A Bayesian Approach to Speaker Recognition Based on GMMs Using Multiple Model Structures

Takafumi Hattori, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda

Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya, Japan

Abstract

This paper proposes a speaker recognition technique using multiple model structures based on the Bayesian approach. In recent speaker recognition, many sophisticated statistical models have been proposed, e.g., Joint Factor Analysis and i-Vector based method. However, since most of them are based on Gaussian Mixture Models (GMMs), therefore improving estimation accuracy of generative models, i.e. GMMs, with limited amount of training data is still an important problem in speaker recognition. For this purpose, a Bayesian approach which marginalizes all possible model parameters has been applied to the GMM based speaker recognition. This paper extends it to the model structure marginalization. The proposed method can improve the estimation accuracy by integrating multiple GMMs with different numbers of mixtures within the Bayesian framework. Experimental results show that the proposed method improved the identification rates from the conventional method using a single model structure.

Index Terms: speaker recognition, GMM, Bayesian approach, model structure

1. Introduction

In speaker recognition systems, users need to record their speech as training data. These systems have been required to achieve high performance with amounts of recorded speech that are as small as possible to reduce the burden on users. To develop these systems, it is important to appropriately represent speaker characteristics with limited amounts of training data. In recent studies on speaker recognition, many sophisticated statistical models have been proposed, e.g., Joint Factor Analysis and i-Vector based method [1, 2]. Most are based on Gaussian Mixture Models (GMMs) [3, 4], therefore, improving estimation accuracy of such generative models, i.e. GMMs, with a limited amount of training data is still an important problem in speaker recognition.

The classical GMM based speaker recognition methods model spectrum features. For improving estimation accuracy, a Bayesian approach has been applied to the GMM based speaker recognition method [5]. The Bayesian approach estimates a reliable predictive distribution that marginalizes all possible model parameters by using posterior distributions for model parameters. Although GMM based methods must determine the number of mixtures in advance, it is difficult to set the optimal number of mixtures for each task because it depends on the training data, e.g., the amount and variety. Even if the optimal number of mixtures is set, a single GMM may not represent speaker characteristics because the true distribution of speech features is not always included in the distribution family of GMMs. These problems are caused by selecting a single model structure. Therefore, improving model structures for representing the true distribution of speech features is necessary.

The Bayesian approach can marginalize model structures by dealing with the model structure as a random variable. Therefore, we propose a speaker recognition technique for marginalizing model structures based on the Bayesian framework. A GMM with a small number of mixtures represents global characteristics and a GMM with a large number of mixtures represents detailed characteristics. That is, GMMs can represent various characteristics according to the number of mixtures. By integrating such GMMs with a different number of mixtures, the proposed method can improve the estimation accuracy. Non-parametric Bayesian methods [6] have been recently applied to many tasks in machine learning. The proposed method is similar to a non-parametric Bayesian method because both methods use multiple model structures and integrate them based on the Bayesian framework. The main difference between them is that model structures are generated for each data sample with a non-parametric Bayesian method whereas model structures are simply prepared with the proposed method. Although the proposed method estimates multiple GMMs independently, it still has the effect of model structure marginalization similar to a non-parametric Bayesian method and can be performed without increasing the complexity of the training process.

The rest of this paper is organized as follows. Section 2 describes speaker recognition based on GMMs. Section 3 describes speaker recognition based on GMMs using multiple model structures. In section 4, experimental results are presented. Finally, conclusions and future work are drawn in section 5.

2. Speaker recognition based on GMM

A GMM is a probability model represented by the linear combination of Gaussian basis functions. Let $\mathbf{O} = (o_1, o_2, ..., o_T)$ be a training data of $D$ dimensional feature vectors. The output probability of the GMM is outlined in Figure 1. The likelihood function is defined by the following equation:

$$P(O|\lambda) = \sum_Z P(O, Z | \lambda)$$

$$= \prod_{t=1}^{T} \sum_{z_t=1}^{M} w_{z_t} N(o_t | \mu_{z_t}, \Sigma_{z_t})$$

(1)

where $Z = (z_1, z_2, ..., z_T)$ is a latent variable sequence representing mixture components, $w_{z_t}$ is a mixture weight of the $z_t$th component, $\mu_{z_t}$ and $\Sigma_{z_t}$ are a mean vector and a covariance matrix of the $m$th Gaussian component, and $\lambda$ is a set of model parameters.
2.1. Maximum Likelihood approach

