# On the Background Model Construction for Speaker Verification Using GMM*

Aleš Padrta and Vlasta Radová

University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
{apadrta, radova}@kky.zcu.cz

**Abstract.** A method of speaker verification based on Gaussian mixture models is presented in this paper. The method works with a background model which is composed of several submodels. Several different approaches for construction of the background model from the submodels are introduced here: the log likelihood of the background model is determined either as the average of the log likelihoods of the particular submodels, or a maximum from the log likelihoods of the particular submodels is selected. A large number of experiments was performed in order to find which of the approaches gives the best result. All experiments show that procedures which use a maximum of the log likelihoods of the background submodels have better performance than the procedure which uses the average log likelihood.

## 1 Introduction

The goal of speaker verification systems is to determine whether a given utterance is produced by a claimed speaker or not. This is performed by comparing a score, which reflects the match between the given utterance and the claimed speaker's model, with a threshold. In verification systems based on stochastic models (such as hidden Markov models and Gaussian mixture models) the simplest score is the likelihood of the utterance given the claimed speaker's model. However, such a score is very sensitive to variations in text, speaking behavior, and recording conditions, especially from the utterances of impostors. The sensitivity causes wide variations in scores, and makes the task of threshold determination a very difficult one. In order to overcome this score's sensitivity, the use of the normalized score based on a background model has been proposed [1]. The problem then arise how to select impostor speakers for the background model. Several methods for solution of this problem have been presented e.g. in [2] and [3].

However there is one more interesting question related to the background model construction which have not been studied so much. The question is how to

* 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.

compute the likelihood produced by the background model when the background model is composed of several submodels.

In this paper, we propose 3 methods that can be used for determination of the log likelihood of the background model composed of several submodels. Since we suppose in this paper that the speaker verification procedure is based on the Gaussian mixture models (GMMs) the basic principle of the GMMs is shortly mentioned in Sect. 2. Next, in Sect. 3, the speaker verification procedure is described and the methods for determination of the log likelihood of the background model are introduced. Section 4 deals with the experiments. Finally, in Sect. 5, a conclusion is given.

## 2 Gaussian Mixture Models

Gaussian mixture models are a type of density model which comprises of a number of Gaussian component functions. These component functions are combined to provide multimodal density [1].

A Gaussian mixture density of a feature vector $\boldsymbol{o}$ given the parameters $\lambda$ 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 Procedure

Suppose that there is a group of $J$ speakers, and each speaker $j$, $j = 1, \ldots, J$, is represented by a Gaussian mixture model $\rho_j$. Further suppose that an utterance $O$ is represented by $I$ feature vectors, i.e. $O = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_I\}$, and that the

speaker $\overline{j}$ claims he is the speaker of the utterance $O$. The goal of the verification procedure is to decide whether the utterance $O$ was spoken by the speaker $\overline{j}$ or not.

The verification procedure consists of two steps. First, each feature vector $\boldsymbol{o}_i$ is tested whether it was spoken by the speaker $\overline{j}$ or not, and then, as the second step, a decision is make about the whole utterance.

Let $D(i) = 1$ in the case when the feature vector $\boldsymbol{o}_i$ was spoken by the speaker $\overline{j}$, and $D(i) = -1$ when the feature vector $\boldsymbol{o}_i$ was not spoken by the speaker $\overline{j}$. The overall decision about the whole utterance can be then determined according to the formula

$$D = \sum_{i=1}^{I} D(i). \tag{5}$$

If the overall decision $D$ is positive then the utterance $O$ is proclaimed as being spoken by the speaker $\overline{j}$. The negative value of the overall decision indicates the utterance $O$ was not spoken by the speaker $\overline{j}$. An undecided result occurs when $D = 0$.

The partial decision $D(i)$ can by obtained from the formula [1]

$$D(i) = \begin{cases} 1 \text{ if} & \log p(\boldsymbol{o}_i|\rho_{\overline{j}}) - \log p(\boldsymbol{o}_i|\Lambda) \geq T \\ -1 \text{ if} & \log p(\boldsymbol{o}_i|\rho_{\overline{j}}) - \log p(\boldsymbol{o}_i|\Lambda) < T \end{cases} \tag{6}$$

where $p(\boldsymbol{o}_i|\rho_{\overline{j}})$ is the likelihood of the claimed speaker represented by the model $\rho_{\overline{j}}$ for the feature vector $\boldsymbol{o}_i$, $p(\boldsymbol{o}_i|\Lambda)$ is the likelihood of the background model for the feature vector $\boldsymbol{o}_i$, and $T$ is an a priori selected threshold. $p(\boldsymbol{o}_i|\rho_{\overline{j}})$ is computed according to (1).

