Journal of ELECTRICAL ENGINEERING, VOL. 65, NO. 1, 2014, 30–36

EVALUATION OF SPECTRAL AND PROSODIC FEATURES OF SPEECH AFFECTED BY ORTHODONTIC APPLIANCES USING THE GMM CLASSIFIER

Jiří Přibil * — **Anna Přibilová** ** — **Daniela Ďuračková** **

The paper describes our experiment with using the Gaussian mixture models (GMM) for classification of speech uttered by a person wearing orthodontic appliances. For the GMM classification, the input feature vectors comprise the basic and the complementary spectral properties as well as the supra-segmental parameters. Dependence of classification correctness on the number of the parameters in the input feature vector and on the computation complexity is also evaluated. In addition, an influence of the initial setting of the parameters for GMM training process was analyzed. Obtained recognition results are compared visually in the form of graphs as well as numerically in the form of tables and confusion matrices for tested sentences uttered using three configurations of orthodontic appliances.

K e y w o r d s: spectral and prosodic features of speech, effect of orthodontic appliances, GMM classifier

1 INTRODUCTION

An orthodontic appliance is a mechanical tool for application of a pressure to the teeth and their supporting tissues to produce changes in the relationship of the teeth and/or the related osseous structures [1]. There are two large categories of these appliances: fixed and removable. Typical types of the fixed orthodontic appliances are braces in which small metal brackets are bonded to the centre of the teeth together with a metal wire running horizontally through the brackets to connect them. The removable appliance consists of active elements which exert orthodontic forces on the teeth, and retentive elements which help to retain the appliance in the mouth; finally a plastic plate holds these two sets of elements together. Wearing of orthodontic appliances as well as dental prostheses causes problems with articulation and speech intelligibility [2, 3]. In this case, articulation together with phonation and respiration are affected by physiological changes similar to the influence of "foreign objects" in the mouth, investigated in the well-known bite-block experiments [4].

In our previous research, we have tried to evaluate the determined spectral and prosodic properties of speech using the spectrograms, ANOVA analysis, and Ansari-Bradley hypothesis tests applied to the power spectral density (PSD) values of the spectrograms [5]. The same method was used for analysis of influence of the fixed and removable orthodontic appliances on spectral properties of emotional speech [6] as an alternative to the standard subjective comparison method the listening tests.

At present, our motivation is the use of the Gaussian mixture models (GMM) for automatic classification of the speech uttered by a person wearing orthodontic appliances. Therefore, this paper is focused on description of an experiment with using the GMM classifier for evaluation of influence of the upper removable plate and the lower conventional fixed orthodontic brackets and their combination on the spectral changes of speech in a neutral style.

2 METHOD

Disadvantage of the evaluation method based on the spectrograms is a necessity of using the same regions of interest (ROI) for comparison of calculated spectrograms. Therefore, normalization of the speech signal in the time domain had to be made first.

Our application of the GMM method to classification uses the features representing the spectral properties and prosodic parameters of the tested speech signal. These features are determined during pitch-asynchronous speech analysis performed in the frames of the fixed length and overlapping. Only statistical properties of these features are used for GMM evaluation. Thanks to this approach, the order of the features and the time duration of the analyzed speech frames have no influence on the correctness of the obtained results and there is no need for any further speech signal pre-processing to be applied.

2.1 Determination of speech spectral features and prosodic properties

Spectral features of speech can be determined in the course of cepstral analysis [7], during which the absolute values of the fast Fourier transform (FFT) are calculated

* Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia, Jiri.Pribil@savba.sk; ** Institute of Electronics and Photonics, Faculty of Electrical Engineering & Information Technology, Slovak University of Technology, Bratislava, Slovakia, {Anna.Pribilova; Daniela.Durackova}@stuba.sk

