 1b) shows visualization multiple comparison of group means applied to the results of ANOVA statistics - each group mean is represented by a symbol and an interval around the symbol. Three means are significantly different if their intervals are disjoint, and two groups ("Neutral" and "Joy") are not significantly different if their intervals overlap. Corresponding values of Ansari-Bradley test are stored in Table 1.



a) b)

Fig. 1. Box plot of basic statistical parameters (a), visualization of multiple comparison of group means applied to the results of ANOVA statistics (b).

h / p	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 1. Results of the Ansari-Bradley hypothesis test of values corresponding to multiple comparison of group means in Fig. 1b).

2.2 Analysis and evaluation of basic spectral properties

Speech spectrum is represented very well by a pole/zero model using cepstral coefficients in comparison with linear predictive coding (LPC) corresponding only to an all-pole approximation of the vocal tract. From the input samples of the speech signal (after segmentation and weighting by a Hamming window) the complex spectrum by the Fast Fourier Transform (FFT) algorithm is calculated. In the next step the powered spectrum is

computed and the natural logarithm is applied. Second application of the FFT algorithm gives the symmetric real cepstrum

$$\{c_n\} = \left\{ c_0, c_1, \ldots, c_{N_{FFT}/2} \mid c_{N_{FFT}/2-1}, \ldots, c_1 \right\} . \tag{2}$$

By limitation to the first N_0+1 coefficients, the Z-transform of the real cepstrum can be obtained

$$C(z) = c_0 + c_1 z^{-1} + c_2 z^{-2} + \ldots + c_{N_0} z^{-N_0}. \tag{3}$$

The truncated cepstrum represents an approximation of a log spectrum envelope

$$E(f) = c_0 + 2\sum_{n=1}^{N_0} c_n \cos(n \cdot 2\pi f). \tag{4}$$

The cepstral speech synthesis is performed by a digital filter implementing approximate inverse cepstral transformation. The system transfer function of this filter is given by an exponential relation where the exponent is the Z-transform of the truncated speech cepstrum and it represents the minimum phase approximation of the real cepstrum. Approximation of the system transfer function can be performed by a cascade connection of N_0 elementary filter structures. Using the Padé approximation of the exponential function it has been found out that the minimum number of N_0 (25/50 at 8/16 kHz sampling frequency) cepstral coefficients is necessary for sufficient approximation error (Vích 2000). As the value range of the cepstral coefficients exponentially falls, only the first eight coefficients are analyzed (the remaining coefficients practically have not influence on the filter stability, structure, and implementation).

The basic cepstral analysis scheme including the spectral features calculation is shown in the block diagram in Fig. 2. Described method of cepstral speech analysis was supplied with determination of the fundamental frequency F0 and the energy En (calculated from the first cepstral coefficient c_0). After removal of the low energy starting and ending frames by the energy threshold (En_{min}) the limited working length (number of frames) for next processing was obtained – see Fig. 3. Cepstral analysis must be preceded by classification and sorting process of the cepstral coefficients in dependence on the voice type (male / female) and the speech style (neutral / emotional). Realization of analysis of the cepstral coefficient properties was processed in following phases:

a. manual (subjective) classification of voice type and emotional speech style, further automatic processing,
b. cepstral analysis of speech signal (from the main two speech databases consisting of short utterances of male/female voice pronounced in neutral and different emotional styles).

As a graphical output, the histogram of cepstral coefficients for every emotional state was also constructed. For objective comparison, the extended statistical parameters of skewness and kurtosis were subsequently calculated. The performed statistical analysis of cepstral coefficients consists of four parts:

1. determination of basic statistical parameters of the cepstral coefficients (minimum, maximum, mean value, and standard deviation),
2. calculation and building of histograms,
3. calculation of extended statistical parameters from histograms (kurtosis and skewness),
4. comparison of the mean values and the ranges of the cepstral coefficients for emotional and neutral states.

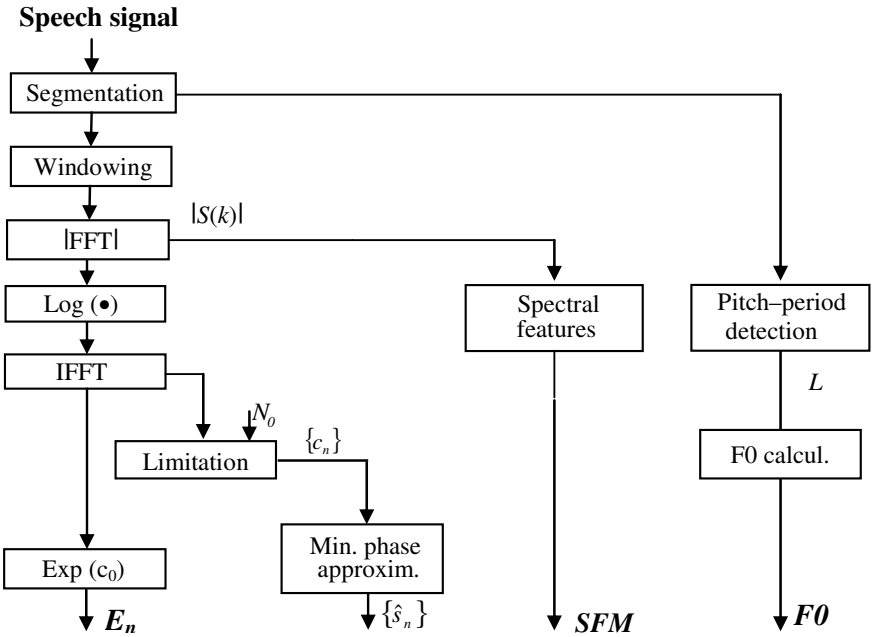


Fig. 2. Block diagram of used cepstral analysis method (N_0 = 50, and the sampling frequency f_s = 16 kHz).

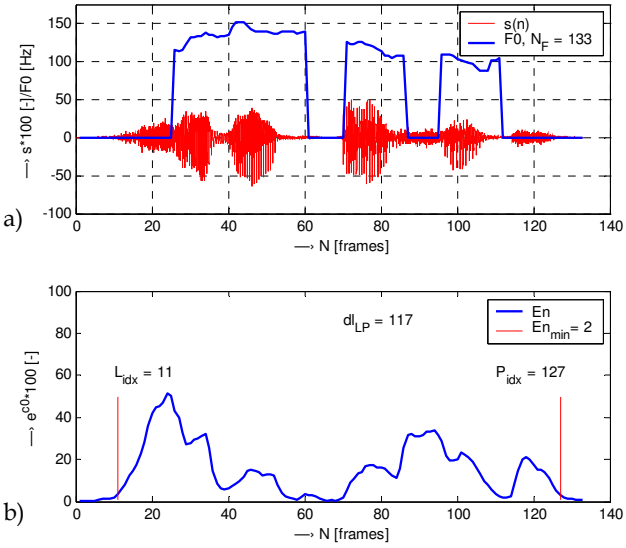


Fig. 3. The processed sentence "Život a řeč" *(Life and speech)*, Czech male speaker: speech signal with F0 contour (a), *En* contour calculated from the first cepstral coefficient c_0 (b).

In frequency domain we analyze the first three formant positions (F1, F2, and F3), and the difference between smoothed spectra for comparison of analyzed speech on segmental or phoneme level. Smoothed spectra are computed from the chosen region of interest (ROI) areas of voiced part of speech by the Welch method (Oppenheim et al. 1999). From these mean periodograms the first three formats are determined as the first three local maxima of the Welch's periodogram where its gradient changes from positive to negative.

From the summary comparison of cepstral speech analysis follows that emotional speech brings about the most significant spectral changes for voiced speech (see spectrogram in Fig. 4) therefore the extended analysis by mean periodograms of sounds was subsequently performed. For this purpose the second database consisting vowels "a", "e", "i", "o", "u" and voiced consonants "m", "n" and "l" was used. The whole spectral analysis with the help of Welch's periodograms was practically performed in five steps:

1. calculation of smoothed spectra in the form of Welch's periodograms from the selected ROIs of voiced part of neutral and emotional speech signal,
2. determination of the first three formant positions (F1, F2, and F3) from the obtained periodograms,
3. calculation of mean emotional-to-neutral formant position ratios,
4. calculation and visual comparison of mean periodograms of sounds from the database of vowels and voiced consonants,
5. numerical matching of results from the calculated spectral distances between corresponding periodograms by the RMS method.

