10.2478/v10048-009-0016-4

MEASUREMENT SCIENCE REVIEW, Volume 9, No. 4, 2009

Statistical Analysis of Spectral Properties and Prosodic Parameters of Emotional Speech

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 paper addresses reflection of microintonation and spectral properties in male and female acted emotional speech. Microintonation component of speech melody is analyzed regarding its spectral and statistical parameters. According to psychological research of emotional speech, different emotions are accompanied by different spectral noise. We control its amount by spectral flatness according to which the high frequency noise is mixed in voiced frames during cepstral speech synthesis. Our experiments are aimed at statistical analysis of cepstral coefficient values and ranges of spectral flatness in three emotions (joy, sadness, anger), and a neutral state for comparison. Calculated histograms of spectral flatness distribution are visually compared and modelled by Gamma probability distribution. Histograms of cepstral coefficient distribution are evaluated and compared using skewness and kurtosis. Achieved statistical results show good correlation comparing male and female voices for all emotional states portrayed by several Czech and Slovak professional actors.

Keywords: cepstral speech analysis, emotional speech, spectral flatness, microintonation

1. INTRODUCTION

EMOTIONAL speech is characterized by prosodic features (F0, energy, duration) and several voice quality features (e.g. jitter, shimmer, glottal-to-noise excitation ratio, Hammarberg index) [1], [2]. The voice quality parameter "jitter" describes pitch perturbations in the context of vocal expression. There exist different approaches to define vocal jitter: the majority of authors use definitions resulting from perturbation in pitch period [1], [3]-[6], some authors define jitter as pitch frequency perturbation [7], [8]. According to [9] jitter is difficult to manipulate for actors and there is only tendency for anger portrayals to show more jitter than sadness portrayals. On the other hand, in [10] an example is reported that a speaker may increase jitter for "happiness" rather than increasing the overall pitch level. For these perturbations also the term "microintonation" is used [9]. We analyze microintonation of male and female emotional speech representing joy, sadness, anger, and a neutral state. Obtained results of spectral analysis can also be used to synthesize FIR digital filter for suppression of the microintonation component of a speech signal prior to decomposition of its virtual melody contour into the sentence melody and the word melody. This prosodic parameter can be applied to the text-to-speech (TTS) system enabling expressive speech production, or it can be used in emotional speech transformation (conversion) method based on cepstral speech description [11].

Different types of emotions are manifested also in the spectral domain [12]. 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. For this reason, we also perform statistical analysis of the cepstral coefficients and spectral flatness values (in voiced speech only) for the mentioned emotional states. There is a necessity for adaptation of the speech synthesizer based on cepstral speech model. The kernel of the cepstral synthesizer represented by the source-filter model based on Padé approximation of a human vocal tract was designed and optimized on the basis of the processed cepstral coefficient properties [13]. The parameters used in the original realization of the cepstral speech synthesizer had been obtained by statistical evaluation of a speech signal in the database of phones uttered by a male speaker in a neutral speech style. Therefore we decided to carry out basic statistical analysis of values and ranges of cepstral coefficients obtained from speech signals expressing different emotional states. As regards spectral flatness, it is a useful measure to distinguish between voiced and unvoiced speech [14]. Its usage in speech processing can be extended to whispered speech recognition in noisy environment [15], or voicing transition frequency determination in harmonic speech modelling [16]. In cepstral speech synthesis the spectral flatness measure was used to determine voiced / unvoiced energy ratio in voiced speech. According to psychological research of emotional speech different emotions are accompanied by different spectral noise [17]. We control its amount by spectral flatness measure according to which the high frequency noise is mixed in voiced frames during cepstral speech synthesis.

2. SUBJECT & METHODS

2.1 Microintonation analysis

Microintonation, together with sentence melody and word melody, represents melody of speech given by a fundamental frequency (F0) contour. Microintonation component of speech melody can be supposed to be a random, band-pass signal described by its spectrum and statistical parameters. Fig. 1 shows the block diagram of our speech processing method of microintonation analysis.

Speech frames classified as voiced are analyzed separately depending on the emotional state (joyous, sad, angry, and

neutral) and the voice type (male, female). Joy, sadness and anger are chosen as three representatives of emotional states as according to [18] they correspond to grouping of emotions according to similarity:

1. anger, rage, disgust, unwillingness;
2. joy, gratitude, happiness, pleasantness, elation;
3. sadness, disconsolation, loneliness, anxiety.

The whole microintonation analysis procedure is divided into four phases:

1. Determination of F0 values, definition of the voiced and unvoiced parts of the processed speech signal.
2. F0 contour analysis, microintonation extraction, calculation of zero crossing parameters in the voiced parts of the speech signal.
3. Microintonation and zero crossing statistical analysis of the concatenated signal.
4. Microintonation signal spectral analysis and 3-dB bandwidth (B_3) determination from the concatenated signal.

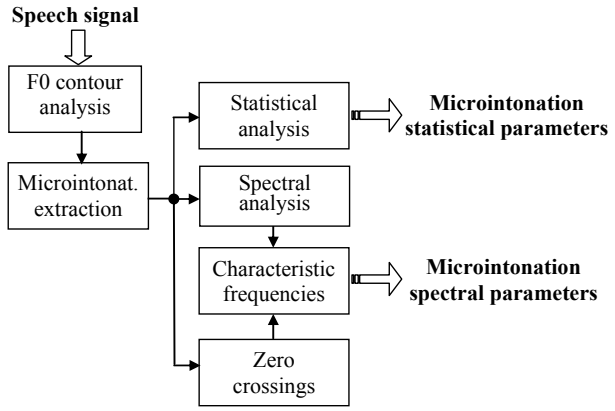

Fig.1 Block diagram of microintonation statistical and spectral parameter estimation.

