Comparison of Several Speaker Verification Procedures Based on GMM

Aleš Padrta, Vlasta Radová

Department of Cybernetics
University of West Bohemia in Pilsen, Czech Republic
apadrta@kky.zcu.cz, radova@kky.zcu.cz

Abstract

In this paper, three speaker verification procedures are tested. All the procedures are based on Gaussian mixture models (GMM), however, they differ in the way, in which they use particular feature vectors of an utterance for speaker verification. A lot of experiments have been performed in a group of 329 speakers. The results showed that there is a procedure that enables to achieve better results than the commonly used procedure based on the log likelihood of the whole utterance – the procedure based on the majority voting rule for single feature vectors.

1. Introduction

In the paper [1], we studied the influence of the amount of test speech data upon the speaker identification performance. The goal of that paper was to find the minimum amount of the test data necessary for a decision about the identity of an unknown speaker. Three identification procedures based on the hidden Markov models of phonemes were described in that paper. It was shown in the experiments, that quite a good performance can be reached with a relatively small amount of data when the procedure based on the majority voting rule for sequences of about 7 phonemes is used.

The presented paper follows up the paper [1]. However, our interest is focused on the speaker verification task based on Gaussian mixture models (GMM) now. The principle of the GMM is briefly introduced in Section 2. Next, in Section 3, the three identification procedures presented in [1] are modified in order they can be used for speaker verification based on GMM. Experiments are described in Section 4 and their results are discussed in Section 5. Finally, a conclusion is given in Section 6.

2. Gaussian mixture models

Gaussian mixture models are a type of density model which consists of a number of Gaussian component functions. These component functions are combined to provide multimodal density [2].

The Gaussian mixture density of a feature vector $\boldsymbol{o}$ given parameters λ is a weighted sum of M component densities, and is given by the equation

$$p(\boldsymbol{o}|\lambda) = \sum_{i=1}^{M} c_i p_i(\boldsymbol{o}), \tag{1}$$

where $\boldsymbol{o}$ is an N-dimensional random vector, $p_i(\boldsymbol{o})$, $i = 1, \ldots, M$, are the component densities, and c_i, $i = 1, \ldots, M$, are the mixture weights. Each component density is an N-variate Gaussian function of the form

$$p_i(\boldsymbol{o}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{(\boldsymbol{o}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{o}-\boldsymbol{\mu}_i)\right\} \tag{2}$$

with the mean vector $\boldsymbol{\mu}_i$ and the covariance matrix $\boldsymbol{\Sigma}_i$. The mixture weights satisfy the constraint

$$\sum_{i=1}^{M} c_i = 1. \tag{3}$$

The complete Gaussian mixture density model is parameterized by the mean vectors, the covariance matrices and the mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{c_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}, \quad i = 1, \ldots, M. \tag{4}$$

3. Speaker verification procedures

The goal of speaker verification systems is to determine whether a given utterance is produced by the claimed speaker or not. This is performed by comparing a score, which reflects the agreement between the given utterance and the model of the claimed speaker, with an a priori given threshold. In verification systems based on GMM the simplest score is the likelihood of the utterance given the model of the claimed speaker.

Assume that there is a group of J reference speakers and that each speaker is represented by a Gaussian mixture model. We denote the model of the j-th speaker as ρ_j, $j = 1, \ldots, J$. Further suppose that a test utterance O consists of I feature vectors $\boldsymbol{o}_i$, $i = 1, \ldots, I$. The score reflecting the agreement between the feature vector $\boldsymbol{o}_i$ and the speaker j is then represented by the likelihood $p(\boldsymbol{o}_i|\rho_j)$ and is computed according to the formula (1).