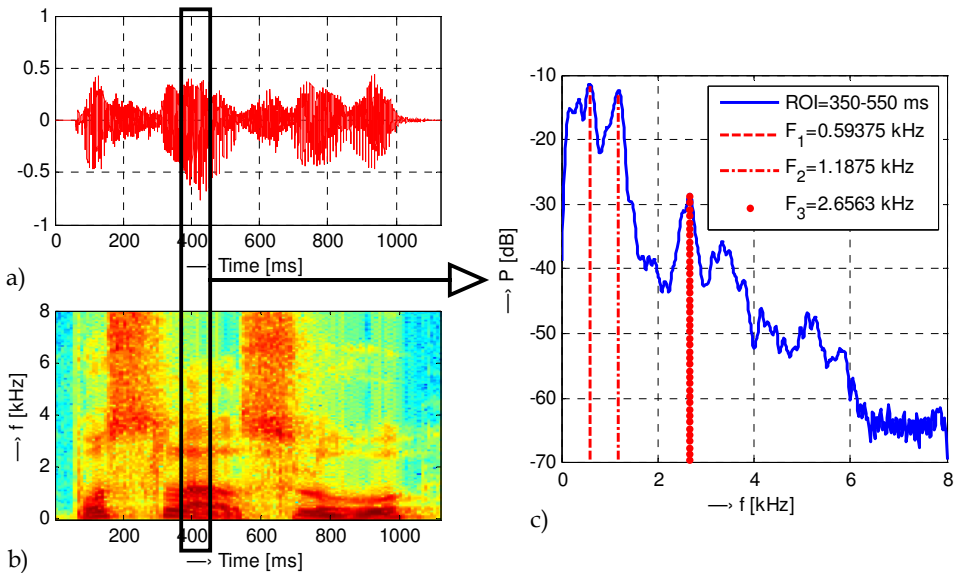


Fig. 4. Processed sentence "Poslal sluhu" (*He sent his servant*), Slovak male speaker: speech signal (a), corresponding spectrogram (b), calculated mean periodogram estimate in [dB] of selected ROI with determined formant positions F1,F2, and F3 (c).

2.3 Analysis of complementary spectral feature

As a complementary spectral feature, the spectral flatness measure (SFM) was analyzed. This spectral feature is calculated during cepstral speech analysis (see block diagram in Fig. 2) using absolute value of the fast Fourier transform denoted as $|S(k)|$

$$SFM = \frac{\left[\prod_{k=1}^{N_{FFT}/2} |S(k)|^2\right]^{\frac{2}{N_{FFT}}}}{\frac{2}{N_{FFT}} \sum_{k=1}^{N_{FFT}/2} |S(k)|^2}. \tag{5}$$

According to psychological research of emotional speech different emotions are accompanied by different spectral noise (Scherer et al. 2003). In cepstral speech synthesis the spectral flatness measure SFM was used to determine voiced/unvoiced energy ratio in voiced speech analysis (Vích 2000). The SFM values lie generally in the range of $(0 \div 1)$ – the zero value represents totally voiced signal (for example pure sinusoidal signal); in the case of SFM = 1, the totally unvoiced signal is classified (for example white noise signal). According to the statistical analysis of the Czech and Slovak words the ranges of SFM = $(0 \div 0.25)$ for voiced speech frames and SFM = $(0 \div 0.75)$ for unvoiced frames were evaluated (Madlová & Přibil 2000). The demonstration example in Fig. 5 shows the input speech signal with detected pitch frequency F0 and calculated SFM values with voiceness classification.

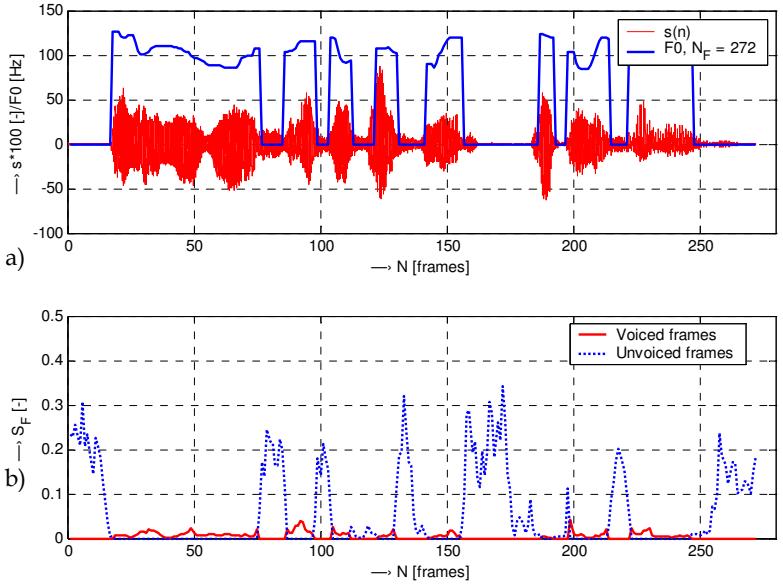


Fig. 5. Demonstration of SFM calculation: input speech signal – sentence "Lenivý si a zle gazduješ" (*You are lazy and you keep your house ill*) pronounced in angry emotional style, male Slovak speaker with F0 contour (a), SFM values for voiced and unvoiced frames (b).

In our algorithm, the values of SFM are obtained from the voiced speech frames and are separately processed in dependence on voice type (male / female). For every voice type the SFM values were subsequently sorted by emotional styles and stored in separate stacks. These classification operations were performed manually, by subjective listening method. Next operations with the stacks were performed automatically – calculation of statistical parameters: minimum, maximum, mean values, and standard deviation. From the mean spectral feature values the ratio between emotional and neutral states is subsequently calculated. As a graphical output used for visual comparison (subjective method) the histogram of sorted spectral features values for each of the stacks is also calculated. Consequently the extended statistical parameters of histograms (skewness and kurtosis) were subsequently calculated. The second approach based on ANOVA was applied to SFM values together with multiple comparison of groups test as an objective evaluation method– see block diagram in Fig. 6.

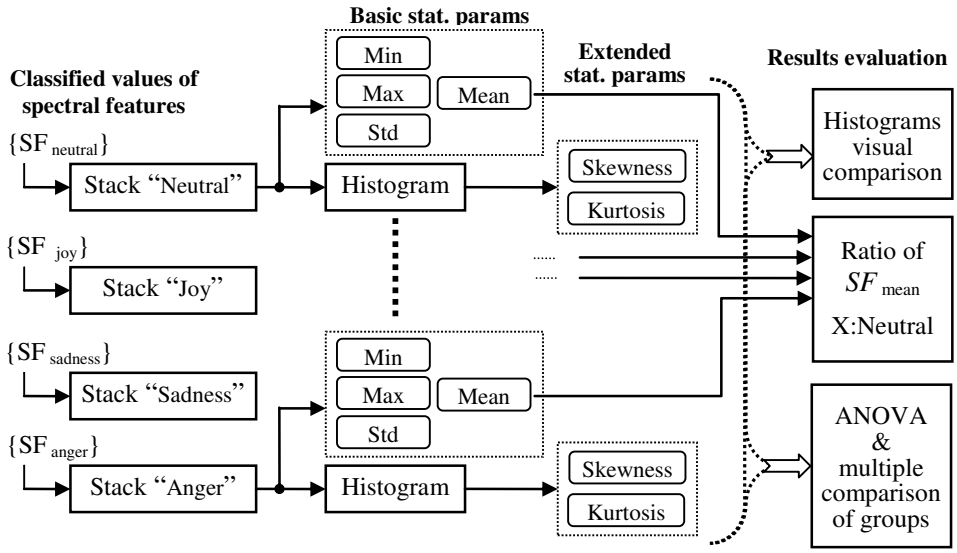


Fig. 6. Block diagram of processed operations with the stacks filled with classified spectral feature values.

2.4 Analysis of prosodic parameters and their comparison

Melody of speech utterances is given by the fundamental frequency (F0) contour. Microintonations as well as jitter together with the sentence melody and the word melody also represent the speech melody. Microintonation can be supposed to be a random, band-pass signal described by statistical parameters.

The whole prosodic parameter analysis procedure is divided into four phases:

1. analysis of the speech signal: determination of F0 and energy contour,
2. analysis of F0 contour, microintonation extraction, determination of pitch periods in the voiced parts of the speech signal – see example in Fig. 7,
3. statistical evaluation of F0, energy, microintonation, zero crossings, and calculation of ratios for emotional / neutral states (see block diagram in Fig. 8),

4. microintonation signal spectral analysis and 3-dB bandwidth (B_3) determination from the concatenated signal.

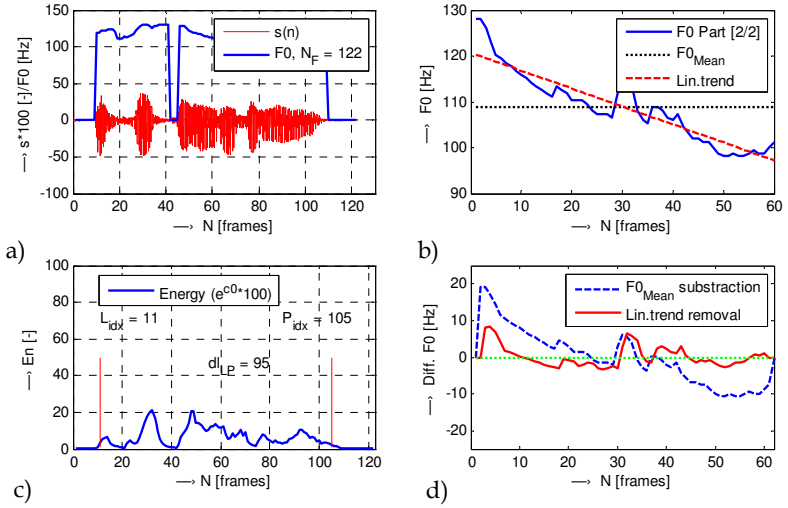


Fig. 7. Demonstration of microintonation analysis: speech signal with F0 contour (a), the second voiced part: original F0, mean F0, and LT (b), energy contour (c), differential signal after $F0_{mean}$ and LT subtraction (d) – the sentence "Řekl Radomil" *(Radomil said)* uttered by a male Czech speaker.

