# Determination of Formant Features in Czech and Slovak for GMM Emotional Speech Classifier

*Jiří PŘIBIL [1], Anna PŘIBILOVÁ [2]*

[1] Inst. of Measurement Science, SAS, Dúbravská cesta 9, SK-841 04 Bratislava, Slovakia
[2] Inst. of Electronics and Photonics, Faculty of Electrical Engineering & Information Technology, SUT, Ilkovičova 3, SK-812 19 Bratislava, Slovakia

Jiri.Pribil@savba.sk, Anna.Pribilova@stuba.sk

**Abstract.** *The paper is aimed at determination of formant features (FF) which describe vocal tract characteristics. It comprises analysis of the first three formant positions together with their bandwidths and the formant tilts. Subsequently, the statistical evaluation and comparison of the FF was performed. This experiment was realized with the speech material in the form of sentences of male and female speakers expressing four emotional states (joy, sadness, anger, and a neutral state) in Czech and Slovak languages. The statistical distribution of the analyzed formant frequencies and formant tilts shows good differentiation between neutral and emotional styles for both voices. Contrary to it, the values of the formant 3-dB bandwidths have no correlation with the type of the speaking style or the type of the voice. These spectral parameters together with the values of the other speech characteristics were used in the feature vector for Gaussian mixture models (GMM) emotional speech style classifier that is currently developed. The overall mean classification error rate achieves about 18 %, and the best obtained error rate is 5 % for the sadness style of the female voice. These values are acceptable in this first stage of development of the GMM classifier that should be used for evaluation of the synthetic speech quality after applied voice conversion and emotional speech style transformation.*

## Keywords

Formant features of speech, emotional speech, statistical analysis.

## 1. Introduction

Emotion identification in speech depends on the chosen set of features extracted from the speech signal. These features are systematically divided into segmental and supra-segmental ones [1]. Short-term segmental features derived from the speech frames with short duration are usually in relation with the speech spectrum. These include traditional features like linear predictive coefficients (LPC), line spectral frequencies, mel-frequency cepstral coefficients (MFCC), or linear prediction cepstral coefficients (LPCC) [2].

During pleasant emotions the larynx and the pharynx are expanded, the vocal tract walls are relaxed, and the mouth corners are retracted upward. The result is falling first formant and raised resonances. For unpleasant emotions the larynx and the pharynx are constricted, the vocal tract walls are tensed, and the mouth corners are retracted downward. The result is more high-frequency energy, rising first formant, and falling second and third formants [3]. We can conclude that the first formant and the higher formants of emotional speech shift in opposite directions. For pleasant emotions the first formant shifts to the left, and the higher formants to the right. For unpleasant emotions the opposite situation occurs: the first formant shifts to the right, and the higher formants to the left.

Spectral features like MFCC or LPCC together with energy and prosodic parameters are most commonly used in voice and emotional speech classification [4]. On the other hand, in automatic speech recognition (ASR) systems based on the hidden Markov models (HMM) approach [5], the acoustic vector comprises such components as the formant central frequencies and bandwidths. Relative position of formants and formant trajectories can be used as the main indicator for speech classification in the voiced parts [6]. Together with complementary spectral features (spectral flatness and spectral entropy), also prosodic parameters (F0, microintonation, jitter, shimmer) will be used for classification of emotional speech types in our classifier based on the Gaussian mixture models (GMM) principle [7] that is currently developed.

In our experiments we performed statistical analysis and comparison of the formant features (FF) of male and female emotional speech representing joy, sadness, anger, and a neutral state in Czech and Slovak languages. It comprises analysis of the first three formant positions together with their bandwidths and the formant tilts [8]. In the case of the first three formant positions the histograms of distribution were also calculated and the extended statistical parameters (skewness and kurtosis) were subsequently determined from these histograms. To confirm the disjunction of obtained data groups for GMM recognition [9],

these histograms were further evaluated by the analysis of variances (ANOVA) approach [10] and the hypothesis tests [11] were used for numerical matching.

# 2. Subject and Method

The formant features consisting of the basic frequency parameters as the first three formant positions and their bandwidths as well as the complementary parameters (the formant tilts defined as directions and angles between the first three spectral maxima of the smoothed envelope) can be calculated by several techniques. In practice two approaches of the basic FF determination are mostly applied: the first one uses calculation from the complex roots of the LPC polynomial; the second one consists in finding of the local maxima of the smoothed spectral envelope where its gradient changes from positive to negative.

## 2.1 Smoothing of Spectral Envelopes

Mostly the formant positions and their bandwidths are determined from the smoothed envelope of the voiced parts of the speech signal. To obtain the smoothed spectral envelope, the mean periodograms of the chosen regions of interest (ROI) in the voiced parts of the speech signal can be computed by the Welch method [12]. By this approach we obtain an estimation of the power spectral density (PSD) of the input speech signal – it means the periodogram that uses an $N_{FFT}$-point FFT to compute the power spectral density as $S(e^{j\omega})/f_s$ where $f_s$ is the sampling frequency.

The smoothed spectral envelope of the speech signal can also be determined during the cepstral analysis. The cepstral analysis of the speech signal is performed in the following way: first, the complex spectrum using the FFT algorithm is calculated from the input samples (after segmentation and weighting by a Hamming window). In the next step, the power spectrum is computed and the natural logarithm is applied. Application of the inverse FFT algorithm gives the symmetric real cepstrum. Limitation to the first $N_0+1$ cepstral coefficients represents an approximation of the log spectrum envelope [13].

