社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74 (32)
2008 / 12 / 10

変分ベイズ法を用いたGMMに基づく話者認識

伊藤 達也† 橋本 佳† 南角 吉彦† 李 晃伸† 徳田 恵一†

† 名古屋工業大学 工学研究科 創成シミュレーション工学専攻 〒466–8555 愛知県名古屋市昭和区御器所町

あらまし ガウス混合モデル (Gaussian Mixture Model; GMM) に基づく話者認識は，これまで尤度最大化 (Maximum Likelihood; ML) 基準や事後確率最大化 (Maximum a Posterior; MAP) 基準が用いられてきた．しかし，ML基準は，モデルパラメータを確定的変数として点推定するため，学習データが十分に得られない場合，モデルの推定精度が低化する可能性がある．この問題に対し，近年，変分ベイズ法が提案され，様々なモデルにベイズ学習が適用可能となった．本研究では，GMMの学習に変分ベイズ法を適用し，ベイズ基準による話者認識の有効性について検討する．

キーワード 話者認識，GMM，変分ベイズ法

Speaker Recognition Based on Gaussian Mixture Models Using Variational Bayesian Method

Tatsuya ITO†, Kei HASHIMOTO†, Yoshihiko NANKAKU†, Akinobu LEE†, and Keiichi TOKUDA†

† Department of Computer Science and Engineering, Nagoya Institute of Technology Gokiso-cho, Showa-ku, Nagoya, 466–8555 Japan

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 robust prediction. To solve complicated integral calculations in the Bayesian approach, the variational Bayesian method has been proposed. This paper investigates the performance of the Bayesian approach in large speaker identification tasks.

Key words 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 random 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 adjustive parameter which determines the degree of influence of UBM in estimating the posterior distribution. To automatically determine this adjustive 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 $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$ be a training data of D dimensional feature vectors. The likelihood function is defined by the following equation:

$$P(\boldsymbol{O}|\boldsymbol{\Lambda}) = \prod_{t=1}^{T} \sum_{z_t} P(\boldsymbol{o}_t, z_t \mid \boldsymbol{\Lambda}) = \prod_{t=1}^{T} \left[\sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_m, \boldsymbol{S}_m^{-1}) \right] \quad (1)$$

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

Given training data $\boldsymbol{O}$, optimal model parameters of the ML method can be written as follows:

$$\boldsymbol{\Lambda}_{ML} = \arg\max_{\boldsymbol{\Lambda}} P(\boldsymbol{O} \mid \boldsymbol{\Lambda}) \quad (2)$$

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

$$k_{\max} = \arg\max_{k} P(\boldsymbol{X} \mid \boldsymbol{\Lambda}^{(k)}) \quad (3)$$

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 $\boldsymbol{\Lambda}$ is obtained with the famous Bayes theorem as follows:

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

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

$$P(\boldsymbol{X} \mid \boldsymbol{O}) = \int P(\boldsymbol{X} \mid \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda} \mid \boldsymbol{O}) d\boldsymbol{\Lambda} \quad (5)$$

From Eq. (5), prior information can be utilized via the estimation of the posterior distribution, which depends on the prior distribution. Therefore, the Bayesian approach is superior to the ML approach.

However, Eq. (4), (5) are difficult to solve analytically in general. Therefore, an effective approximation technique is required.

3.1 Maximum A Posterior (MAP) approximation

In a simple approximation for Bayesian approach, the MAP method can usually be evaluated. An appropriate model structure approached by the MAP method can be written as follows:

$$\boldsymbol{\Lambda}_{MAP} = \arg\max_{\boldsymbol{\Lambda}} P(\boldsymbol{\Lambda} \mid \boldsymbol{O}) = \arg\max_{\boldsymbol{\Lambda}} P(\boldsymbol{O} \mid \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda}). \quad (6)$$

The MAP method can be seen as a regularization of the ML method. Therefore, it also uses a point estimate of parameters. While we can utilize prior information which is represented by the prior distribution $P(\boldsymbol{\Lambda})$ in MAP method, the integral calculation is not employed to estimate the predictive estimation $P(\boldsymbol{X} \mid \boldsymbol{\Lambda}_{\mathrm{MAP}})$ as same as the ML method. Thus, it still has the effect of the overtraining due to a point estimate.