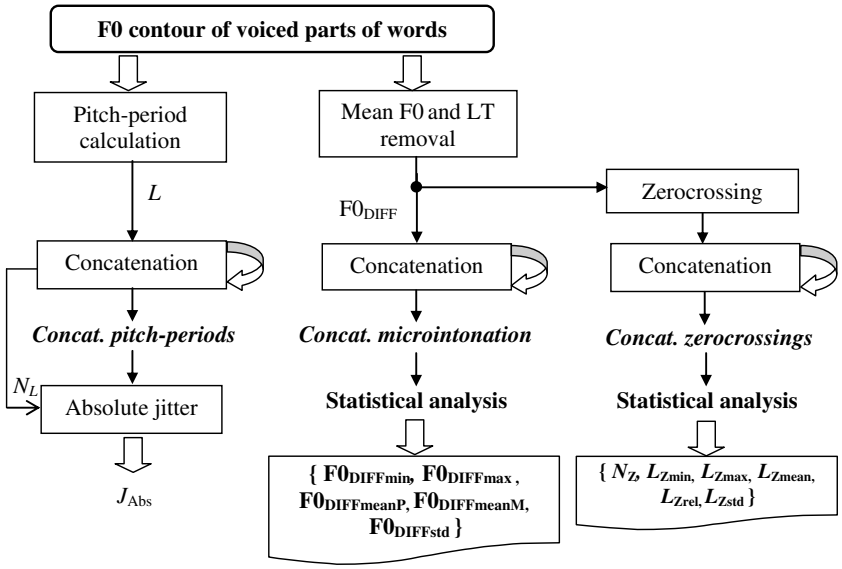


Fig. 8. Block diagram of microintonation signal analysis including basic and zero crossing statistical analysis.

The introductory microintonation processing phase consists of the following steps:

1. determination of the melody contours from the voiced parts of speech smoothed by a median filter,
2. determination of F0 mean values, $F0_{range}$ (as difference of minimum and maximum) and calculation of the linear trend (LT) by the mean square method,

$$LT\ (a,b) = a + b\ n, \tag{6}$$

where $n = 1,2,...N_F$ and N_F is number of frames of the F0 contour. The best linear fit to a given set of F0 values is solved by least squares fitting technique of linear regression yielding

$$a = \frac{\sum_{n=1}^{N_F} F0(n) \sum_{n=1}^{N_F} n^2 - \sum_{n=1}^{N_F} n \sum_{n=1}^{N_F} n\ F0(n)}{N_F \sum_{n=1}^{N_F} n^2 - \left(\sum_{n=1}^{N_F} n\right)^2}, \quad b = \frac{N_F \sum_{n=1}^{N_F} n\ F0(n) - \sum_{n=1}^{N_F} n \sum_{n=1}^{N_F} F0(n)}{N_F \sum_{n=1}^{N_F} n^2 - \left(\sum_{n=1}^{N_F} n\right)^2}, \tag{7}$$

3. calculation of differential microintonation signal $F0_{DIFF}$ by subtraction of these values from the corresponding F0 contours ($F0_{mean}$ and LT removal) – see Fig. 7b)

$$F0_{DIFF}(n) = \left(F0(n) - F0_{mean}\right) - LT(n), \tag{8}$$

4. calculation of the absolute jitter values J_{Abs}, as the average absolute difference between consecutive pitch periods L measured in samples (Farrús et al. 2007)

$$J_{Abs} = \frac{1}{f_s (N_L - 1)} \sum_{n=1}^{N_L - 1} \left| L_n - L_{n+1} \right|, \tag{9}$$

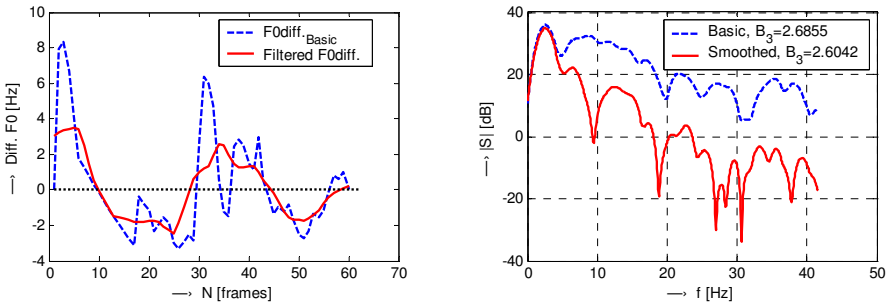
where f_s is the sampling frequency and N_L is the number of extracted pitch periods,

5. detection of zero crossings, calculation of zero crossing periods L_Z.

Spectral analysis of concatenated differential microintonation signal is also carried out for all emotions. This analysis phase is divided into three steps:

1. Calculation of the frequency parameters from the zero crossing periods $L_{Zx} = \{L_{Zmin}, L_{Zmax}, L_{Zmean}, L_{Zrel}, L_{Zstd}\}$ as $F_{Zxl} = f_F / (2\ L_{Zx})$, where f_F is the frame frequency.
2. Microintonation signal spectral analysis by periodogram averaging using the Welch method.
3. Determination of B_3 values from these spectra for each of the emotion types.

To obtain spectrum of smoothed microintonation signal (see Fig. 9b), the concatenated differential F0 signal is filtered by a moving average (MA) filter of the length M_F (voiced parts shorter than M_F+2 frames are not processed in further analysis) – see Fig. 9a).



a) b)

Fig. 9. Demonstration of microintonation smoothing and spectrum determination (obtained from the same sentence as in Fig. 7): basic differential F0 signal and the one filtered by moving average (a), corresponding spectra and their 3-dB bandwidths B_3 (b).

3. Material, experiments and results

As follows from our previous experiments, the basic spectral properties (cepstral coefficients and formant positions) as well as the complementary spectral features depend on a speaker but they do not depend on nationality (it was confirmed that it holds for the Czech and Slovak languages). Therefore, the created speech database consists of neutral and emotional sentences uttered by several speakers (extracted from the Czech and Slovak stories performed by professional actors). The speech material was collected in two databases (separately from male – 134 sentences, and female voice – 132 sentences, 8+8 speakers altogether) consisting of sentences with duration from 0.5 to 5.5 seconds, resampled at 16 kHz representing four emotional states (sad, joyful, angry, and neutral for comparison). Classification of emotional states was carried out manually by subjective listening method.

The F0 vales (pitch contours) were given by autocorrelation analysis method (Oppenheim 1999) with experimentally chosen pitch-ranges by visual comparison of testing sentences (one typical sentence from each of emotions and voice classes) as follows: 35÷250 Hz for male, and 105÷350 Hz for female voices. The F0 values were next compared and corrected by results obtained with the help of the PRAAT program (Boersma & Weenink 2008) with similar internal settings of F0 values.

Speech signal analysis was performed for total number of 25988 frames (8 male speakers) and 24017 frames (8 female speakers). The formant positions and the spectral flatness values were determined only for the voiced frames (totally 11639 of male and 13464 of female voice). In the case of prosodic parameters analysis, the minimum length of the processed voiced parts was set to 10 frames and the corresponding length of $M_F = 8$ for moving average filter was chosen. Number of analyzed voiced parts / voiced frames) was in total:

a. Male: neutral - 112/2698, joy - 79/1927, sadness - 128/3642, anger - 104/ 2391.
b. Female: neutral - 86/2333, joy - 87/2541, sadness - 92/2203, anger - 91/2349.

3.1 Results of analysis of basic spectral properties

Results of determined basic statistical parameters of the first 8 cepstral coefficients for different speech styles are shown in the form of box plot graph in Fig. 10 (male voice).

Summary histograms of cepstral coefficients (c_1-c_8) are shown in Fig. 11 and comparison of histogram contours for different emotions of cepstral coefficients (c_1-c_4) is shown in Fig. 12 (both male voice). Table 2 contains values of kurtosis parameters and Table 3 contains values of skewness obtained from the compared histograms of c_1-c_4 (for male and female voices).

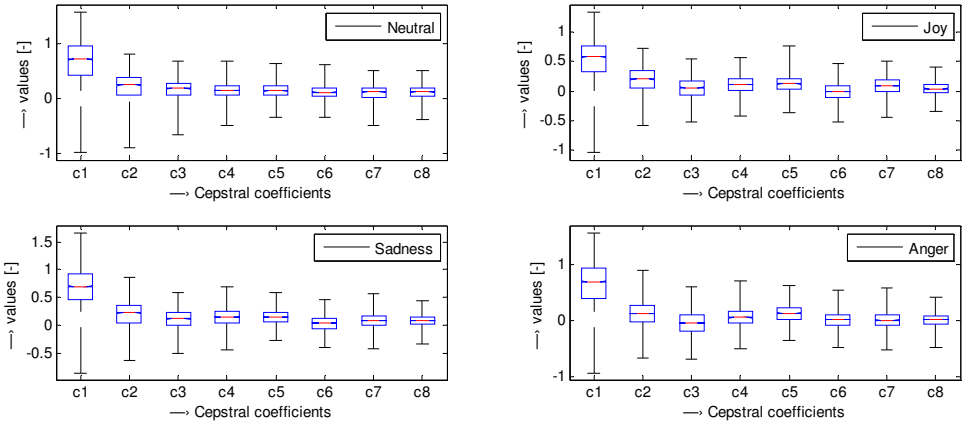


Fig. 10. Box plot of basic statistical parameters of the first 8 cepstral coefficients for different speech styles - male voice.