Given training data $O$, optimal model parameters of the Maximum Likelihood (ML) method can be written as follows:

$$\lambda_{ML} = \arg \max_{\lambda} P(O | \lambda)$$  \hspace{1cm} (2)

The identification system is a straightforward maximum-likelihood classifier. For a reference group of $K$ speakers represented by models $\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$, the objective is to find the speaker model that has the maximum posterior probability for the input feature vector sequence $X$. The decision rule is

$$k_{max} = \arg \max_{k} P(X | \lambda_k)$$  \hspace{1cm} (3)

2.2. Bayesian approach

The Bayesian approach deals with model parameters as random variables and estimates posterior distributions for model parameters $\lambda$, while the ML method estimates constant model parameters. The posterior distributions for $\lambda$ is obtained with the Bayes theorem as follows:

$$P(\lambda | O) = \frac{P(O | \lambda)P(\lambda)}{P(O)}$$  \hspace{1cm} (4)

where $P(\lambda)$ is a prior distribution for $\lambda$, and $P(O)$ is evidence. Once the posterior distribution $P(\lambda | O)$ is estimated, the predictive distribution for $X$ is given as follows:

$$P(X | O) = \int P(X | \lambda)P(\lambda | O)d\lambda$$  \hspace{1cm} (5)

The model parameters are integrated in Eq. (5) so that the effect of over-fitting is mitigated. Therefore, the Bayesian approach exhibits higher generalization ability than the ML method. However, Eq. (4) and (5) are generally difficult to solve analytically. Therefore, an effective approximation technique is required.

2.2.1. Maximum A Posterior (MAP) approximation

In a simple approximation for the Bayesian approach, the MAP method is widely used. The optimal model parameters of the MAP method can be written as follows:

$$\lambda_{MAP} = \arg \max_{\lambda} P(O | \lambda)P(\lambda)$$

$$= \arg \max_{\lambda} P(O | \lambda)p(\lambda).$$  \hspace{1cm} (6)

The MAP method can be seen as a regularization of the ML method. Therefore, it also uses a point estimate of parameters. While the prior distribution representing prior information can be used in the MAP method, it does not use an integral calculation to estimate the predictive distribution $P(X | \lambda_{MAP})$ similar to the ML method. Thus, it is still affected by the over-fitting problem.

2.2.2. Variational Bayesian approximation

Given a training data $O$, the Bayesian approach is aimed at optimizing the log marginal likelihood $L(O)$ as follows:

$$L(O) = \log \sum Z P(O, Z, \lambda)d\lambda$$  \hspace{1cm} (7)

Using Jensen’s inequality, a lower bound of log marginal likelihood $F$ is defined by:

$$L(O) = \log \sum Z Q(Z)Q(\lambda)\frac{P(O, Z, \lambda)}{Q(Z)Q(\lambda)}d\lambda$$

$$\geq \sum Z Q(Z)Q(\lambda)\log \frac{P(O, Z, \lambda)}{Q(Z)Q(\lambda)}d\lambda$$

$$= F$$  \hspace{1cm} (8)

In the Variational Bayesian (VB) method [7], the VB posterior distributions $Q(\lambda)$ and $Q(Z)$ are introduced to approximate the true corresponding posterior distributions. The optimal VB posterior distributions can be obtained by maximizing $F$ with the variational method as follows:

$$Q(\lambda) \propto P(\lambda)\exp\left\{\sum Z Q(Z)\log P(O, Z | \lambda)\right\}$$  \hspace{1cm} (9)

$$Q(Z) \propto \exp\left\{\int Q(\lambda)\log P(O, Z | \lambda)d\lambda\right\}$$  \hspace{1cm} (10)

These optimizations can be effectively performed using iterative calculations similar to those of the Expectation Maximization (EM) algorithm [8] in the ML method.

