Abstract

This paper presents a speaker identification system based on Gaussian Mixture Models (GMM) using the variational Bayesian method. Maximum Likelihood (ML) and Maximum A Posterior (MAP) are well-known methods for estimating GMM parameters. However, the overtraining problem occurs with insufficient data due to a point estimate of model parameters. The Bayesian approach estimates a posterior distribution of model parameters and achieves a more robust prediction than ML and MAP approach. To solve complicated integral calculations in the Bayesian approach, the variational Bayesian approach has been proposed and applied to many classification problems using latent variable models. However, the performance of the Bayesian approach has not been extensively investigated in large speaker identification tasks. The experimental results shows that the VB method improves the overtraining problem than the conventional ML and MAP methods.

Index Terms: speaker recognition, GMM, variational bayesian method

1. Introduction

In speaker identification systems, users are required to enroll by recording their speech as training data. However, it is desired that the amount of recorded speech be as small as possible. To develop such a system, it is important to reliably estimate statistical models, i.e., Gaussian mixture models (GMMs) [1, 2] from limited amounts of training data.

The current successes in speaker recognition are based on pattern recognition techniques which use statistical learning theory. The Maximum Likelihood (ML) and Maximum A Posterior (MAP) methods have become the standard techniques for constructing speaker models in speaker recognition. However, those methods use a point estimate of model parameters. Therefore, insufficient training data leads to the overtraining problem. In order to avoid this problem, the Bayesian approach [3] has been employed. The Bayesian approach deals with model parameters as probabilistic variables and marginalizes them for constructing prediction distribution of observations. Based on this posterior distribution estimation, the Bayesian approach can generally achieve a more robust prediction than the ML approach. However, the Bayesian approach requires complicated integral calculations to obtain posterior distributions in GMMs.

Recently, the Variational Bayesian (VB) approach [4] which employs the variational approximation technique [5, 6] has been proposed and applied to many classifications using latent variable models. However, the performance of this approach has not been extensively investigated in large speaker recognition tasks. In this paper, we propose speaker recognition based on the VB approach and investigate its effectiveness.

In the Bayesian approach, the determination of prior distribution is an important problem for estimating appropriate models, because prior distributions affect the estimation of posterior distributions. In the MAP approach, an Universal Background Model (UBM) [7] has been widely used. This model is typically constructed by using training data of all speakers, and GMM parameters of each speaker are estimated by adapting the UBM trained with sufficient training data. In this paper, we utilize an UBM as the prior distribution of Bayesian approach. However, there is an adjustable parameter which determines the degree of influence of UBM in estimating the posterior distribution. To automatically determine this adjustable parameter, we evaluate an optimization technique based on a Bayesian criterion which maximizes the marginal likelihood in a speaker identification experiment. The rest of this paper is organized as follows. Section 2 describes speaker identification based on the ML approach. Section 3 describes speaker identification based on the Bayesian approach. In section 4, experimental results on the ATR Japanese dataset are presented. Finally, conclusions and future works are drawn.

2. Speaker identification based on the Maximum Likelihood (ML) approach

GMM is a probability model which is represented by the linear combination of Gaussian basis functions. Let \( O = (o_1, o_2, ..., o_T) \) be a training data of \( D \) dimensional feature vectors. The likelihood function is defined by the following equation:

\[
P(O | \Lambda) = \prod_{t=1}^{T} \sum_{S_t} P(o_t, z_t | \Lambda)
\]

where \( S = (z_1, z_2, ..., z_T) \) is a latent variable sequence representing mixture components, \( \Lambda \) is a set of model parameters which consists of the mixture weights \( w = \{w_m\}_{m=1}^{M} \) and a Gaussian \( \mathcal{N}(\mu_m, S_m^{-1}) \) with the mean vector \( \mu_m \) and the covariance matrix \( S_m^{-1} \).

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

\[
\Lambda_{ML} = \arg \max_{\Lambda} P(O | \Lambda)
\]

The identification system is a straightforward maximum-likelihood classifier. For a reference group of \( K \) speakers represented by models \( \{\Lambda^{(1)}, \Lambda^{(2)}, ..., \Lambda^{(K)}\} \), the objective is to find the speaker model which has the maximum posterior probability for the input feature vector sequence \( X \). The decision rule is

\[
k_{\max} = \arg \max_{k} P(X | \Lambda^{(k)})
\]
The ML method uses a point estimate of GMM parameters, thus the overtraining problem can occur.