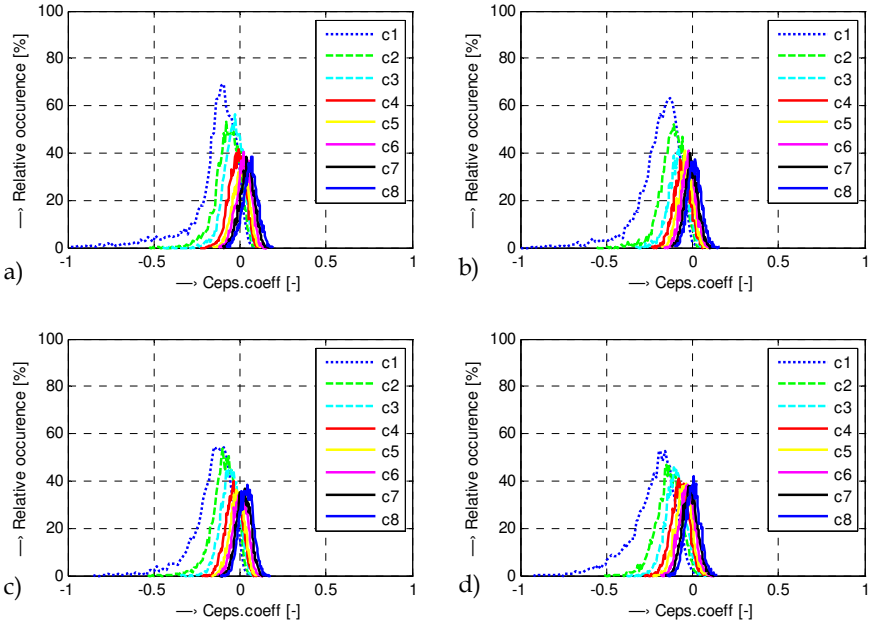


Fig. 11. Histograms of the first 8 cepstral coefficients for different speech styles (male voice): "neutral" speech (a), "joy" (b), "sadness" (c), and "anger" (d).

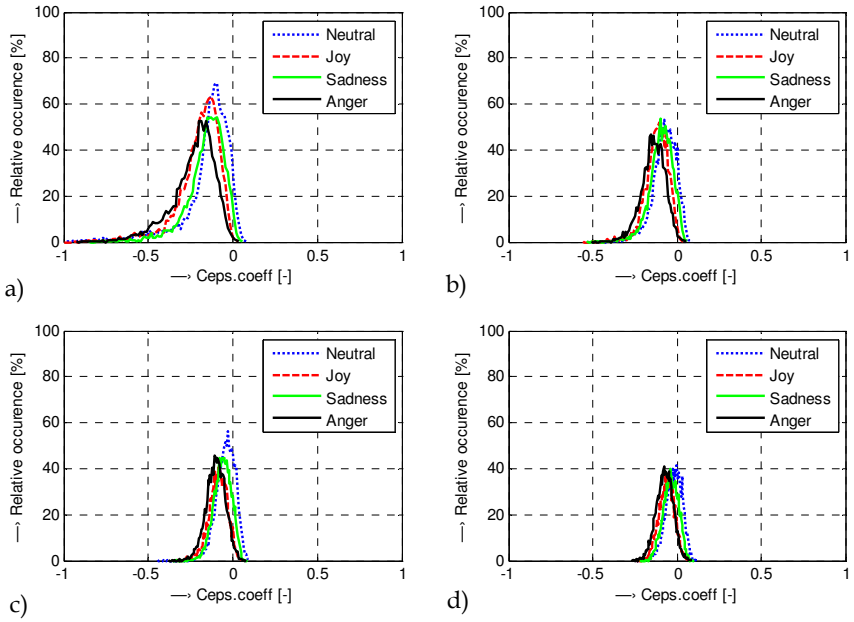


Fig. 12. Histogram comparison for different speech styles (male voice): for cepstral coefficients c_1 (a), coefficients c_2 (b), coefficients c_3 (c), and coefficients c_4 (d).

Emotion	Male voice				Female voice			
	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
Neutral	3.93	1.36	1.17	0.65	7.82	5.86	1.42	-0.26
Joy	2.53	0.91	0.31	0.01	6.16	3.86	0.11	-0.09
Sadness	1.72	0.82	0.29	-0.06	4.03	2.81	-0.09	-0.24
Anger	1.12	0.01	0.11	0.04	2.78	1.45	0.11	-0.04

Table 2. Kurtosis parameters determined from histograms of c_1-c_4 cepstral coefficients for male and female voices.

Emotion	Male voice				Female voice			
	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
Neutral	-1.79	-0.99	-0.93	-0.73	-2.47	-1.54	-0.63	-0.14
Joy	-1.20	-0.64	-0.42	-0.22	-2.00	-1.65	-0.33	-0.21
Sadness	-1.03	-0.75	-0.46	-0.13	-1.64	-1.45	-0.26	-0.13
Anger	-0.84	-0.36	-0.12	0.09	-1.42	-1.09	-0.29	-0.16

Table 3. Skewness parameters determined from histograms of c_1-c_4 cepstral coefficients for male and female voices.

Results of basic statistical parameters for the first three formant positions F1, F2, and F3 of male and female voice in neutral speech are shown are shown in Fig. 13, detailed histograms of distribution are shown in Fig. 14. Comparison of histograms of F1, F2, and F3 for different

speech styles are introduced in Fig. 15 (male voice) and Fig. 16 (female voice). Table 4 contains values of kurtosis parameters and Table 5 contains values of skewness obtained from the compared histograms of F1, F2, and F3 (for male and female voices). Summary results of mean neutral-to-emotional formant position ratios for both voices can be seen in Table 6.

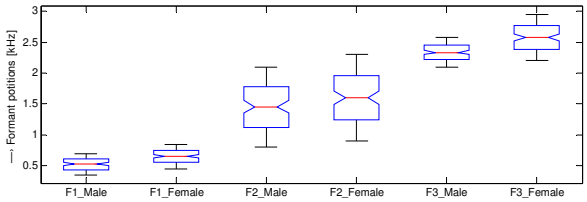


Fig. 13. Box plot of basic statistical parameters of analysis of the first three format positions (male and female voice, neutral speech style).

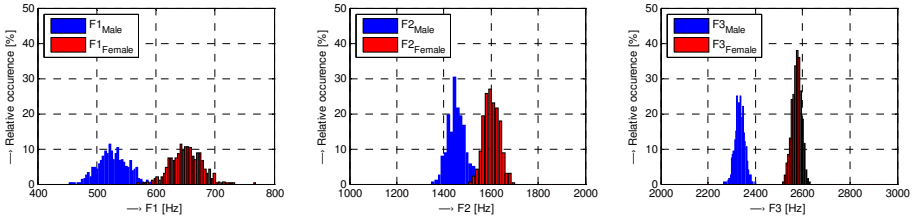


Fig. 14. Detailed histograms of the first three format positions – male and female voice, neutral speech style.

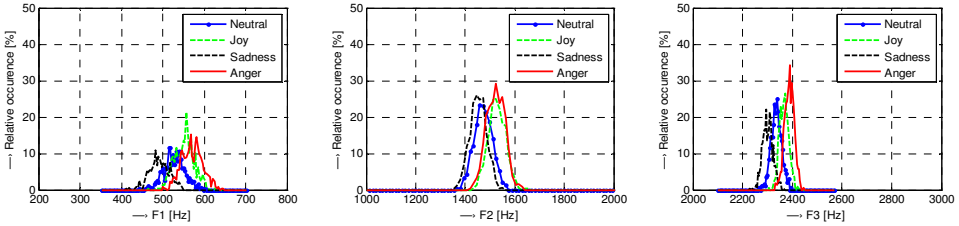


Fig. 15. Comparison of histograms of the first three formant positions for different speech styles – male voice.

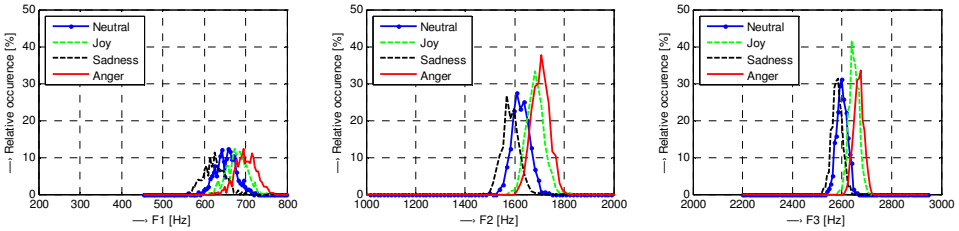


Fig. 16. Comparison of histograms of the first three formant positions for different speech styles – female voice.

Emotion	$F_{1\ male}$	$F_{2\ male}$	$F_{3\ male}$	$F_{1\ female}$	$F_{2\ female}$	$F_{3\ female}$
Neutral	2.702	9.257	4.884	2.771	11.440	9.823
Joy	3.298	10.457	5.253	2.613	12.155	10.376
Sadness	2.095	10.655	5.406	2.594	13.919	10.257
Anger	1.694	8.263	5.389	2.054	11.121	10.476

Table 4. Kurtosis parameters determined from histograms of the first three formant positions for male and female voices.