3.2 Variational approximation

Given a training data $\boldsymbol{O}$, the Bayes approach aims at optimizing the log marginal likelihood $\mathcal{L}(\boldsymbol{O})$ as follows:

$$\mathcal{L}(\boldsymbol{O}) = \log P(\boldsymbol{O})$$

$$= \log \sum_{\boldsymbol{Z}} \int P(\boldsymbol{O}, \boldsymbol{Z}, \boldsymbol{\Lambda}) d\boldsymbol{\Lambda} \tag{7}$$

Using Jensen's inequality, a lower bound of log marginal likelihood $\mathcal{F}$ is defined as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{O}) &= \log \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) \frac{P(\boldsymbol{O}, \boldsymbol{Z}, \boldsymbol{\Lambda})}{Q(\boldsymbol{Z}, \boldsymbol{\Lambda})} d\boldsymbol{\Lambda} \\ &\geqq \sum_{\boldsymbol{Z}} \int Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) \log \frac{P(\boldsymbol{O}, \boldsymbol{Z}, \boldsymbol{\Lambda})}{Q(\boldsymbol{Z}, \boldsymbol{\Lambda})} d\boldsymbol{\Lambda} \\ &= \mathcal{F}\end{aligned} \tag{8}$$

where $P(\boldsymbol{\Lambda})$ is a prior distribution. In the VB approach, the VB posterior distributions $Q(\boldsymbol{\Lambda})$ and $Q(\boldsymbol{Z})$ are introduced to approximate the true corresponding posterior distribution. By setting $Q(\boldsymbol{Z}, \boldsymbol{\Lambda}) = Q(\boldsymbol{Z})Q(\boldsymbol{\Lambda})$ to make integral calculations possible, $\mathcal{F}$ is obtained as follows:

$$\begin{aligned}\mathcal{F} = \sum_{\boldsymbol{Z}} \int &\{Q(\boldsymbol{Z})Q(\boldsymbol{\Lambda}) \log \{P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda})P(\boldsymbol{\Lambda})\} \\ &- Q(\boldsymbol{Z})Q(\boldsymbol{\Lambda}) \log \{Q(\boldsymbol{Z})Q(\boldsymbol{\Lambda})\}\} d\boldsymbol{\Lambda}\end{aligned} \tag{9}$$

The optimal VB posterior distributions over $\boldsymbol{\Lambda}$ and $\boldsymbol{Z}$ can be obtained by maximizing $\mathcal{F}$ with respect to $Q(\boldsymbol{\Lambda})$ and $Q(\boldsymbol{Z})$ with the variational method. The optimal VB posterior distributions $Q(\boldsymbol{\Lambda})$, $Q(\boldsymbol{Z})$ are obtained as follows:

$$Q(\boldsymbol{\Lambda}) = C_{\boldsymbol{\Lambda}} P(\boldsymbol{\Lambda}) \exp \left\{ \sum_{\boldsymbol{Z}} Q(\boldsymbol{Z}) \log P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda}) \right\} \tag{10}$$

$$Q(\boldsymbol{Z}) = C_{\boldsymbol{Z}} \exp \left\{ \int Q(\boldsymbol{\Lambda}) \log P(\boldsymbol{O}, \boldsymbol{Z} \mid \boldsymbol{\Lambda}) d\boldsymbol{\Lambda} \right\} \tag{11}$$

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

In speaker identification using the VB method, the predictive distribution for the unknown data $\boldsymbol{X}$ is given as follows:

$$P(\boldsymbol{X} \mid \boldsymbol{O}) = \sum_{\boldsymbol{Z}_{\boldsymbol{X}}} \int P(\boldsymbol{X}, \boldsymbol{Z}_{\boldsymbol{X}} \mid \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda} \mid \boldsymbol{O}) d\boldsymbol{\Lambda} \tag{12}$$

where $\boldsymbol{Z}_{\boldsymbol{X}}$ is a latent variable of the unknown data. By the approximation $P(\boldsymbol{\Lambda} \mid \boldsymbol{O}) \propto Q(\boldsymbol{\Lambda})$, the lower bound of predictive distribution is obtained as follows:

$$\begin{aligned}\log P(\boldsymbol{X} \mid \boldsymbol{O}) &\propto \log \sum_{\boldsymbol{Z}_{\boldsymbol{X}}} \int P(\boldsymbol{X}, \boldsymbol{Z}_{\boldsymbol{X}} \mid \boldsymbol{\Lambda}) Q(\boldsymbol{\Lambda}) d\boldsymbol{\Lambda} \\ &= \log \sum_{\boldsymbol{Z}_{\boldsymbol{X}}} \int Q(\boldsymbol{Z}_{\boldsymbol{X}}) Q(\boldsymbol{\Lambda}) \log \frac{P(\boldsymbol{X}, \boldsymbol{Z}_{\boldsymbol{X}} \mid \boldsymbol{\Lambda})}{Q(\boldsymbol{Z}_{\boldsymbol{X}})} d\boldsymbol{\Lambda} \\ &\geqq \sum_{\boldsymbol{Z}_{\boldsymbol{X}}} \int Q(\boldsymbol{Z}_{\boldsymbol{X}}) Q(\boldsymbol{\Lambda}) \log \frac{P(\boldsymbol{X}, \boldsymbol{Z}_{\boldsymbol{X}} \mid \boldsymbol{\Lambda})}{Q(\boldsymbol{Z}_{\boldsymbol{X}})} d\boldsymbol{\Lambda} \\ &= \sum_{\boldsymbol{Z}_{\boldsymbol{X}}} \int Q(\boldsymbol{Z}_{\boldsymbol{X}}) Q(\boldsymbol{\Lambda}) \log P(\boldsymbol{X}, \boldsymbol{Z}_{\boldsymbol{X}} \mid \boldsymbol{\Lambda}) \\ &\quad - \sum_{\boldsymbol{Z}_{\boldsymbol{X}}} Q(\boldsymbol{Z}_{\boldsymbol{X}}) \log Q(\boldsymbol{Z}_{\boldsymbol{X}}) \\ &= \bar{\mathcal{F}}(\boldsymbol{X} \mid \boldsymbol{O})\end{aligned} \tag{13}$$

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

$$\begin{aligned}k_{\max} &= \arg\max_k P(k \mid \boldsymbol{X}, \boldsymbol{O}) \\ &= \arg\max_k P(\boldsymbol{X} \mid k, \boldsymbol{O}) P(k) \\ &= \arg\max_k \bar{\mathcal{F}}(\boldsymbol{X} \mid \boldsymbol{O}^{(k)})\end{aligned} \tag{14}$$

3.3 Prior distribution

In this paper, a conjugate prior distribution is utilized as the prior distribution $P(\boldsymbol{\Lambda})$. 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 $\boldsymbol{w}$, and Gauss-Wishart distribution for the mean vector $\boldsymbol{\mu}_m$ and the precision matrix $\boldsymbol{S}_m$.

$$P(\boldsymbol{w}) = \mathcal{D}(\{w_m\}_{m=1}^M \mid \{\phi_m\}_{m=1}^M) \tag{15}$$

$$\begin{aligned}P(\boldsymbol{\mu}_m, \boldsymbol{S}_m) &= \mathcal{N}(\boldsymbol{\mu}_m \mid \boldsymbol{\nu}_m, (\xi_m \boldsymbol{S}_m)^{-1}) \\ &\quad \times \mathcal{W}(\boldsymbol{S}_m \mid \eta_m, \boldsymbol{B}_m)\end{aligned} \tag{16}$$

where $\{\phi_m, \xi_m, \eta_m, \boldsymbol{\nu}_m, \boldsymbol{B}_m\}_{m=1}^M$ represents a set of parameters of prior distributions and these parameters are called hyper-parameters in the Bayesian approach.

In this paper, we assume that the prior distribution is set as $P(\boldsymbol{\Lambda}) = P(\boldsymbol{\Lambda} \mid \bar{\boldsymbol{O}})$ by using the data $\bar{\boldsymbol{O}}$ 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:

$$\begin{aligned}P(\boldsymbol{\Lambda}) &\simeq \frac{1}{C_{\boldsymbol{\Lambda}}} \exp \left[\sum_{\bar{\boldsymbol{Z}}} Q(\bar{\boldsymbol{Z}}) \log P(\bar{\boldsymbol{O}}, \bar{\boldsymbol{Z}} \mid \boldsymbol{\Lambda}) \right] \\ &= \mathcal{D}(\{w_m\}_{m=1}^M \mid \{T_m\}_{m=1}^M) \\ &\quad \times \prod_{m=1}^M \left[\mathcal{N}(\boldsymbol{\mu}_m \mid \bar{\boldsymbol{o}}_m, (T_m \boldsymbol{S}_m)^{-1}) \right. \\ &\quad \left. \times \mathcal{W}(\boldsymbol{S}_m \mid T_m + D, (T_m \bar{\boldsymbol{C}}_m)) \right]\end{aligned} \tag{17}$$

