Spectral Flatness Analysis for Emotional Speech Synthesis and Transformation

Jiří Přibil[1] and Anna Přibilová[2]

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

Abstract. 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 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. Obtained statistical parameters and emotional-to-neutral ratios of their mean values show good correlation for both male and female voices and all three emotions.

Keywords: spectral flatness, speech analysis and synthesis, emotional speech.

1 Introduction

Spectral flatness (S_F) is a useful measure to distinguish between voiced and unvoiced speech [1]. Its usage in speech processing can be extended to empty speech pauses identification [2], whispered speech recognition in noisy environment [3], or voicing transition frequency determination in harmonic speech modelling [4]. In cepstral speech synthesis the spectral flatness measure was used to determine voiced/unvoiced energy ratio in voiced speech [5]. According to psychological research of emotional speech different emotions are accompanied by different spectral noise [6]. We control its amount by spectral flatness measure according to which the high frequency noise is mixed in voiced frames during cepstral speech synthesis [7]. We perform the statistical analysis of spectral flatness values in voiced speech for four emotional states: joy, sadness, anger, and a neutral state. Obtained statistical results of the spectral flatness ranges and values are shown also in the form of histograms in a way similar to that used by other authors for prosodic emotional features [8], [9].

A. Esposito and R. Vích (Eds.): Cross-Modal Analysis, LNAI 5641, pp. 106–115, 2009.

2 Subject and Method

As follows from the experiments, the S_F 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 each of several speakers (extracted from the Czech and Slovak stories performed by professional actors). Analysis must be preceded by classification and sorting process of the S_F values in dependence on voice type (male / female) and speaking style (neutral / emotional). The performed statistical analysis of spectral flatness values consists of the 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:

1. determination of mean ratio between neutral and emotional states,
2. visual comparison of histogram figures,
3. histograms fitting and modelling by Gamma distribution – comparison of parameters α, λ and Root Mean Square (RMS) approximation error.

2.1 Spectral Flatness Calculation Overview

The spectral flatness measure S_F calculated during the cepstral speech analysis is defined as

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

where the values $\left|S_k\right|^2$ represent the magnitude of the complex spectrum, and N_{FFT} is number of points of the Fast Fourier Transform (FFT) [10]. 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 – see Fig. 1. 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 demonstration example in Fig. 2 shows the input speech signal with detected pitch frequency F0 (pitch period reciprocal) and calculated S_F values with voiceness classification. The influence of median filtration applied to the L values is documented in Fig. 3.

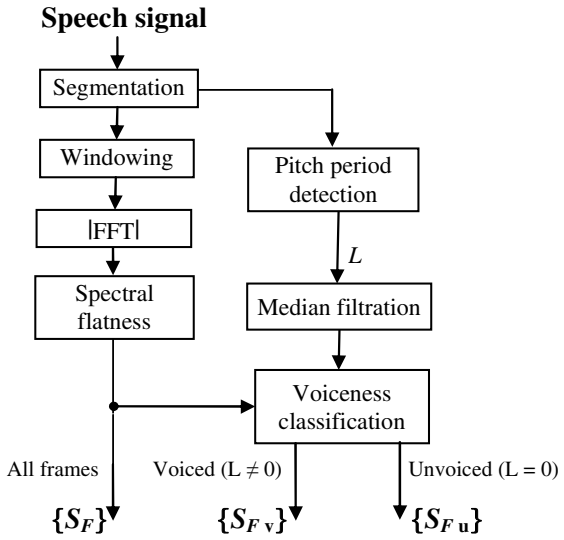

Fig. 1. Partial block diagram of speech analysis with spectral flatness values calculation