An autoregressive (AR) model is well known in speech processing as an LPC model being an all-pole model of a vocal tract. The autocorrelation method uses the Levinson-Durbin recursion to compute the parameters $\{a_k\}$ describing the speech spectral envelope in dependence on the chosen order $N_A$ of the AR model.

## 2.2 Calculation of Formant Features

Although the formant frequencies differ to some extent for different speakers and their ranges are overlapped [14] the male voice vowel formant areas without overlap can be determined: $F_1 \approx 250 \div 700$ Hz, $F_2 \approx 700 \div 2000$ Hz, $F_3 \approx 2000 \div 3200$ Hz [15]. Using the general knowledge of [14] that females have on average 20 % higher formant frequencies than males, the female voice vowel formant areas without overlap will be: $F_1 \approx 300 \div 840$ Hz, $F_2 \approx 840 \div 2400$ Hz, $F_3 \approx 2400 \div 3840$ Hz. We apply two methods for determination of the basic formant features:

1. Estimation of the formant frequencies and their bandwidths directly from the complex roots of the LPC polynomial $A(z)$ – poles of the LPC transfer function. Using the sampling frequency $f_s$, the formant frequency $F_n$ and the 3-dB bandwidth $B_n$ in [Hz] are determined as

$$F_n = \frac{f_s}{2\pi}\theta_n, \quad B_n = -\frac{f_s}{\pi}\ln|z_n| \tag{1}$$

where $\theta_n$ is the angle in [rad] of the complex root pairs $z_n = |z_n|e^{j\pm\theta_n}$.

2. For the formant positions as the first three local maxima of the smoothed spectral envelope where its gradient changes from positive to negative, the corresponding bandwidths are obtained as the frequency intervals between the points of the 3-dB decrease of the magnitude spectrum from the formant amplitudes.

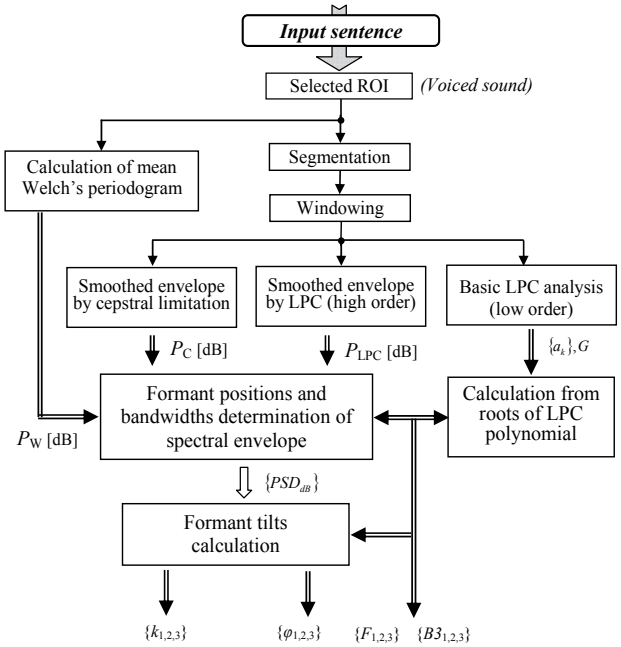


**Fig. 1.** Block diagram of used formant features determination method.

The indirect determination of the basic FF is realized using all three mentioned approaches to spectral envelope calculation and smoothing. In the case of the LPC envelope calculation the higher order is applied; in the case when the FF are calculated directly from the roots of the LPC polynomial, the lower order is applied – see the block diagram in Fig. 1. Correctness of the basic FF values obtained by all three indirect methods as well as by direct calculation from the roots is assessed by two criteria: the resulting values of 3-dB bandwidths must be less than 500 Hz [16], and the

found values of the first three formant positions must fall within the corresponding frequency interval in dependence on the voice type (male/female).

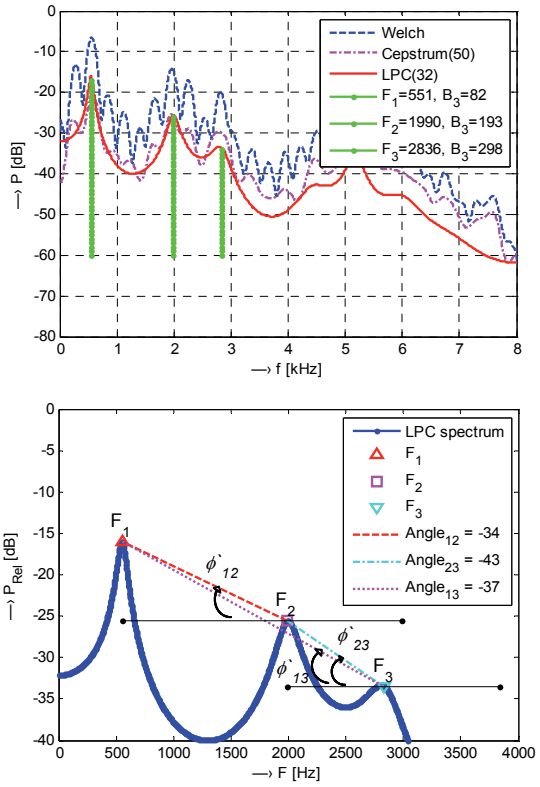


**Fig. 2.** Example of the FF determination from a long vowel "e" (female voice, F0 = 191 Hz, $f_s$ = 16 kHz): comparison of resulting smoothed spectral envelopes (upper), determination of formant tilts from LPC spectral envelope (lower); the complementary angles are calculated as $\varphi` = \varphi - 180$.