where D is the dimension of a feature vector, $C_{\boldsymbol{\Lambda}}$ is a normalization term and $Q(\bar{\boldsymbol{Z}})$ is an approximate distribution of $P(\bar{\boldsymbol{Z}} \mid \bar{\boldsymbol{O}}, \boldsymbol{\Lambda})$ which can be estimated via EM algorithm using prior data $\bar{\boldsymbol{O}}$. Statistics $T_m, \bar{\boldsymbol{o}}_m, \bar{\boldsymbol{C}}_m$ denote the amount, the mean vector, and the covariance matrix of prior data in the m-th mixture component, respectively.

Typically, a Universal Background Model (UBM) is used as prior information in MAP approach based speaker recognition. The UBM is trained by using training data of all speakers. Therefore, even if 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:

$$\phi_m = \xi_m = \bar{T} w_m^{(\mathrm{UBM})}, \quad \eta_m = \bar{T} w_m^{(\mathrm{UBM})} + D,$$

Table 1 Experimental condition

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

$$\boldsymbol{\nu}_m = \bar{\boldsymbol{o}}_m^{(\mathrm{UBM})}, \quad \boldsymbol{B}_m = \bar{T}\bar{\boldsymbol{C}}_m^{(\mathrm{UBM})} \tag{18}$$

where $\bar{T}$ corresponds to the amount of prior data $\bar{\boldsymbol{O}}$. By adjusting $\bar{T}$, 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 three approaches "ML," "MAP," and "VB" were compared. The experimental conditions are summarized in Table 1. Two sets of the training data consist of 216 and 5 words were prepared from ATR Japanese database c-set. The test set consists of 520 words which are not included in the training data. In the ML method, the parameters of UBM are used for the initial value of model parameters. From the results of preliminary experiments, the value of the adjustive parameter $\bar{T}$ is set to 100 in the MAP and VB methods.

4.2 Experiments Results

4.2.1 Number of mixtures and identification error rate

Figure 1, 2 show identification error rates for the text independent speaker identification using 216 and 5 words as training data, respectively. Among the three methods, "VB" achieved the best results. Especially in Fig.1, when the number of mixtures 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.

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

In the previous experiment, the adjustive parameter $\hat{T}$ was set to 100 for all speaker models. However, it is not an optimal value for each speaker model. Using the VB approach, it might be possible to obtain a more appropriate posterior distribution by setting $\hat{T}$. In this paper, we optimized the parameter $\hat{T}$ so as to maximize the lower bound $\mathcal{F}$ for each

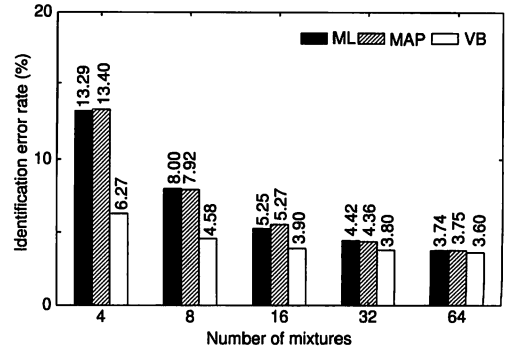

Fig 1 Identification error rate (216 words)

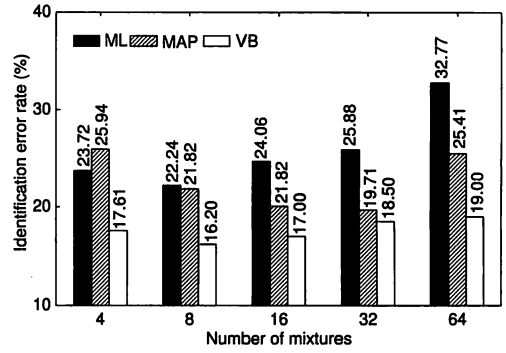

Fig 2 Identification error rate (5 words)

