DE GRUYTER OPEN

Journal of ELECTRICAL ENGINEERING, VOL. 66, NO. 4, 2015, 194–202

STORYTELLING VOICE CONVERSION: EVALUATION EXPERIMENT USING GAUSSIAN MIXTURE MODELS

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

In the development of the voice conversion and personification of the text-to-speech (TTS) systems, it is very necessary to have feedback information about the users' opinion on the resulting synthetic speech quality. Therefore, the main aim of the experiments described in this paper was to find out whether the classifier based on Gaussian mixture models (GMM) could be applied for evaluation of different storytelling voices created by transformation of the sentences generated by the Czech and Slovak TTS system. We suppose that it is possible to combine this GMM-based statistical evaluation with the classical one in the form of listening tests or it can replace them. The results obtained in this way were in good correlation with the results of the conventional listening test, so they confirm practical usability of the developed GMM classifier. With the help of the performed analysis, the optimal setting of the initial parameters and the structure of the input feature set for recognition of the storytelling voices was finally determined.

K e y w o r d s: storytelling voice conversion, spectral and prosodic features of speech, evaluation of speech quality, GMM classifier

1 INTRODUCTION

Storytelling speaking style may be used for narration of stories for children [1, 2] or in special book reading software for blind users [3]. Prosodic variations corresponding to fairy tale speech can improve not only storytelling quality but naturalness as well [4]. Our previous research improvement of text-to-speech (TTS) synthesis was aimed at storytelling speaking style in addition to its multivoice realization [5] and expression of emotional states [6]. The achieved results of performed storytelling voice transformation had shown existence of audible differences between the basic sentences generated by the TTS system and those modified by storytelling voice transformation, however, not all cases of applied storytelling voices had good naturalness [7]. For that reason, it is very necessary for us to get feedback informati on about the users' opinion on the resulting synthetic speech quality. Several subjective and objective methods are used to verify the quality of the produced synthetic speech [8, 9]. The most often used subjective method is the listening test; among the objective methods, the automatic speech recognition system yielding the final evaluation in the form of a recognition score can be used [10]. These recognition systems are often based on neural networks [11], hidden Markov models [12], or Gaussian mixture models (GMM) [13]. As these statistical evaluation methods work automatically (without any human interaction) and the achieved results can be subsequently numerically compared, we decided to use the GMM-based approach in our evaluation experiment.

Motivation of this work was to verify that the quality of synthetic speech produced by a TTS system after applied storytelling voice conversion can be evaluated by identification of the original storytelling voice and whether the identification score depends on the used method of speech modelling and production. Spectral features like mel frequency cepstral coefficients together with energy and prosodic parameters are most commonly used in GMM speaker or speech recognition [13] as well as environmental sound classification [14, 15]. However, because of correspondence with applied storytelling voice transformation method [7], the speech features determined from the spectral envelopes, as well as the supplementary spectral parameters, and the prosodic parameters were used in the described GMM identification. The performed experiments were next oriented on analysis of the influence of initial settings in the GMM creation and training phases (number of used mixture components and processed iterations) and different types of used speech features on correctness of the GMM identification. In addition, the computational complexity (computing times) in dependence on the number of used mixture components is analysed in the paper.

2 SUBJECT AND METHOD

2.1 Applied Method for Storytelling Voice Conversion

The storytelling voice conversion method is based on spectral transformation and prosodic parameters modification according to the prototype in a similar way as it had been used for emotional style conversion [16]. More

[*] Department of Imaging Methods, Institute of Measurement Science, Slovak Academy of Sciences in Bratislava, jiri.pribil@savba.sk

[**] Institute of Electronics and Photonics, Faculty of Electrical Engineering and Information Technology STU, Ilkovičova 3, SK-812 19 Bratislava

DOI: 10.2478/jee-2015-0032, Print ISSN 1335-3632, On-line ISSN 1339-309X © 2015 FEI STU

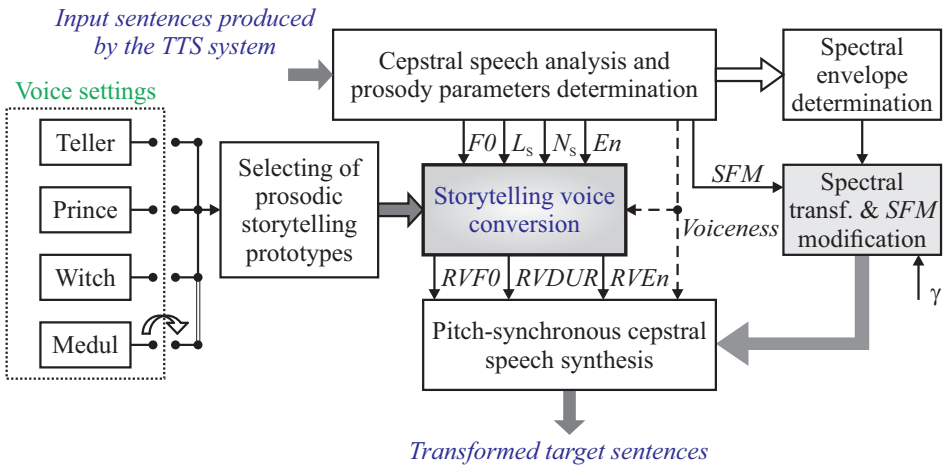

Fig. 1. Block diagram of post-processing application for storytelling speech conversion