The introductory microintonation processing phase consists of the following steps:

– Determination of the melody contours from the voiced parts of speech smoothed by a median filter.
– Determination of $F0_{mean}$ values and calculation of the sentence melody declination of the F0 contour given by the linear trend (LT)

$$LT(a,b) = a + b\,n, \tag{1}$$

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

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

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

– Calculation of differential microintonation signal $F0_{DIFF}$ by subtraction of these values from the corresponding F0 contours ($F0_{mean}$ and LT removal)

$$F0_{DIFF}(n) = (F0(n) - F0_{Mean}) - LT(n)\ . \tag{3}$$

– Detection of zero crossings, calculation of zero crossing periods L_Z.

Demonstration example of microintonation analysis processing phases is shown in Fig.4.

Basic statistical analysis phase is performed in two steps:

– Statistical analysis of microintonation signal: minimum, maximum, and standard deviation (mean value of microintonation signal approaches to zero). For both positive and negative microintonation values the mean parameters are determined (see Fig.2).
– Statistical analysis of the zero crossing periods: the minimum, maximum, mean values, standard deviation and relative value, defined as $L_{Zrel} = N_F / N_Z$ where N_Z is the total number of zero crossings for one speaker.

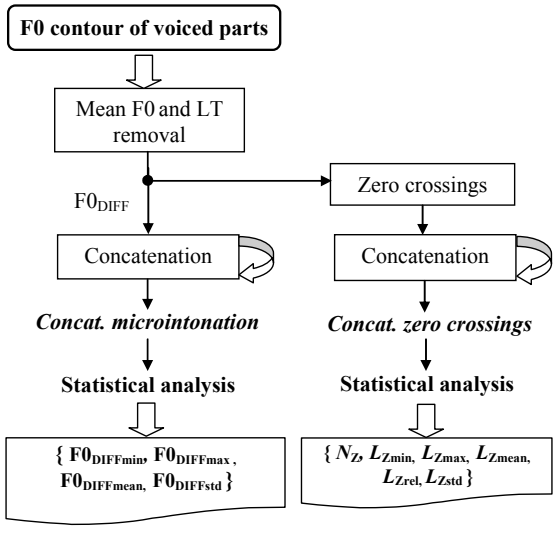

Fig.2 Block diagrams of microintonation basic and zero crossing statistical analysis.

Spectral analysis of the concatenated differential microintonation signal is also carried out for all emotions. This analysis phase is divided into three steps (see Fig.3):

– Calculation of the frequency parameters $F_{Zx} = f_F / (2\,L_{Zx})$ from the zero crossing periods $L_{Zx} = \{L_{Zmin}, L_{Zmax}, L_{Zmean}, L_{Zstd}, L_{Zrel}\}$, where f_F is the frame sampling frequency.
– Microintonation signal spectral analysis by periodogram averaging using the Welch method [19].
– Determination of B_3 values from these spectra for each of the emotional types.

To obtain the spectrum of the smoothed microintonation signal (see Fig.5b), 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).

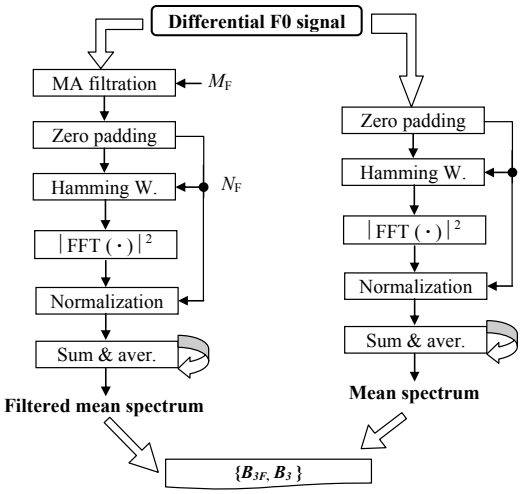

Fig.3 Block diagrams of microintonation signal spectral analysis.

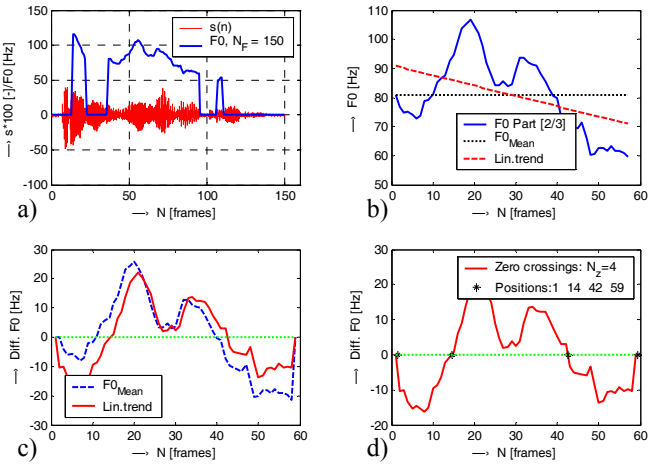

Fig.4 Demonstration of microintonation analysis: speech signal with F0 contour (a), the second voiced part: original F0, mean F0, and linear trend (b), differential signal after $F0_{mean}$ and LT subtraction (c), zero crossing of differential F0 signal (d) – the sentence "Prosím, nehnevajte sa" ("Please, don't be angry") uttered in sad emotional style by a male Slovak speaker.

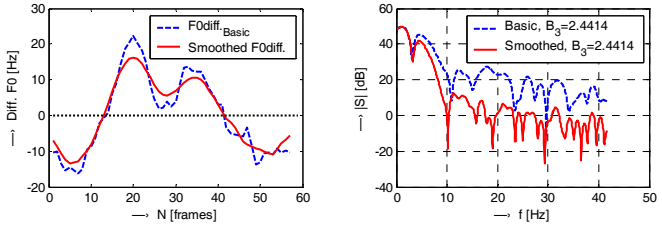

Fig.5 Microintonation smoothing and spectrum determination (obtained from the same sentence's second voiced part as in Fig.4): basic differential F0 signal and the one filtered by moving average (left), corresponding spectra and their 3-dB bandwidths B_3 (right).