Emotion	$F_{1\ male}$	$F_{2\ male}$	$F_{3\ male}$	$F_{1\ female}$	$F_{2\ female}$	$F_{3\ female}$
Neutral	-1.081	0.207	-0.554	-0.999	0.532	0.266
Joy	-1.044	0.372	-0.439	-1.029	0.596	0.369
Sadness	-1.195	0.404	-0.458	-1.101	0.777	0.352
Anger	-1.297	0.100	-0.435	-1.141	0.471	0.349

Table 5. Skewness parameters determined from histograms of the first three formant positions for male and female voices.

Formant ratio	F_{1male}	F_{2male}	F_{3male}	$F_{1female}$	$F_{2female}$	$F_{3female}$
Joyous: neutral	0.712	1.025	1.038	0.898	1.082	1.049
Sad: neutral	1.043	0.813	0.899	1.353	0.948	0.938
Angry: neutral	1.123	0.795	0.762	1.282	0.885	0.887

Table 6. Summary results of mean emotional-to- neutral formant position ratios.

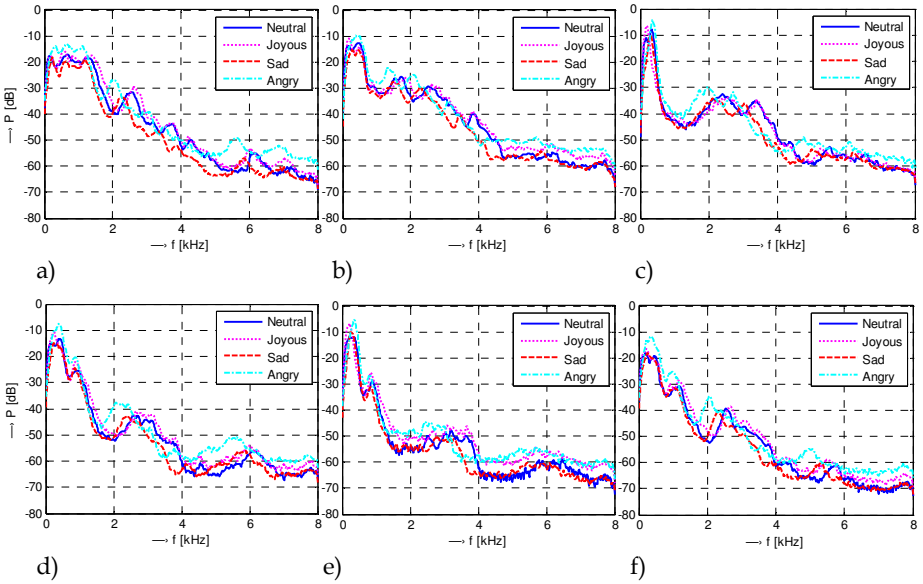


Fig. 17. Mean periodograms of analyzed voiced speech parts corresponding to sounds: "a" (a), "e" (b), "i" (c), "o" (d), "u" (e), and "l" (f) – male voice.

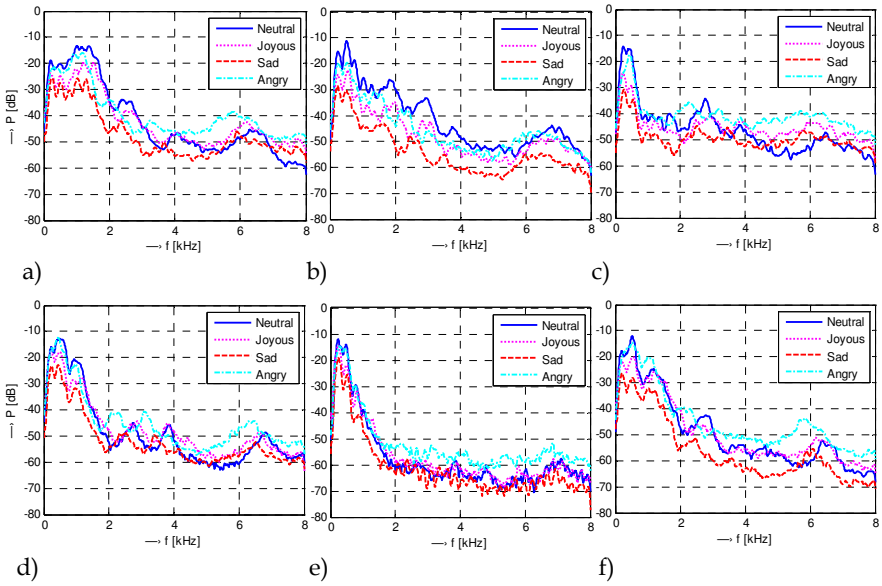


Fig. 18. Mean periodograms of analyzed voiced speech parts corresponding to sounds: "a" (a), "e" (b), "i" (c), "o" (d), "u" (e), and "l" (f) – female voice.

Results of extended analysis of formant positions by Welch's periodograms of sounds from the database of vowels and voiced consonants (detailed periodograms corresponding to sounds "a", "e", "i", "o", "u" and "l") are shown in Fig. 17 (male voice) and Fig. 18 (female voice). The spectral distances calculated between mean periodograms in "neutral" and emotional styles are summarized in Table 7.

Neutral – to:	Male voice			Female voice		
	joyous	sad	angry	joyous	sad	angry
D_{RMS} of "a" [dB]	2.517	4.516	5.845	4.598	7.223	8.382
D_{RMS} of "e" [dB]	2.608	3.551	4.862	5.708	6.599	13.012
D_{RMS} of "i" [dB]	2.427	3.769	5.279	5.794	7.236	8.060
D_{RMS} of "o" [dB]	2.710	3.639	6.110	3.841	5.866	6.370
D_{RMS} of "u" [dB]	3.509	4.596	5.846	2.692	4.973	6.298
D_{RMS} of "m"[dB]	2.205	4.809	6.595	4.186	4.724	4.774
D_{RMS} of "n"[dB]	2.160	3.852	4.615	3.985	6.597	8.909
D_{RMS} of "l" [dB]	2.839	3.408	6.063	3.207	6.929	8.064

Table 7. Summary results of spectral distances of analyzed sounds (D_{RMS} are calculated between periodograms of "neutral" and emotional styles) for male and female voices.

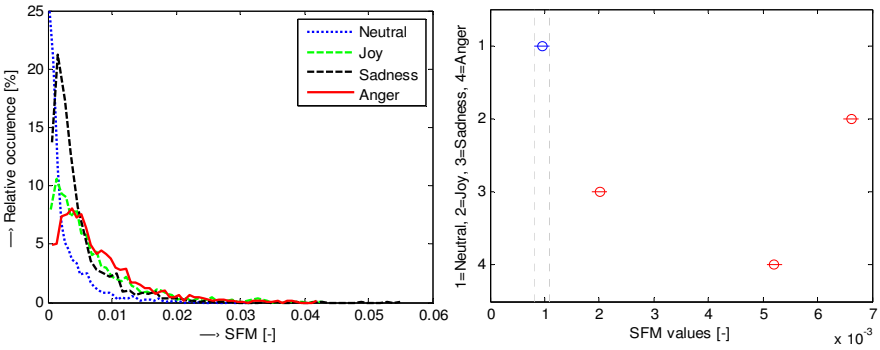
3.2 Results of analysis of a complementary spectral feature

The results of basic statistical parameters of the spectral flatness values for male and female voice analysis determined only from the voiced frames are summarized in Table 8. Histograms of *SFM* values for different emotions together with visualization of the difference between group means calculated using ANOVA statistics are shown in Fig. 19 (male voice) and Fig. 20

(female voice). Corresponding values of Ansari-Bradley test are stored in Table 9 (male voice) and Table 10 (female voice). The main result – mean spectral flatness value ratios between different emotional states and a neutral state – is given in Table 11.

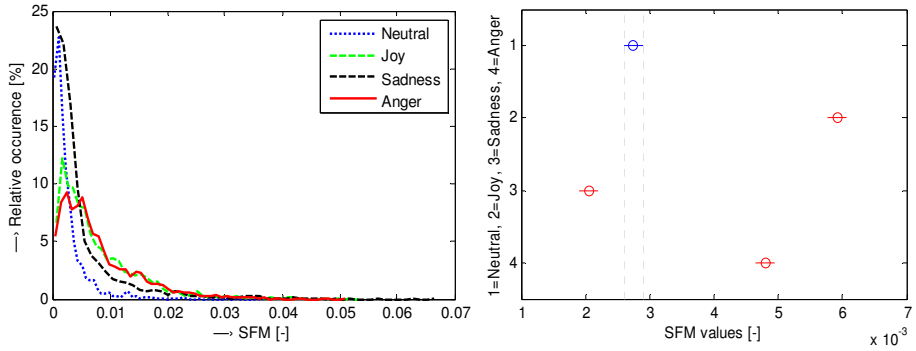
Emotion	Male voice				Female voice			
	mean	min	max	std	mean	min	max	std
Neutral	0.00286	$3.78 \cdot 10^{-5}$	0.03215	0.00364	0.00274	$3.15 \cdot 10^{-5}$	0.03731	0.00346
Joy	0.00662	$1.36 \cdot 10^{-4}$	0.04327	0.00650	0.00784	$2.07 \cdot 10^{-4}$	0.05414	0.00726
Sadness	0.00444	$1.12 \cdot 10^{-4}$	0.05540	0.00462	0.00506	$9.48 \cdot 10^{-5}$	0.06694	0.00674
Anger	0.00758	$2.28 \cdot 10^{-4}$	0.04228	0.00614	0.00807	$1.41 \cdot 10^{-4}$	0.05129	0.00692