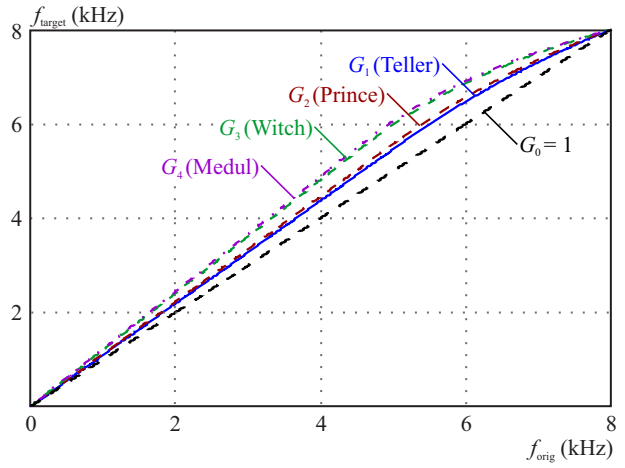

Fig. 2. Visualization of the non-linear frequency scale mapping functions for different values of γ parameter used for spectral envelope transformation in correspondence to the transformed voice type for $f_s = 16$ kHz

Table 1. Summary results of mean significant frequencies ratios between different storytelling voices and the TTS (for all vowels)

	Teller:TTS	Prince:TTS	Witch:TTS	Medul:TTS
mean ratio γ	1.0924	1.1146	1.2214	1.2026

detailed description of the applied cepstral analysis and synthesis method can be found in [17], and the preparation process of storytelling voice prototypes is similar to the method of emotional style prototype creation [7]. The storytelling voice conversion method was realized as the post-processing operation on the speech utterances produced by the TTS system — see the block diagram in Fig. 1. The cepstral analysis of the prototype as well as the source sentences is performed with segmentation in correlation with the used storytelling voice (male / female), and speech signal processing is performed in dependence on the determined type of voiceness (voiced/unvoiced segment). The developed conversion method consists of the following steps.

- Analysis of significant frequency positions, calculation of mean ratios between different storytelling voices and TTS.
- Evaluation of the mean values of the spectral flatness measure (SFM) [18].
- Prosodic parameter analysis — determination of the pitch frequency (F0), energy (En), and time duration (DUR) contours from the original sentences.
- Cepstral analysis of source sentences generated by the TTS system with basic prosody by rules.
- Linear time scale mapping from the original to the target sentence prosodic contours applying the compression or expansion of the virtual contours of VF0, VEn, and VDUR.
- Building of the storytelling prototypes for modification of prosodic parameters as the relative RVF0, RVEn, and RVDUR contours.
- Resynthesis of the target sentences by the cepstral speech model with transformed significant frequency positions, and with applied modification of the SFM values and prepared prosodic prototypes.

The stationary parts of vowels ‘a’, ‘e’, ‘i’, ‘o’, and ‘u’ for each of four storytelling voices and the TTS synthetic voice were extracted from utterances in the same phonetic context. Subsequently, the analysis of positions of the first six significant frequencies $f_1 \div f_6$ in the frequency interval from 80 Hz to 5.5 kHz was carried out on the spectral envelope obtained by inverse B-spline filtering. From these results, the mean $f_1 \div f_6$ values for all vowels and all voices were determined. In the next step, the mean ratios γ of all these frequencies between different voices and the TTS were calculated — see the values in Table 1 and visualization of the used non-linear frequency transformation functions in Fig. 2. The spectral flatness measure values were determined only from the voiced frames, and subsequently the SFM ratios between different voices and TTS were calculated. Each of the storytelling voice prosodic prototypes consists of five relative virtual contour files RVF0, VRDUR, and VREn separately for voiced and unvoiced frames — see the mean values in Tab. 2.

Table 2. Mean values of SFM and prosodic parameter ratios in dependence on the frames voiceness

Voice type/ X:TTS ratio	F0	DUR (voiced/ unvoiced)	En (voiced/ unvoiced)	SFM (voiced only)
Teller	1.05	1.02/0.91	0.99/0.85	1.6336
Prince	1.18	0.97/0.64	1.57/0.99	1.8577
Witch	0.88	1.94/1.75	2.15/2.17	1.8922
Medul	1.28	0.79/1.18	0.75/1.04	1.9883

2.2 Basic Description of the Used GMM Classification Method

The Gaussian mixture models [13] can be defined as a linear combination (mixture) of multiple Gaussian probability distribution functions $P_k(x)$ of the input data vector x

$$f(x) = \sum_{k=1}^{M} \alpha_k P_k(x)\,, \tag{1}$$

where α_k is the weighting parameter and M is the number of these functions with dimension d expressed as

$$P_k(x) = \frac{1}{\sqrt{(2\pi)^d \det(\sum_k)}} \exp\Big(-\frac{(x-\mu_k)^\top \sum^{-1}(x-\mu_k)}{2}\Big), \tag{2}$$

where $\sum_k$ is the covariance matrix and μ_k is the vector of mean values. For the GMM creation it is necessary to determine $\sum_k$, μ_k, and α_k from the input training data. Using the expectation-maximization (EM) iteration algorithm the maximum likelihood function of the GMM is found. 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. The GMM classifier returns the probability (so called score) that the tested utterance belongs to the GMM model. The resulting score i^* is given by the maximum overall likelihood for the given class using the $score(T, n)$ representing the likelihood value of the GMM classifier for the models trained for the current n-th class in the evaluation process, and the input vector T of the features obtained from the tested sentence. This relatively simple and robust approach cannot achieve the best recognition accuracy in all cases. In our experiment the developed GMM-based identification algorithm using the accumulated score was used for final decision about the classified original storytelling voice. The accumulated score can be expressed by the relation

$$i_{ACC} = \underset{1\le n\le M}{\arg\max} \bigcup_{p=1}^{P} (i^*(n,p) \equiv n)\,, \tag{3}$$

where $i^*(n,p)$ represents the resulting score for the current p-th window, P is the number of windows in the sentence, and the union operator represents the occurrence rate of the n-th class. Practical realization of the described identification algorithm resulted in an experimental one-level structure of the GMM classifier for classification of the original storytelling voice as shown in the block diagram in Fig. 3. The original storytelling voice identification block uses the GMM models that were created and trained on the data of the feature vectors obtained from the original sentences with different storytelling voices performed by actors. This proposed identifier architecture expects that input feature vectors from the tested sentences with the transformed storytelling voices are processed along with the feature vectors from the sentences resynthesized in a neutral style by the TTS system (for comparison). The obtained individual values of $score(T, n)$ are further used for calculation of the accumulated score i_{ACC} and depending on the used discrimination level the M output classes of original storytelling voices are finally determined.