2.2 Cepstral coefficient analysis

Cepstral speech analysis is performed in the frequency domain as follows: From the input samples of the speech signal (after segmentation and weighting by a Hamming window) the complex spectrum by the Fast Fourier Transformation (FFT) algorithm is calculated. In the next step the powered spectrum is computed and the natural logarithm is applied – see the block diagram in Fig. 6. Second application of the FFT algorithm gives the symmetric real cepstrum

$$\{c_n\} = \{c_0, c_1, \ldots, c_{N_{FFT}/2} \mid c_{N_{FFT}/2-1}, \ldots, c_1\} \tag{4}$$

By limitation to the first N_0+1 coefficients, the Z-transform of the real cepstrum in the form

$$C(z) = c_0 + c_1 z^{-1} + c_2 z^{-2} + \ldots + c_{N_0} z^{-N_0} \tag{5}$$

can be obtained.

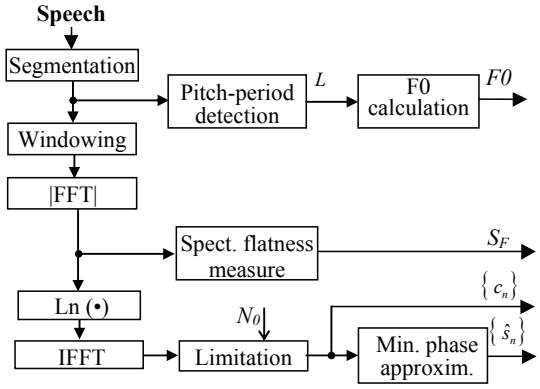

Fig.6 Cepstral speech analysis method.

Cepstral speech synthesis is performed by a digital filter implementing approximate inverse cepstral transformation. In general, the system transfer function is given by an exponential relation

$$G(z) = e^{\hat{S}(z)}, \tag{6}$$

where the exponent is the Z-transform of the truncated speech cepstrum

$$\hat{S}(z) = \sum_{n=0}^{N_0} \hat{s}_n z^{-n}, \tag{7}$$

where $\{\hat{s}_n\}$ represents the minimum phase approximation of the real cepstrum

$$\begin{aligned} \hat{s}_n &= c_n, && n = 0,\ N_0/2, \\ \hat{s}_n &= 2c_n, && 1 \le n < N_0/2, \\ \hat{s}_n &= 0\ , && N_0/2 < n \le N_0 - 1. \end{aligned} \tag{8}$$

The system transfer function of the synthesis filter is defined as

$$G_F(z) = e^{c_0} \prod_{i=1}^{25} G_i(z), \tag{9}$$

and 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 [13]. The cepstral synthesis block structure is given by a cascade of digital filters (of the 1st, 2nd or 3rd order in the second canonic form) that perform the inverse transformation of N_0 cepstral components – see Fig.7.

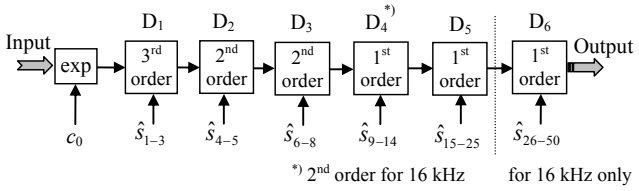

Fig.7 Cascade realization of Padé approximation filter.

Cepstral coefficient analysis must be preceded by classification and sorting process of the cepstral coefficients in dependence on voice type (male / female) and speech style (neutral / emotional). The performed statistical analysis of cepstral coefficients consists of three 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).

Realization of statistical analysis of the cepstral coefficient properties was processed in the following phases: manual (subjective) classification of voice type and emotional speech style, further automatic processing – cepstral analysis of speech signal, computation of the basic statistical parameters of determined cepstral coefficients, comparison of cepstral coefficient mean values and ranges for emotional and neutral states. As the graphical output, the histogram of cepstral coefficients for every emotional state is also constructed. Extended statistical parameters are subsequently calculated from these histograms. The skewness y and kurtosis k of a distribution is defined as

$$y = \frac{E(x-\mu)^3}{\sigma^3}, \qquad k = \frac{E(x-\mu)^4}{\sigma^4}, \tag{10}$$

where μ is the mean of x, σ is the standard deviation of 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.

2.3 Spectral flatness analysis

The spectral flatness measure S_F calculated during the cepstral speech analysis (see Fig. 6) is defined as

$$S_F = \frac{\exp\left[\frac{2}{N_{FFT}}\sum_{k=1}^{N_{FFT}/2}\ln|S_k|^2\right]}{\frac{2}{N_{FFT}}\sum_{k=1}^{N_{FFT}/2}|S_k|^2}, \tag{11}$$

where the values $|S_k|^2$ represent the magnitude of the complex spectrum, and N_{FFT} is the number of points of the fast Fourier transform (FFT). The S_F 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 $S_F = 1$, 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 $S_F = (0 \div 0.25)$ for voiced speech frames and $S_F = (0 \div 0.75)$ for unvoiced frames were evaluated.

For voiceness frame classification, the value of detected pitch-period L was used. If the value $L \neq 0$, the processed speech frame is determined as voiced, in the case of $L = 0$ the frame is marked as unvoiced. On the border between voiced and unvoiced part of speech signal a situation can occur that the frame is classified as voiced, but the S_F value corresponds to the unvoiced class. For correction of this effect, the output values of the pitch-period detector are filtered by a 3-point recursive median filter.