DOI: 10.2478/jee-2014-0004, Print ISSN 1335-3632, On-line ISSN 1339-309X,

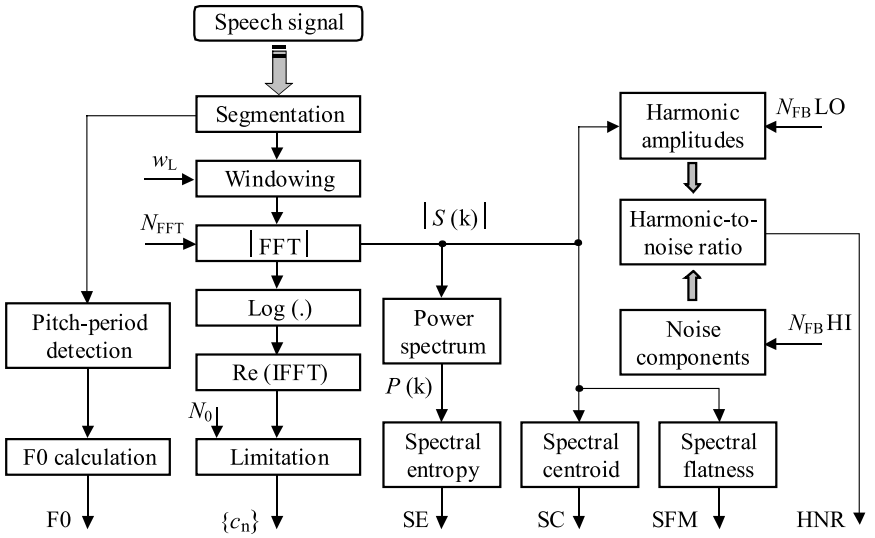

Fig. 1. Block diagram of spectral and prosodic analysis of the speech signal

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. Second application of the FFT algorithm gives the symmetric real cepstrum $\{c_n\}$. The cepstral speech analysis can be further used for determination of the complementary spectral features (CSF) [8]. These CSF include also the harmonics-to-noise ratio (HNR) providing an indication of the overall periodicity of the speech signal. Specifically, it quantifies the ratio between periodic and aperiodic components in the signal. Noise at harmonic locations is typically estimated as an average of the noise estimates at both sides of the harmonic locations. The harmonic portion of the spectrum is summed from low frequencies corresponding to the interval about 70 – 4500 Hz; the noise portion is calculated from high frequencies corresponding to the interval about 1500 – 4500 Hz. The spectral centroid (SC) is a centre of gravity of the power spectrum and it represents an average frequency weighted by the values of the normalized energy of each frequency component in the spectrum. The spectral flatness measure (SFM) can be used to determine the degree of periodicity in the signal. This spectral feature is calculated as a ratio of the geometric and the arithmetic mean values of the PSD. The spectral entropy (SE) as a measure of spectral distribution quantifies a degree of randomness of spectral probability density represented by normalized frequency components of the spectrum. Depending on the type of the feature, the resulting values are calculated either from voiced frames of the analyzed utterance or from both voiced and unvoiced frames. Therefore, determination of the spectral features is supplemented with detection of the pitch period L [9] in samples and calculation of the fundamental frequency F0 in Hertz — see block diagram in Fig. 1. The detected pitch period L is used for preliminary classification of voicing of the frames. 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. For the special purposes, the auxiliary speech spectral properties consisting of the formant positions, their 3 dB bandwidths, and the spectral tilts are determined [10]. The estimation of the formant frequencies and their bandwidths can be determined directly from the linear prediction coding (LPC) polynomial complex roots corresponding to the poles of the LPC transfer function using the Newton-Raphson or the Bairstow algorithm [11].

As regards supra-segmental speech properties, wearing of the orthodontic appliances has influence only on the microintonation component which can be supposed to be a random, band-pass signal described by its spectrum and statistical parameters. For analysis we use the jitter, the shimmer, and the relative number of signal zero crossings (ZCR). The absolute jitter values are calculated as the average absolute difference between consecutive pitch periods measured in samples. In the case of the shimmer measure determination, a period-to-period variability of amplitudes of a speech signal is used.