Suppose now that the background model $\Lambda$ consists of $B$ submodels $\lambda_b$, $b = 1, \ldots, B$, i.e. $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_B\}$. Further suppose that each submodel $\lambda_b$ is represented by a Gaussian mixture model, so its likelihood for the feature vector $\boldsymbol{o}_i$ can be evaluated according to (1). $\log p(\boldsymbol{o}_i|\Lambda)$ in (6) can be then determined in several ways. In this paper we introduce the three presented in the following subsections.

### 3.1 Average Log Likelihood

In this approach, the log likelihood of the background model $p(\boldsymbol{o}|\Lambda)$ is defined as an average of the log likelihoods $p(\boldsymbol{o}|\lambda_b)$ of the individual submodels $\lambda_b$, $b = 1, \ldots, B$, i.e.

$$\log p(\boldsymbol{o}|\Lambda) = \frac{1}{B} \sum_{b=1}^{B} \log p(\boldsymbol{o}|\lambda_b), \tag{7}$$

where $\boldsymbol{o}$ represents a feature vector.

### 3.2 Maximum Log Likelihood Based on Single Feature Vectors

This approach uses the best submodel (i.e. the submodel with the highest log likelihood) for each separate feature vector. The log likelihood $p(\boldsymbol{o}|\Lambda)$ of the

background model for a feature vector $\boldsymbol{o}$ is then determined as

$$\log p(\boldsymbol{o}|\Lambda) = \max_{b=1,\ldots,B} \log p(\boldsymbol{o}|\lambda_b). \tag{8}$$

It means, the log likelihood of the background model for different feature vectors (even from one speaker) can be computed using different submodels.

### 3.3 Maximum Log Likelihood Based on the Whole Utterance

This approach uses the whole utterance for determination which of the submodels is the best one for all feature vectors. If an utterance $O$ consists of the feature vectors $\boldsymbol{o}_k$, $k = 1, \ldots, K$, the best matching submodel for the whole utterance $\lambda^*$ is selected according to the formula

$$\lambda^* = \operatorname*{argmax}_{\substack{\lambda_b \\ b=1,\ldots,B}} \sum_{k=1}^{K} \log p(\boldsymbol{o}_k|\lambda_b). \tag{9}$$

The log likelihood of the background model $\Lambda$ for a feature vector $\boldsymbol{o}$ is then computed as

$$\log p(\boldsymbol{o}|\Lambda) = \log p(\boldsymbol{o}|\lambda^*). \tag{10}$$

## 4 Experimental Setup

### 4.1 Speech Data

A part of the UWB_S01 corpus was used in our experiments. The UWB_S01 corpus is a read-speech corpus originally designed for training and testing of speech recognition systems [4]. It consists of the speech of 100 speakers (64 male and 36 female). 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. The corpus was recorded in an office room where only the speaker was present. Each utterance was recorded by two different microphones simultaneously. A close-talking microphone (Sennheisser HMD 410-6) recorded utterances of a high-quality, whereas a desk microphone (Sennheisser ME65) recorded utterances including common office noise. Signals from both microphones were sampled at 44.1 kHz with 16-bit resolution.

Only the utterances of each speaker which correspond to the 40 sentences identical across all speakers and which were recorded by the close-talking microphone were used in the experiments described in this paper. They were divided into three parts: 20 utterances of each speakers were used for the training of the GMMs of the reference speakers, 10 other utterances of each speaker were reserved for training of the background model, and the remaining 10 utterances of each speaker were used for tests.

### 4.2 Feature Vectors, Acoustic Modelling

The voice activity detector described in [5] was used for elimination of non-speech parts of the utterances before parameterization. All utterances were resampled to 8 kHz and parameterized using a 25 ms-long Hamming window with a 15 ms overlap. Feature vectors consist of energy and 12 mel-frequency cepstral coefficients, i.e. the dimension of each feature vector is 13.

All models in the experiments are the Gaussian mixture models. Each speaker model consists of 32 Gaussian densities and the background submodels consist of 128 Gaussian densities.

### 4.3 Description of Experiments

In order to simulate various speaker verification situations we divided the whole set of 100 speakers into two groups. One group contained speakers 1–50, the other group consisted of speakers 51–100. The background model was always trained using only the data of the speakers 1–50. We used the background model composed of two background submodels – one model for female speech ($\lambda_1$) and one model for male speech ($\lambda_2$). So, according to the notation introduced in Sect. 3, we had

$$\Lambda = \{\lambda_1, \lambda_2\}. \tag{11}$$

The data used for the training of the models of the reference speakers and the data for the tests changed according to the experiment. The overview of the speakers employed in individual experiments is given in Table 1, a detailed description of the experiments follows.