The performed statistical analysis of spectral flatness values consists of two parts:

1. determination of basic statistical parameters of the S_F values,
2. calculation and building of histograms.

Practical evaluation of obtained results is further processed in three ways:

– determination of mean ratio between neutral and emotional states,
– visual comparison of histogram figures,
– histogram fitting and modelling by Gamma distribution – comparison of parameters α, λ and root mean square (RMS) approximation error.

We compute the S_F values of the sentences in the basic ("Neutral") speech style and the S_F values of the sentences pronounced in the emotional states ("Joy", "Sadness", and "Anger") and perform statistical analysis of these values. In our algorithm, the S_F values obtained from the speech frames classified as voiced are separately processed in dependence on voice type (male/female). For every voice type the S_F values are subsequently sorted by emotional styles and stored in separate stacks. These classification operations are 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 (STD). From the mean S_F values the ratio between emotional and neutral states is subsequently calculated. As the graphical output used for visual comparison

(subjective method), the histogram of sorted S_F values for each of the stacks is also calculated. These histograms can also be fitted and modelled by the Gamma distribution (objective evaluation method).

The generalized Gamma distribution of the random variable X is given by the probability density function (PDF) [20]

$$f(x)=\frac{\lambda^a}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} \quad x\geq 0, \quad \alpha>0, \quad \lambda>0, \qquad (11)$$

where α is a shape parameter and λ is a scale parameter. The Gamma function is defined by

$$\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}dx. \qquad (12)$$

The graphs of the PDFs for different parameters α, λ are shown in Fig.8.

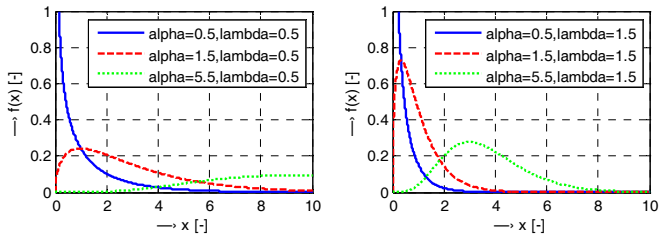

Fig.8 Example of the Gamma probability density functions for $\lambda = 0.5$ (left), $\lambda = 1.5$ (right).

The shape and scale parameters of the Gamma distribution enable easy and rather accurate modelling of obtained histograms of S_F values. It means the finding of α and λ parameters for minimum RMS error between the histogram envelope curve and the Gamma PDF. Simultaneous control of two parameters represents a two- dimensional regulation process. Its practical realization with sufficient precision is a difficult task. Therefore, a simplified control method was used – only one parameter is changed and the second one has a constant value. The developed algorithm can be divided into three phases:

1. Initialization phase:
   - fitting the histogram bars by the envelope curve
   - rough estimation of α, λ parameters
   - calculation of the Gamma PDF
   - calculation of the RMS error, storing this value to the memory.
2. Finding the RMS minimum by change of α parameter:
   - modification of α parameter with constant value of λ parameter (estimated in phase 1)
   - calculation of the Gamma PDF and the RMS error, storing to the memory
   - comparison of the current RMS error with the last value from the memory

(repeating the steps in this phase until the minimum of RMS).

3. Finding the RMS minimum by change of λ parameter:
   - modification of λ parameter with constant value of α parameter (determined in phase 2)
   - calculation of the Gamma PDF and the RMS error, storing to the memory
   - comparison of the current RMS error with the last value from the memory

(repeating the steps in this phase until the minimum of RMS).

3. Material, Experiments, and Results

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. The sentences of four emotional states (sad, joyful, angry, and neutral for comparison) were obtained from multimedia CDs containing recordings of stories in Czech and Slovak languages uttered by professional actors.

Classification of emotional states was carried out manually, by subjective listening method. The frame length depends on the mean pitch period of the processed signal. In our experiment, we had chosen 24-ms frames for male voice, and 20-ms frames for female voice. It corresponds to the frame frequency $f_F = 83.3$ Hz for males, and $f_F = 125$ Hz for females when the sampling frequency $f_s = 16$ kHz is used. Pitch contours were given with the help of the PRAAT program [21]. The PRAAT internal settings for F0 value determination were experimentally chosen by visual comparison of testing sentences (one typical sentence from each of emotions and voice classes) as follows: cross-correlation analysis method [22], pitch-range 35÷250 Hz for male and 105÷350 Hz for female voices.

In the case of microintonation analysis, the minimum length of the processed voiced parts was experimentally set to 10 frames and the corresponding filter length of $M_F = 8$ was chosen. Number of analyzed voiced parts / voiced frames was:

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

As follows from the experiments, the cepstral coefficients, as well as the spectral flatness values, 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 described method of cepstral speech analysis was supplied with determination of the fundamental frequency F0 and energy *En* contours (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 in frames for next processing was obtained – see demonstration example in Fig.9.

Cepstral analysis of the speech signal was performed for total number of 25988 frames (8 male speakers) and 24017 frames (8 female speakers):

- 6031 frames (neutral), 4128 frames (joy), 7085 frames (sadness), 5648 frames (anger) – Male.
- 4472 frames (neutral), 5333 frames (Joy), 5753 frames (sadness), 5601 frames (anger) – Female.

As the value range of the cepstral coefficients exponentially falls, analysis only of the first 16 coefficients is performed (the remaining coefficients practically have no influence on the filter stability, structure, and implementation). The spectral flatness values were determined only from the voiced frames (totally 11639 of male and 13464 of female voice).