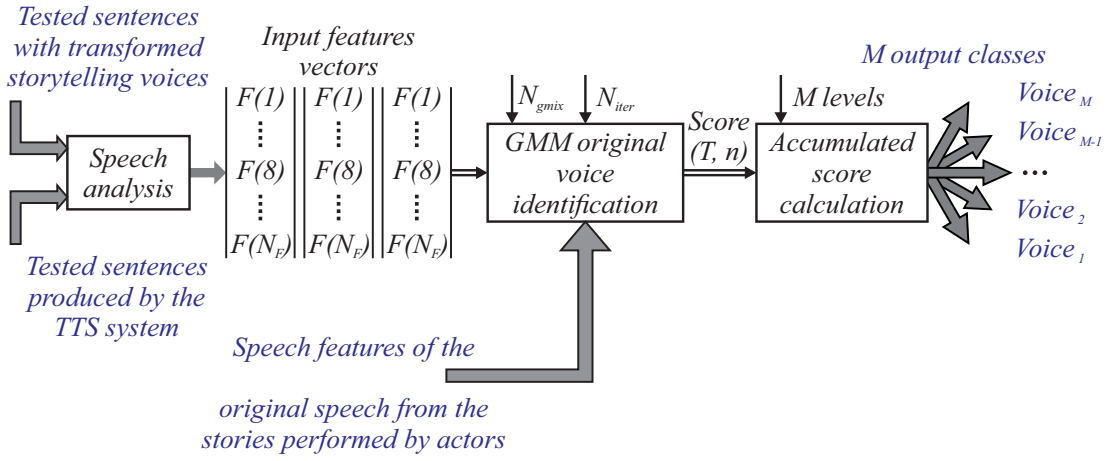

Fig. 3. Block diagram of the developed GMM-based classifier for identification of the original storytelling voice from converted synthetic speech produced by the TTS system

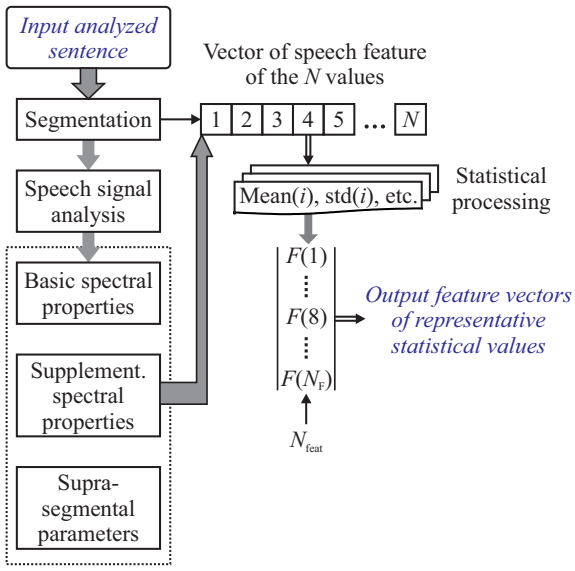

Fig. 4. Block diagram of determination of the feature vectors from the speech spectral properties and supra-segmental parameters

2.3 Determination of Speech Spectral Features and Prosodic Parameters

The speech signal analysis is performed in the following way: in the first step the fundamental frequency F0 is determined from the input sentence after segmentation and weighting. The obtained F0 contour is subsequently used for determination of the supra-segmental parameters describing the microintonation component of speech melody. As the next operation, the smoothed spectral envelope and the power spectral density (PSD) are computed from the speech frames. These basic parameters are used for further processing — determination of supra-segmental features, basic and supplementary spectral properties. The obtained values are subsequently processed statistically to determine representative values for the feature vectors with the length N_{feat} used in the GMM classifier — see block diagram in Fig. 4.

Spectral features of speech can be determined in the course of cepstral analysis. To eliminate the frames with very low energy of the analysed speech signal, the energy contour is calculated from the first cepstral coefficient c_0 [17] and the frames in the beginning and at the end of the sentences with the energy lower than the threshold $En_{\min}$ are removed — see the demonstration example in Fig. 5. After this limitation, the rest frames are used for processing — computing of the smoothed spectral envelope, the PSD values, and other spectral properties determination.

Cepstral coefficients $\{c_n\}$ obtained during the cepstral analysis bring information about spectral properties of the human vocal tract [17]. It means that these coefficients can be directly used in the feature vector for GMM classification. The basic spectral properties describe the shape of the spectrum obtained from the analysed speech segment. They include the first two formant frequencies F_1, F_2, and their ratios (F_1/F_2) together with the spectral centroid (SC) representing an average frequency weighted by the values of the normalized energy of each frequency component in the spectrum, and the spectral decrease (tilt). 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 [19].