2.2 Basic principles of GMM classification

The Gaussian mixture models [12] can be defined as a linear combination of multiple Gaussian probability distribution functions (GPDF) of the input data vector x

$$f(x)=\sum_{k=1}^{K}\frac{\alpha_k}{\sqrt{(2\pi)^d \det \Sigma}}\exp\Big(-\frac{(x-\mu)^\top\Sigma^{-1}(x-\mu)}{2}\Big), \tag{1}$$

where d is the dimension of the GPDF, and K is the number of these distribution functions in a mixture. The covariance matrix Σ, the vector of the mean values μ, and the weighting parameters α_k must be determined from the input training data. Using the expectation-maximization (EM) iteration algorithm [13] the maxi-

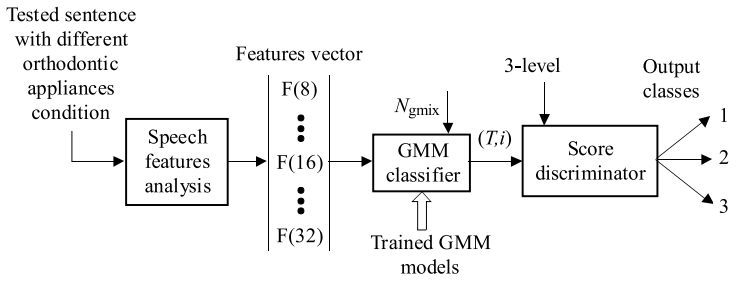

Fig. 2. Block diagram of currently developed GMM classifier

Table 1. Used types of values in the basic feature vector set for GMM classifier

No	Feature name	Frame type	Value type
1	HNR	Voiced	Mean
2	HNR	Voiced	Std
3	Spectral tilt	Voiced	Min
4	Spectral spread	All	Mean
5	SC	Voiced	Mean
6	SC	Voiced	Std
7	SFM	Voiced	Mean
8	SFM	Voiced	Std
9	SE	All	Mean
10	SE	All	Std
11	Signal ZCR	All	Median
12	Signal ZCR	All	Std
13	Jitter	Voiced	Median
14	Jitter	Voiced	Rel. max
15	Shimmer	All	Median
16	Shimmer	All	Rel. max

mum likelihood function of the GMM is defined as

$$\log L\big(\Theta\big|x\big)= \log \prod_{m=1}^{M} \sum_{k=1}^{K} \alpha_k P_k\big(x_m\big|\Theta_k\big), \qquad (2)$$

where $Pk(\ )$ are the GPDFs, M is the number of the trained vectors, and the term $\Theta = (\mu, \Sigma)$ represents the parameters of the Gaussian probability distribution. For control of the EM algorithm, the N_{gmix} parameter represents the number of used mixtures in each of the GMM models, and the N_{iter} corresponds to the number of iteration steps. The iteration stops when the difference between the previous and the current probabilities fulfils the internal condition or the predetermined maximum number of iterations is reached. In the evaluation phase, the GMM classifier returns the partial $score(T, i)$ representing the probability value of the models trained for the i-th evaluated class, where T is the input vector of the features obtained from the tested sentence. The resulting class i^* is given by its maximum overall probability using the relation

$$I^* = \arg \max_{1 \le i \le N} score(T, i)\,, \qquad (3)$$

where N is the number of all partial scores corresponding to the number of the classes.

3 EXPERIMENTS AND RESULTS

Our experiments were aimed at analysis of:

- influence of the used number of mixtures on GMM classification error rate, $1 \le N_{\mathrm{gmix}} \le 7$,
- influence of the used number of training iterations on GMM classification correctness, $100 \le N_{\mathrm{iter}} \le 1000$,
- influence of different length of the feature vector on GMM classification error rate, $N_{\mathrm{FEAT}} \in \{8, 16, 32\}$,
- influence of different length of the feature vector on the computational time (complexity) of GMM creation, training, and classification.

For analysis we use the sentences uttered under three types of conditions:

1. without orthodontic appliances (NO OA),
2. with the lower fixed orthodontic brackets (LF OB),
3. with the upper removable plate and the lower fixed orthodontic brackets (UP LB).

The speech material was recorded using the Behringer professional Podcastudio USB with the dynamic cardioid microphone Ultravoice XM8500 and the mixing console Xenyx 502 connected to a personal computer through the UCA200 high-performance audio interface. The collected speech database consists of 300 records with mean duration of 5 seconds uttered in a neutral speaking style. Every record consists of five concatenated words with a similar phonetic sound in Czech but often totally different meaning (*eg* "pes", "nes", "ves" – in English: "dog", "carry", "village") usually used in the rhythm test for evaluation by the automatic speech recognition systems (ASR) [14]. These speech records were uttered by a female speaker with F0$\approx$ 200 Hz, recorded at 32 kHz, and subsequently resampled to $f_s = 16$ kHz. Setting of the parameters for spectral analysis was chosen in correspondence with the speaker's mean F0 in this way: window length $L_W = 180$ samples, window overlapping $L_O = 40$ samples, and $N_{FFT} = 1024$ points. In our algorithm, the values of the complementary spectral features SC and SFM are obtained only from the voiced speech frames. In the case of the SE and the HNR parameters the values are determined from the voiced as well as unvoiced frames. The supra-segmental parameter jitter is calculated from the voiced frames only, opposite to the shimmer when both types of the frames are used for determination.