Table 8. Summary results of basic statistical analysis of the spectral flatness values for male and female voice, voiced frames.



a) b)

Fig. 19. Histograms of SFM values for different speech styles (a), the difference between group means with the help of ANOVA statistics (b) - male voice, voiced frames.



a) b)

Fig. 20. Histograms of SFM values for different speech styles (a), the difference between group means with the help of ANOVA statistics (b) - female voice, voiced frames.

h / p	Neutral	Joy	Sadness	Anger
Neutral	**0/1**	1/3.60 10^{-20}	1/0.002	1/2.10 10^{-11}
Joy		**0/1**	1/3.95 10^{-46}	1/1.21 10^{-35}
Sadness			**0/1**	1/1.52 10^{-39}
Anger				**0/1**

Table 9. Results of the Ansari-Bradley hypothesis test of SFM values corresponding to Fig. 19b).

h / p	Neutral	Joy	Sadness	Anger
Neutral	**0/1**	1/2.78 10^{-12}	1/0.0015	1/8.34 10^{-9}
Joy		**0/1**	1/5.57 10^{-26}	1/2.08 10^{-15}
Sadness			**0/1**	1/4.27 10^{-11}
Anger				**0/1**

Table 10. Results of the Ansari-Bradley hypothesis test of SFM values corresponding to Fig. 20b).

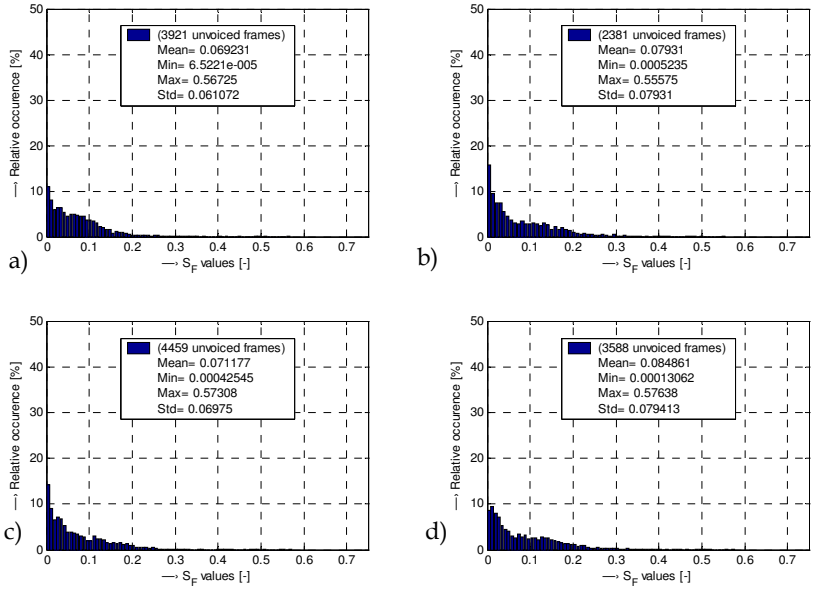


Fig. 21. Histograms of spectral flatness values calculated from the unvoiced frames (male voice): "neutral" style (a), and emotions - "joy" (b), "sadness" (c), and "anger" (d).

mean *SFM* **ratio**	**joy**:neutral	**sadness**:neutral	**anger**:neutral
male voice	1.349	1.725	1.321
female voice	1.455	1.795	1.377

Table 11. Mean spectral flatness values ratios between different emotional states and a neutral state (for voiced frames only).

3.3 Results of analysis of prosodic parameters

F0 contour was created from the frames with energy exceeding a chosen threshold. Histograms of F0 values distribution are shown in Fig. 22, the basic statistical parameters of $F0_{DIFF}$ and L_Z values for male and female voices of neutral and emotional speech as the box plots are presented in Fig. 23. Results of statistical analysis of energy contours (calculated from the first cepstral coefficient c_0) are shown in Fig. 24 and Table 12 consists of energy and absolute jitter ratios for emotional speech styles and neutral style for both voices. In Table 13, there are stored together mean $F0_{DIFF}$ values and absolute jitter values. Resulting summary emotional-to-neutral ratios of mean $F0_{DIFF}$ and $F0_{RANGE}$ for male and female voice are in Table 14.

Results of basic statistical analysis of zero crossing periods L_Z are shown in Table 15. For objective matching of L_Z the ANOVA and multiple comparison of group means together with the Ansari-Bradley test were performed – see Fig. 25 and results in Tables 16 to 18. Zero crossing periods were next used to calculate microintonation signal spectral analysis. Summary results including the 3-dB bandwidth values for male and female voices are shown in Table 19. The average microintonation spectra can be seen in Fig. 26 (male voice) and Fig. 27 (female voice).

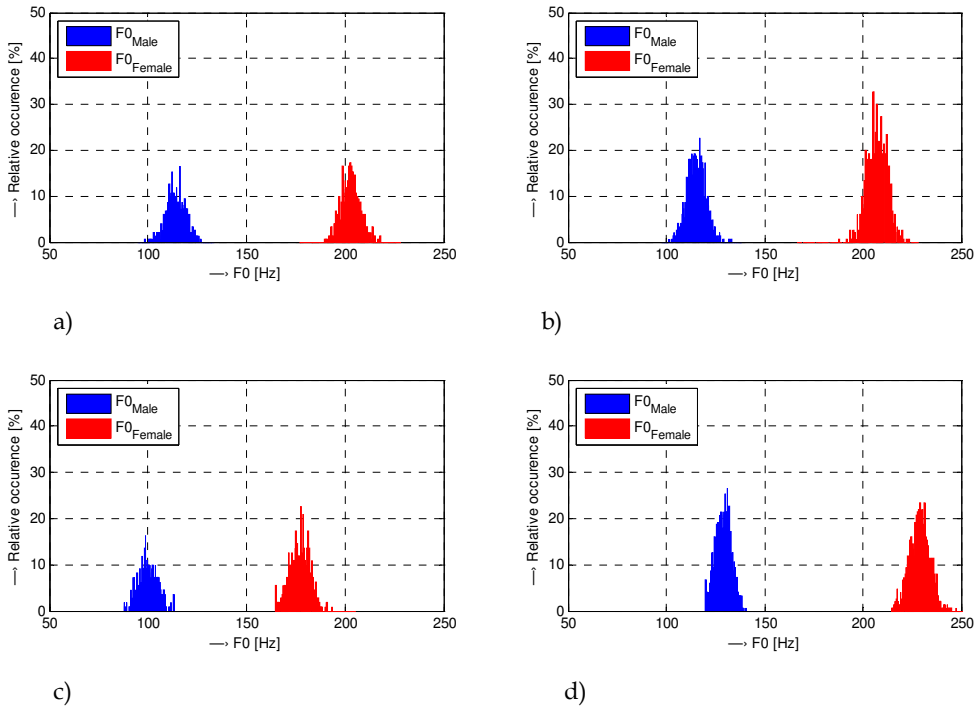
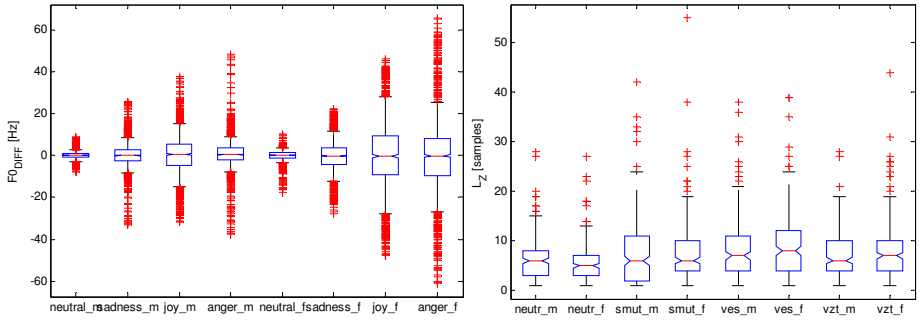


Fig. 22. Histograms of F0 values for male and female voices in "neutral" style (a), and emotions "joy" (b), "sadness" (c), and "anger" (d).



a) b)

Fig. 23. Box plot of basic statistical parameters of $F0_{DIFF}$ (a) and L_Z values (b) for male and female voices.

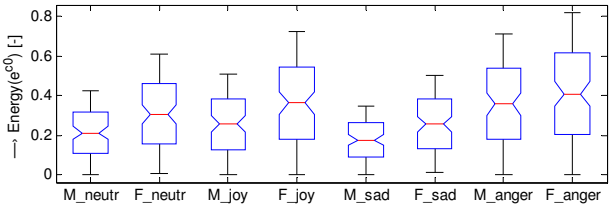


Fig. 24. Results of statistical analysis of energy contours (calculated from the first cepstral coefficient c_0): male / female voice, neutral / emotional states.

Voice	energy ratio X:neutral			J_{Abs} ratio X:neutral		
	joy	sadness	anger	joy	sadness	anger
Male	1.32	0.95	1.70	2.45	1.55	2.77
Female	1.50	0.73	1.84	1.94	1.41	2.06

Table 12. Summary male and female energy and absolute jitter ratios between different emotional states and a neutral state.