However, such a score is very sensitive to variations in text, speaking behavior, and recording conditions, especially in the utterances of impostors. The sensitivity causes wide variations in scores, and makes the task of the threshold determination a very difficult one. In order to overcome this sensitivity, the use of the normalized score based on a background model has been proposed [2]. The score is then determined as the normalized log likelihood $\overline{p}(\boldsymbol{o}_i|\rho_j, \Lambda)$,

$$\overline{p}(\boldsymbol{o}_i|\rho_j, \Lambda) = \log(p(\boldsymbol{o}_i|\rho_j)) - \log(p(\boldsymbol{o}_i|\Lambda)), \tag{5}$$

where $p(\boldsymbol{o}_i|\Lambda)$ is the likelihood of the background model computed again using the formula (1). In this paper, we use two background models – one for male speech (Λ_1), and the other for female speech (Λ_2). For the normalization only the background model which is in better conformity with the speech data is used, i.e.

$$p(\boldsymbol{o}_i|\Lambda) = \max(p(\boldsymbol{o}_i|\Lambda_1), p(\boldsymbol{o}_i|\Lambda_2)). \tag{6}$$

Having the log likelihood $\overline{p}(\boldsymbol{o}_i|\rho_j,\Lambda)$ we can now use several procedures for the verification of the whole test utterance.

3.1. Verification based on the log likelihood of the whole utterance

The verification procedure widely used in speaker verification systems [4] is based on the sum of the particular log likelihoods, i.e the log likelihood of the whole utterance $\overline{p}(O|\rho_j,\Lambda)$ is computed according to the equation

$$\overline{p}(O|\rho_j,\Lambda)=\frac{1}{I}\sum_{i=1}^{I}\overline{p}(\boldsymbol{o}_i|\rho_j,\Lambda). \tag{7}$$

In order to make the sum independent of the length of the utterance, it is divided by the number of feature vectors I in the utterance.

The decision whether the speaker j can be accepted as the speaker of the utterance O or not is then determined by a comparison of the $\overline{p}(O|\rho_j,\Lambda)$ with the threshold T, i.e. according to the formula

$$D=\begin{cases} 1 & \text{if} \quad \overline{p}(O|\rho_j,\Lambda)\geq T \\ -1 & \text{if} \quad \overline{p}(O|\rho_j,\Lambda)<T \end{cases}, \tag{8}$$

where $D=1$ means "accept the speaker j" and $D=-1$ means "reject the speaker j".

3.2. Verification based on the majority voting rule for single feature vectors

Since the majority voting rule proved to be a good tool for performance enhancement in our previous speaker identification experiments [5], we tried to use it also in the verification task. Therefore we designed a verification procedure that is based on the majority voting rule.

In that case, first the log likelihood of each feature vector $\overline{p}(\boldsymbol{o}_i|\rho_j,\Lambda)$ is compared with the threshold to receive the partial decisions D_i, $i=1,\ldots,I$. The partial decision D_i is computed according to the formula

$$D_i=\begin{cases} 1 & \text{if} \quad \overline{p}(\boldsymbol{o}_i|\rho_j,\Lambda)\geq T \\ -1 & \text{if} \quad \overline{p}(\boldsymbol{o}_i|\rho_j,\Lambda)<T \end{cases}. \tag{9}$$

Then the partial decisions D_i, $i=1,\ldots,I$ are used to form a final decision D. The final decision is computed according to the formula

$$D=\sum_{i=1}^{I}D_i. \tag{10}$$

The negative value of the decision D represents the rejection of the speaker j, and the positive value imply that the speaker pronounced the test utterance. The zero value signalizes that no decision can be made.

3.3. Verification based on the majority voting rule for sequences of feature vectors

The speaker verification procedures described in Sections 3.1 and 3.2 can be regarded as boundary cases of a general verification procedure: in one case we use all feature vectors together in order to obtain a verification decision, in the other case we use each feature vector separately. So let us try now to design a general procedure that will be able to use also parts larger than one feature vector but shorter than the whole utterance for speaker verification.

Denote the sequence of K successive feature vectors of the test utterance which starts with the feature vector $\boldsymbol{o}_l$ and ends with the feature vector $\boldsymbol{o}_{l+K-1}$ as S_l. It means $S_l=[\boldsymbol{o}_l,\boldsymbol{o}_{l+1},\ldots,\boldsymbol{o}_{l+K-1}], l=1,\ldots,I-K+1$, where I is the number of feature vectors in the test utterance. The log likelihood that the sequence S_l was spoken by the speaker j is

$$\overline{p}(S_l|\rho_j,\Lambda)=\frac{1}{K}\sum_{k=0}^{K-1}\overline{p}(\boldsymbol{o}_{l+k}|\rho_j,\Lambda). \tag{11}$$

Analogically to the equation (9), the partial decision $D(S_l)$ for the sequence S_l is computed according to the formula

$$D(S_l)=\begin{cases} 1 & \text{if} \quad \overline{p}(S_l|\rho_j,\Lambda)\geq T \\ -1 & \text{if} \quad \overline{p}(S_l|\rho_j,\Lambda)<T \end{cases}. \tag{12}$$

Now we form the final decision from the partial decisions $D(S_l)$ according to the formula

$$D=\sum_{l=1}^{I-K+1}D(S_l). \tag{13}$$

Similarly as in (10), the negative value of the final decision D corresponds to the rejection of the speaker j, the positive value denotes the acceptance of the speaker, and the zero value means no decision.

