Experiment with Evaluation of Quality of the Synthetic Speech by the GMM Classifier*

Jiří Přibil[1,2], Anna Přibilová[3], and Jindřich Matoušek[1]

[1] University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
jmatouse@kky.zcu.cz

[2] SAS, Institute of Measurement Science, Dúbravská cesta 9, SK-841 04 Bratislava, Slovakia
Jiri.Pribil@savba.sk

[3] Slovak University of Technology, Faculty of Electrical Engineering & Information Technology, Institute of Electronics and Photonics, Ilkovičova 3, SK-812 19 Bratislava, Slovakia
Anna.Pribilova@stuba.sk

Abstract. This paper describes our experiment with using the Gaussian mixture models (GMM) for evaluation of the speech quality produced by different methods of speech synthesis and parameterization. In addition, the paper analyzes and compares influence of different types of features and different number of mixtures used for GMM evaluation. Finally, the GMM evaluation scores are compared with the results obtained by the conventional listening tests based on the mean opinion score (MOS) evaluations. Results of evaluations obtained by these two ways are in correspondence.

Keywords: GMM classifier, spectral and prosodic features of speech, synthetic speech evaluation.

1 Introduction

At present, the requirements on the quality of produced synthetic speech are rapidly increasing because the proper, quick, and easy understanding is a basic condition for effectivity and suitable strategy of dialogue management in the voice communication systems with the human-machine interface. For that reason the evaluation of the synthetic speech quality –first of all, intelligibility and naturalness – must often be performed. Several subjective and objective methods are used to verify the quality of produced synthetic speech. As regards subjective approaches [1], [2] the listening tests are usually used for giving the feedback information about user's opinion. On the other hand, especially in the case of intelligibility, the objective method based on automatic speech recognition (ASR) system can be used, where the final result represents the recognition score [3]. These recognition systems are often based on neural networks, hidden Markov models [4], or Gaussian mixture models (GMM) [5] in the case of speech emotion classification. The main advantage of this statistical evaluation method is that it works automatically without human interaction and the obtained results can be numerically judged.

* The work has been supported by the Technology Agency of the Czech Republic, project No. TA01030476, 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).

I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, pp. 241–248, 2013.

This paper describes our experiment with using the GMM for evaluation of the speech quality produced by different methods of speech synthesis and parametric description that are often used for the speech production by the text-to-speech (TTS) systems. The GMM were trained with a corpus consisting of the original male and female Czech speech. The original sentences were resynthesized with different setting of segmentation method, type of speech modelling, as well as with different number of parameters of the vocal tract model. For the GMM evaluation, the basic and complementary spectral properties [6] including the supra-segmental parameters [7], were determined from the synthesized sentences and used in the input feature vector. The paper next analyzes and compares influence of different types of spectral features and supra-segmental parameters used for GMM evaluation as well as the influence of the number of used GMM mixture components.

Motivation of our work was to find an alternative approach to the standard listening tests, especially when audible differences were too small or hardly recognizable by listeners, in problems with their collective realization, etc.

2 Subject and Method

The Gaussian mixture models can be defined as a linear combination of multiple Gaussian probability distribution functions of the input data vector. For GMM creation, it is necessary to determine the covariance matrix, the vector of mean values, and the weighting parameters from the input training data. Using the expectation-maximization (EM) iteration algorithm, the maximum likelihood function of GMM is found [8]. For control of the EM algorithm, the N_{gmix} represents the number of used mixtures in each of the GMM models. In standard use of the GMM classifier, the resulting score is given by the maximum overall probability for the given class using the $score(T, i)$ representing the probability value of the GMM classifier for the models trained for current i-th class in the evaluation process, and an input vector T of the features obtained from the tested sentence [5]. For our purpose, only one model is created and trained in dependence on the speaker voice (male/female) with the help of the input feature vectors from the original sentences. In the classification phase, we obtain the scores using the input feature vectors from the tested sentences synthesized by various methods. These scores are sorted by the absolute size and quantized to N levels corresponding to N output classes. It means that the obtained highest score represents the synthesized sentences having the speech features that are most similar to those obtained from the original sentences used for GMM model training; the minimum score corresponds to the tested sentence with the greatest differences in comparison to the originals. To obtain correspondence (comparable values) with the mean opinion score (MOS) evaluation method where the perceived quality is scaled from "5" representing the best quality to "1" corresponding to poor quality; finally we use five output classes: in the score discriminator block (see Fig. 1) the highest obtained score is assigned to the value 5, the lowest score to the value 1. To obtain speaker independent GMM classification, the data k-fold cross-validation method [8] can be applied during the training and the