The complementary FF can be defined as formant tilts – angles between spectrum peaks in the place of the determined first three format positions (see documentary Fig. 2). The general bisector formula in the parametric form can be used for calculation

$$y - y_1 = k(x - x_1)\ ,\quad k = \frac{y_2 - y_1}{x_2 - x_1},\quad k = tg(\varphi) \qquad (2)$$

where $k$ is a bisector direction, $y_{1,2}$ represent values of PSD in [dB] of the determined formants, and $x_{1,2}$ are positions of the formants on the frequency axis in [Hz]. When $k < 0$ the formants have declining trend, when $k > 0$ the formants have ascending trend. The resulting angle $\varphi$ in degrees is defined as $\varphi = (\mathrm{Arctg}(k)/\pi)\cdot 180$.

## 2.3 Statistical Analysis and Comparison of Formant Feature Values

Obtained basic and complementary FF values are processed separately in dependence on the voice type (male / female), and sorted by emotional styles. The whole process of statistical analysis of FF values consists of six steps:

1. calculation of the basic statistics of the formant frequencies and their 3-dB bandwidths, and formant tilt parameters (directions and angles),
2. calculation and building of the histograms for $F_{1,2,3}$ frequencies,
3. calculation of extended statistical parameters from histograms (kurtosis and skewness),
4. calculation of the mean emotional-to-neutral $F_1$, $F_2$, $F_3$ formant position ratios,
5. evaluation of histograms by the ANOVA supplemented with multiple comparison of group means,
6. numerical matching by the hypothesis test.

Skewness is a measure of the asymmetry of the data around the sample mean. If the skewness is negative, the data are spread out more to the left of the mean than to the right. If the 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. We use these parameters together with other types of FF in the feature vector for GMM classification.

# 3. Material and Experiments

The main FF analysis was carried out on the speech corpus obtained from multi-medial CDs containing the Czech and Slovak stories performed by professional actors. At present, our database consists of sentences with duration from 0.5 to 5.5 seconds (resampled at 16 kHz), with different contents expressed in four emotional styles: "neutral", "joy", "sadness", and "anger" uttered by several speakers (134 sentences spoken by male voices and 132 sentences spoken by female voices, 8+8 speakers altogether). From the main speech signal database of sentences, the next one consisting of manually selected speech segments corresponding to the stationary parts of the vowels "a", "e", "i", "o", "u", and consonants "m" and "n" was consequently created for detailed analysis. Number of analyzed voiced frames was in total:

a) Male: neutral - 5103, joy - 4927, sadness - 4642, anger - 4391.

b) Female: neutral - 5223, joy - 4541, sadness - 4203, anger - 4349.

The frame length for spectral analysis depends on the mean pitch period of the processed signal. In our experiment, we had chosen 24-ms frames for the male voices, and 20-ms frames for the female voices. Calculation of the FF values was supplemented with determination of the fundamental frequency F0 by autocorrelation analysis method with experimentally chosen pitch ranges as fol-

lows: 55÷250 Hz for the male voices, and 105÷350 Hz for the female ones. Then, the F0 values were compared and corrected by the results obtained using the PRAAT program [17] with similar internal settings of F0 values. The obtained mean F0 values for all eight male and eight female speakers are shown in Tab. 1.

| Speaker / F0 [Hz] | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 |
|---|---|---|---|---|---|---|---|---|
| Male | 133 | 127 | 98 | 132 | 99 | 101 | 88 | 118 |
| Female | 228 | 177 | 207 | 198 | 215 | 205 | 201 | 219 |

**Tab. 1.** Speakers' mean F0 values for male and female voices.

At present, the developed GMM emotional speech classifier has only one-level structure as it can be seen in Fig. 3. This simple architecture expects that the gender of the voice (male/female) was correctly recognized in the previous process (manually, by listening tests, etc.) as a pre-processing phase that is usually used in speech recognition systems [18], [19]. Subsequently, the emotional speech style is identified for each of two gender classes. In our first GMM emotional style classification test, we use the feature set consisting of 16 values as the input data vector for GMM training and classification containing the basic spectral parameters: skewness and kurtosis from the histograms of $F_1$, $F_2$ values, the formant tilts, the complementary spectral parameters (harmonic-to-noise ratio, spectral flatness, and entropy), and supra-segmental parameters (F0, jitter, and shimmer) – see Tab. 2. For this experiment the number of mixtures ($N_{\text{gmix}}$) for every emotion model was set to four, and for control of the expectation-maximization algorithm training [18] the number of iteration steps ($N_{\text{iter}}$) was set to 1000.

| No | Feature name | Feature type | Value type |
|---|---|---|---|
| 1 | F1h*) | Basic | Skewness |
| 2 | F2h*) | Basic | Skewness |
| 3 | F1h*) | Basic | Kurtosis |
| 4 | F2h*) | Basic | Kurtosis |
| 5 | F12 formant tilt | Basic | Rel. Min |
| 6 | F12 formant tilt | Basic | Std |
| 7 | Harmonic-to-noise ratio | Complementary | Mean |
| 8 | Harmonic-to-noise ratio | Complementary | Std |
| 9 | Spectral flatness | Complementary | Mean |
| 10 | Spectral flatness | Complementary | Std |
| 11 | Spectral entropy | Complementary | Mean |
| 12 | Spectral entropy | Complementary | Std |
| 13 | F0 | Supra-segmental | Median |
| 14 | $F0_{\text{DIFF}}$ | Supra-segmental | Rel. max |
| 15 | Jitter | Supra-segmental | Rel. max |
| 16 | Shimmer | Supra-segmental | Median |

*) Calculated from histogram

**Tab. 2.** Used types of values in the feature vector set for GMM emotional speech style classifier.

To obtain speaker independent GMM classification, the data k-fold cross-validation method [18] was applied during the training and the testing processes. For our extracted data from the Czech and Slovak database the groups of speakers were divided by the ratio of 7:1 (seven for training, one for testing/classification – both voices together). For practical implementation of the GMM model creation, data training, and classification the basic functions from the Ian T. Nabney "Netlab" pattern analysis toolbox [20] were used.