Note that we get the rule described in Section 3.2 for $K=1$ and the rule described in Section 3.1 for $K=I$.

4. Description of experiments

4.1. Speech data

Utterances from 329 speakers (199 male and 130 female) were used in our experiments. They were recorded in the same way as in the UWB_S01 corpus [6]. Each speaker read 150 sentences that were divided into 2 groups: 40 sentences were identical for all speakers, and the remaining 110 sentences were different for each speaker. Only the utterances which are identical across all speakers were used in the experiments. They were divided into three parts: 20 sentences of each speaker were used for training of the GMM of the speaker, 10 sentences were used for the construction of the background model, and 1 sentence was used for the tests.

4.2. Acoustic modelling

The voice activity detector described in [7] was used for elimination of the non-speech parts of the utterances (both training and test) before the parametrization. All utterances were resampled to 8 kHz and parametrized using a 25 ms-long Hamming window with a 15 ms overlap. The feature vectors consist of energy and 12 mel-frequency cepstral coefficients, i.e. the dimension of each feature vector is 13.

The models of the speakers and the background model were represented by Gaussian mixture models created using the HTK toolkit. The model of each speaker consists of 32 Gaussian mixtures. Two background models were employed, one for female and the other for male. Each of them consists of 128 Gaussian mixtures. The number of the trained models and the number of the Gaussian mixtures per model are given once more in Table 1.

Table 1: *Overview of the employed models.*

Type of model	Number of models	Number of mixtures
Speaker	329	32
Background	2	128

4.3. Description of tests

In order to find the dependence of the speaker verification performance upon the amount of test data, the number of feature vectors I in the test utterances was gradually changed from 1 to I_{max}. It means that at first only the first feature vector of each test utterance was used for speaker verification, then first two feature vectors were used, and so on. The shortest test utterance consisted of 100 feature vectors, therefore we set $I_{max} = 100$. Using the 25ms-long window with the 15ms overlap during the parametrization (see Section 4.2) the above stated means that the amount of the test speech data changed from 25ms to 1,025ms.

Further, in order to find the dependence of the speaker verification performance upon the size of the sequence in (13), the size of the sequence K was gradually changed from 1 to K_{max}. The maximal size of the sequence K_{max} is equal to the actual value I, because the size of the sequence cannot exceed the number of available feature vectors.

The experiments were performed for each possible combination of the number of the feature vectors and the size of the sequence, it means that there were 5,050 experiments in total.

Each test consisted of a set of verification trials. In each trial, a test utterance was verified against each speaker model. Since we had 329 test utterances and 329 models of speakers, there were $329 \cdot 329 = 108{,}241$ verification trials in one test. 329 of the trials were the trials of the true speaker, the remaining 107,912 trials were impostor trials.

The performance of the tests can be measured by the detection error trade-off (DET) curve, which shows the value of false acceptance and the value of false rejection for various operating points of the verification system. At the point of the DET curve where the false rejection rate and the false acceptance rate are equal so-called equal error rate (EER) is defined. The EER values are used for evaluation of our tests, because EER is more suitable for the comparison of large amount of tests.

5. Experimental results

The values of EER for all experiments are depicted in Fig. 1. Each point of the surface of the 3D graph corresponds to one test. The number of the feature vectors and the size of the sequence are given on the y and x axes, respectively, and the corresponding EER is shown on the z axis. The depicted area has the triangular shape, because the size of the sequence cannot exceed the number of the used feature vectors.