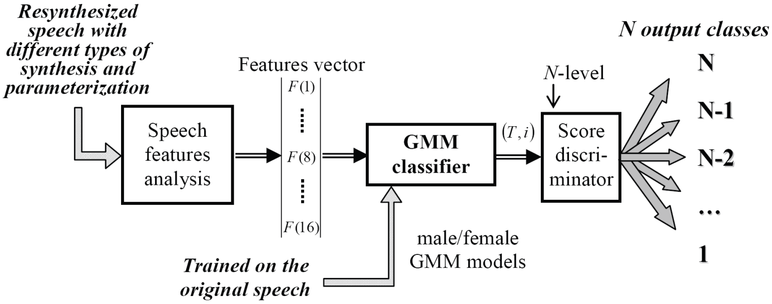

Fig. 1. Block diagram of the GMM classifier used for speech quality evaluation

testing processes. In our case this approach is not currently applied, so the classification works speaker-dependently. In the first stage of development – the GMM speech quality classifier has only one-level structure. This simple architecture expects that the gender of the voice (male/female) was correctly recognized in the previous process – manually, by listening tests, etc. Subsequently, the quality of the speech is identified for each of two gender classes.

2.1 Spectral Features and Prosodic Properties Determination

Spectral analysis of the speech signal is performed in the following way: from the input samples, after segmentation and weighting by a Hamming window, the absolute values of the fast Fourier transform, the power spectrum, and the smoothed spectral envelope (using the mean Welch's periodogram) are calculated for further use in the feature determination process. The spread of the signal around the higher harmonics has consequences on spectral tilt measurements. On the other hand, the harmonics-to-noise ratio (HNR) provides an indication of the overall periodicity of the speech signal. Specifically, it quantifies the ratio between the periodic and aperiodic components in the signal. The spectral centroid (SC) is a centre of gravity of the power spectrum and 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 can be calculated as a ratio of the geometric and the arithmetic mean values of the power spectrum. The spectral entropy (SE) is a measure of spectral distribution. It quantifies a degree of randomness of spectral probability density represented by normalized frequency components of the spectrum.

Microintonation component of speech melody given by F0 contour can be supposed to be a random, band-pass signal described by its spectrum and statistical parameters. We use differential microintonation signal $F0_{\mathrm{DIFF}}$ obtained by subtraction of mean F0 values and linear trends. The jitter related to microvariations of a pitch curve is computed as a relative number of zero crossings (ZCR) of a derivative pitch curve normalized by utterance duration. For calculation of the absolute jitter values, the average absolute difference between consecutive pitch periods measured in samples was used. In the case of the shimmer measure determination, a period-to-period variability of amplitudes of the speech signal was used.

3 Material, Experiments, and Results

In our experiment we use the speech material which consists of sentences with duration from 1.5 to 10.5 seconds, resampled at 16 kHz, representing speech in a neutral (declarative) state, and yes/no questions uttered by one male and one female speaker. The original sentences were analyzed and subsequently resynthesized using five different types of speech modelling (cepstral [9], harmonic [10]), segmentation, and with different number of applied coefficients N_{coeff} – see detailed specifications in Table 1. These synthesis methods are currently used in the Czech TTS system based on the HMM approach for synthetic speech personification or expressive speech production [11]. The whole test speech corpus includes $180 + 180$ sentences consisting of the originals and five types of resynthesis. For the listening test evaluation only $80 + 80$ sentences (male + female) were selected from the main test corpus. The processed speech material originates from speakers with different mean F0 values so different parameter settings for analysis – frame (window) length W_L and window overlapping W_O – were applied. The F0 values (pitch contours) were determined by autocorrelation analysis method with experimentally chosen pitch ranges as follows: $55 - 250$ Hz for male voices, and $105 - 350$ Hz for female ones.

Table 1. Detailed specification of used speech analysis and synthesis method

Type	Model	Specification	N_{coeff}
1 (h48)	harmonic	Spectral envelope smoothed by B-Splines	$48^{*)}$
2 (i80)	cepstral	Impulse response of the real cepstrum	$80^{*)}$
3 (k64)	cepstral	Minimum phase of real cepstrum, mixed excitation	$64^{*)}$
4 (s64)	cepstral	Real cepstrum, excitation by Hilbert impulse	$64^{**)}$
5 (o50)	cepstral	Optimized structure of cascade approximation filter	$50^{*)}$

$^{*)}$ W_L=12/10 ms, W_O=W_L/2 for male/female voices. $^{**)}$ W_L=24/20 ms, W_O=W_L/4.