3. Speaker identification based on the Bayesian approach

The Bayesian approach is based on posterior distribution instead of a constant model parameters in the ML approach. The posterior distribution for a model $\Lambda$ is obtained with the famous Bayes theorem as follows:

$$P(\Lambda | O) = \frac{P(O | \Lambda)P(\Lambda)}{P(O)}, \quad (4)$$

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

$$P(X | O) = \int P(X | \Lambda)P(\Lambda | O)d\Lambda \quad (5)$$

From Eq. (5), prior information can be utilized via the estimation of the posterior distribution, which depends on the prior distribution. In VB, the VB posterior distributions $Q(\Lambda)$, $Q(Z)$ is obtained as follows:

$$Q(\Lambda) \propto P(\Lambda) \exp \left\{ \sum_{Z} Q(Z) \log P(O, Z | \Lambda) \right\} \quad (9)$$

$$Q(Z) \propto \int Q(\Lambda) \log P(O, Z | \Lambda)d\Lambda \quad (10)$$

These optimizations can be effectively performed by iterative calculations as same as the Expectation and Maximization (EM) algorithm [8], which increase $\mathcal{F}$ at each iteration until convergence.

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

$$P(X | O) = \sum_{Z_X} \int P(X, Z_X | \Lambda)P(\Lambda | O)d\Lambda \quad (11)$$

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

$$h_{max} = \arg \max \mathcal{F}(X | O^{(k)}) \quad (13)$$

3.3. Prior distribution

In this paper, a conjugate prior distribution is utilized as the prior distribution $P(\Lambda)$. The definition of the conjugate prior distribution is that the posterior belongs to the same functional family as the prior. In GMM, the conjugate distributions become Dirichlet distribution for mixture weights $w$, and Gauss-Wishart distribution for the mean vector $\mu_m$ and the precision matrix $S_m$.

$$P(w) = \mathcal{D}(\{w_m\}_{m=1}^M, [\{\phi_m\}_{m=1}^M]) \quad (14)$$

$$P(\mu_m, S_m) = \mathcal{N}(\mu_m | \nu_m, (\xi_m S_m)^{-1}) \times \mathcal{W}(S_m | \eta_m, B_m) \quad (15)$$

where $\{\phi_m, \xi_m, \eta_m, \nu_m, B_m\}_{m=1}^M$ represents prior distributions and are called hyper-parameters in the Bayesian approach.

We set the prior distribution as $P(\Lambda) = P(\Lambda | O)$ by using the data $O$ given in advance (we call this prior data). By using the same approximation techniques as the VB method, the prior distribution is obtained as follows:

$$P(\Lambda) \simeq \frac{1}{C_\Lambda} \exp \left[ \sum_{Z_X} Q(Z) \log P(O, Z | \Lambda) \right]$$

$$= \mathcal{D}(\{w_m\}_{m=1}^M | T_m) \prod_{m=1}^M \left[ \mathcal{N}(\mu_m | \bar{o}_m, (T_m S_m)^{-1}) \times \mathcal{W}(S_m | T_m + D, (T_m C_m)) \right] \quad (16)$$
where $D$ is the dimension of a feature vector, $C_A$ is a normalization term and $Q(Z)$ is an approximate distribution of $P(Z|O, \Lambda)$ which can be estimated via EM algorithm using prior data $O$. Statistics $T_m, \alpha_m, C_m$ denote the amount, the mean vector, and the covariance matrix of prior data with the mixture component, respectively.

Typically, a Universal Background Model (UBM) is used as prior information in MAP approach. The UBM is trained by using training data of all speakers. Therefore, especially since training data is limited, each speaker model can be derived from adapting the parameters of UBM. Using the UBM as prior information in the VB, the hyper-parameters are given as follows:

\[
\phi_m = \xi_m = \hat{T}u_m^{(UBM)}, \quad \eta_m = \hat{T}w_m^{(UBM)} + D, \\
\nu_m = \alpha_m^{(UBM)}, \quad \beta_m = \hat{T}c_m^{(UBM)}
\]

where $\hat{T}$ corresponds to the amount of prior data $O$. By adjusting $\hat{T}$, we can control the degree of influence of the prior distribution on the posterior distribution.