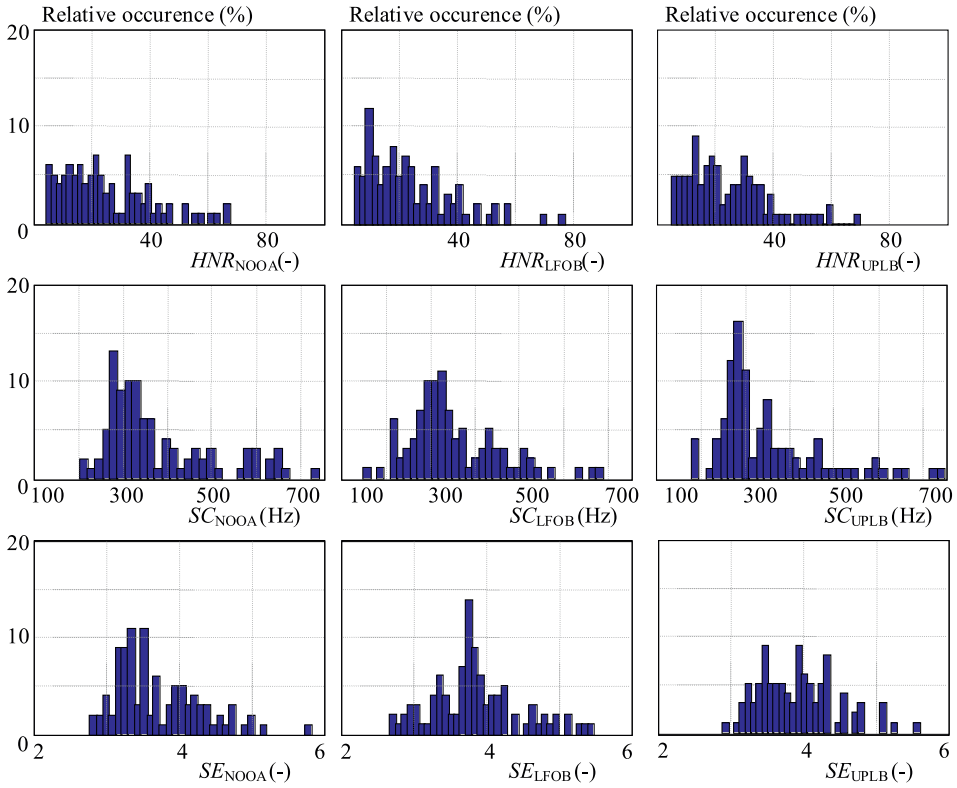

Fig. 3. Histograms of selected spectral features (HNR, SC, and SE) for different configurations of the orthodontic appliances