speaker model. To evaluate this approach, we performed speaker recognition experiments. Figure 3, 4 show the average of adjustive parameter $\hat{T}$ obtained by maximizing $\mathcal{F}$ and the sum of lower bound $\mathcal{F}$, respectively. From these figures, it can be seen that when the number of mixtures increases, larger adjustive parameters were obtained. This is because sufficient data is required for each mixture component and training data is compensated by using larger adjustive parameters. Figure 5, 6 compare the identification error rates of the optimization methods ($\hat{T} = 100$ for all speaker models and varied $\hat{T}$) in 216 and 5 words, respectively. In 216 words case, no notable difference was observed between the 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. Figure 7, 8 show the lower bound of log marginal likelihood $\mathcal{F}$ in 216 and 5 words, respectively. And, Fig. 9, 10 show identification rate in 216 and 5 words, respectively. In both 216 and 5 words cases, the values of adjustive parameter pf prior informaiton $\hat{T}$ which give the highest $\mathcal{F}$ are different from that achieve the highest identification rates. 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. Therefore, another criterion which represent the classification performance directly is requied. Figure 11, 12 focus on the lower bound of log marginal likelihood $\mathcal{F}$

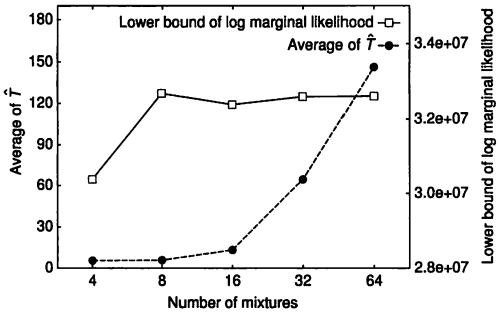

Fig 3 Average of adjustive parameter $\hat{T}$ and lower bound of log marginal likelihood $\mathcal{F}$ (216 words)

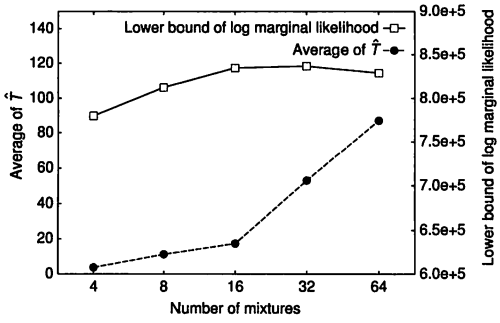

Fig 4 Average of adjustive parameter $\hat{T}$ and lower bound of log marginal likelihood $\mathcal{F}$ (5 words)

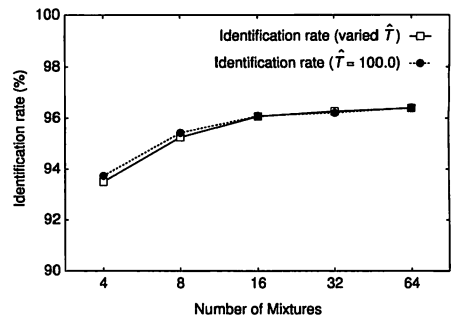

Fig 5 Identification error rates with fixed $\hat{T}$ and varied $\hat{T}$ (216 words)

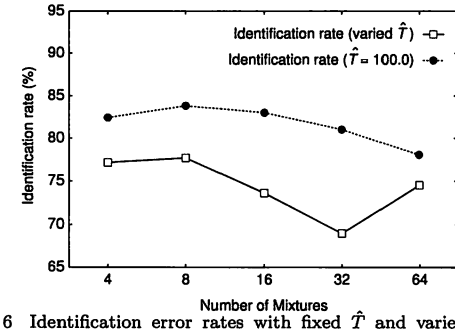

Fig 6 Identification error rates with fixed $\hat{T}$ and varied $\hat{T}$ (5 words)

with smaller value of adjustive parameter of prior information ($10 \leqq \hat{T} \leqq 100$) . In 216 words case, the value of $\hat{T}$ which maximizes the lower bound $\mathcal{F}$ is around 50 with 32 mixture models. In 5 words case, the value of $\hat{T}$ which maximizes the lower bound $\mathcal{F}$ is around 40 with 32 mixture models. From these figures, it can be seen that the number of mixtures M and the adjustive parameter $\hat{T}$ should be determined simultaneously.

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 adjustive 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 investigate the prior distribution determination techniques and other criteria which represent the classification performance directly to construct speaker models.

Reference

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. on Speech and Audio Processing, vol.3, no.1, pp.72–83, Jan. 1995.

[2] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, vol.17, no.1–2, pp.91–108, Aug. 1995.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, "Bayesian data analysis, " Chapman & Hall, 1995.

[4] Z. Ghahramani and M. J. Beal, "Variational inference for Bayesian mixture of factor analysers," Advances in Neural Information Processing Systems, MIT Press, pp.449-455, 2000.