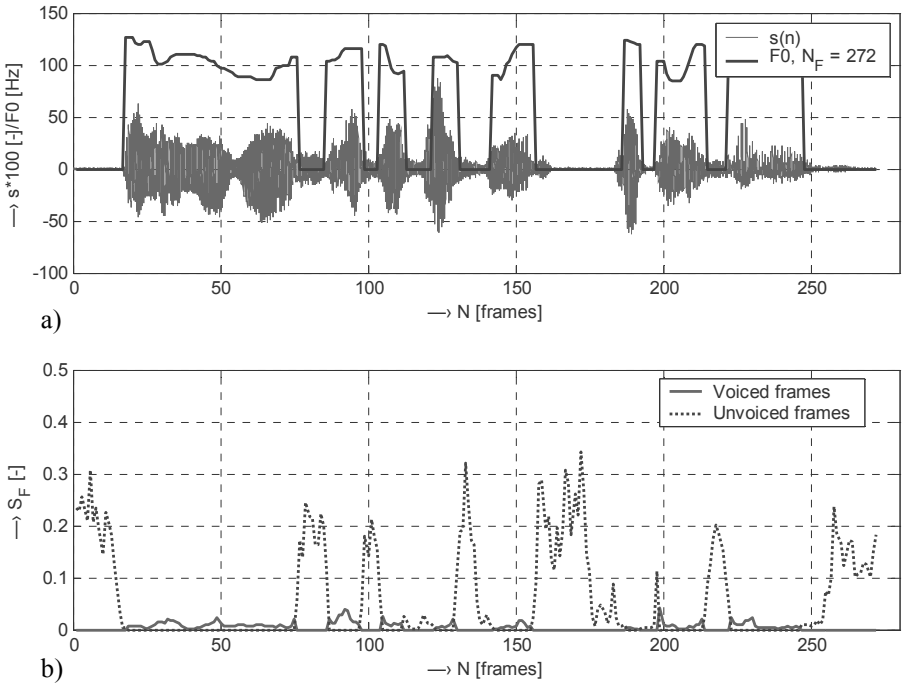

Fig. 2. Demonstration of spectral flatness calculation process: input speech signal – sentence "Lenivý si a zle gazduješ" ("You are lazy and you keep your house ill") pronounced in angry emotional style, male Slovak speaker with F0 contour (a), spectral flatness for voiced and unvoiced frames (b)

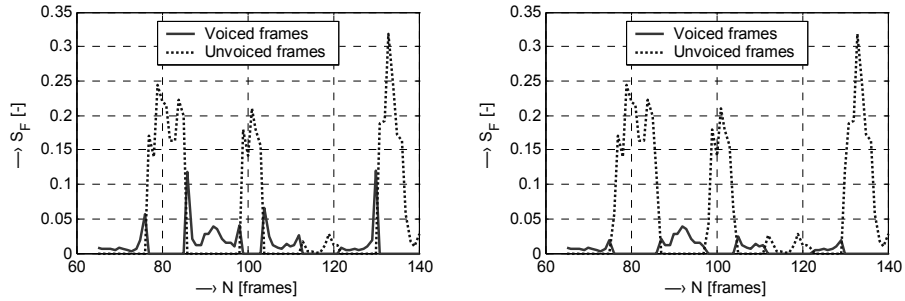

Fig. 3. Influence of median filtration pitch-period values on the S_F voiceness classification process (detail of frames 65÷140): without filtration (left), applied median filtration (right)

2.2 Statistical Analysis of Spectral Flatness Values

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 – see the block diagram in Fig. 4. 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). For the summary comparison the stack with all emotional styles is filled and processed – see the block diagram in Fig. 5.