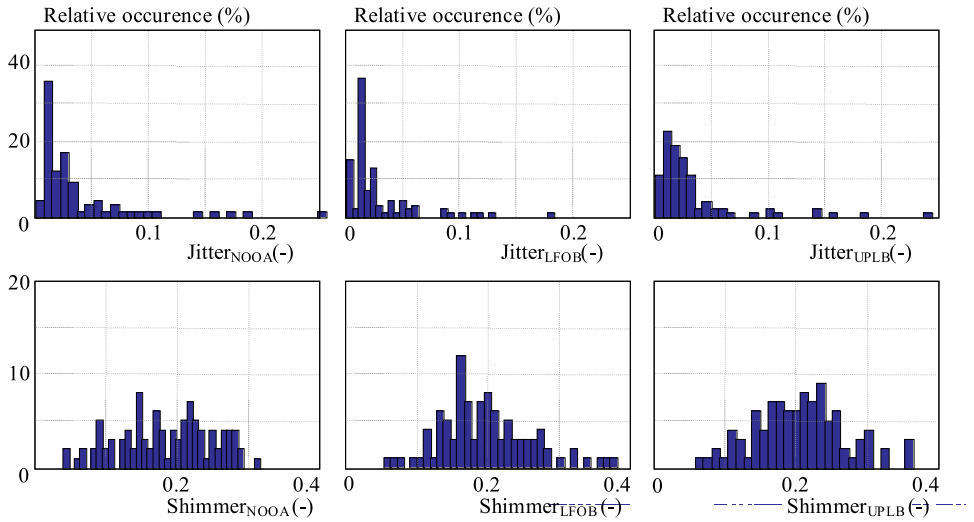

Fig. 4. Histograms of supra-segmental parameters (jitter and shimmer) for different configurations of the orthodontic appliances

At present, the developed GMM classifier has only one-level structure and three-level score discriminator for three final output classes as it can be seen in the block diagram in Fig. 2.

The simple architecture of the classifier expects that GMM models used for final classification were trained with speech data in any order from the collected database, without any pre-processing or auxiliary signal operations that are usually used in the speech recognition systems [15]. For practical implementation of the input feature vector for GMM evaluation, the representative values in the form of the basic statistical parameters — mean value and standard deviation (std) — were used for description of the speech basic and auxiliary spectral features. In the case of the spectral features represented by the real cepstral coefficients, the histograms of distribution were used to determine the extended statistical parameters — skewness and kurtosis. For implementation of the supra-segmental parameters the statistical types of median values, range of values, std, and/or relative maximum and minimum we used in the feature vectors.

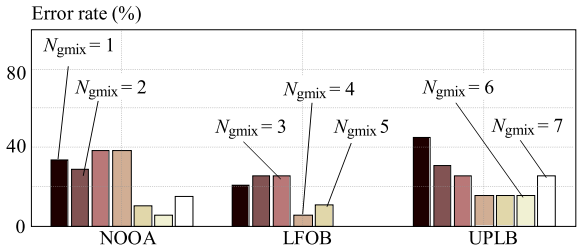

Fig. 5. Influence of the number of used mixtures on the GMM error rate; $N_{\text{iter}} = 600$, feature set P16

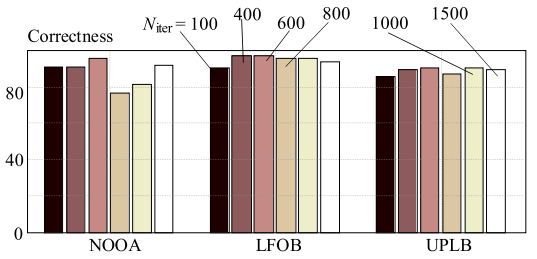

Fig. 6. Influence of the number of training iterations on the correctness of GMM classification; $N_{\text{gmix}} = 6$, feature set P1

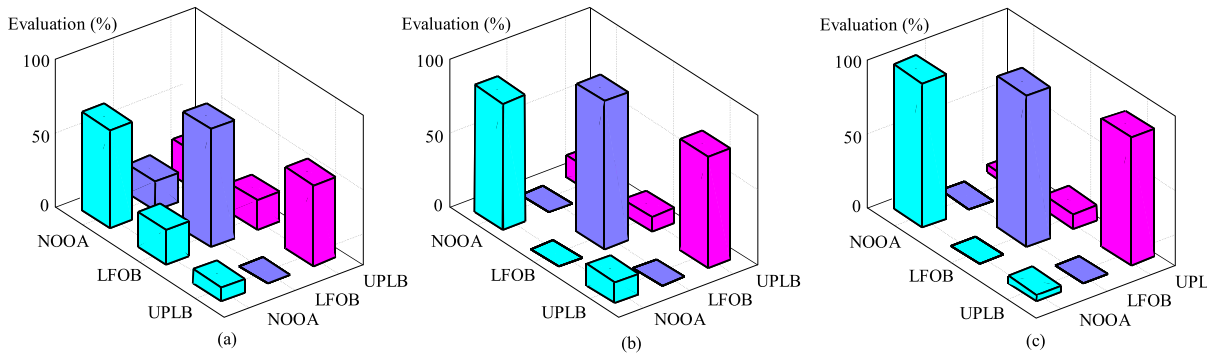

Fig. 7. Comparison of GMM classification results by the confusion matrices, $N_{\text{gmix}} = 6$, $N_{\text{iter}} = 600$, feature set: (a) — P8, (b) — P16, and (c) — P32