**Table 1.** Overview of the speakers used in the experiments

| experiment no. | background model speakers | reference speakers | test set speakers |
|---|---|---|---|
| 1 | 1–50 | 1–50 | 1–50 |
| 2 | 1–50 | 1–50 | 1–100 |
| 3 | 1–50 | 51–100 | 1–100 |
| 4 | 1–50 | 51–100 | 51–100 |

As Table 1 shows, the experiments cover all possible combinations that may happen. Experiment 1 can be regarded as an ideal one because all speakers from the test set (i.e. all possible impostors) are included in the background model[1]. Experiments 2 and 3 are more real, the test sets contain both the speakers that are and the speakers that are not included in the background model. It could seem a little bit strange for someone that the test sets in Experiments 2 and 3 cover all the speakers instead of covering only the speakers 51–100 in

[1] Recall from Sect. 4.1 that in spite of the fact that the reference speakers, the speakers in the test set, and the speakers included in the background model are the same, the utterances used for training of the models differ.

Experiment 2 or the speakers 1–50 in Experiment 3. It is because we use the equal error rate (EER) for the evaluation and the test set composed of the speakers 51–100 in Experiment 2 would cause zero values of false rejectance rate for all thresholds. Similar situation would occur also in Experiment 3 if the test set consisted only of the speakers 1–50. Experiment 4 is rather theoretical because it is not common that no one of the reference speakers is included in the background model.

### 4.4 Experimental Results

All experiments specified in Table 1 were carried out for each of the methods for the determination of the log likelihood of the background model presented in Sect. 3. In addition, the influence of the amount of the test data upon the speaker verification performance was tested. This was implemented in such a way that the number of feature vectors $I$ used for evaluation of (5) 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, next first two feature vectors were used, and so on. The shortest utterance consisted of 150 feature vectors, therefore we set $I_{max} = 150$.

The results of the experiments are presented in Table 2. The information about the number of feature vectors used in the experiment is shown in the first row of the table. The first column specifies the experiment, the second column refers to the method used for the determination of the log likelihood of the background model.

In order to obtain a single result which can be used for comparison of the methods of the computation of the log likelihood of the background model, the results of all experiments were averaged. The average equal error rates for various numbers of feature vectors used in the experiments are shown in Table 3 and depicted in Figure 1. The solid line represents the method which uses the maximal log likelihood based on single feature vectors, the dashed line represents the method which uses the maximal log likelihood based on the whole utterance, and the line with the crosses represents the method which computes the average log likelihood of all background submodels.

It can be seen from the results that the method which uses the average of the log likelihoods has the worst performance in all performed experiments. It is caused by the fact that the log likelihood of the bad matching background submodel is also included into the log likelihood of the whole background model, because both the female background submodel and the male background submodel are always used regardless of whether the claimed speaker is a female or a male.

The difference in the performance between the method which uses the maximal log likelihood based on single feature vectors and the method which uses the maximal log likelihood based on the whole utterance is not very significant. Therefore we can say that both methods working with the maximal log likelihood of the background submodels are suitable for computing the log likelihood of the background model in speaker verification tasks. However, the method

**Table 2.** EER in % achieved for various methods of determination of the log likelihood of the background model in various experiments

| | $I$ | 15 | 30 | 45 | 60 | 75 | 90 | 105 | 120 | 135 | 150 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Exp. 1 | max – feature vectors | 16.20 | 9.43 | 6.58 | 5.63 | 4.00 | 3.88 | 4.00 | 4.00 | 4.00 | 2.30 |
| | max – whole utterance | 16.67 | 8.95 | 7.23 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 |
| | average | 14.33 | 12.00 | 8.82 | 7.80 | 5.97 | 5.63 | 6.00 | 6.37 | 6.00 | 4.85 |
| Exp. 2 | max – feature vectors | 16.41 | 10.07 | 7.44 | 6.00 | 4.59 | 4.00 | 4.00 | 4.00 | 4.00 | 3.20 |
| | max – whole utterance | 16.64 | 9.93 | 8.00 | 4.50 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 |
| | average | 15.63 | 12.42 | 9.48 | 8.11 | 6.31 | 5.82 | 6.00 | 6.39 | 6.00 | 4.98 |
| Exp. 3 | max – feature vectors | 14.85 | 9.68 | 9.47 | 7.69 | 4.67 | 2.84 | 2.70 | 2.00 | 1.47 | 1.52 |
| | max – whole utterance | 15.44 | 9.91 | 9.90 | 6.00 | 4.84 | 3.64 | 2.30 | 2.25 | 1.65 | 1.36 |
| | average | 16.98 | 14.60 | 12.07 | 10.00 | 8.61 | 7.11 | 6.00 | 4.55 | 3.76 | 2.88 |
| Exp. 4 | max – feature vectors | 13.68 | 8.29 | 8.04 | 6.25 | 3.59 | 2.00 | 2.00 | 1.33 | 0.93 | 0.93 |
| | max – whole utterance | 14.74 | 8.63 | 8.68 | 6.00 | 4.00 | 2.32 | 2.00 | 1.68 | 1.02 | 0.82 |
| | average | 15.23 | 13.03 | 10.96 | 10.00 | 7.39 | 6.07 | 5.13 | 3.17 | 3.09 | 2.02 |