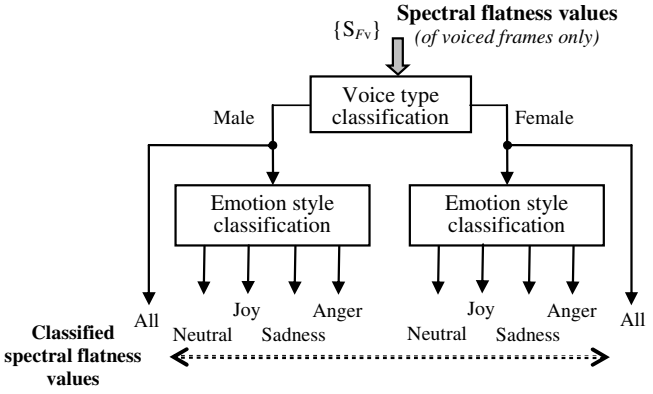

Fig. 4. Block diagram of used manual classification method of the S_F values

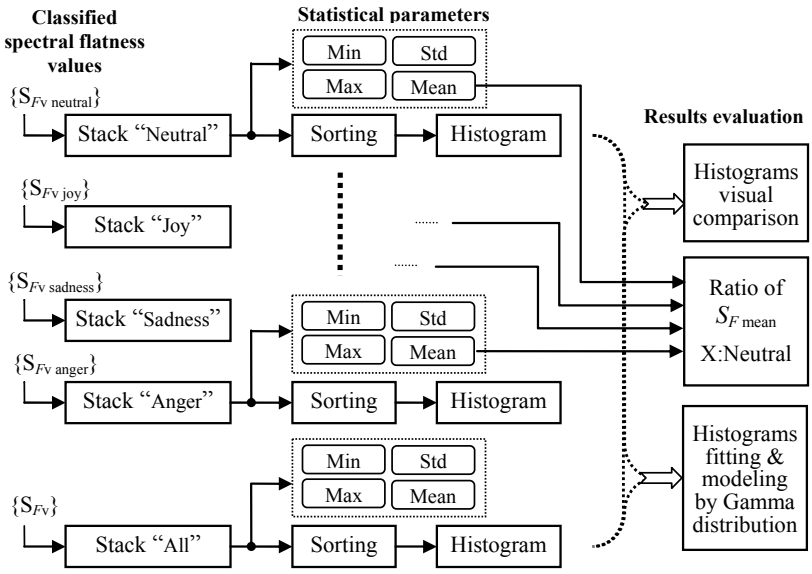

Fig. 5. Block diagram of used automatically processed operations with the stack filled with classified S_F values of voiced frames

2.3 Histograms Fitting and Modelling by Gamma Distribution

The generalized Gamma distribution of the random variable X is given by the probability density function (PDF) [11], [12]

$$f(x) = \frac{\lambda^a}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \qquad x \geq 0, \quad \alpha > 0, \quad \lambda > 0, \tag{2}$$

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

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx. \tag{3}$$

The graphs of the PDFs for different parameters α, λ are shown in Fig. 6.

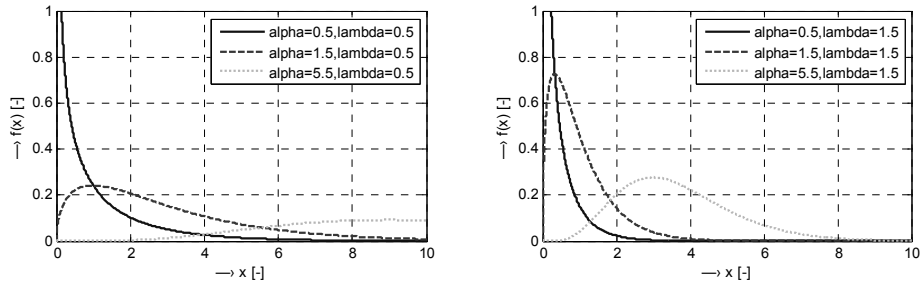

Fig. 6. Example of the Gamma probability density functions for $\lambda = 0.5$ (left), $\lambda = 1.5$ (right)

The shape and scale parameters of the Gamma distribution enable easy and rather accurate modelling of obtained histograms of S_F values. It means 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
   - 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 of 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
   - 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 of steps in this phase until the minimum of RMS)

3 Material, Experiments and Results

The speech material for spectral flatness analysis was collected in two databases (separately of male – 134 sentences, and female voice – 132 sentences) consisting of sentences with duration from 0.5 to 5.5 seconds. The sentences of four emotional states – "Neutral", "Joy", "Sadness", and "Anger" were extracted from the Czech and Slovak stories narrated by professional male and female actors. Pitch-contours given with the help of the PRAAT program [13] were used for segment determination as voiced or unvoiced. The PRAAT internal settings for F0 values 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 [14], pitch-range 35÷250 Hz for male and 105÷350 Hz for female voices.