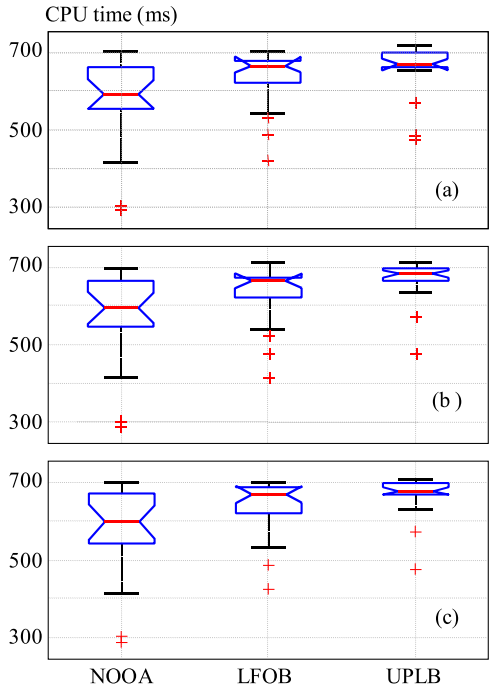

Fig. 8. Results of the basic statistical parameters of the CPU times during the classification phase with $N_{\text{gmix}} = 6$, feature set: (a) — P8, (b) — P16, (c) — P32

Three types of vectors with different lengths of N_{FEAT} = 8, 16, and 32 were used for analysis of different number of values in the feature vector. In the case of the shortest one with the length of 8 we used the parameters {1, 5, 7, 9, 11, 12, 14, 16} of the original feature vector with the length $N_{FEAT} = 16$ — see Table 1. The longest feature set consisting of 32 values includes an extended selection of the spectral parameters: the skewness and the kurtosis of the first six cepstral coefficients, the formant ratios of the first three formant frequencies F_1, F_2, F_3, the values of all types of complementary spectral features (HNR, SC, SFM, and SE), and the extended statistical values of the supra-segmental parameters.

The basic functions from the Ian T. Nabney "Netlab" pattern analysis toolbox [16] were used for creation of the GMM models, data training, and classification. The computational complexity for two algorithmic phases (the first consisting of model creation and training, and the second containing classification) was tested using the obtained mean CPU times on the PC with the processor Intel(R) i3-2120 at 3.30 GHz, 8 GB RAM, and Windows 7 professional OS.

Histograms of selected spectral features (HNR, SC, and SE) calculated from the speech signal data for different configurations of the orthodontic appliances are shown in Fig. 3, and histograms of chosen supra-segmental parameters (jitter and shimmer) are presented in Fig. 4. The obtained results of the experiment with the GMM classifier describing the influence of the number of used mixtures on the GMM error rate are shown in Fig. 5; the results of detailed numerical comparison are given in Table 2. The results of influence of the used number of training iterations on the correctness of GMM classification are presented in Fig. 6 as a bar graph; the detailed numerical results can be seen in Table 3. The achieved classification results in the form of the confusion matrices for three configurations of the orthodontic

Table 2. Influence of N_{gmix} parameter on the GMM classification mean error rate in (%) – summary results for all three configurations of orthodontic appliances; $N_{\text{iter}} = 600$, feature set P16

Value/N_{gmix}	1	2	3	4	5	6	7
Minimum	20.02	25.05	25.04	5.26	9.52	1.14	2.28
Maximum	45.01	30.02	38.09	33.85	15.59	15.84	21.03
Mean	32.78	27.85	29.36	19.36	11.51	6.58	13.09
Std	12.51	2.58	7.56	16.97	3.03	7.66	12.54

Table 3. Influence of N_{iter} parameter on the correctness of GMM classification in (%) – summary results for all three configurations of orthodontic appliances; $N_{\text{gmix}} = 6$, feature set P16