The cepstral speech analysis can be further used for determination of the complementary spectral features [20] including the spectral entropy (SE) as a measure of spectral distribution quantifying a degree of randomness of spectral probability density represented by normalized frequency components of the spectrum, the SFM which can be calculated as a ratio of the geometric and the

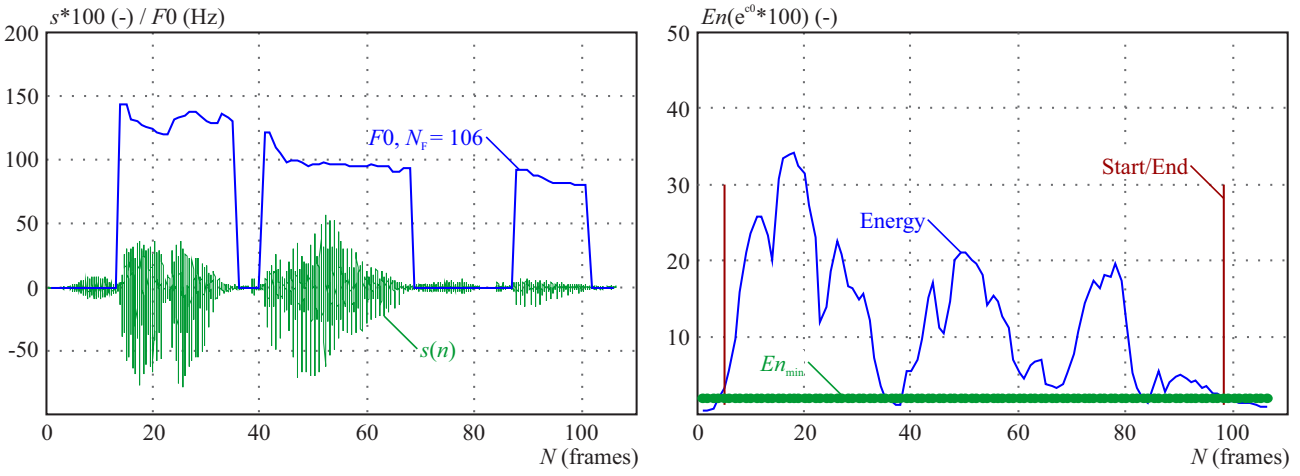

Fig. 5. Pre-processing of the analysed speech signal: the input sentence generated by the TTS system ($f_s = 16$ kHz, frame length = 24 ms) together with F0 contour (left), En contour calculated from the first cepstral coefficient c_0, the determined threshold $En_{\min} = 0.02$, and eliminated beginning and ending parts (right)

arithmetic mean values of the power spectrum and was also used for the storytelling voices transformation, and harmonics-to-noise ratio (HNR) providing an indication of the overall periodicity of the speech signal by the energy ratio between periodic and aperiodic components in the signal [21].

Variations in the pitch and energy contours can also be used for indication of supra-segmental changes in the speech signal [22]. For this reason, the F0 is used for determination of the supra-segmental parameters describing the microintonation component of speech melody. The differential contour $F0_{DIFF}$ can be obtained by subtraction of mean F0 values and linear trends (including the zero crossings $F0_{ZCR}$). Further parameters represent microvariations of F0 (jitter) and the variability of the peak-to-peak amplitude (shimmer).

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. The pitch period L recalculated from the F0 values 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.

3 EXPERIMENTS AND RESULTS

For evaluation of applied storytelling speech conversion the listening test called *Determination of storytelling speech type* had been processed in 2008 [7]. In this test, eighteen listeners (10 Slovaks and 8 Czechs, 13 men and 5 women) had chosen the storytelling speech type from "Teller", "Prince", "Witch", or "Cannot be recognized". The received results of this subjective evaluation approach in the form of a confusion matrix are presented in Fig. 6. The main goal of our experiment was to make a comparison with the results obtained by an objective approach based on the GMM classifier.

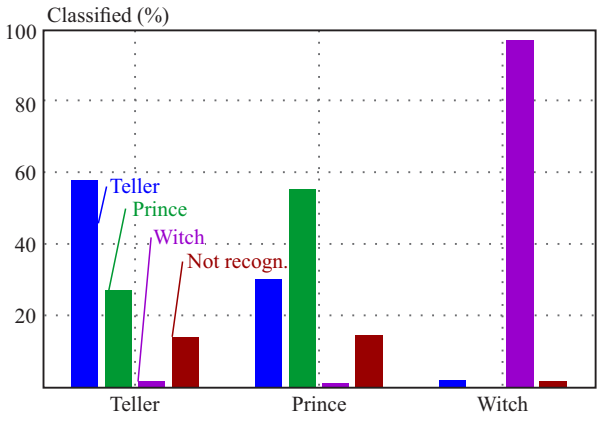

Fig. 6. 2D representation of the confusion matrix of perceptual result of the Determination of storytelling speech type listening tests performed in 2008 [7]

The original storytelling speech corpus used for GMM models creation, training, and testing consists of 65 sentences from the story *Witch's Garden* containing four storytelling voices called "Teller", "Prince", "Witch", and "Medulienka" (a girl who likes a honey very much) performed by a professional actor. Sentences with time duration from 2.2 to 15.5 seconds were resampled from 44.1 to 16 kHz with an anti-aliasing filter. This speech material was compared with the synthetic speech generated by the Czech and Slovak TTS system based on the diphone inventory with cepstral description realized as the speech engine for MS SAPI 5 standard [7, 16]. Synthesis parameters were set as follows: male voice, $f_s = 16$ kHz, $F0_{basic} = 110$ Hz, speech rate = 130 %. The testing corpus was created with the help of the above-mentioned TTS system. It consists of 200 sentences altogether in the Slovak language (with the average time duration of 3.54 seconds) — every evaluation set contains one basic sentence (for comparison with its sound before application of storytelling speech conversion) and four representative sentences for each of four storytelling speech types (it means sixteen sentences with converted storytelling speech). The performed experiment with GMM recognition of applied transformation of the storytelling voices, the analysis and comparison was aimed at investigation of