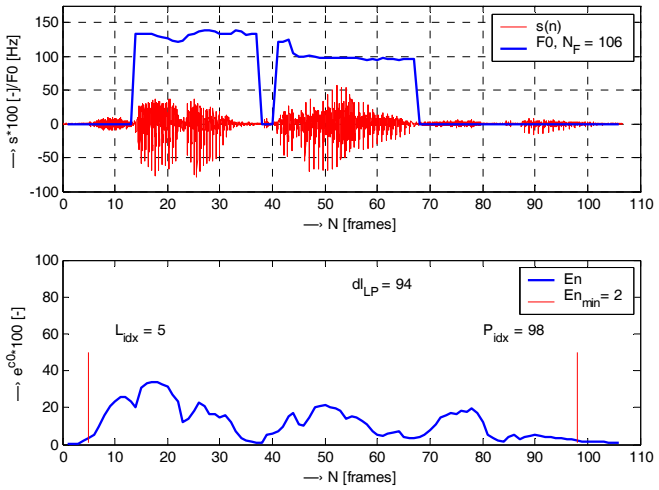

Fig.9 Processed sentence "Dcera královská" *(King's daughter)*, Czech male speaker, f_s = 16 kHz: speech signal with F0 contour (top), *En* contour (bottom).

3.1 Results of microintonation analysis

Results of basic statistical microintonation analysis for all four emotional states are summarized in Tab.1 (male voice) and Tab.2 (female voice). Results of performed zero crossing analysis for male / female voices are shown in Tab.3 / Tab.4. Zero crossing periods were used to calculate microintonation signal spectral analysis. Summary results including the 3-dB bandwidth values are shown in Tab.5 for male voice, and in Tab.6 for female voice. The average microintonation spectra (with and without smoothing by moving average) can be seen in Fig.10 (male voice), and Fig.11 (female voice).

Table 1. Summary results of microintonation basic statistical analysis (differential F0 parameters in [Hz]) – male voice.

Emotion	$F0_{DIFFmin}$	$F0_{DIFFmax}$	$F0_{DIFFmean}$	$F0_{DIFFstd}$
Neutral	-16.75	22.71	2.66	3.92
Joy	-32.68	34.48	7.27	9.71
Sadness	-34.13	25.27	4.02	5.57
Anger	-56.32	63.88	9.62	14.23

Table 2. Summary results of microintonation basic statistical analysis – female voice.

Emotion	$F0_{DIFFmin}$	$F0_{DIFFmax}$	$F0_{DIFFmean}$	$F0_{DIFFstd}$
Neutral	-23.49	23.98	3.67	6.06
Joy	-32.88	33.02	8.49	10.72
Sadness	-30.28	34.56	6.29	8.43
Anger	-44.35	42.95	10.16	13.07

Table 3. Summary results of zero crossing analysis (zero crossing period L_Z parameters in [frames]) – male voice.

Emotion	N_Z	L_{Zmax} *)	L_{Zmean}	L_{Zstd}
Neutral	592	26	6.04	4.19
Joy	403	59	8.26	6.52
Sadness	681	57	6.82	5.69
Anger	521	23	6.74	4.57

*) $L_{Zmin} = 1$

Table 4. Summary results of zero crossing analysis – female voice.

Emotion	N_Z	L_{Zmax} *)	L_{Zmean}	L_{Zstd}
Neutral	546	28	5.26	3.78
Joy	468	40	6.64	5.23
Sadness	524	40	6.69	5.43
Anger	478	30	6.32	4.43

*) $L_{Zmin} = 1$

Table 5. Summary results of spectral analysis (frequency parameters in [Hz] derived from concatenated differential F0 signal) – male voice.

Emotion	F_{Zmean}	F_{Zrel}	B_3	B_{3F} *)
Neutral	6.89	8.83	6.75	4.56
Joy	5.04	6.45	4.56	3.82
Sadness	6.11	7.78	4.39	2.69
Anger	6.18	8.00	5.37	4.07

*) 3-dB bandwidth for signal smoothed by MA filter with $M_F = 8$

Table 6. Summary results of spectral analysis – female voice.

Emotion	F_{Zmean}	F_{Zrel}	B_3	B_{3F} *)
Neutral	11.88	14.60	11.59	6.71
Joy	9.41	11.94	9.03	5.61
Sadness	9.33	11.66	7.20	3.17
Anger	9.88	12.59	10.74	5.86

*) 3-dB bandwidth for signal smoothed by MA filter with $M_F = 8$

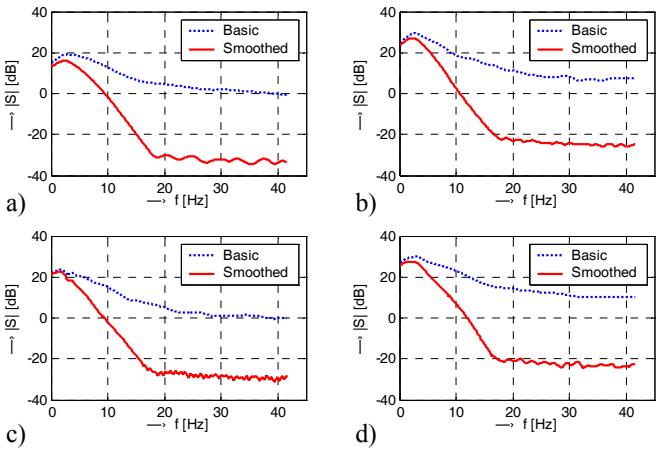

Fig.10 Spectra of microintonation used for 3-dB bandwidth determination for emotions (with and without smoothing by moving average): "neutral" (a), "sadness" (b), "joy" (c), and "anger" (d) - male voice, f_F = 83.3 Hz.