Value/N_{iter}	100	400	600	800	1000	1500
Minimum	89.59	89.44	84.16	89.15	86.41	82.16
Maximum	90.48	91.31	98.86	91.74	94.87	95.26
Mean	90.16	90.24	93.42	90.40	88.66	87.80
Std	0.28	0.78	5.89	2.55	7.12	8.73

Table 4. Comparison of the GMM classification mean error rate in (%) for different lengths of the feature vector; $N_{\text{gmix}} = 6$, $N_{\text{iter}} = 600$

N_{FEAT}	Mean classification error rate		
	NO OA	LF OB	UP LB
8	33.32	20.68	55.89
16	14.28	5.71	25.14
32	9.82	3.66	15.46

Table 5. Comparison of the GMM classification mean error rate in (%) for different lengths of the feature vector; $N_{\text{gmix}} = 6$, $N_{\text{iter}} = 600$

N_{FEAT}	Creation	Classification of			Summarized
	and training	NO OA	LF OB	UP LB	CPU time
8	469	599	643	675	1108
16	651	605	649	682	1296
32	781	609	652	684	1429

appliances are presented in Fig. 7; the numerical results of the mean classification error rate are shown in Table 4. Table 5 summarizes the mean CPU times for different lengths of the feature vectors (8/16/32 values) calculated as the duration of creation and training phase summed with the mean duration of the classification phase averaged over all the tested configurations of the orthodontic appliances. The results of the basic statistical parameters of the measured CPU times during the classification phase for different lengths of the feature vector are presented in the form of box-plot graphs in Fig. 8.

4 DISCUSSION AND CONCLUSION

The performed experiments have successfully confirmed that the chosen conception of on-level architecture of the GMM classifier is correct and the system is functional and usable for classification of spectral and prosodic changes in the speech signal produced with different configurations of the orthodontic appliances.

From the analysis of the influence of the initial parameters on creation and training of the GMM model follows that there is a substantial relationship between the number of the used mixtures and the number of the classes that are to be recognized — in our case, the types of the speech signal uttered with/without wearing of the orthodontic appliances. The number of the mixtures should be greater or equal to twice the number of the output recognized classes. Contrary to it, choice of the number of iterations has not great weight when its order is more than hundreds, therefore the optimum value about six hundred was chosen. As is documented by histograms of selected spectral properties and supra-segmental parameters (see Figs. 3 and 4), obtained values have good differentiation for all three analyzed configurations of the orthodontic appliances, so they can be statistically matched and subsequently used for classification based on GMM approach. For improvement of classification correctness, a comparative experiment with different types of features in the feature vector should be realized in the future.

From next comparison follows that the obtained GMM classification error rate using only 8 parameters in the feature vector gives the mean value of 37 %. However, error rates for condition UP LB were more than 50 %, what makes the whole classifier practically unusable. Comparison of the attained mean error rates between classification with the help of the feature vector consisting of 32 values and with the basic length of 16 values gave ambiguous results (see Table 4). While the extension to 32 values brought a little improvement in the overall mean error rate of 10 % when compared to 15 % error rate for the length of 16 values. On the other hand, the summarized results of the achieved CPU times shown in Table 5 are in correspondence with general expectancy: the maximum corresponds to the feature vector of 32 values, the minimum corresponds to the length of 8 features, and using the feature vector of 32 values (in comparison with the basic one consisting of 16 values) causes increasing of the mean CPU time only by 9 %, which is relatively negligible. Detailed analysis of the basic statistical parameters (minimum, maximum, and mean values) of the CPU times during the classification phase show great similarity (see Fig. 8), so only the mean values are used for next comparison.

In near future, we plan to collect a larger database of speech records from more speakers (male/female), especially the children or the young, who very often wear the orthodontic appliances. We would also like to evaluate influence of the time duration of the orthodontic appliances wearing upon the spectral changes of speech signals. For

this purpose, speech recordings of the same person must be done in the periods of about half a year corresponding also to the time of the vimprovement of the teeth position.

Acknowledgment

The work has been supported by the Grant Agency of the Slovak Academy of Sciences (VEGA VEGA 2/0013/14) and the Ministry of Education of the Slovak Republic (VEGA 1/0987/12).

References

[1] HILTON, L.: Orthodontic Appliances Information on Healthline, Gale Encyclopedia of Nursing and Allied Health, The Gale Group Inc., Gale, Detroit, 2002.