Emotion	$F0_{DIFFmean}$ male	$F0_{DIFFmean}$ female	J_{Abs} male	J_{Abs} female
Neutral	2.66	3.67	0.29	0.17
Joy	7.27	8.49	0.71	0.33
Sadness	4.02	6.29	0.45	0.24
Anger	8.62	10.16	0.60	0.35

Table 13. Mean values of differential F0 in [Hz] (calculated from positive microintonation values) together with absolute jitter values (in [ms]).

F0 ratio	$F0_{mean}$ joy	$F0_{mean}$ sadness	$F0_{mean}$ anger	$F0_{range}$ joy	$F0_{range}$ sadness	$F0_{range}$ anger
male voice	1.18	0.81	1.16	1.25	0.62	1.30
female voice	1.32	0.79	1.27	1.52	0.65	1.68

Table 14. Summary male and female F0 parameters modification ratio values between emotional and neutral speech.

Emotion	Male voice			Female voice		
	L_{Zmax}	L_{Zmean}	L_{Zstd}	L_{Zmax}	L_{Zmean}	L_{Zstd}
Neutral	57	6.82	5.69	40	6.64	5.23
Joy	23	6.74	4.57	28	5.26	3.78
Sadness	59	8.26	6.52	40	6.69	5.43
Anger	26	6.04	4.19	30	6.32	4.43

Table 15. Summary results of zero crossing basic statistical analysis (zero crossing period L_Z parameters in [frames]) – male and female voice, $L_{Zmin} = 1$.

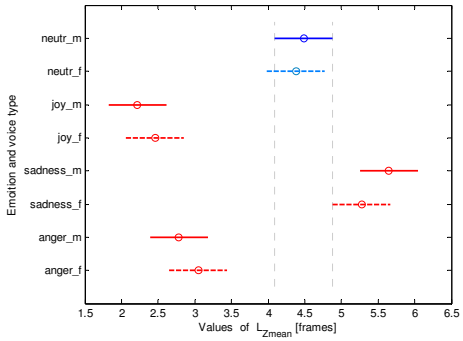


Fig. 25. Graphical results of zero crossing periods L_Z multiple comparison of ANOVA (male and female voice groups) for corresponding emotions.

h / p	Neutral	Joy	Sadness	Anger
Neutral	**0/1**	1 / $3.37\ 10^{-7}$	1 / 0.035	1 / 0.066
Joy		**0/1**	1 / $8.99\ 10^{-7}$	1 / 0.006
Sadness			**0/1**	1 / 0.021
Anger				**0/1**

Table 16. Partial results of zero crossing periods L_Z Ansari-Bradley hypothesis test based on comparison of distributions – male voice group.

h / p	Neutral	Joy	Sadness	Anger
Neutral	0/1	$1 / 4.8810^{-15}$	1 / 0.006	1 / 0.017
Joy		**0/1**	1 / 0.002	$1 / 4.01.10^{-8}$
Sadness			**0/1**	1 / 0.002
Anger				**0/1**

Table 17. Partial results of zero crossing periods L_Z hypothesis test – female voice group.

	Neutral	Joy	Sadness	Anger
h / p	0 / 0.4397	0 / 0.8926	0 / 0.6953	0 / 0. 5773

Table 18. Summary results of zero crossing periods L_Z hypothesis test (comparison male vs. female voice group) between particular emotions.

Emotion	Male voice					Female voice				
	F_{Zmin} [1)]	F_{Zmean}	F_{Zrel}	B_3	B_{3F} [2)]	F_{Zmin} [1)]	F_{Zmean}	F_{Zrel}	B_3	B_{3F} [2)]
Neutral	1.60	6.89	8.83	6.75	4.56	2.23	11.88	14.60	11.59	6.71
Joy	0.71	5.04	6.45	4.56	3.82	1.56	9.41	11.94	9.03	5.61
Sadness	0.73	6.11	7.78	4.39	2.69	1.56	9.33	11.66	7.20	3.17
Anger	1.81	6.18	8.00	5.37	4.07	2.08	9.88	12.59	10.74	5.86

[1)] $L_{Zmin} = 1 \Rightarrow F_{Zmax} = f_F / 2$
[2)] 3-dB bandwidth for signal smoothed by MA filter with $M_F = 8$

Table 19. Summary results of spectral analysis (frequency parameters in [Hz] derived from concatenated differential F0 signal) – male and female voice.

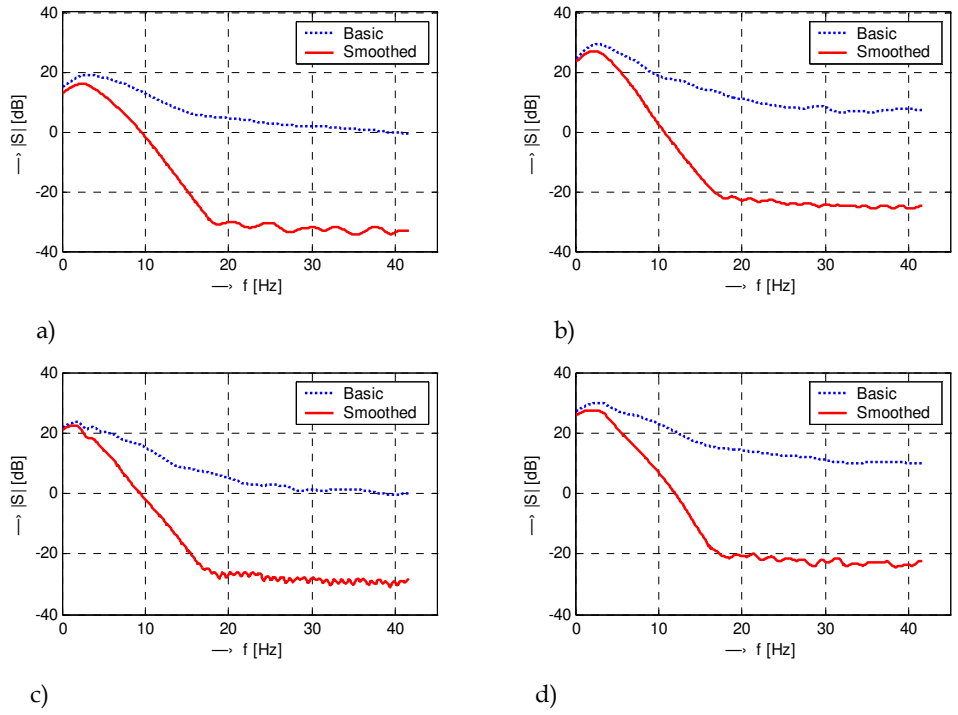


Fig. 26. Spectra of microintonation used for 3-dB bandwidth determination for emotions (with and without smoothing by moving average): "neutral" (a), "sadness" (b), "joy" (c), and "anger" (d) - male voice.

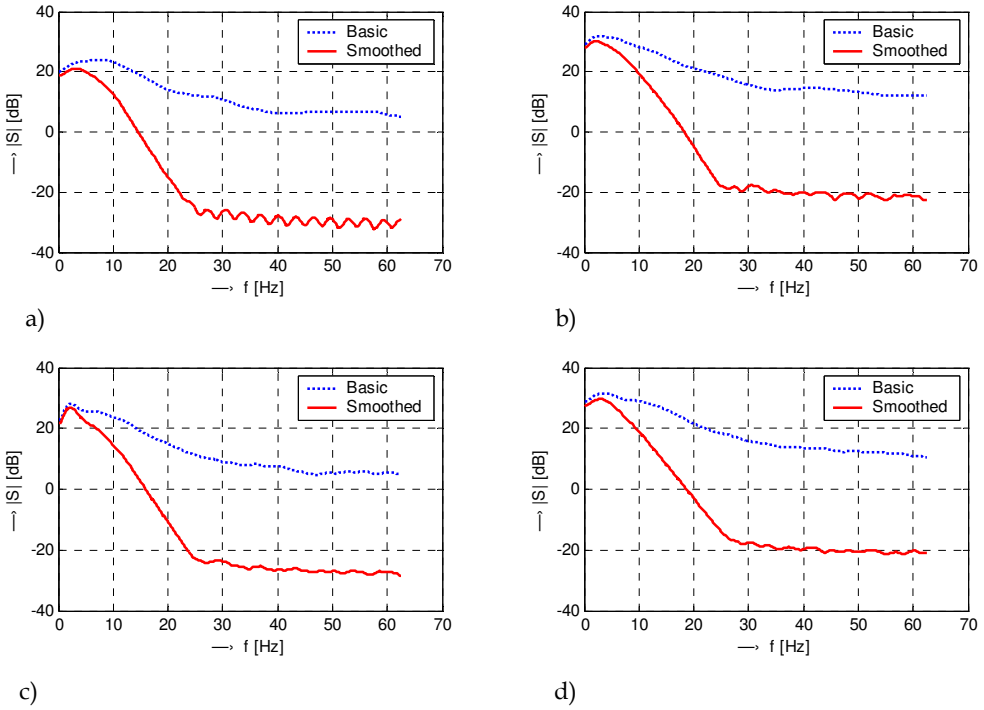


Fig. 27. Spectra of microintonation used for 3-dB bandwidth determination for emotions (with and without smoothing by moving average): "neutral" (a), "sadness" (b), "joy" (c), and "anger" (d) - female voice.