**Table 3.** Average EER in %

| $I$ | 15 | 30 | 45 | 60 | 75 | 90 | 105 | 120 | 135 | 150 |
|---|---|---|---|---|---|---|---|---|---|---|
| max – feature vectors | 15.28 | 9.37 | 7.88 | 6.39 | 4.21 | 3.18 | 3.18 | 2.83 | 2.60 | 1.99 |
| max – whole utterance | 15.87 | 9.35 | 8.45 | 5.13 | 4.21 | 3.49 | 3.08 | 2.98 | 2.67 | 2.54 |
| average | 15.54 | 13.01 | 10.33 | 8.98 | 7.07 | 6.16 | 5.78 | 5.26 | 4.71 | 3.68 |

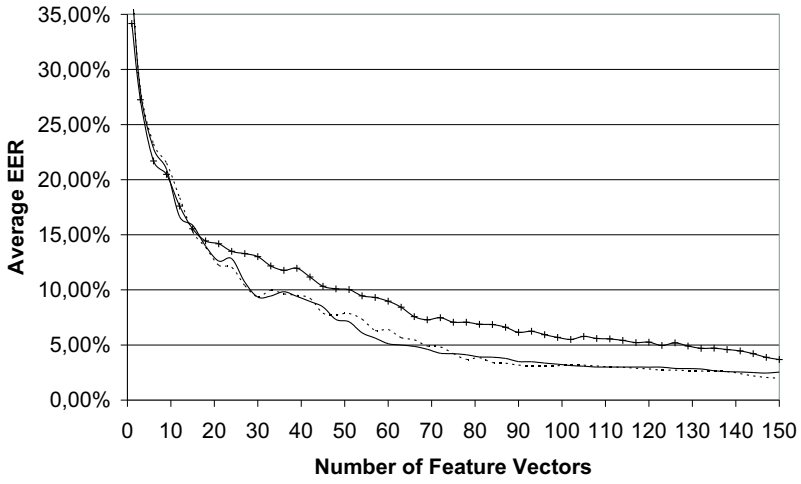


**Fig. 1.** The average EER for particular methods of the computation of the log likelihood of the background model. Solid line = maximal log likelihood based on single feature vectors; Dashed line = maximal log likelihood based on the whole utterance; Line with the crosses = average log likelihood.

which is based on single feature vectors is less time-consuming, therefore it can be regarded as the better one.

## 5 Conclusion

The goal of this paper was to study the dependence of the speaker verification performance on the method which is used for the determination of the log likelihood of the background model composed of several submodels. Three methods for the determination of the log likelihood of the background model were tested. All of them were based on the Gaussian mixture models. The methods were tested in various speaker verification situations using different amount of test data. The results show quite logically that more test data always lead to a better performance of the speaker verification system. However, the procedures which use a maximum of the log likelihoods of the background submodels allow to achieve better results than the procedure which uses the average of the log likelihoods of the background submodels. Therefore they can be regarded as a useful methods for the determination of the background model's likelihood regardless of the amount of test data.

## References


1. Reynolds, D. A.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communication **17** (1995) 91–108
2. Sivakumaran, P., Furtuna, J., Ariyaeeinia, A. M.: Score Normalization Applied to Open-Set, Text-Independent Speaker Identification. EUROSPEECH 2003 Geneva (2003) 2669–2672
3. Zigel, Y., Cohen, A.: On Cohort Selection for Speaker Verification. EUROSPEECH 2003 Geneva (2003) 2977–2980
4. Radová, V., Psutka, J.: UWB_S01 Corpus – A Czech Read-Speech Corpus. Proc. ICSLP 2000 Beijing China (2000) 732–735
5. Prcín, M., Müller, L., Šmídl, L.: Statistical Based Speech/Non-speech Detector with Heuristic Feature Set. SCI 2002 – World Multiconference on Systemics, Cybernetics and Informatics Orlando FL-USA (2002) 264–269