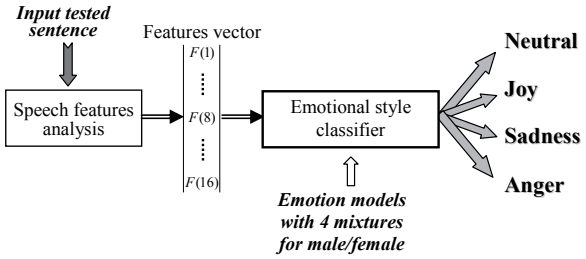


**Fig. 3.** Block diagram of currently developed GMM emotional speech style classifier for Czech and Slovak.

# 4. Obtained Results

Partial results of analysis of all voiced frames of the main speech corpus are presented in the form of the boxplot graphs of basic statistical parameters of the $F_{1,2,3}$ values determined from the neutral speech of the male and the female voices together with the bar graph of the $F_{1,2,3}$ mean frequencies (see Fig. 4). The values of the first three formant 3-dB bandwidths for both voices are presented in Fig. 5. Two diagrams of bisectors with directions given by formant tilts from the male and the female voices are shown in Fig. 6. Summary histograms of the first three formant frequencies for the male and female speech in different emotional styles are shown in Fig. 7 and 8.

The graphs with visualization of the difference between group means calculated using ANOVA statistics for different speech styles and separately for the male and the female voice can be seen in Figs. 9-11. These graphs are supplemented with the merged tables containing the null hypothesis/probability results for 5% significance level of the Ansari-Bradley test (see Tab. 3 and 4).

Results of extended statistical analysis of the FF values – skewness and kurtosis parameters determined from the histograms for male and female voice in neutral and emotional states are given in Tab. 5 and 6. Obtained results of the additional FF parameter analysis (mean values of the formant tilts) are presented in Tab. 7. Detailed results of the mean $F_{1,2,3}$ frequencies of the selected voiced sounds in neutral speaking style are shown in the common Tab. 8 for male and female voices. The summary results – the FF value ratios between different emotional states and a neutral state for male and female voice are given in Tab. 9.

The first results of the experiment with the GMM emotional style classifier are presented in the form of the bar graph of the confusion matrix for both voices and four emotional speaking states – see Fig. 12. Tab. 10 summarizes the achieved mean values of the GMM emotion recognition error rate in [%].

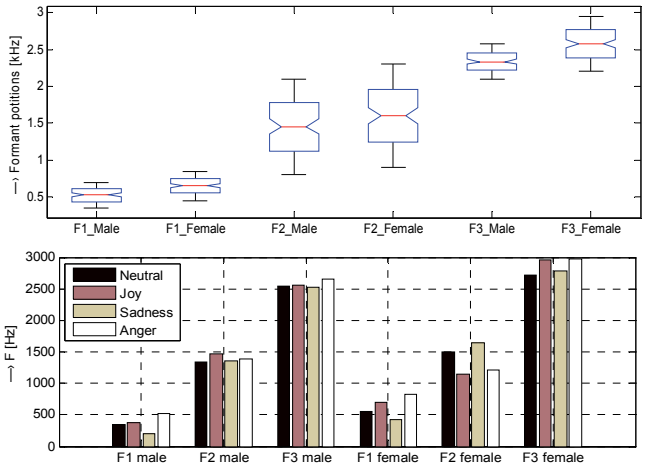


**Fig. 4.** Partial results of basic statistical parameters of the $F_{1,2,3}$ values for the male and female speech in a neutral style (upper), and bar graphs of mean values of the first three formant frequencies for different emotional states of male and female voices (lower).

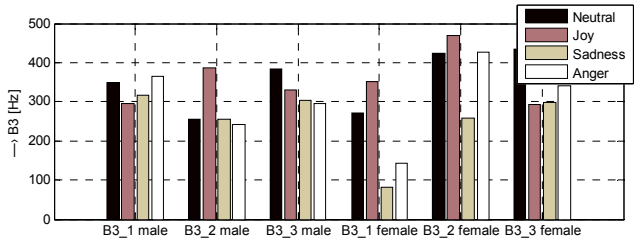


**Fig. 5.** Bar graphs of the 3-dB bandwidth mean values of the first three formant frequencies for different emotional states of male and female voices.

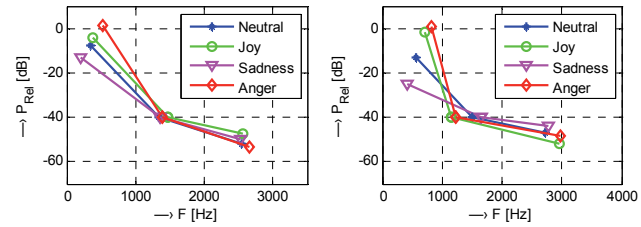


**Fig. 6.** Summary diagrams of bisectors with directions given by formant tilts for different emotional states: male (left), and female (right) voices.