Speech signal analysis was performed for total number of 25988 frames (8 male speakers) and 24017 frames (8 female speakers). The spectral flatness values were determined only from the voiced frames (totally 11639 of male and 13464 of female voice) – see statistical results in Tab. 1 (male), Tab. 2 (female). The main result – mean spectral flatness values ratios between different emotional states and a neutral state – is given in Tab. 3. Summary histograms of S_F values for different emotions in dependence on the speaker gender are shown in Fig. 7 (male) and Fig. 8 (female). For comparison, the histograms of unvoiced frames (male voice) are shown in Fig. 9. Tab. 4 (male) and Tab. 5 (female) contain parameters α, λ of the Gamma distribution for histogram fitting and modelling together with the resulting RMS approximation errors.

Table 1. 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 2. 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 3. Mean spectral flatness values 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 4. Evaluated parameters α, λ of Gamma distribution for histogram fitting and modelling together with resulting RMS approximation 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 5. 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

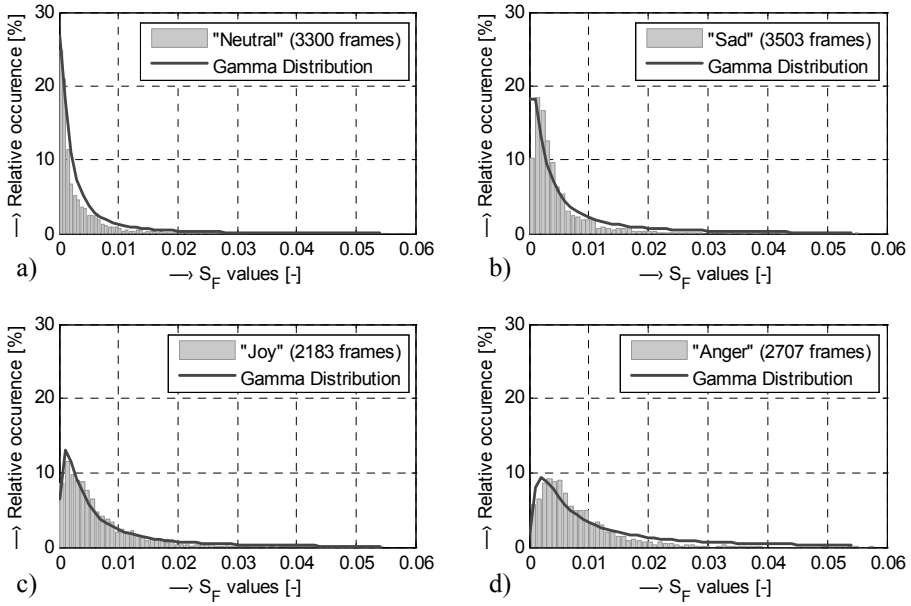

Fig. 7. Histograms of spectral flatness values together with fitted and modelled curves of Gamma distribution - determined from the speech signal with emotions: "neutral" (a), "sadness" (b), "joy" (c), and "anger" (d) - male voice, voiced frames