In speaker identification with the model learned by the VB method, the predictive distribution for the unknown data $X$ is given as follows:

$$P(X | O) = \sum Z P(X, Z | \lambda)P(\lambda | O)d\lambda$$  \hspace{1cm} (11)

where $Z$ is a latent variable sequence of the unknown data. By the approximation $P(\lambda | O) \approx Q(\lambda)$, the lower bound of predictive distribution is obtained as follows:

$$\log P(X | O) \approx \log \sum Z P(X, Z | \lambda)Q(\lambda)d\lambda$$

$$\geq \sum Z Q(\lambda)\log \frac{P(X, Z | \lambda)}{Q(\lambda)}d\lambda$$

$$= F(X | O)$$  \hspace{1cm} (12)

Using this lower bound as an approximate predictive distribution, the decision rule becomes as follows:

$$k_{max} = \arg \max_{k} F(X | O_k)$$  \hspace{1cm} (13)

Based on the posterior distribution estimation, the Bayesian approach can generally achieve a more robust prediction than the ML method. However, since the optimal number of mixtures depends on several factors (e.g., the amount of training data, the word for training), these approaches are difficult to set the optimal number of mixtures for each speaker model. Even if the optimal number of mixtures is set, a single GMM may not represent the speaker characteristics because the true distribution of speech features is not always included in a distribution family of GMMs.
3. Speaker recognition based on GMMs using multiple model structures

The Bayesian approach can marginalize model structures by dealing with model structures as a random variable. We propose a speaker recognition technique for marginalizing model structures based on the Bayesian approach. The proposed method is performed based on a model integrating multiple GMMs with a different number of mixtures. The likelihood function of the model integrating GMMs \( \{\lambda^{(1)}, \ldots, \lambda^{(N)}\} \) is defined by the following equation:

\[
P(O | \Lambda) = \sum_{N} \sum_{Z_N} P(O, N, Z_N | \Lambda)
\]

\[
= \prod_{t=1}^{T} \sum_{n_t=1}^{M \text{(m)}} \sum_{x_t=1}^{p(\omega^{(n_t)})} P(\omega^{(n_t)}, \omega^{(n_t)} | o_t, z_t | n_t, \Lambda)
\]

\[
= \prod_{t=1}^{T} \sum_{n_t=1}^{M \text{(m)}} \sum_{x_t=1}^{p(\omega^{(n_t)})} N \left( o_t | \mu^{(n_t)}_{x_t}, \Sigma^{(n_t)}_{x_t} \right)
\]

\[
(14)
\]

where \( N \) is a latent variable sequence representing indexes of GMMs, \( Z_N \) is a latent variable sequence representing mixture components, and \( p(m) \) is an integration weight satisfying the following condition:

\[
\sum_{n_t=1}^{N} p(m) = 1
\]

(15)

Figure 2 shows the output probability of the model integrating multiple model structures. If we assume that a new parameter of weights \( \omega^{(n_t)} \) is equal to \( p(m) \omega^{(n_t)} \), Eq. (14) is rewritten as follows:

\[
P(O | \Lambda)
\]

\[
= \prod_{t=1}^{T} \sum_{n_t=1}^{M \text{(m)}} \sum_{x_t=1}^{p(\omega^{(n_t)})} N \left( o_t | \mu^{(n_t)}_{x_t}, \Sigma^{(n_t)}_{x_t} \right)
\]

\[
= \prod_{t=1}^{T} \sum_{n_t=1}^{M \text{(m)}} \sum_{x_t=1}^{p(\omega^{(n_t)})} N \left( o_t | \mu^{(n_t)}_{x_t}, \Sigma^{(n_t)}_{x_t} \right)
\]

(16)

Equation (16) can be regarded as the likelihood function of the GMM in which the number of mixtures is \( n_t \). That is, this integrated model has a similar structure to the conventional model and can be used on the same speaker identification system as conventional GMM based systems. The difference with the conventional GMM is that the proposed model integrates multiple GMMs trained independently. By integrating the GMMs which represent different characteristics, the proposed method can improve the estimation accuracy. In speaker recognition with the integrated model, GMMs with the optimum mixture can be selected by the posterior probability, and multiple GMMs can simultaneously be used for each frame.