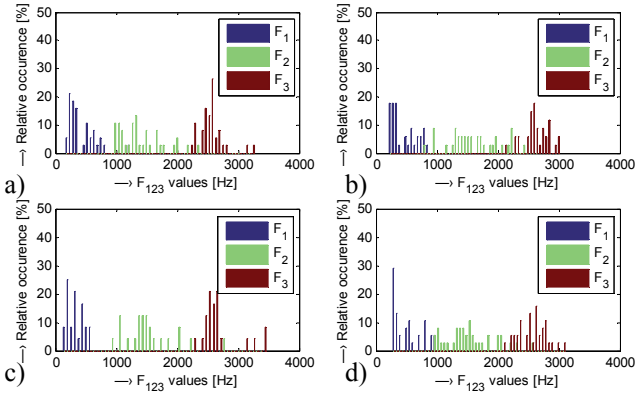


**Fig. 7.** Histograms of $F_{1,2,3}$ values for different emotional states: neutral (a), joy (b), sadness (c), and anger (d) – male voices.

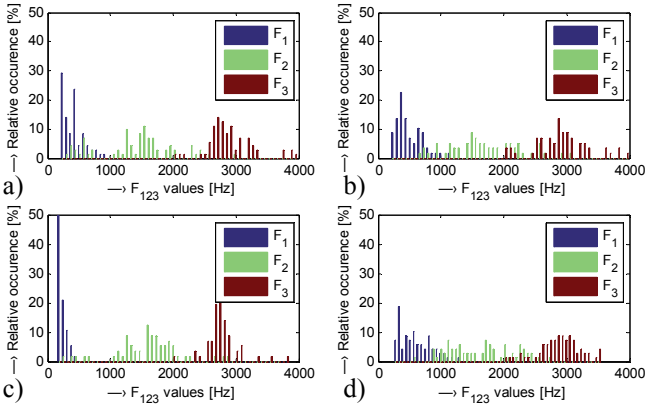


**Fig. 8.** Histograms of $F_{1,2,3}$ values for different emotional states: neutral (a), joy (b), sadness (c), and anger (d) – female voices.

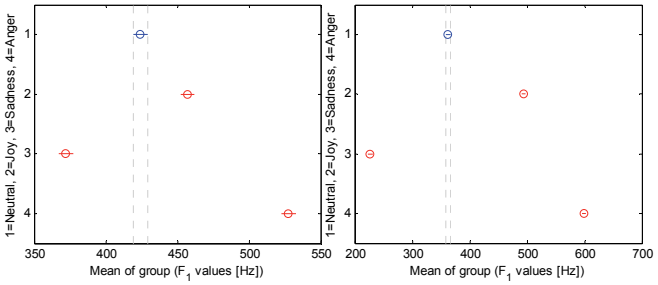


**Fig. 9.** Visualization of multiple comparison of group means applied to ANOVA results of histograms of $F_1$ positions for neutral and different emotional speech styles: male voice (left), female voice (right).

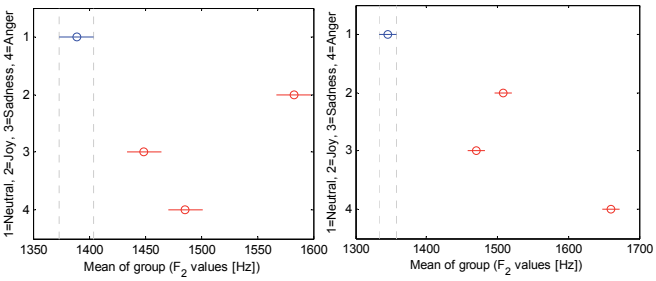


**Fig. 10.** Visualization of multiple comparison of group means applied to ANOVA results of histograms of $F_2$ positions for neutral and different emotional speech styles: male voice (left), female voice (right).

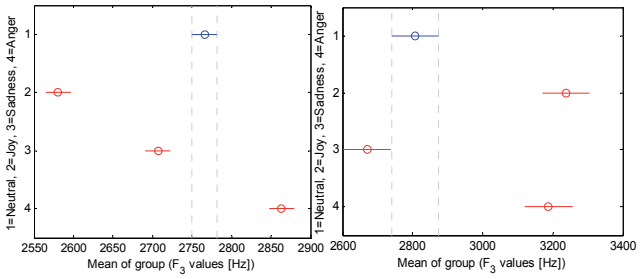


**Fig. 11.** Visualization of multiple comparison of group means applied to ANOVA results of histograms of $F_3$ positions for neutral and different emotional speech styles: male voice (left), female voice (right).

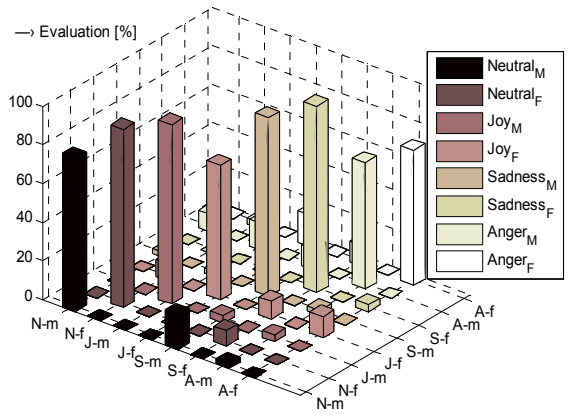


**Fig. 12.** Comparison of obtained GMM classification results in the form of the integrated confusion matrix for male and female voices; $N_{gmix} = 4$, $N_{iter} = 1000$.