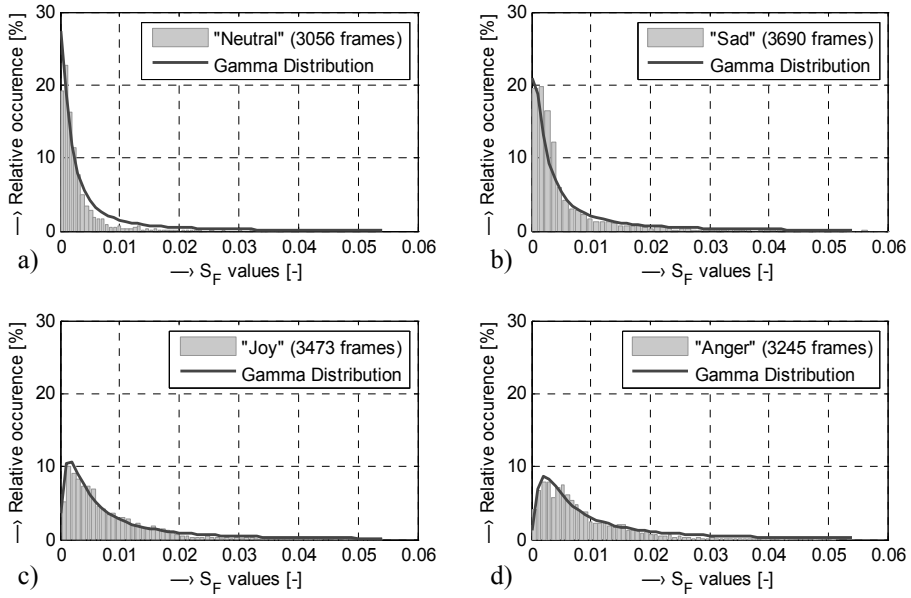

Fig. 8. Histograms of spectral flatness values together with fitted and modelled curves of Gamma distribution - determined from the speech signal with emotions: "neutral" (a), "sadness" (b), "joy" (c), and "anger" (d) - female voice, voiced frames

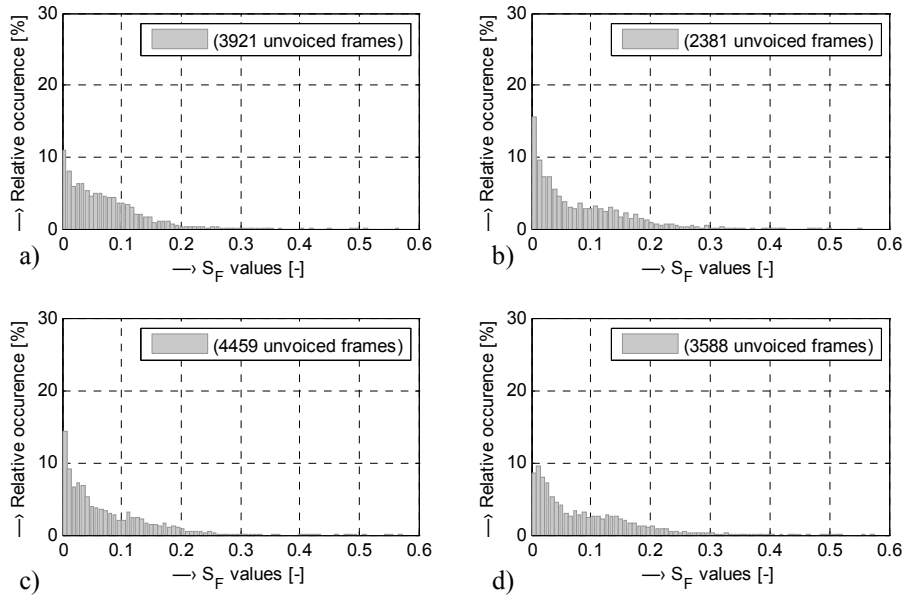

Fig. 9. Histograms of spectral flatness values calculated from the unvoiced frames (male voice): "neutral" style (a), and emotions - "joy" (b), "sadness" (c), and "anger" (d)

4 Conclusion

The statistical analysis of spectral flatness values was performed. Obtained statistical results of the spectral flatness ranges and values 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. 1 and Tab. 2. From Tab. 3 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. 7 and Fig. 8 comparing corresponding emotions for male and female voices. This subjective result is confirmed by the objective method – histogram modelling with the help of the Gamma distribution. Given vales of α and λ parameters – showed in Tab. 4 and Tab. 5 – are also in correlation with previously obtained results. On the other hand, it was confirmed that only S_F values calculated from voiced frames of speech give sufficient information – in Fig. 9 it is evident that the histograms are practically the same for all three emotions.