[2] HOHOFF, A.—SEIFERT, E.—FILLION, D.—STAMM, T.—HEINECKE, A.—EHMER, U.: American Journal of Orthodontics and Dentofacial Orthopedics **123** (2003), 146-152.

[3] KONG, H. J.—HANSEN, C. A.: Customizing Palatal Contours of a Denture to Improve Speech Intelligibility, The Journal of Prosthetic Dentistry **99** (2008), 243-248.

[4] LANE, H.—DENNY, M.—GUENTHER, F. H.—MATTHIES, M. L.—MÉNARD, L.—PERKELL, J. S.—STOCKMANN, E. —TIEDE, M.—VICK, J.—ZANDIPOUR, M.: Effects of Bite Blocks and Hearing Status on Vowel Production, Journal of Acoustical Society of America **118** No. 3 (2005), 1636-1646.

[5] PŘIBIL, J.—PŘIBILOVÁ, A.: An Experiment with Evaluation of Emotional Speech Conversion by Spectrograms, Measurement Science Review **10** No. 3 (2010), 72–77.

[6] PŘIBIL, J.—PŘIBILOVÁ, A.—ĎURAČKOVÁ, D.: An Experiment with Spectral Analysis of Emotional Speech Affected by Orthodontic Appliances, Journal of Electrical Engineering **63** No. 5 (2012), 296-302.

[7] VÍCH, R.: Cepstral Speech Model, Padé Approximation, Excitation, and Gain Matching in Cepstral Speech Synthesis, In Proceedings of the 15th Biennial EURASIP Conference Biosignal 2000, Brno, Czech Republic, 2000, pp. 77-82.

[8] HOSSEINZADEH, D. KRISHNAN, S.: On the Use of Complementary Spectral Features for Speaker Recognition, EURASIP Journal on Advances in Signal Processing (2008), Article ID 258184.

[9] VEPREK, P.—SCORDILIS, M. S.: Analysis, Enhancement and Evaluation of Five Pitch Determination Techniques, Speech Communication **37** (2002), 249-270.

[10] KOOLAGUDI, S. G.—KROTHAPALLI, R. S.: Two Stage Emotion Recognition Based on Speaking Rate, International Journal of Speech Technology **14** (2011), 35–48.

[11] SHAH, N. H.: Numerical Methods with C++ Programming, Prentice-Hall of India Learning Private Limited, New Delhi, 2009.

[12] REYNOLDS, D. A.: Speaker Identification and Verification using Gaussian Mixture Speaker Models, Speech Communication **17** (1995), 91-108.

[13] MOON, T. K.: IEEE Signal Processing Magazine (Nov 1996), 47–60.

[14] VÍCH, R.—NOUZA, J.—VONDRA, M.: Automatic Speech Recognition used for Intelligibility Assessment of Text-to-Speech Systems, In Verbal and Nonverbal Features of Human-Human and Human-Machine Interactions (Esposito A., Bourbakis N., Avouris N., Hatrzilygeroudis I., eds.), LNAI vol. 5042, Springer-Verlag Berlin Heidelberg, 2008, pp. 136–148.

[15] REYNOLDS, D. A.—ROSE, R. C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, IEEE Transactions on Speech and Audio Processing **3** (1995), 72–83.

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

Received 24 April 2013

Jiří Přibil (Ing, PhD), born in 1962 in Prague, Czechoslovakia. He received his MSc degree in computer engineering in 1991 and his PhD degree in applied electronics in 1998 from the Czech Technical University in Prague. At present, he is a scientific worker at the Department of Imaging Methods Institute of Measurement Science, Slovak Academy of Sciences in Bratislava. His research interests are signal and image processing, speech analysis and synthesis, and text-to-speech systems.

Anna Přibilová (Ing, PhD) 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. 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.

Daniela Ďuračková (Prof, Ing, PhD) received her MSc and PhD degrees from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology (FEEIT SUT) in 1974 and 1981, respectively. Since 1991 she has been an associate professor and since 2005 a professor at the Microelectronics Department (since 2011 the Institute of Electronics and Photonics) of the FEEIT SUT in Bratislava. The main field of her research and teaching activities has moved from semi-conductor devices towards the design of analog and digital ASICs and neural network implementation on chip.