| $F_x$ -Emotion | h/p: joy | h/p: sadness | h/p: anger |
|---|---|---|---|
| $F_1$ - Neutral | 1/2.35 $10^{-7}$ | 1/3.44 $10^{-15}$ | 1/2.52 $10^{-25}$ |
| $F_1$ - Joy | **0/1** | 1/8.17 $10^{-11}$ | 1/5.69 $10^{-18}$ |
| $F_1$ - Sadness | | **0/1** | 1/3.29 $10^{-34}$ |
| $F_2$ - Neutral | 1/1.68 $10^{-25}$ | 1/8.54 $10^{-9}$ | 1/4.26 $10^{-13}$ |
| $F_2$ - Joy | **0/1** | 1/2.17 $10^{-16}$ | 1/1.86 $10^{-11}$ |
| $F_2$ - Sadness | | **0/1** | 1/5.33 $10^{-3}$ |
| $F_3$ - Neutral | 1/2.63 $10^{-18}$ | 1/6.45 $10^{-3}$ | 1/4.27 $10^{-12}$ |
| $F_3$ - Joy | **0/1** | 1/1.13 $10^{-16}$ | 1/6.46 $10^{-36}$ |
| $F_3$ - Sadness | | **0/1** | 1/2.25 $10^{-20}$ |

**Tab. 3.** Merged hypothesis/probability values as results of the Ansari-Bradley hypothesis test of $F_{1,2,3}$ positions – male voice.

| $F_x$ -Emotion | h/p: joy | h/p: sadness | h/p: anger |
|---|---|---|---|
| $F_1$ - Neutral | 1/6.38 $10^{-8}$ | 1/7.24 $10^{-14}$ | 1/2.02 $10^{-22}$ |
| $F_1$ - Joy | **0/1** | 1/2.17 $10^{-12}$ | 1/3.28 $10^{-6}$ |
| $F_1$ - Sadness | | **0/1** | 1/4.52 $10^{-32}$ |
| $F_2$ - Neutral | 1/3.83 $10^{-11}$ | 1/3.72 $10^{-9}$ | 1/5.15 $10^{-28}$ |
| $F_2$ - Joy | **0/1** | 1/7.46 $10^{-4}$ | 1/3.73 $10^{-14}$ |
| $F_2$ - Sadness | | **0/1** | 1/4.66 $10^{-16}$ |
| $F_3$ - Neutral | 1/2.49 $10^{-15}$ | 1/8.36 $10^{-3}$ | 1/1.89 $10^{-15}$ |
| $F_3$ - Joy | **0/1** | 1/5.35 $10^{-28}$ | **0/4.52** |
| $F_3$ - Sadness | | **0/1** | 1/3.28 $10^{-16}$ |

**Tab. 4.** Merged hypothesis/probability values as results of the Ansari-Bradley hypothesis test of $F_{1,2,3}$ positions – female voice.

| Emotion type | Male voice | | | Female Voice | | |
|---|---|---|---|---|---|---|
| | $F_1$ | $F_2$ | $F_3$ | $F_1$ | $F_2$ | $F_3$ |
| Neutral | 0.405 | 0.166 | 0.074 | 0.379 | 0.085 | -0.009 |
| Joy | 0.451 | 0.374 | 0.429 | 0.426 | 0.305 | 0.326 |
| Sadness | 0.181 | -0.095 | -0.065 | -0.142 | -0.163 | -0.166 |
| Anger | 0.528 | 0.532 | 0.453 | 0.472 | 0.530 | 0.466 |

**Tab. 5.** Skewness parameters determined from the histograms of $F_{1,2,3}$ frequencies for male and female voice in neutral and emotional states.

| Emotion type | Male voice | | | Female Voice | | |
|---|---|---|---|---|---|---|
| | $F_1$ | $F_2$ | $F_3$ | $F_1$ | $F_2$ | $F_3$ |
| Neutral | 1.5176 | -0.741 | 3.269 | 2.477 | -0.308 | 5.332 |
| Joy | 0.349 | 1.878 | 0.706 | 0.493 | 2.248 | 0.427 |
| Sadness | 1.938 | -0.129 | 9.672 | 4.134 | 0.989 | 11.836 |
| Anger | -0.381 | 2.828 | 5.257 | -0.679 | 3.386 | 7.402 |

**Tab. 6.** Kurtosis parameters determined from the histograms of $F_{1,2,3}$ frequencies for male and female voice in neutral and emotional states.

| Emotion type | Male voice | | | Female Voice | | |
|---|---|---|---|---|---|---|
| | $\varphi'_{12}$ | $\varphi'_{23}$ | $\varphi'_{13}$ | $\varphi'_{12}$ | $\varphi'_{23}$ | $\varphi'_{13}$ |
| Neutral | -41 | -37 | -44 | -44 | -54 | -29 |
| Joy | -29 | -20 | -34 | -30 | -23 | -33 |
| Sadness | -44 | -49 | -40 | -52 | -64 | -19 |
| Anger | -34 | 10 | -47 | -15 | 14 | -25 |

**Tab. 7.** Mean values of formant tilts; complementary angles in [deg].

| Sound type | Male / female | | | |
|---|---|---|---|---|
| | $N_F$ [-] | $F_1$ [Hz] | $F_2$ [Hz] | $F_3$ [Hz] |
| "a" | 782 / 758 | 631 / 756 | 1363 / 1545 | 2529 / 2725 |
| "e" | 802 / 720 | 451 / 521 | 1590 / 1754 | 2490 / 2744 |
| "i" | 576 / 684 | 297 / 371 | 1479 / 1622 | 2415 / 2814 |
| "o" | 618 / 608 | 514 / 541 | 1158 / 1371 | 2514 / 2606 |
| "u" | 684 / 735 | 393 / 438 | 1113 / 1428 | 2531 / 2943 |
| "m" | 696 / 784 | 259 / 314 | 1186 / 1557 | 2544 / 2653 |
| "n" | 945 / 934 | 271 / 357 | 1211 / 1592 | 2557 / 2690 |

**Tab. 8.** Detailed results of the mean $F_{1,2,3}$ frequencies together with the number of processed frames; neutral speaking style, 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 |
| sadness: 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 |

**Tab. 9.** Mean emotional-to-neutral $F_{1,2,3}$ formant position ratios.

| Error rate /emotion | Neutral | Joy | Sadness | Anger |
|---|---|---|---|---|
| Male | 21.948 | 8.571 | 9.948 | 34.432 |
| Female | 9.094 | 29.272 | 4.827 | 25.769 |
| Total | 15.521 | 18.921 | 7.387 | 30.105 |

**Tab. 10.** Summarized mean values of GMM emotion recognition error rate in [%] for the emotional speech style classifier.