Several vertical sections parallel to the y axis are depicted in Fig. 2 as the dependence of EER upon the number of feature vectors for several sizes of the sequence of feature vectors. For the sake of clarity, only the sizes 1, 25, 50, 75, and 100 are picked out. We can say after the inspection of the results in Fig. 2 that more feature vectors causes higher speaker verification performance. We can see, that the procedures that use larger sequences have worse performance than the procedures that use the shorter sequences for the same number of the feature vectors. The pure majority voting rule (i.e. the size of the sequence is 1) shows the best performance. This is a different result than that was achieved in our previous research [1] for speaker identification, where the best results were achieved using the procedure based on the sequences of speech segments. We suppose that the contrast is caused by different units of the speech signal used in the tests – our previous research dealt with the phonemes (one phoneme consists of several feature vectors), whereas the experiments presented in this paper use the feature vectors. Another reason can be the different number of the models per speaker – the current experiments employ one model per speaker, however the previous research exploited a set of the models for each speaker.

Further, several vertical sections of Fig. 1 parallel to the x axis are depicted as the dependence of EER upon the size of the sequence for several numbers of feature vectors in Fig. 3. For the sake of clarity, only the numbers 1, 25, 50, 75, and 100 are picked out. We can say after the inspection of the results in Fig. 3 that the higher size of the sequence the lower performance for any number of the feature vectors. This conclusion supports the results acquired from Fig. 2, i.e. the pure majority voting rule outperforms the other verification procedures tested in this paper.

6. Conclusions

In this paper, three procedures for speaker verification based on GMM has been described. The procedures differ in the way in which they use particular feature vectors of the utterance for verification of the speaker of the whole utterance: one procedure uses the likelihood of the whole utterance, another one is based on the majority voting rule for single feature vectors, and the remaining one exploits the majority voting rule for sequences of feature vectors. The procedures were tested in the group of 329 speakers for various amount of speech data. The achieved results showed quite logically that the more test data the higher speaker verification performance. However, the procedure based on the majority voting rule for single feature vectors outperformed the other two. Such a result enables us to say that the majority voting rule is better for speaker verification than the procedure based on the log likelihood for the whole utterance which is commonly used in many speaker verification systems.

7. Acknowledgements

The work was supported by the Grant Agency of the Czech Republic, project no. 102/02/0124, and by the Ministry of Education of the Czech Republic, project no. MSM 235200004.

8. References

[1] Padrta, A., Radová, V., “On the Amount of Speech Data Necessary for Successful Speaker Identification”, Proc. of the Eurospeech’03, pp. 3021–3024, Geneva, Switzerland, 2003.

[2] Reynolds, D. A., “Speaker identification and verification using Gaussian mixture speaker models”, Speech Communication, 17, pp. 91–108, 1995.

[3] Stapert, R., Mason, J. S., “Speaker Recognition and the Acoustic Speech Space”, Proc. of the 2001: A Speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece, 2001.

[4] Meuwly, D., Drygajlo, A., “Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modeling (GMM)”, Proc. of the 2001: A Speaker