- influence of the initial parameter during the GMM creation on the resulting identification score: the number of applied mixtures of the Gaussian probability density functions $N_{gmix} = \{4, 8, 16, 32, 64, 128, 256\}$ — see the summarized mean values in the bar-graph and the detailed basic statistical parameters in the box-plot in Fig. 7,
- influence of the used number of training iterations on the GMM classification accuracy for $N_{iter} = \{10, 50, 100, 500, 1000, 1500\}$ is demonstrated in the graphical form in Fig. 8,
- influence of different types of speech parameters used in the three sets P0-P2 of the input feature vectors — see the results for all four types of transformed storytelling voices in the numerical form presented by the Table 4 accompanied with the bar graphs in Fig. 9,
- comparison of obtained results of the original storytelling voice recognition from sentences with the pure TTS synthesis (using the cepstral source-filter model) and sentences with applied storytelling voice transformation — see the 3D representation of the confusion matrices in Fig. 10,
- comparison of computational complexity: CPU times for the GMM creation and training phases as well as mean values of the GMM classification accuracy for different number of used mixtures; summarized for all four types of transformed storytelling voices presented by Table 5.

The length of the input feature vector $N_{feat} = 16$ was experimentally chosen in correspondence with the obtained results of previous research [23, 24]. These feature sets contain the features determined from the spectral

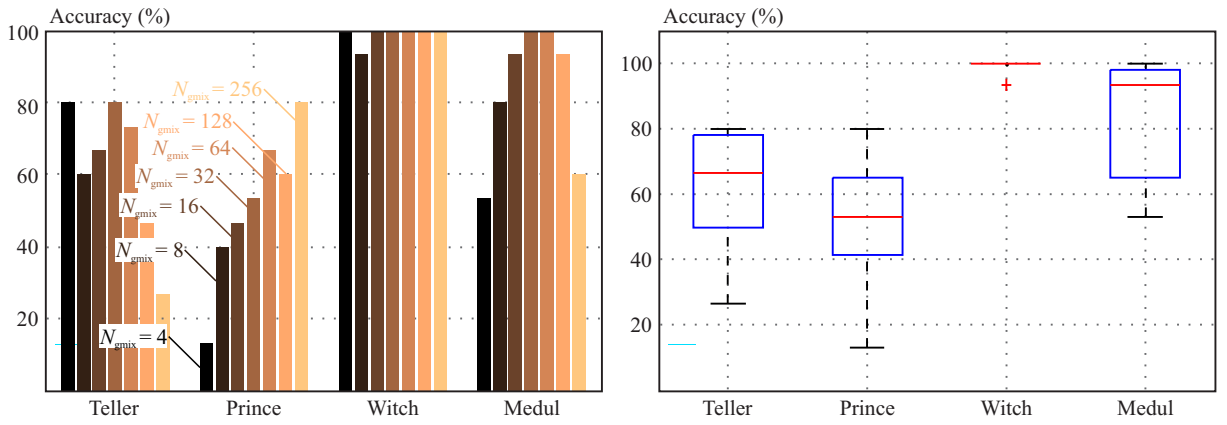

Fig. 7. Influence of the number of used mixtures on the GMM recognition accuracy; bar-graph of the mean values (left), box-plot of the basic statistical parameters (right)

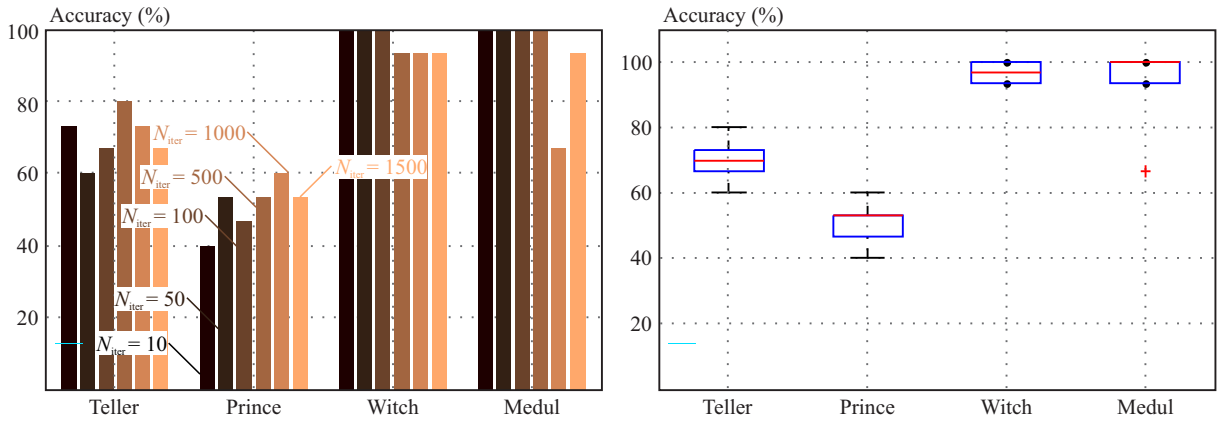

Fig. 8. Influence of Niter parameter on the GMM recognition accuracy: bar-graph of the mean values (left), box-plot of the basic statistical parameters (right)

envelopes, as well as the supplementary spectral parameters, and the prosodic parameters as described in Table 3. The table shows the structure of the three different feature sets P0-P2 that were used in detailed analysis of influence on the GMM recognition accuracy. They comprise the basic spectral features, the higher-order statistic parameters (skewness, kurtosis) of the first four cepstral coefficients, the first two formant positions and values of their ratios. In the case of the supplementary spectral features, the first and the second-order statistics — mean values and standard deviations (std) — were used as the representative values in the feature vectors for GMM classification. For implementation of the prosodic speech parameters the basic statistical types (median, range of values, std, and/or relative maximum and minimum) were used in the feature vectors.