Two basic types of experiments were performed for comparison in this paper:

1. Objective – automatic evaluation of the synthetic speech quality using the statistical method based on the GMM approach,
2. Subjective – manually performed listening test using the MOS evaluation method.

In GMM evaluation the analysis and comparison was aimed at investigation of:

- influence of the number of used mixtures on GMM evaluation (in the range of $N_{gmix} \in < 1 - 3 >$),
- influence of the used type of the feature set on GMM evaluation score (sets P1-3),
- influence of the feature order in the input data vector on GMM score; the set P1 was used with the reversed order of features giving thus the set called P4.

The feature set of 16 values as the input data vector for GMM training and classification containing the features determined from the spectral envelopes (skewness, kurtosis, spread, and tilt), the complementary spectral parameters (HNR, SC, SFM, SE), and supra-segmental parameters (F0, jitter, and shimmer) was used, as shown Table 2. In the case of the spectral features, the basic statistical parameters-mean values and standard deviations (std)-were used as the representative values in the feature vectors for GMM evaluation. For implementation of the supra-segmental parameters of speech, the statistical types – median values, range of values, std, and/or relative maximum and minimum we used in the feature vectors. A simple diagonal covariance matrix of the GMM was applied in this first evaluation experiment. The basic functions from the Ian T. Nabney "Netlab" pattern analysis toolbox [12] were used for creation of the GMM models, data training, and classification.

Table 2. Detailed specification of used speech analysis and synthesis method

No	Feature name			Statistical value		
	P1	P2	P3	P1	P2	P3
1	HNR	Spec. envelope	F0	Mean	Skewness	Median
2	HNR	Spec. envelope	F0	Std	Kurtosis	Std
3	Spec. tilt	Spec. centroid	$F0_{DIFF}$	Min	Mean	Median
4	SC	Spec. spread	$F0_{DIFF}$	Mean	Std	Std
5	SC	Spec. tilt	$F0_{DIFF}$	Std	Min	Rel. max
6	SFM	SFM	$F0_{DIFF}$	Mean	Mean	Rel. min
7	SFM	F0	$F0_{ZCR}$	Std	Std	Median
8	SE	F0	$F0_{ZCR}$	Mean	Rel. max	Std
9	SE	$F0_{DIFF}$	$F0_{ZCR}$	Std	Median	Rel. max
10	Signal ZCR	$F0_{DIFF}$	$F0_{ZCR}$	Median	Std	Rel. min
11	Signal ZCR	$F0_{ZCR}$	Jitter	Std	Median	Median
12	$F0_{DIFF}$	$F0_{ZCR}$	Jitter	Rel. max	Rel. max	Std
13	Jitter	Jitter	Jitter	Median	Median	Rel. max
14	Shimmer	Jitter	Shimmer	Max	Rel. max	Median
15	Shimmer	Shimmer	Shimmer	Median	Median	Std
16	Shimmer	Shimmer	Shimmer	Rel. max	Rel. max	Rel. max

Subjective evaluation was realized by the conventional listening test called "Speech quality evaluation – male/female voice" located on the web page `http://www/lef.um.savba.sk/scripts/itstposl2.dll`. This listening test program in the form of MS ISAPI DLL script including the testing speech runs on the server PC and communicates with the user via the HTTP protocol by WEB pages with frames in the HTML language. The currently used type of the listening test is based on the MOS evaluation for naturalness and intelligibility of the synthetic speech example. By reason of differentiation for creation and training of the GMM models, the evaluation must be performed separately for male and female voices. The complete test consists of 10 evaluation sets using sentences selected randomly from the speech corpus. In addition, the listening test program generates also the test protocol with time marks, so we can determine the duration of the performed test.

3.1 Obtained Results

Obtained results of performed GMM evaluation experiment are presented in graphical form for visual comparison by the boxplot of basic statistical parameters (see Fig. 2), or by the bar graphs (see Figures 3 and 4) separately in dependence on the speaker's voice. Twenty two listeners (5 women and 17 men) took part in our listening test experiment: 20 listening tests of male voice and 20 tests of female voice were executed, 40 tests in total. Evaluation of the listening test results was realized in dependence on listener's sex for all synthesis method categories, see Fig. 5. The final comparison of evaluation results based on GMM approach and MOS listening tests separately for male/female voice are shown in Table 3.