[5] Z. Ghahramani and M. I. Jordan, "On structured variational approximations," University of Toronto Technical Report, 1997, revised 2002.

[6] M. I. Jordan, Z. Ghahramani, T, S. Jaakkola, and L. K. Saul, "An Introduction to Variational Methods for Graphical Models," Machine Learning, vol. 37, pp.183–233, 1999.

[7] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," Eurospeech, pp.963–966, 1997.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," J. Royal Statist. Coc. Ser. B (methodological), vol.39, pp.1–38, 1977.

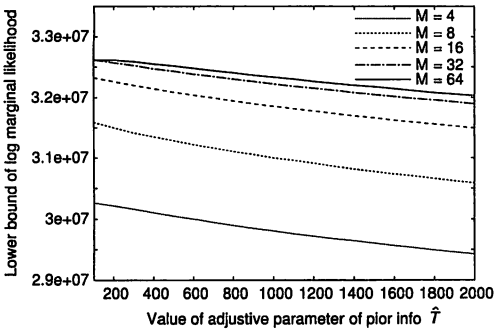

Fig 7 Lower bound of log marginal likelihood $\mathcal{F}$ (216 words)

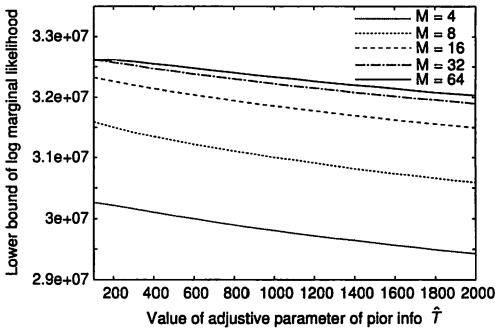

Fig 8 Lower bound of log marginal likelihood $\mathcal{F}$ (5 words)

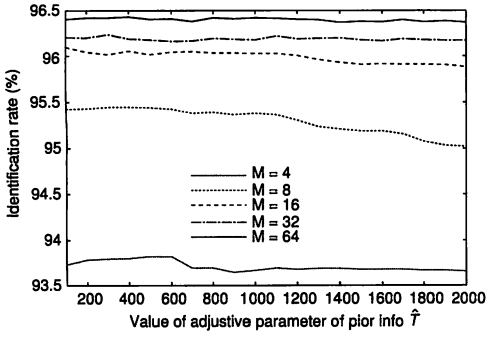

Fig 9 Identification rate (216words)

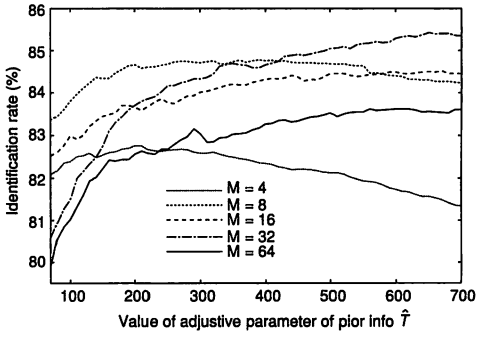

Fig 10 Identification rate (5words)

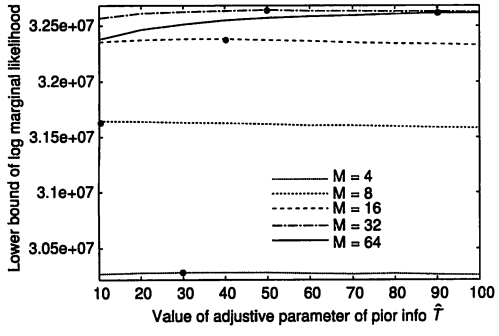

Fig 11 Lower bound of log marginal likelihood $\mathcal{F}$ (216 words)

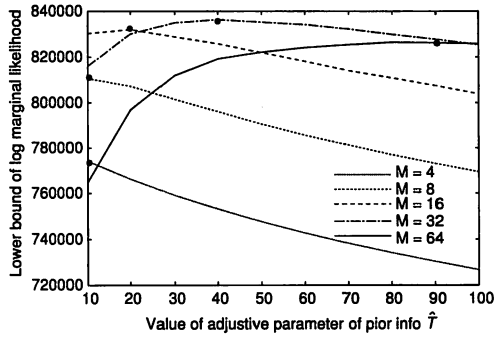

Fig 12 Lower bound of log marginal likelihood $\mathcal{F}$ (5 words)