For all presented comparison results holds that if not defined otherwise, the used parameter setting in tables or figure captions was: the feature set P0; $N_{gmix} = 32$, $N_{iter} = 500$. The obtained results are presented for visual comparison using the graphical form (the confusion matrices and/or the bar graphs of the identification accuracy in (%)) as well as numerical matching of the mean values stored in tables. The GMM identification accuracy was calculated from X_A sentences with correctly identified original storytelling voice and the total number N_U of the tested sentences as $(X_A/N_U)^*100\,\%$. The values in the confusion matrices were calculated in a similar way.

Table 3. Structure of the feature sets used for GMM identification

Feature set	Feature type	Statistical value	N_{feat}
P0	{HNR, spect. decrease, SC, SFM, SE, F0_{DIFF}, jitter, shimmer}	{min, rel. max, mean, median, std}	16
P1	{$F_1, F_2, F_1/F_2$, spect. decrease, HNR, SFM, SE, F0_{DIFF}, jitter, shimmer}	{rel.max, mean, median, std, skewness, kurtosis}	16
P2	{c1, - c4, spect. decrease, centroid, flatness, SE, F0_{DIFF}, jitter, shimmer}	{mean, median, std, skewness}	16

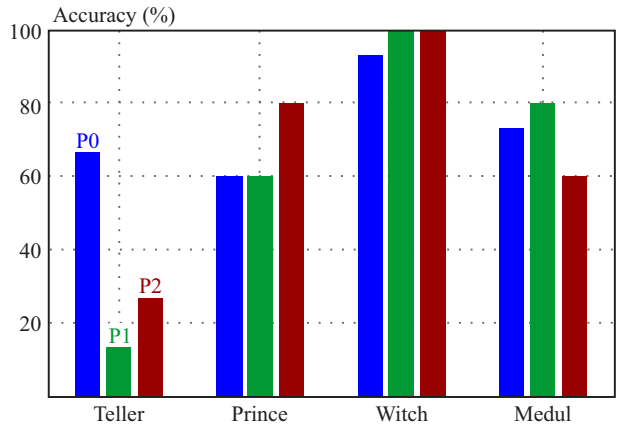

Fig. 9. Influence of the used feature set on the GMM recognition accuracy; results for all four types of transformed voices

Table 4. Basic statistical values of the GMM recognition accuracy in (%) documenting influence of the used type of the feature vectors P0-P2; results for all four types of transformed voices

Value/feature set	P0	P1	P2
Minimum	60.0	13.3	26.7
Maximum	93.3	100	100
Mean	73.3	63.3	66.7
Std	14.4	37.1	31.3

In the tasks of the overall evaluation for all four voices (results shown in graphs of confusion matrices in Fig. 10) the tested files are stored in a common directory and the same method of classification accuracy calculation was applied as described above.

As regards the developed GMM-based classifier, the simple diagonal covariance matrix of mixture models was applied in this identification experiment. The basic functions from the Ian T. Nabney "Netlab" pattern analysis toolbox [25] were used for creation of the GMM models, data training, and classification. For determination of the spectral features and the prosodic parameters of the synthesized as well as the original speech, the functions from Matlab ver. 2010b environment with the help of "Signal Processing Toolbox" and "Statistics Toolbox" were applied. The computational complexity for two algorithmic phases (the first one consisting of model creation and training, and the second one 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.

4 DISCUSSION AND CONCLUSION

The performed experiments have confirmed that the proposed GMM-based evaluation method is practically usable for recognition of different storytelling voices that were transformed from the sentences generated by the Czech and Slovak TTS system producing the synthetic speech using the diphone speech inventory, the source-filter speech model, and with the neutral (flat) prosody generation by the rules [5, 7, 16]. It is also in correlation with previous findings that the quality of synthetic speech produced by a TTS system can be evaluated by the GMM classifier and the identification accuracy depends on the used method of speech modelling and synthetic speech production [26]. In addition, unlike the subjective evaluation approaches based on the listening tests, the objective evaluation methods can work automatically without human interaction and the obtained results can be numerically judged.

The secondary aim of our experiments — to find the suboptimal setting of the initial GMM training parameters and the structure of the input feature set for the storytelling voice recognition - was also fulfilled successfully. It is well known that higher number of Gaussian mixtures can increase the recognition accuracy [27]. For this reason, we perform an analysis of influence of the used number of GMM mixtures the interval from 4 to 256 mixtures. The comparison of obtained results shows that a relatively maximum of the summarized mean recognition accuracy was observed for all four transformed voices together: 85 % for the best case of $N_{gmix} = 32$ in comparison with 62 % accuracy for the minimum number of 4 mixtures as shown in Fig. 7. Therefore, 32 Gaussian mixtures were subsequently applied in the identification process. On the other hand, a choice of the number of iterations has not great influence, so the optimum value $N_{iter} = 500$ was chosen for use in next experiments.

The right choice of the input feature set has a significant influence on the achieved GMM identification accuracy: the best results are obtained in the case of the set P0 which represents a mix of supra-segmental, basic, and supplementary spectral features. This analysis also shows that some types of speech features are not suitable for this identification task; it holds especially for the features based on formant frequencies and their bandwidths — see the worst results for the set P2 which was affected mainly in the case of the "Teller" voice. On the other hand, the type of the used representative value is not critical.