The log marginal likelihood of the integrated model can be written as follows:

\[
\hat{L}(O) = \log \sum_{Z_N} \sum_{n_t=1}^{N} \int P(O, N, Z_N, \Lambda) d\Lambda
\]

(17)

If we assume a new parameter of \( Z \) is equal to the index of GMMs \( N \) and the latent variable sequence representing mixture components \( Z_N \), \( Z \) can be defined as the latent variable sequence of the model integrating multiple GMMs at the time. Therefore, Eq. (17) is rewritten as follows:

\[
\hat{L}(O) = \log \sum_{Z} \int P(O, Z, \Lambda) d\Lambda
\]

(18)

Equation (18) is the new log marginal likelihood of the integrated model, and the lower bound of this log marginal likelihood can also be obtained from this equation by using the VB method. In this method, the posterior distributions for each model structures are independently trained. Since the prior distributions of each structure are used, it can select the appropriate prior distribution for each input feature. In addition, the proposed method has the effect of model parameter marginalization. Due to prior distribution, model parameter marginalization, and model structure marginalization, the proposed method can perform higher generalization ability than the ML method and can mitigate the over-fitting problem.

4. Experiments

4.1. Experimental Conditions

To confirm the effectiveness of the proposed method, text independent speaker identification experiments were performed. Two sets of training data consisting of 10 and 50 words per speaker were prepared from the ATR Japanese database c-set which consists of 80 people (male/female 40/40). The test set consisted of 520 words per each speaker not included in the training data. Speech signals were sampled at a 10-kHz rate and windowed at a 10-ms frame rate using a 25.6-ms Blackman window. In this experiment, the following two methods were compared.

- GMM: model using the single GMM
- UNI: integrated model with uniform integration weights

GMM was represented by GMM-\( \{\# \text{ of mixtures}\} \), and the UNI was represented by UNI-\( \{A, B\} \) using the GMM sets. In the proposed method, GMMs with \( 2^N (N = 0, 1, \ldots, 8) \) mixture components were prepared, and the following two sets of GMMs were used for integration. “A set” used all GMMs, and “B set” used GMMs with the top five identification rates in the prepared GMMs in order to evaluate the difference of GMMs to be integrated. In addition, to examine the dependency of the models used in the integrated model, “C set” which used GMMs with the top three identification rates was prepared in the 50 words experiment. The GMMs were trained using ML and MAP, VB estimation methods. In the ML method, LBG codebook was used for the initial value of model parameters. In the MAP and VB methods, a Universal Background Model (UBM) [9], which was trained from the training data of all speakers,
was used as prior information. The integrated model can set the distribution profile by using integration weights. The integration weights based on a uniform distribution and Bayesian Information Criterion (BIC) [10, 11] were used in a preliminary experiment. However, since difference of the identification rate by the integration weights was small, we used only the uniform distribution for determining the integration weights in this experiment.

4.2. Experimental Results

Figure 3 and 4 show the identification rates for the conventional method and the proposed integration method trained with 10 and 50 words, respectively. From Fig. 3, it can be seen that the proposed method achieved a higher identification rate than the conventional one. The conventional method may not represent speaker characteristics because the true distribution of speech features is not always included in the distribution family of GMMs. On the other hand, the proposed method can effectively represent the features of speaker because the integrated model was able to use multiple GMMs with a different number of mixtures for each frame. In addition, because these GMMs with the appropriate number of mixtures based on the posterior probabilities of GMM index $t_k$ were stochastically selected. This indicates that the integrated model was more effective than the conventional GMM. By comparing the VB method and other methods, the proposed method with the VB method was more effective than with the ML and MAP methods. This is because the integrated GMMs based on the VB method can more flexibly select the prior distribution than the single GMM since it can use the prior distribution estimated by each GMM. In addition, the integrated models based on the VB method can be adequately estimated by mitigating the over-fitting problem. In the 50 words experiment, the identification rate of UNI-A was lower than that of the conventional model. This is because the A set in 50 words included many GMMs with a low identification rate. In particular, the integrated models with the VB method were more affected by the prepared GMMs with a small number of mixtures. When the GMM to be integrated were B and C sets, the identification rate with 50 words improved. This suggests that we need to investigate an optimal method for determining the identification weights including the model selection method such as setting the integration weights to 0.

5. Conclusions

We proposed a Bayesian approach to speaker recognition based on GMMs using multiple model structures. The proposed method can improve estimation accuracy by integrating multiple GMMs with a different number of mixtures within the Bayesian framework. The experimental results showed that the proposed method improved the identification rates from the conventional method. For future work, we will investigate a determination method of the optimal integration weights, and extend the integrated model to Joint Factor Analysis and i-Vector based method.

6. Acknowledgements

The research leading to these results was partly funded by the Core Research for Evolutional Science and Technology (CREST) from Japan Science and Technology Agency (JST).