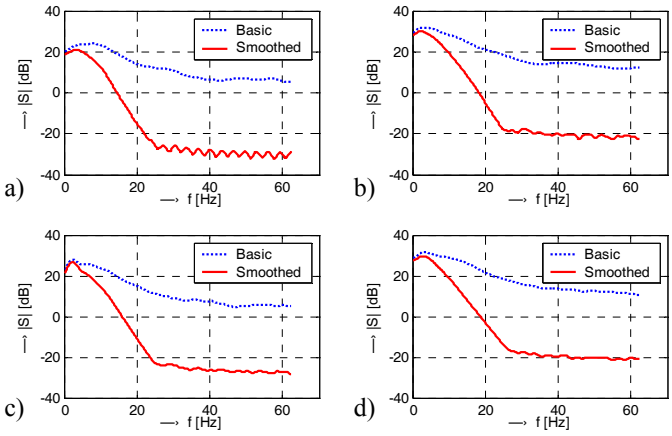

Fig.11 Spectra of microintonation used for 3-dB bandwidth determination for emotions (with and without smoothing by moving average): "neutral" (a), "sadness" (b), "joy" (c), and "anger" (d) - female voice, f_F = 125 Hz.

3.2 Results of cepstral coefficient analysis

Results of determined mean values of the first six cepstral coefficients are shown in Table 7 (male voice) and Table 8 (female voice). Summary histograms of cepstral coefficients (c_1-c_8) can be seen in Fig.12 and histogram contour comparison for different emotions of cepstral coefficients (c_1-c_4) is shown in Fig.13 (both male voice). Table 9 contains values of kurtosis parameters, and Table 10 contains values of skewness obtained from the compared histograms of c_1-c_6 (both male voice). In contrast to kurtosis definitions in (10) we subtract 3 from the computed value, so that the normal distribution has kurtosis of zero.

Table 7. Mean values of cepstral coefficients c_1-c_6, male voice.

Emotion	c_1 mean	c_2 mean	c_3 mean	c_4 mean	c_5 mean	c_6 mean
Neutral	-0.079	0.024	0.082	0.132	0.179	0.237
Joy	-0.165	-0.053	0.014	0.073	0.131	0.196
Sadness	-0.110	-0.015	0.047	0.098	0.151	0.215
Anger	-0.215	-0.088	-0.019	0.040	0.095	0.156

Table 8. Mean values of cepstral coefficients c_1-c_6, female voice.

Emotion	c_1 mean	c_2 mean	c_3 mean	c_4 mean	c_5 mean	c_6 mean
Neutral	-0.079	0.002	0.051	0.102	0.160	0.224
Joy	-0.127	-0.036	0.019	0.068	0.124	0.188
Sadness	-0.175	-0.079	-0.021	0.028	0.078	0.139
Anger	-0.174	-0.083	-0.025	0.028	0.081	0.144

Table 9. Kurtosis parameters determined from histograms of c_1-c_6 cepstral coefficients, male voice.

Emotion	c_1	c_2	c_3	c_4	c_5	c_6
Neutral	3.93	1.36	1.17	0.65	0.58	0.29
Joy	2.53	0.91	0.31	0.01	-0.35	-0.19
Sadness	1.72	0.82	0.29	-0.06	-0.16	-0.18
Anger	1.12	0.01	0.11	0.04	-0.07	-0.08

Table 10. Skewness parameters determined from histograms of c_1-c_6 cepstral coefficients, male voice.

Emotion	c_1	c_2	c_3	c_4	c_5	c_6
Neutral	-1.79	-0.99	-0.93	-0.73	-0.56	-0.39
Joy	-1.20	-0.64	-0.42	-0.22	-0.03	0.16
Sadness	-1.03	-0.75	-0.46	-0.13	-0.07	0.04
Anger	-0.84	-0.36	-0.12	0.09	0.25	0.35

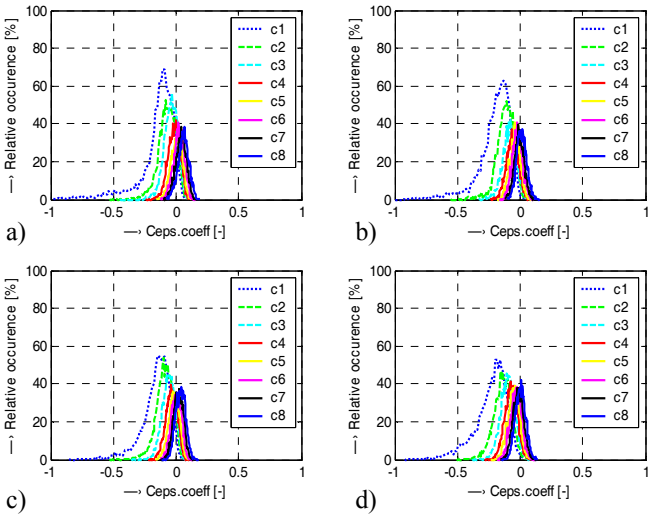

Fig.12 Histograms of the first 8 cepstral coefficients for different speech styles (male voice): neutral speech (a), joy (b), sadness (c), and anger (d).

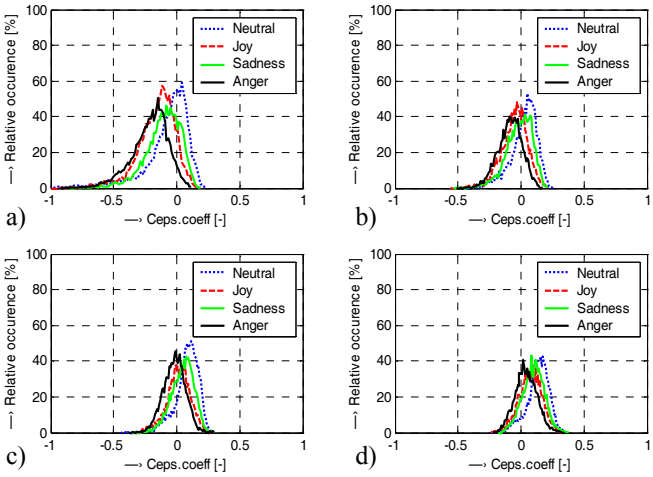

Fig.13 Histogram comparison for different speech styles (male voice): for cepstral coefficients c_1 (a), coefficients c_2 (b), coefficients c_3 (c), and coefficients c_4 (d).