# 5. Discussion

Generally, it can be said that statistical distribution of the analyzed $F_{1,2,3}$ frequencies for male speech is better for our purpose than that for female speech. Values obtained from the female voices have higher standard deviation (compare box-plot graphs of basic statistical parameters in Fig. 4) and the frequencies of the formants are approximately about 15 % higher than that of the male voices. Contrary to it, the values of the formant 3-dB bandwidths have no correlation with the type of speaking style or the type of the voice (see the bar graph in Fig. 5). On the other hand, comparison of the formant tilts shows good differentiation between neutral and emotional styles (see graphs in Fig. 6) for both voices. The "anger" emotion has the greatest $\varphi'_{12}$ angle (the greatest ratio of PSD in [dB] at $F_1$ and $F_2$ frequencies) and the "sadness" emotion has the lowest $\varphi'_{12}$ angle for both voices. The complementary angles between PSD at frequencies $F_1$ and $F_2$ ($\varphi'_{12}$) and the complementary angles between PSD at frequencies $F_1$ and $F_3$ ($\varphi'_{13}$) have always negative values. The complementary angles between PSD at frequencies $F_2$ and $F_3$ ($\varphi'_{23}$) can have also positive values (the formants have ascending trend – see values in Tab. 7). Results of detailed analysis of seven voiced sounds represent differences between $F_{1,2,3}$ positions as it is documented in Tab. 8. However, in the case of the consonants "m" and "n" the differences of the $F_{1,2,3}$ values are lower due to smaller absolute amplitudes of the speech signal than for the vowels and they cannot be compared with sufficient accuracy.

Extended statistical parameters – skewness and kurtosis – subsequently calculated from histograms of $F_{1,2,3}$ frequencies also show correlation between the corresponding types of emotions for both voices (compare values in Tab. 5-6). Values of these histograms were next evaluated by the ANOVA approach. From the analysis of difference between group means calculated using ANOVA statistics follows that there also exists some "similarity" between individual groups. It is mainly expressed for the emotion groups "sadness" and "anger" for the male voice, and "joy" and "sadness" for the female voice in the case of $F_2$ positions, and the groups "neutral" and "sadness" for the female voice in the case of $F_3$ positions – the groups "joy" and "anger" are even overlapping.

The results of the first experiment with our GMM classifier show that this realization is applicable to emotional style classification (see confusion matrix in Fig. 12). From the basic spectral parameters only statistical values of the first two formant frequencies $F_1$ and $F_2$ and spectral tilts were used (see Tab. 2) since the frequencies $F_3$ have lower significance (differentiation) of the ANOVA statistics results (see Fig. 11 and values in Tab. 3-4). The obtained recognition error of the GMM classifier presented in Tab. 10 achieves acceptable values (the mean error rate for all four emotions and both voices is about 18 %). From the detailed results per emotions follows that recognition problems occur in the "anger" state of the male voice and in the "joy" state of the female voice.

# 6. Conclusion

The realized statistical comparison of the first three formant frequencies shows correlation of the results for the male and female voices inside the currently analyzed speech corpus, and significant differences between the data groups in the emotional and neutral styles. Therefore, these parameters can be used together with the values of the basic spectral properties and the prosodic parameters of speech in Czech and Slovak for creation of the database of values for the emotional speech classifier based on statistical approach that is currently being developed.

Alternatively, the anticipated application is in the voice communication systems with the human-machine (computer) interface [21] where emotion recognition (current emotional state of a user) helps to make communication more effective by selection of the suitable strategy of dialogue management. On the other hand, this speech material with achieved statistical properties can be used in the Czech TTS systems working on the statistical approach (based on the HMM) for synthetic speech personification [22] or expressive speech production [23]. However, at first we plan to use this GMM classifier for objective evaluation of emotional speech synthesis as an option to manually performed listening tests.

Our future aim will be to test the influence of the used values in the input feature vector and the initial parameter setting on creation and training of the GMM model, and further influence on the obtained emotion recognition score. It means to find out the best (optimal) feature set, the optimal number of used mixtures, and number of iterations for GMM emotion classification and voice recognition. In near future we would also like to supplement our Czech and Slovak speech database with another three emotions (boredom, surprise, etc.) and carry out extension of the GMM classifier for these emotional states. Considering the fact that our current database consists of speech only with acted emotional styles, the analysis of FF properties considering also speech material spoken under real emotions should be performed. Last but not least, we would like to use broader comparison with other databases in different languages (e.g. the German speech database Emo-DB [24], or international COST 2102 Italian Database of Emotional Speech [9]).

# Acknowledgements

This work has been supported by the Grant Agency of the Slovak Academy of Sciences (VEGA 2/0090/11) and the Ministry of Education of the Slovak Republic (VEGA 1/0987/12).

# References


[1] CHETOUANI, M., MAHDHAOUI, A., RINGEVAL, F. Time-scale feature extractions for emotional speech characterization. *Cognitive Computation,* 2009, vol. 1, p. 194-201.

[2] LUENGO, I., NAVAS, E., HERNÁEZ, I. Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia*, 2010, vol. 12, p. 490-501.