4. Discussion and conclusion

Results of performed analysis and comparison (consequently computed parameter ratios between emotional and neutral states) will be applied for extension of the text-to-speech (TTS) system enabling expressive speech production of voices (male / female) or it be also used in emotional speech transformation (conversion) method based on cepstral speech description (Přibil & Přibilová 2008). The main advantage of this approach consists in a fact that only new cepstral description must be created and the original speech database is applied as a common area for all voices.

Statistical analysis of cepstral coefficients and the first three formant positions has shown that different emotional states are manifested in a speech signal in observed parameters. Spectrograms, histogram envelopes together with other parameters may be used for identification of individual emotions. This method can also be used for evaluation of emotional synthetic speech as a supplementary approach parallel to the listening tests.

From visual comparison of these spectrograms and histograms follows 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. 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 approach (Vích et al. 2008) can be used here (e.g. in the form of a simple phoneme alignment procedure) to get these ROIs automatically.

Results of the spectral flatness ranges and values statistical analysis show good correlation for both types of voices and all three emotions. The greatest mean SFM value is observed in "anger" style for both voices. Similar shape of SFM histograms can be seen in Fig. 19a) and Fig. 20a) comparing corresponding emotions for male and female voices. On the other hand, it was confirmed that only SFM values calculated from voiced frames of speech give sufficient information – in Fig. 21 it is evident that the histograms are practically the same for all three emotions. This subjective result is confirmed by the objective method – multiple comparison of groups based on results of ANOVA statistics and hypothesis test. Our final aim was to obtain the ratio of mean values which can be used to control the high frequency noise component in the mixed excitation during cepstral speech synthesis of voiced frames (Vích 2000). From summary results follows that the ratio of mean values is 1.18 times higher for female voice than for male voice.

From comparison of basic statistical microintonation analysis follows that absolute jitter values are in accordance with the human vocal tract properties. But there should be a problem with accuracy of jitter measurements, caused by the fact that jitter estimation on running speech (in contrast to steady vowels) is very difficult (Sun et al. 2009). Female shorter pitch periods are accompanied with shorter values of the absolute jitter but higher relative changes in the frequency domain (mean $F0_{DIFF}$ values). The highest values of jitter correspond to "joy" and the lowest ones correspond to "sadness" for both voices. Similar results are shown in (Tao et al. 2006). The same tendency can be observed for statistical results of zero crossing analysis. Although different frame lengths were used in microintonation frequency analysis for male and female voices, we can see matched similar values for all corresponding emotions. Visual comparison of histograms of zero crossing periods L_Z is not significant but higher relative occurrence of low L_Z values can be noticed in "neutral" style for both voices. On the other hand, as regards visual comparison of average spectra, similar curves can be matched in Fig. 26 and Fig. 27 for male and female voice for all corresponding emotions. Obtained results of microintonation spectral analysis (especially the B_3 values) can be used to synthesize a digital filter for suppression of microintonation component of a speech signal. From objective statistical comparison of zero crossing periods by Ansari-Bradley hypothesis test follows that the null hypotheses were rejected at the 5% significance level for each emotion type inside the gender group and simultaneously, the null hypotheses between the corresponding emotions of both types of voices are in all cases fulfilled at the same significance level. The result of final multiple comparison of ANOVA also confirms good correlation between particular emotions.

In the next future, we plan to use results of ANOVA and hypothesis test for creation of the database of values for emotional speech classifier based on statistical evaluation approach (Iriondo et al. 2009), or it can be used for identification of speaker emotional states or in real-time emotion recognition systems (Attasi & Smékal 2008).

5. Acknowledgment

The work has been done in the framework of the COST 2102 Action "Cross-Modal Analysis of Verbal and Non-Verbal Communication". It has also been supported by the Grant Agency of the Czech Republic (GA102/09/0989), by the Grant Agency of the Slovak Academy of Sciences (VEGA 2/0090/11) and the Ministry of Education of the Slovak Republic (VEGA 1/0903/11).

6. References

Atassi, H. & Smékal, Z. (2008). Real-Time Model for Automatic Vocal Emotion Recognition. In: *Proceedings of 31st International Conference on Telecommunications and Signal Processing*, Parádfürdö, Hungary, pp.21-25, 2008

Boersma, P. & Weenink, D. (2008). Praat: Doing Phonetics by Computer (Version 5.0.32) [Computer Program]. Retrieved August 12, 2008 from http://www.praat.org/
Everitt, B. S. (2006). *The Cambridge Dictionary of Statistics*. Third Edition. Cambridge University Press, 2006

Fant, G. (2004). *Speech Acoustics and Phonetics*. Kluwer Academic Publishers, Dordrecht Boston London, 2004

Farrús, M., Hernando, J. & Ejarque, P. (2007). Jitter and Shimmer Measurements for Speaker Recognition. In: *Proceedings of Interspeech 2007*, Antwerp, Belgium, pp. 778-781, 2007

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.

Hosseinzadeh, D. & Krishnan S. (2008). On the Use of Complementary Spectral Features for Speaker Recognition. *EURASIP Journal on Advances in Signal Processing,* Vol. 2008, Article ID 258184, 10 pages, doi:10.1155/2008/258144, Hindawi Publishing Corp.

Iriondo, I., Planet, S., Socoro, J.C, Martínez, E., Alías, F. & Monzo, C. (2009). Automatic Refinement of an Expressive Speech Corpus Assembling Subjective Perception and Automatic Classification. *Speech Communication* Vol. 51, pp. 744–758, Elsevier, 2008.

Kuwabara, H. & Sagisaka, Y. (1995). Acoustic characteristics of speaker individuality: Control and conversion. *Speech Communication* Vol. 16, pp. 165-173, Elsevier, 1995

Li, X., Tao, J., Johnson, M.T., Soltis, J., Savage, A., Kirsten M., Leong, K.M. & Newman, J.D. (2007). Stress and Emotion Classification Using Jitter and Shimmer Features. In: *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASPP '07)*, Honolulu, HI, pp. IV-1081-IV-1084, 2007

Madlová, A. & Přibil, J. (2000). Comparison of Two Approaches to Speech Modelling Based on Cepstral Description. In: *Proceedings of the 15th Biennial International EURASIP Conference Biosignal 2000*, Brno, Czech Republic, pp. 83-85, June 21-23, 2000

Nwe, T. L., Foo, S. W. & De Silva, L. (2003). Speech Emotion Recognition Using Hidden Markov Models. *Speech Communication* Vol. 41, pp. 603–623, Elsevier, 2003

Oppenheim, A.V. Schafer, R.W. & Buck, J.R. (1999). *Discrete-Time Signal Processing*. Second Edition. Prentice Hall, 1999

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 Interactions: *Lecture Notes in Artificial Intelligence 5042*, pp. 201-213, Springer-Verlag: Berlin Heidelberg, 2008

Scherer, K.R. (2003). Vocal Communication of Emotion: A Review of Research Paradigms. *Speech Communication*, Vol. 40, pp. 227–256, Elsevier, 2003

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

Stevens, K.N. (1997). Models of speech production. In: Crocker, M.J. (Ed.), *Encyclopedia of Acoustics*, pp. 1565-1578, John Wiley & Sons, Inc. 1997

Suhov, Y. & Kelbert, M. (2005). *Probability and Statistics by Example*. *Volume* I. Basic Probability and Statistics. Cambridge University Press 2005

Sun, R., Moore, E. & Torres, J.F. (2009). Investigating Glottal Parameters for Differentiating Emotional Categories with Similar Prosodics. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009*), Taipei, Taiwan, pp. 4509-4512, 2009

Tao, J., Kang, Y. & Li, A. (2006). Prosody Conversion from Neutral Speech to Emotional Speech. *IEEE Transactions on Audio, Speech, and Language Processing* Vol. 14, pp. 1145–1154

Vích, R. (2000). Cepstral Speech Model, Padé Approximation, Excitation, and Gain Matching in Cepstral Speech Synthesis. In: *Proceedings of the 15th Biennial EURASIP Conference Biosignal 2000*, Brno, Czech Republic, pp. 77–82, 2000

Vích, R., Nouza, J. & Vondra, M. (2008). Automatic speech recognition used for intelligibility assessment of text-to-speech systems. In: Esposito, A., et al. (eds.) Verbal and Nonverbal Features of Human-Human and Human-Machine Interactions: *Lecture Notes in Artificial Intelligence 5042*. pp. 136-148. Springer-Verlag: Berlin Heidelberg, 2008

Volaufova J. (2005). Statistical Methods in Biomedical Research and Measurement Science. *Measurement Science Review*, Vol. 5, No. 1, Section 1, pp. 1-10, Versita, 2005

Speech and Language Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-322-4
Hard cover, 344 pages
Publisher InTech
Published online 21, June, 2011
Published in print edition June, 2011

This book addresses state-of-the-art systems and achievements in various topics in the research field of speech and language technologies. Book chapters are organized in different sections covering diverse problems, which have to be solved in speech recognition and language understanding systems. In the first section machine translation systems based on large parallel corpora using rule-based and statistical-based translation methods are presented. The third chapter presents work on real time two way speech-to-speech translation systems. In the second section two papers explore the use of speech technologies in language learning. The third section presents a work on language modeling used for speech recognition. The chapters in section Text-to-speech systems and emotional speech describe corpus-based speech synthesis and highlight the importance of speech prosody in speech recognition. In the fifth section the problem of speaker diarization is addressed. The last section presents various topics in speech technology applications like audio-visual speech recognition and lip reading systems.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821