Table 3. Final comparison of evaluation results based on GMM approach and MOS listening test separately for the male/female voice and the applied type of the synthesis method

Evaluation method	Male voice					Female voice				
	h48	i80	k64	s64	o50	h48	i80	k64	s64	o50
GMM-full corpus$^{*)}$	2.45	4.75	3.05	2.95	1.50	2.75	4.05	3.75	3.20	1.25
GMM-limit. corpus$^{*)}$	2.06	4.87	3.01	3.13	1.95	2.86	4.06	3.32	3.12	1.44
MOS listening test	2.75	3.65	2.70	3.01	2.34	2.27	3.55	3.06	2.31	2.19

$^{*)}$ N_{gmix}=1, feature set P1

4 Discussion and Conclusion

As follows from our first comparison experiment, increasing the number of the used mixtures brings not always positive effect to the GMM evaluation – it holds especially in the case of female speech, where differentiation of the discriminated GMM score is worse for two or three mixtures than for only one mixture – see Fig. 3. Therefore, in further analysis we use setting $N_{gmix} = 1$; but in this case there are actually compared averages of features in the input vector from natural and synthesized speech using the Mahalanobis distance measure. Application of the proper type of the input features for GMM evaluation is very important – as demonstrated by the results of our second experiment (see Fig. 4): the best results are produced by the feature set P1 consisting of a mix of spectral and prosodic features, the worst results correspond to the set P3 when only supra-segmental features were used. These results consider the fact, that the resynthesized sentences have the similar distribution of F0 values as the original ones. The differences can be caused only by the used type of pitch-period detection and the segmentation method (processing) for signal analysis as it is documented by Table 1.

On the other hand, the order of the parameters in the input feature vector has minimal influence on the GMM score – values for the sets P1 and P4 are practically the same. Comparison of the GMM scores obtained with the help of the full speech corpus and using only the sentences applied in the listening test shows that they are similar. The summary results of the MOS test are also in correspondence – compare values in Table 3 and Fig. 5. Due to disproportion in the groups of listener's gender (the female group is smaller than the male one) the score differentiation in the case of the female listeners is not as objective as in the male listeners.

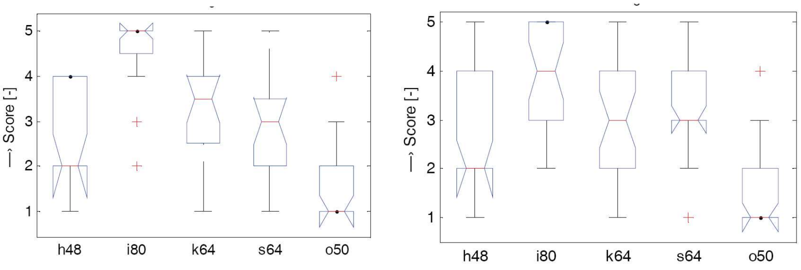

Fig. 2. The boxplot of the basic statistical parameters of the discriminated score values: for male (left) and female (right) voices; $N_{gmix} = 1$, feature set P1

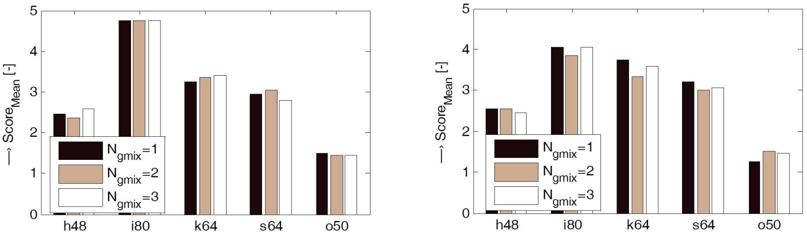

Fig. 3. Influence of the number of used mixtures on the discriminated GMM score: mean values for male (left) and female (right) voices; feature set P1