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. This parameter can be applied directly to the text-to-speech system enabling expressive speech production [15], or it can be used in emotional speech transformation (conversion) method based on cepstral speech description for modification of degree of voicing in voiced frames [4], [16].

Out next aim will be to find out how to use obtained statistical parameters of spectral flatness for evaluation of different emotional states in speech. Further these results can be used for determination of voicing transition frequency (for speech synthesis based on the harmonic speech model) [4].

Acknowledgments. 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), and the Ministry of Education of the Slovak Republic (COST2102/STU/08).

References

1. Gray Jr., A.H., Markel, J.D.: 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 (1974)
2. Esposito, A., Stejskal, V., Smékal, Z., Bourbakis, N.: The Significance of Empty Speech Pauses: Cognitive and Algorithmic Issues. In: Proceedings of the 2nd International Symposium on Brain Vision and Artificial Intelligence, Naples, pp. 542–554 (2007)
3. Ito, T., Takeda, K., Itakura, F.: Analysis and Recognition of Whispered Speech. Speech Communication 45, 139–152 (2005)
4. Přibil, J., Přibilová, A.: Voicing Transition Frequency Determination for Harmonic Speech Model. In: Proceedings of the 13th International Conference on Systems, Signals and Image Processing, Budapest, pp. 25–28 (2006)
5. Přibil, J., Madlová, A.: Two Synthesis Methods Based on Cepstral Parameterization. Radioengineering 11(2), 35–39 (2002)
6. Scherer, K.R.: Vocal Communication of Emotion: A Review of Research Paradigms. Speech Communication 40, 227–256 (2003)
7. Vích, R.: Cepstral Speech Model, Padé Approximation, Excitation, and Gain Matching in Cepstral Speech Synthesis. In: Proceedings of the 15th Biennial International EURASIP Conference Biosignal, Brno, pp. 77–82 (2000)
8. Paeschke, A.: Global Trend of Fundamental Frequency in Emotional Speech. In: Proceedings of Speech Prosody, Nara, Japan, pp. 671–674 (2004)
9. Bulut, M., Lee, S., Narayanan, S.: A Statistical Approach for Modeling Prosody Features Using POS Tags for Emotional Speech Synthesis. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Honolulu, Hawai, pp. 1237–1240 (2007)
10. Markel, J.D., Gray Jr., A.H.: Linear Prediction of Speech. Springer, Heidelberg (1976)
11. Suhov, Y., Kelbert, M.: Probability and Statistics by Example. Basic Probability and Statistics, vol. I. Cambridge University Press, Cambridge (2005)
12. Everitt, B.S.: The Cambridge Dictionary of Statistics, 3rd edn. Cambridge University Press, Cambridge (2006)
13. Boersma, P., Weenink, D.: Praat: Doing Phonetics by Computer (Version 5.0.32) [Computer Program], http://www.praat.org/ (retrieved August 12, 2008)
14. Boersma, P., Weenink, D.: Praat - Tutorial, Intro 4. Pitch analysis (September 5, 2007), http://www.fon.hum.uva.nl/praat/manual/Intro_4__Pitch_analysis.html
15. Přibil, J., Přibilová, A.: Application of Expressive Speech in TTS System with Cepstral Description. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction. LNCS (LNAI), vol. 5042, pp. 200–212. Springer, Heidelberg (2008)
16. Přibilová, A., Přibil, J.: Spectrum Modification for Emotional Speech Synthesis. In: Esposito, A., et al. (eds.) Multimodal Signals: Cognitive and Algorithmic Issues. LNCS (LNAI), vol. 5398, pp. 232–241. Springer, Heidelberg (2009)