3.3 Results of cepstral coefficient analysis

Summary results of statistical analysis of the spectral flatness values are shown in Tab.11 (male), Tab.12 (female). The main result – mean spectral flatness value ratios between different emotional states and a neutral state – is given in Tab.13. Summary histograms of S_F values for different emotions in dependence on the speaker's gender are shown in Fig.14 (male) and Fig.15 (female). Tab.14 (male) and Tab.15 (female) contain parameters α, λ of the Gamma distribution for histogram fitting and modelling together with the resulting RMS approximation errors.

Table 11. Summary results of statistical analysis of the spectral flatness values: male voice, voiced frames.

Emotion	frames	mean	min	max	std
Neutral	3300	0.00286	$3.78 \cdot 10^{-5}$	0.03215	0.00364
Joy	2183	0.00662	$1.36 \cdot 10^{-4}$	0.04327	0.00650
Sadness	3503	0.00444	$1.12 \cdot 10^{-4}$	0.05540	0.00462
Anger	2707	0.00758	$2.28 \cdot 10^{-4}$	0.04228	0.00614

Table 12. Summary results of statistical analysis of the spectral flatness values: female voice, voiced frames.

Emotion	frames	mean	min	max	std
Neutral	3056	0.00274	$3.15 \cdot 10^{-5}$	0.03731	0.00346
Joy	3473	0.00784	$2.07 \cdot 10^{-4}$	0.05414	0.00726
Sadness	3690	0.00506	$9.48 \cdot 10^{-5}$	0.06694	0.00674
Anger	3245	0.00807	$1.41 \cdot 10^{-4}$	0.05129	0.00692

Table 13. Mean spectral flatness value ratios between different emotional states and a neutral state (for voiced frames only).

mean S_F ratio	joy:neutral	sadness:neutral	anger:neutral
Male voice	2.31	1.55	2.65
Female voice	2.86	1.85	2.94
Female to Male ratio	**1.24**	**1.19**	**1.11**

Table 14. Evaluated parameters α, λ of Gamma distribution for histogram fitting and modelling together with resulting RMS error: male voice, voiced frames.

Emotion	α *)	λ *)	RMS
Neutral	2.05	0.48	0.70
Joy	4.15	0.50	0.67
Sadness	2.55	0.54	1.35
Anger	5.40	0.56	0.84

*) Values for minimum RMS error

Table 15. Evaluated parameters α, λ of Gamma distribution for histogram fitting and modelling together with resulting RMS error: female voice, voiced frames.

Emotion	α *)	λ *)	RMS
Neutral	1.95	0.51	1.48
Joy	4.85	0.51	0.54
Sadness	2.35	0.54	0.75
Anger	6.15	0.51	0.67

*) Values for minimum RMS error

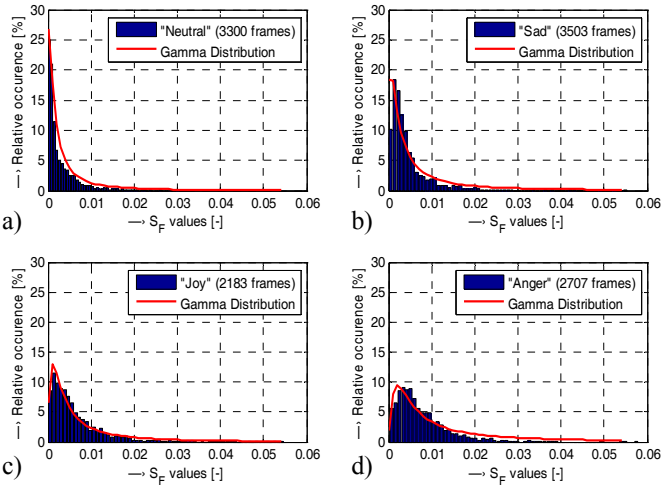

Fig.14 Histograms of spectral flatness values together with fitted and modelled curves of Gamma distribution - determined from the speech signal with emotions: "neutral" (a), "sadness" (b), "joy" (c), and "anger" (d) - male voice, voiced frames.

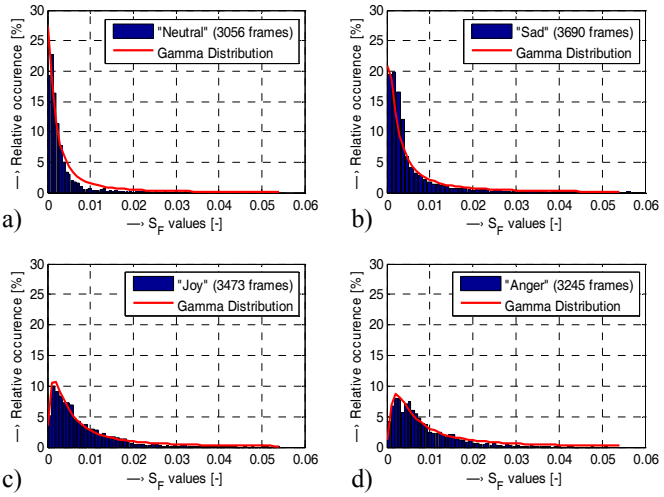

Fig.15 Histograms of spectral flatness values together with fitted and modelled curves of Gamma distribution - determined from the speech signal with emotions: "neutral" (a), "sadness" (b), "joy" (c), and "anger" (d) - female voice, voiced frames.