[3] SCHERER, K. R. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 2003, vol. 40, p. 227 to 256.

[4] HE, L., LECH, M., MADDAGE, N. C, ALLEN, N. B. Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. *Biomedical Signal Processing and Control*, 2011, vol. 6, p. 139-146.

[5] CHALOUPKA, Z., UHLÍŘ, J. Speech defect analysis using hidden Markov models. *Radioengineering,* April 2007, vol. 16, no. 1, p. 67-72.

[6] BOZKURT, E., ERZIN, E., ERDEM, Ç. E., ERDEM, A. T. Formant position based weighted spectral features for emotion recognition. *Speech Communication*, 2011, vol. 53, p. 1186-1197.

[7] YUN, S., YOO, C. D. Loss-scaled large-margin Gaussian mixture models for speech emotion classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, vol. 20, no. 2, p. 585-598.

[8] KOOLAGUDI, S. G., KROTHAPALLI, R. S. Two stage emotion recognition based on speaking rate. *International Journal of Speech Technology*, 2011, vol. 14, p. 35-48.

[9] ATASSI, H., RIVIELLO, M. T., SMÉKAL, Z., HUSSAIN, A., ESPOSITO, A. Emotional vocal expressions recognition using the COST 2102 Italian database of emotional speech. In Esposito, A. et al. (eds.) *Development of Multimodal Interfaces: Active Listening and Synchrony*, LNCS 5967, Springer-Verlag Berlin Heidelberg, 2010, p. 255-267.

[10] RENCHER, A. C., SCHAALJE, G. B. *Linear Models in Statistics*. Second edition. John Wiley & Sons, 2008.

[11] MIZUSHIMA, T. Multisample tests for scale based on kernel density estimation. *Statistics & Probability Letters*, 2000, vol. 49, p. 81-91.

[12] STOICA, P., MOSES, R. L. *Introduction to Spectral Analysis*. Prentice-Hall, 1997, p. 52-54.

[13] VÍCH, R., PŘIBIL, J., SMÉKAL, Z. New cepstral zero-pole vocal tract models for TTS synthesis. In *Proceedings of the IEEE Region 8 EUROCON'2001, Vol. 2, Section S22-Speech Compression and DSP*. Bratislava (Slovakia), 2001, p. 458-462.

[14] FANT, G. Acoustical analysis of speech. In Crocker, M.J. (ed.) *Encyclopedia of Acoustics*. John Wiley & Sons, 1997, p. 1589 to 1598.

[15] FANT, G. *Speech Acoustics and Phonetics*. Dordrecht: Kluwer Academic Publishers, 2004.

[16] ILK, H. G., EROĞUL, O., SATAR, B., ÖZKAPTAN, Y. Effects of tonsillectomy on speech spectrum. *Journal of Voice*, 2002, vol. 16, p. 580-586.

[17] BOERSMA, P., WEENINK, D. *Praat: Doing Phonetics by Computer* (Computer Program, Version 5.2.20). [Online] Cited 2011-03-25. Available at: http://www.praat.org/

[18] REYNOLDS, D. A., ROSE, R. C. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 1995, vol. 3, p. 72-83.

[19] BURGET, R., KARÁSEK, J., SMÉKAL, Z. Recognition of emotions in Czech newspaper headlines. *Radioengineering,* April 2011, vol. 20, no. 1, p. 39-47.

[20] NABNEY, I. T. *Netlab Pattern Analysis Toolbox*. Copyright (1996-2001). Retrieved 16 February 2012, from http://www.mathworks.com/matlabcentral/fileexchange/2654-netlab

[21] LEE, C. M. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, March 2005, vol. 13, no. 2, p. 293-303.

[22] HANZLÍČEK, Z., MATOUŠEK, J., TIHELKA, D. First experiments on text-to-speech system personification. In *Text, Speech and Dialogue*, LNCS vol. 5729, Springer-Verlag Berlin Heidelberg, 2009, p. 186-193.

[23] GRŮBER, M., HANZLÍČEK, Z. Czech expressive speech synthesis in limited domain comparison of unit selection and HMM-based approaches. In P. Sojka et al. (Eds.) *TSD 2012*, LNCS vol. 7499, Springer-Verlag Berlin Heidelberg, 2012, p. 656–664.

[24] BURKHARDT, F., PAESCHKE, A., ROLFES, M., SENDL-MEIER, W., WEISS, B. A. Database of German emotional speech. In *Proceedings of INTERSPEECH 2005*. ISCA, Lisbon (Portugal), 2005, p.1517-1520.


# About Authors ...

**Jiří PŘIBIL** was born in 1962 in Prague, Czechoslovakia. He received Ing (MSc) degree in Computer Engineering and Dr. (PhD) degree in Applied Electronics from the Electrotechnical Faculty, CTU Prague in 1991 and 1998, respectively. In the years 1994 to 2011 he had been working as a researcher at the Department of Digital Signal Processing of the Institute of Photonic and Electronics v.i.i. in Prague. Now, he is a scientific worker at the Department of Imaging Methods of the Institute of Measurement Science, Slovak Academy of Sciences in Bratislava. His research interests are speech analysis and synthesis, text-to-speech systems, NMR image processing.

**Anna PŘIBILOVÁ** received her MSc and PhD degrees from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology (FEEIT SUT) in 1985 and 2002, respectively. For six years she had been with Chirana Research Centre for Medical Equipment as a research assistant. Since 1992 she has been working as a university teacher at the Radioelectronics Department and since 2011 at the Institute of Electronics and Photonics of the FEEIT SUT in Bratislava. The main field of her research and teaching activities is audio and speech signal processing.