The main recognition experiment confirms that the obtained recognition accuracy corresponds to the degree of the voice transformation. Detailed analysis of obtained values per storytelling voices shows that the best results are achieved for the voices of "Witch" and "Medul" that had been converted with greater changes of spectral properties (modification of formant positions — practically similar to male → female voice conversion [5]) as well as the prosodic parameters (energy and duration modification) — see the bar-graph in Fig. 9 and the confusion matrix in Fig. 10. The results obtained in this way are in good correspondence with the values given by the classical listening test approach applied in the framework of our previous experiments [7] as is documented by comparison with the confusion matrix in Fig. 6. Although we expected even distribution in all four output classes when

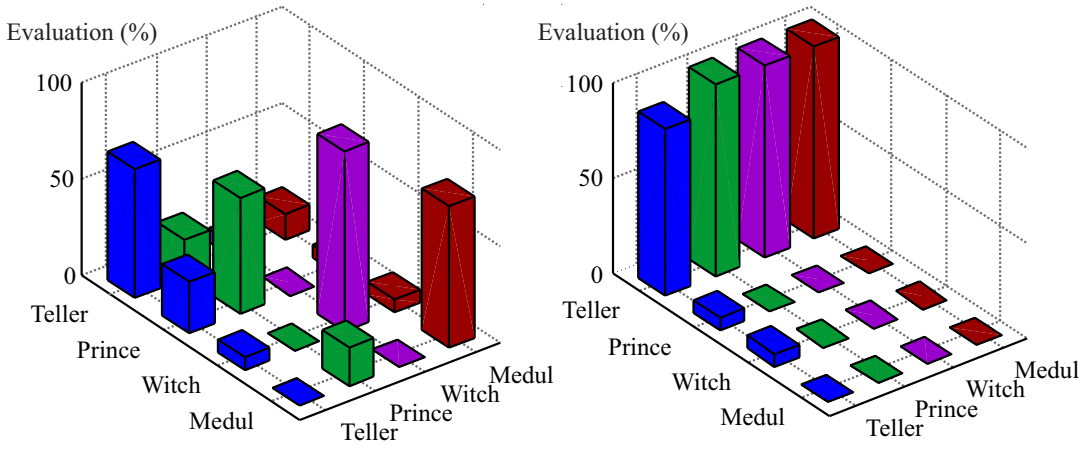

Fig. 10. 3-D representation of confusion matrices of the GMM recognition: sentences with applied storytelling voice transformation (left), from sentences with the pure TTS synthesis (right)

Table 5. Comparison of the computational complexity (CPU time in (s)) for different number of used mixtures; summarized for all transformed storytelling voices

Phase/N_{gmix}	4	8	16	32	64	128	256
Creation and training*	7.3	10.5	16.3	30.9	61.7	124.6	266.2
Identification**	0.60 (25.2)	0.61 (27.8)	0.64 (27.2)	0.70 (27.4)	0.83 (33.4)	0.98 (39.9)	1.33 (50.9)
Total time	7.80	10.51	16.84	31.60	62.53	125.58	267.53

*) Summary values for all transformed storytelling voices (4 models).
**) Mean values per sentence including the standard deviation values in [ms] (in parentheses).

using the pure output of the TTS system, the GMM identification of sentences without storytelling voice conversion gives the classification "Teller" for all voices — see the confusion matrix in Fig. 10. In principle this result is correct: in this case there were only minimal changes in spectral and prosodic parameters of the original TTS.

From the final analysis of the computation complexity follows that use of the maximum number of 256 mixtures increases the CPU time more than 8 times when compared with 32 mixtures (and approximately 36 times higher CPU time than for 4 mixtures) especially in the GMM creation and training phase as documented by the values in Table 5. Therefore, use of 32 mixtures seems to be the best choice: in comparison with the minimum number of 4 mixtures it causes increase in the mean CPU time only four times, which is acceptable. Moreover, the maximum value of $N_{gmix} = 256$ does not bring the best results of the recognition accuracy.

Increase of the GMM identification accuracy can be expected if the full covariance matrix or the probabilistic PCA (Principal Component Analysis) [28] is used for GMM model creation, training, and employment in the classification process. Further, we will try to use these methods for GMM identification although at the expense of higher computational complexity. Finally we plan to perform a larger comparison using the higher number of storytelling voices as well as testing the sentences from stories in other languages (English, German, Italian, *etc.*).

Acknowledgment

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

References

[1] LEE, H. J.: Fairy Tale Storytelling System: Using Both Prosody and Text for Emotional Speech Synthesis, In: Convergence and Hybrid Information Technology (Lee, G., Howard, D., Ślęzak, D., Hong, Y.S., eds.), Communications in Computer and Information Science, vol. 310, Springer, Berlin Heidelberg, 2012, pp. 317–324.

[2] ALCANTARA, J. A.—LU, L. P.—MAGNO, J. K.—SORIANO, Z.—ONG, E.—RESURRECCION, R.: Emotional Narration of Children's Stories,, In: Theory and Practice of Computation (Nishizaki, S.Y., Numao, M., Caro, J., Suarez, M.T., eds.), Proceedings in Information and Communication Technology, vol. 5, Springer, Japan, 2012, pp. 1–14.

[3] DOUKHAN, D.—ROSSET, S.—RILLIARD, A.—D'ALESSANDRO, C.—ADDA-DECKER, M.: Text and Speech Corpora for Text-to-Speech Synthesis of Tales,, In: Proceedings of the 8-th International Conference on Language Resources and Evaluation, Istanbul, Turkey, 2012, pp. 1003–1010.

[4] MAENO, Y.—NOSE, T.—KOBAYASHI, T.—KORIYAMA, T.—IJIMA, Y.—NAKAJIMA, H.—MIZUNO, H.—YOSHIOKA, O.: Prosodic Variation Enhancement Using Unsupervised Context Labeling for HMM-based Expressive Speech Synthesis, Speech Communication **57** (2014), 144–154.