### 4. Experiments

#### 4.1. Experiments Condition

To confirm the effectiveness of the proposed method, speaker identification experiments were performed. In this experiment, the following three approaches were compared: “ML,” “MAP,” and “VB.” The experimental conditions are summarized in Table 1. Two sets of the training data consist of 216 and 5 words respectively. In 216 words case, no notable difference was observed between fixed and varied $\hat{T}$. However, the identification error rate of each optimization method in 216 and 5 words, respectively. In 216 words case, no notable difference was observed between fixed and varied $\hat{T}$. However, the identification error rate of each optimization method in 216 and 5 words, respectively. In 216 words case, no notable difference was observed between fixed and varied $\hat{T}$. However, the identification error rate of each optimization method in 216 and 5 words, respectively. 

<table>
<thead>
<tr>
<th>Database</th>
<th>ATR Japanese database c-set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Speaker</td>
<td>80 (Male/Female 40/40)</td>
</tr>
<tr>
<td>Training data</td>
<td>216 words, 5 words</td>
</tr>
<tr>
<td>Test data</td>
<td>520 words</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>10kHz</td>
</tr>
<tr>
<td>Frame size</td>
<td>25.6ms</td>
</tr>
<tr>
<td>Frame shift</td>
<td>10ms</td>
</tr>
<tr>
<td>Window</td>
<td>Blackman</td>
</tr>
<tr>
<td>Feature vector</td>
<td>12 Mel-Cepstrum Coefficients</td>
</tr>
</tbody>
</table>

#### 4.2. Experiments Results

##### 4.2.1. Number of mixtures and identification error rate

Fig.1, 2 show identification error rates for text independent speaker identification trained by 216 and 5 words, respectively. Among the three methods, “VB” achieved the best results. Especially in Fig.1, when the number of mixture is small, “VB” is more effective than “ML” and “MAP”. In Fig.2, in the case of 64 mixture models, the identification error rates of “ML” and “MAP” significantly increased because of the overtraining problem. On the other hand, the identification error rate of “VB” did not increase unlike “ML” and “MAP”. This indicates the VB method can improve the overtraining problem.

![Figure 1: Identification error rate (216 words)](image1)

![Figure 2: Identification error rate (5 words)](image2)

**4.2.2. Lower bound of log marginal likelihood $\mathcal{F}$ and identification error rate**

In the previous experiment, the adjustable parameter $\hat{T}$ was set to 100 for all speaker models. However, it would not be an optimal value. Using the variational approximation, it would be possible to obtain a more optimal posterior distribution by setting $\hat{T}$ for each speaker. We optimized parameter $\hat{T}$ so as to maximize the lower bound of log marginal likelihood $\mathcal{F}$ for each speaker. Fig. 3, 4 show the average of adjustable parameter $\hat{T}$ and the sum of lower bound of log marginal likelihood $\mathcal{F}$. These figures show that when the number of mixtures increases, the larger adjustable parameters were obtained. This is because sufficient data would be required for each mixture component. Fig. 5, 6 compare the identification error rate of each optimization method in 216 and 5 words, respectively. In 216 words case, no notable difference was observed between fixed and varied $\hat{T}$. However, the identification error rate with fixed $\hat{T}$ is lower than that with varied $\hat{T}$ in 5 words. This indicates that the hyper-parameters which maximize the lower bound of log marginal likelihood $\mathcal{F}$ is over adapted to the training data, therefore the generalization ability for test data was reduced.

### 5. Conclusions

This paper has evaluated speaker recognition based on variational Bayesian method. Experimental results show that the VB approach improves overtraining problem than the conventional ML and MAP approach. We also evaluated an optimization technique of an adjustable parameter in prior distributions based on the Bayesian criterion. However, the generalization ability was degraded because of over adaptation to the training data. As a future work, we will also investigate the prior distribution...
by non prior information or the hierarchical prior distribution.

6. References