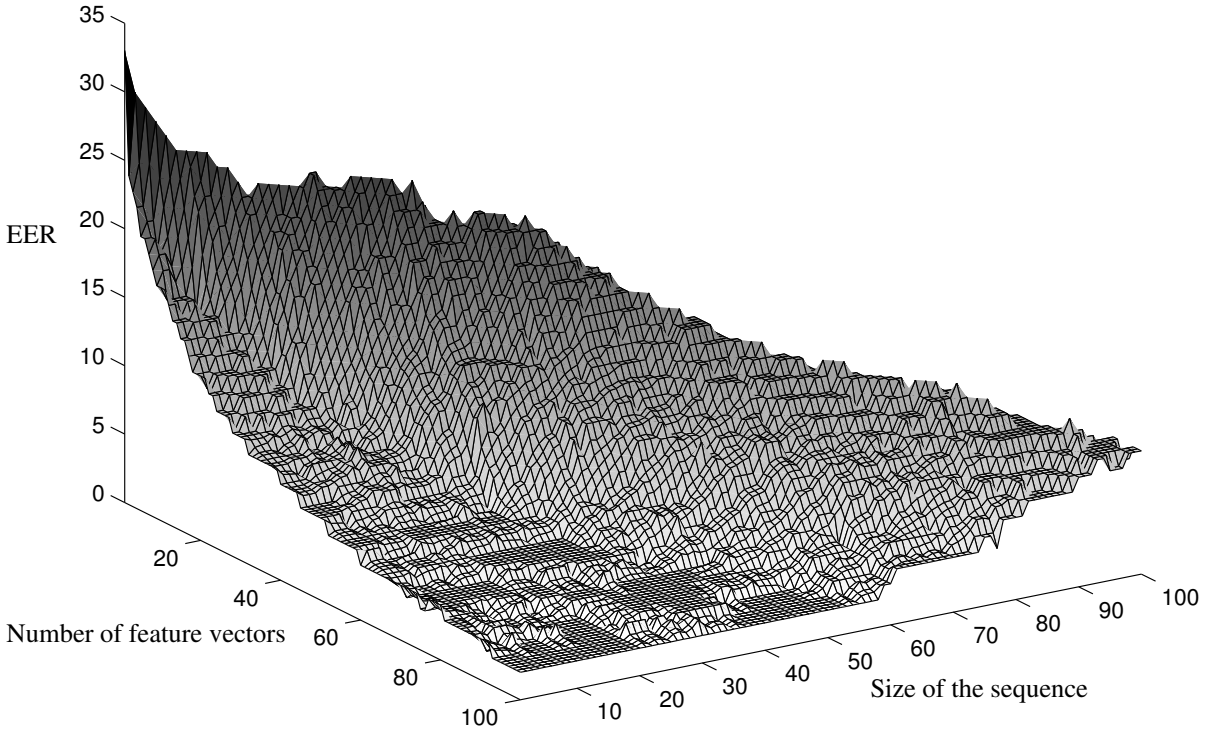

Figure 1: *EER's for all possible combinations of the size of the sequence and the number of feature vectors. The higher EER the darker color of the surface. Values of EER are given in percents.*

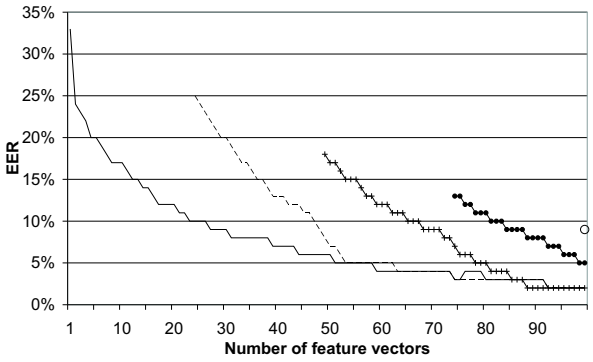

Figure 2: *The dependence of EER upon the number of feature vectors for various sizes of the sequence: 1 = solid line; 25 = dashed line, 50 = solid line with crosses, 75 = solid line with circles, 100 = empty circle (one point)*

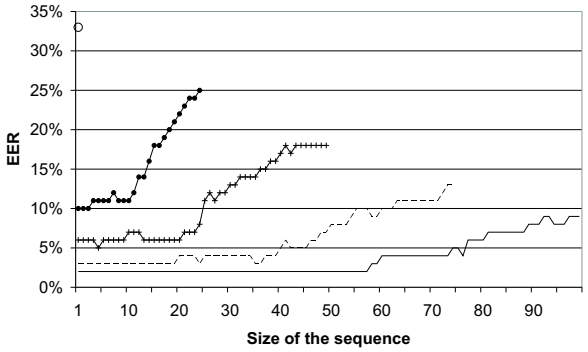

Figure 3: *The dependence of EER upon the size of sequence for various number of feature vectors used: 100 = solid line; 75 = dashed line, 50 = solid line with crosses, 25 = solid line with circles, 1 = empty circle (one point)*

Odyssey, The Speaker Recognition Workshop, Crete, Greece, 2001.

[5] Radová, V., Psutka, J., "An Approach to Speaker Identification Using Multiple Classifiers", Proc. of the ICASSP'97, pp. 1135–1138, Munich, Germany, 1997.

[6] Radová, V., Psutka, J., "UWB_S01 Corpus – A Czech Read-Speech Corpus", Proc. of the ICSLP 2000, pp. 732–735, Beijing, China, 2000.

[7] Prcín, M., Müller, L., Šmídl, L., "Statistical Based Speech/Non-speech Detector with Heuristic Feature Set", Proc. of the SCI 2002 - World Multiconference on Systemics, Cybernetics and Informatics, pp. 264–269, Orlando, FL-USA, 2002.