[5] PŘIBIL, J.—PŘIBILOVÁ, A.: Czech TTS Engine for Braille Pen Device Based on Pocket PC Platform, Proc. of the 16th

Conference Electronic Speech Signal Processing ESSP 05 joined with the 15th Czech-German Workshop Speech Processing (Vch, R., ed.), 2005, pp. 402–408.

[6] PŘIBILOVÁ, A.—PŘIBIL, J.: Spectrum Modification for Emotional Speech Synthesis, In: Multimodal Signals: Cognitive and Algorithmic Issues (Esposito, A., Hussain, A., Marinaro, M., Martone, R., eds.), LNAI 5398, Springer-Verlag Berlin Heidelberg, 2009, pp. 232–241.

[7] PŘIBIL, J.—PŘIBILOVÁ, A.: Application of Expressive Speech in TTS System with Cepstral Description., In: Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction (Esposito, A., Bourbakis, N., Avouris, N., Hatzilygeroudis, I., eds.), LNAI 5042, Springer-Verlag, Berlin Heidelberg, 2008, pp. 201–213.

[8] BLAUERT, J.—JEKOSCH, U.: A Layer Model of Sound Quality, Journal of the Audio Engineering Society **60** (2012), 4–12.

[9] LEGÁT, M.—MATOUŠEK, J.: Design of the Test Stimuli for the Evaluation of Concatenation Cost Functions, In: Text, Speech and Dialogue 2009 (MATOUšEK, V. *et al*, eds.), LNCS 5729, Springer, Heidelberg, 2009, pp. 339–346.

[10] BELLO, C.—RIBAS, D.—CALVO, J. R.—FERRER, C. A.: From Speech Quality Measures to Speaker Recognition Performance., In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (Bayro-Corrochano, E., Hancock, E., eds.), LNCS 8827, Springer International Publishing Switzerland, 2014, pp. 199–206.

[11] ROMPORT, J.—MATOUŠEK, J.: Formal Prosodic Structures and Their Application in NLP, In: Text, Speech and Dialogue 2005 (Matouek, V. *et al*., eds.), LNCS 3658, Springer-Verlag, Berlin Heidelberg, 2005, pp. 371–378.

[12] JEONG, Y.: Joint Speaker and Environment Adaptation Using TensorVoice for Robust Speech Recognition, Speech Communication **58** (2014), 1–10.

[13] 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.

[14] MUHAMMAD, G.—ALGHATHBAR, K.: Environment Recognition for Digital Audio Forensics Using MPEG-7 and Mel Cepstral Features, Journal of Electrical Engineering **62** No. 4 (2011), 199–205.

[15] PISHRAVIAN, A.—SAHAF, M. R. A.: Application of Independent Component Analysis for Speech-Music Separation Using An Efficient Score Function Estimation, Journal of Electrical Engineering **63** No. 6 (2012), 380–385.

[16] PŘIBIL, J.—PŘIBILOVÁ, A.: Emotional Style Conversion in the TTS System with Cepstral Description, In: Verbal and Nonverbal Communication Behaviours (Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M., eds.), LNAI 4775, Springer-Verlag, Berlin Heidelberg New York, 2007, pp. 65–73.

[17] VCH, R.—PŘIBIL, J.—SMÉKAL, Z.: New Cepstral Zero-Pole Vocal Tract Models for TTS Synthesis, Proc. of IEEE Region 8 EUROCON'2001, vol. 2, 2001, pp. 458–462.

[18] MADHU, N.: Note on Measures for Spectral Flatness, Electronics Letters **45** No. 23 (2009), 1195–1196.

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

[20] 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.

[21] SOUSA. R.—FERREIRA, A.—ALKU, P.: The Harmonic and Noise Information of the Glottal Pulses, Speech, Biomedical Signal Processing and Control **10** (2014), 137–143.

[22] LECLERC, I.—DAJANI, H. R.—GIGUERE, C.: Differences in Shimmer Across Formant Regions, Journal of Voice **27** No. 6 (2013), 685–690.

[23] PŘIBIL, J.—PŘIBILOVÁ, A.—ĎURAČKOVÁ, D.: Evaluation of Spectral and Prosodic Features of Speech Affected by Orthodontic Appliances using the GMM Classifier, Journal of Electrical Engineering **65** (2014), 30–36.

[24] PŘIBIL, J.—PŘIBILOVÁ, A.: Determination of Formant Features in Czech and Slovak for GMM Emotional Speech Classifier, Radioengineering **22** (2013), 52–59.

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

[26] PŘIBIL, J.—PŘIBILOVÁ, A.—MATOUŠEK, J.: Experiment with Evaluation of Quality of the Synthetic Speech by the GMM Classifier, In: Text, Speech and Dialogue, Proc. of the 16th International Conference TSD 2013, Plzen, Czech Republic September 2013 (Habernal, I., Matoušek, V., eds.), LNAI 8082, Springer-Verlag, Berlin Heidelberg, 2013, pp. 241–248.

[27] DILEEP, A. D.—SEKHAR, C. CH.: Class-Specific GMM Based Intermediate Matching Kernel for Classification of Varying Length Patterns of Long Duration Speech Using Support Vector Machines, Speech Communication **57** (2014), 126–143.

[28] ZHAO, J.—JIANG, Q.: Probabilistic PCA for t-Distributions, Neurocomputing **69** No. 16-18 (2006), 2217–2226.

Received December 2014

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 senior scientist 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 (Assoc. 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 1985 and 2002, respectively. Since 1992 she has been working as a university teacher at the Radioelectronics Department, and in 2014 she has become an associate professor 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 analogue and digital ASICs and neural network implementation on chip.