4. CONCLUSION

Statistical and spectral analysis of microintonation signal component of speech melody for several speakers and four emotional states (joy, sadness, anger, neutral state) was performed. Summary results of basic statistical microintonation analysis stored in Tab.1 and Tab.2 show good correlation for both types of voices and all three emotions compared with a neutral state. The same tendency can be observed also for statistical results of zero crossing analysis (see Tab.3 and Tab.4). Comparing visually the average spectra, we can see that similar curves can be matched in Fig.10 and Fig.11 for male and female voice for all corresponding emotions in spite of the fact that different frame lengths were used in microintonation frequency analysis for male and female voices.

Statistical analysis of cepstral coefficients has shown that different emotional states are manifested in a speech signal in observed parameters of cepstral coefficients, histogram envelopes and together with other parameters, they may well be used for identification of individual emotions. The values given by numerical evaluation of obtained statistical parameters will be used for modification of the cepstral synthesizer digital approximation filter structure, including possible implementation in the Czech and Slovak TTS system based on cepstral description of speech inventory enabling expression of basic emotional speech styles. Results of the cepstral coefficient ranges and values statistical analysis are shown also in the form of histograms in a similar way as the spectral flatness ranges and values. This method can also be used for evaluation of emotional synthetic speech as a supplementary approach parallel to the listening tests [23].

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 S_F value is observed in "Anger" style for both voices – compare Tab.11 and Tab.12. From Tab.13 follows that the ratio of mean values is 1.18 times higher for female voice than for male voice. Similar shape of S_F histograms can be seen in Fig.14 and Fig.15 comparing corresponding emotions for male and female voices. This subjective result is confirmed by the objective method – histogram modelling with the help of Gamma distribution. Given values of α and λ parameters – showed in Tab.14 and Tab.15 – are also in correlation with previous results. Our final aim was to obtain the ratio of mean values (see Tab.13), which can be used to control the high frequency noise component in the mixed excitation during cepstral speech synthesis of voiced frames. This parameter can be also applied for modification of degree of voicing in voiced frames [16].

ACKNOWLEDGMENT

The work has been done in the framework of the COST 2102 Action. It has also been supported by the Ministry of Education, Youth, and Sports of the Czech Republic (OC08010), the Grant Agency of the Czech Republic (GA102/09/0989), the Ministry of Education of the Slovak Republic (COST2102/STU/08), and by the Grant Agency of the Slovak Academy of Sciences, project VEGA No. 2/0142/08.

REFERENCES

[1] Iriondo, I., et al. (2009). Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification. *Speech Communication*, 51 (9), 744-758.

[2] Gobl, C., Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication,* 40 (1-2), 189–212.

[3] d'Alessandro, C., et al. (1998). Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. *IEEE Transactions on Speech and Audio Processing*, 6, 12–23.

[4] Schoentgen, J. (2003). Decomposition of vocal cycle length perturbations into vocal jitter and vocal microtremor, and comparison of their size in normophonic speakers. *Journal of Voice,* 17, 114–125.

[5] Shahnaz, C., et al. (2006). A new technique for the estimation of jitter and shimmer of voiced speech signal. In *Proceedings of the Canadian Conference on Electrical and Computer Engineering, CCECE 2006.* IEEE, 2112–2115.

[6] Farrús, M., et al. (2007). Jitter and shimmer measurements for speaker recognition. In *Proceedings of the International Conference Interspeech 2007.* Curran Associates, 778–781.

[7] Perrot, P., et al. (2007). Voice disguise and automatic detection: review and perspectives. In Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (eds.) *Progress in Nonlinear Speech Processing (Lecture Notes in Computer Science / Image Processing, Computer Vision, Pattern Recognition, and Graphics).* Springer, 101-117.

[8] Murphy, P. (2008). Source-filter comparison of measurements of fundamental frequency perturbation and amplitude perturbation for synthesized voice signals. *Journal of Voice,* 22, 125–137.

[9] Juslin, P.N., Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological Bulletin,* 129, 770–814.

[10] Tao, J., et al. (2009). Realistic visual speech synthesis based on hybrid concatenation method. *IEEE Transactions on Audio, Speech, and Language Processing*, 17, 469–477.

[11] 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.

[12] Přibilová, A., Přibil, J. (2009). Spectrum modification for emotional speech synthesis. In Esposito, A., Hussain, A., Marinaro, M., Martone, R. (eds.) *Multimodal Signals: Cognitive and Algorithmic Issues (Lecture Notes in Artificial Intelligence).* Springer, 232–241.

[13] 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: University of Technology, 77–82.

[14] Gray, A.H., Jr., Markel, J.D. (1974). A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing,* ASSP-22, 207–217.

[15] Ito, T., et al. (2005). Analysis and recognition of whispered speech. *Speech Communication,* 45, 139–152.

[16] Přibil, J., Přibilová, A. (2006). Voicing transition frequency determination for harmonic speech model. In *Proceedings of the 13th International Conference on Systems, Signals and Image Processing*, 25–28.

[17] Scherer, K.R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication,* 40, 227–256.

[18] Iida, A., et al. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication*, 40, 161–187.

[19] Oppenheim, A.V., Schafer, R.W. (1989). *Digital Signal Processing*. New Jersey: Prentice Hall.

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

[21] 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/

[22] Boersma, P., Weenink, D. (2007). *Praat – tutorial. Intro 4. Pitch analysis*. Retrieved September 5, 2007, from http://www.fon.hum.uva.nl/praat/manual/Intro_4__Pitch_analysis.html

[23] Vich, 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)*. Springer, 136-148.