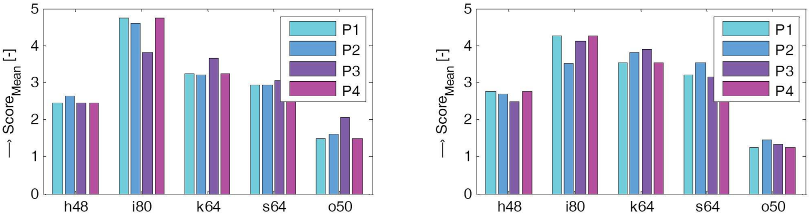

Fig. 4. Influence of the type of the feature vector on the discriminated GMM score: mean values for male (left) and female (right) voices; $N_{gmix} = 1$

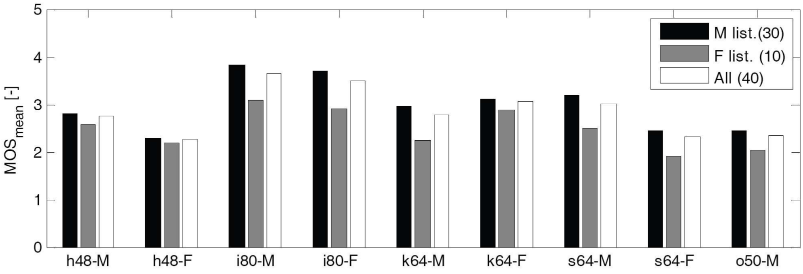

Fig. 5. Summary results of the listening test MOS evaluation in dependence on listener's categories joined for male and female voices

The observed quality of the synthetic speech with the female voice was lower than that with the male voice, so the score values were also lower and worse differentiated. At the same time, recognizable differences were observed in the resynthesis quality of the declarative sentences and yes/no questions, which reached the lowest MOS score. From the listener's feedback information follows that the listening test constructed by this way is very difficult: several listeners executed only one type of the test; some of the addressed persons tried the test but they could not recognize differences so the test was not finished. Therefore, this is the right case when the GMM based evaluation can be applied more effectively and it can bring more accurate results than the conventional listening tests.

In near future, we plan testing of the GMM evaluation using the larger speech databases as well as the ones spoken in other languages, and also make a comparison with other widely applied quality assessment schemes used for telephony speech signals and speech codecs evaluation. Increase of the GMM score can be expected if the full covariance matrix is used, so both approaches will be compared in future. We would also like to try to use the GMM classification for quality evaluation of the voice conversion (male/female/child) or the expressive style transformation of the synthesized speech.

References

1. Audibert, N., Vincent, D., Aubergé, V., Rosec, O.: Evaluation of Expresive Speech Resynthesis. In: Proceedings of LREC 2006 Workshop on Emotional Corpora, Gènes, pp. 37–40 (2006)
2. Iriondo, I., Planet, S., Socoró, J.C., Martínez, E., Alías, F., Monzo, C.: Automatic Refinement of an Expressive Speech Corpus Assembling Subjective Perception and Automatic Classification. Speech Communication 51, 744–758 (2009)
3. Takano, Y., Kondo, K.: Estimation of Speech Intelligibility Using Speech Recognition Systems. IEICE Transactions on Information and Systems E93D(12), 3368–3376 (2010)
4. Vích, R., Nouza, J., Vondra, M.: Automatic Speech Recognition Used for Intelligibility Assessment of Text-to-Speech Systems. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction. LNCS (LNAI), vol. 5042, pp. 136–148. Springer, Heidelberg (2008)
5. Yun, S., Yoo, C.D.: Loss-Scaled Large-Margin Gaussian Mixture Models for Speech Emotion Classification. IEEE Transactions on Audio, Speech, and Language Processing 20(2), 585–598 (2012)
6. 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, 10 pages (2008)
7. Lu, Y., Cooke, M.: The Contribution of Changes in F0 and Spectral Tilt to Increased Intelligibility of Speech Produced in Noise. Speech Communication 51(12), 1253–1262 (2009)
8. Reynolds, D.A., Rose, R.C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Transactions on Speech and Audio Processing 3, 72–83 (1995)
9. 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, pp. 77–82 (2000)
10. Madlová, A.: Autoregressive and Cepstral Parametrization in Harmonic Speech Modelling. Journal of Electrical Engineering 53, 46–49 (2002)
11. Grůber, M., Hanzlíček, Z.: Czech Expressive Speech Synthesis in Limited Domain Comparison of Unit Selection and HMM-Based Approaches. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS, vol. 7499, pp. 656–664. Springer, Heidelberg (2012)
12. Bishop, C.M., Nabney, I.T.: NETLAB Online Reference Documentation (accessed February 16, 2012), http://www.fizyka.umk.pl/netlab/