論題(和文)	話者同定のための複数事前分布を用いたＭＡＰ適応
Title(English)	MAP Adaptation Using Multiple Priors for Speaker Verication
著者(和文)	サンギータビスワス, ヨハンロディーン, 篠田浩一, 古井貞熙
Authors(English)	Sangeeta Biswas, Johan Rohdin, Koichi Shinoda, Sadaoki Furui
出典(和文)	日本音響学会2012年春季研究発表会講演論文集, Vol. , No. , pp. 79-82
Citation(English)	2012 Spring Meeting ASJ, Vol. , No. , pp. 79-82
発行日 / Pub. date	2012, 3

MAP Adaptation Using Multiple Priors for Speaker Verification

◎Sangeeta Biswas, Johan Rohdin, Koichi Shinoda and Sadaoki Furui
(Tokyo Institute of Technology) *

1 Introduction

In automatic speaker verification, Gaussian mixture models (GMMs) [1] have often been used. In order to estimate their parameters robustly by using maximum likelihood (ML) estimation, a large amount of training data are necessary; a small amount of training data generate a non-representative GMM for the acoustic space of the speaker. To deal with this data-sparseness problem, the maximum a posteriori (MAP) adaptation [2] is a well-established method. In MAP adaptation for speaker verification, a prior distribution for each parameter of the GMM is utilized in the training process. How to choose the prior, however, is still a problem.

Reynolds et al. [1] proposed to use parameters of a well-trained GMM called universal background model (UBM) as priors. In this paper, we refer to these priors as UBM priors. Instead of UBM priors, hierarchical priors proposed in SMAP adaptation [3], are used in [4], [5], [6] and [7]. The main difference between the UBM prior and the hierarchical prior is that the UBM prior comes from the average characteristics of many other speakers whereas the hierarchical prior comes from the global characteristics of the same speaker. As shown in Fig. 1, for estimating parameters of subset C of the acoustic space of Speaker-S, UBM priors come from the subset C of other speakers whereas hierarchical priors come from the larger subset, B of the same Speaker-S. The combination of these two priors could improve the performance of speaker verification. In this paper, we propose a technique to combine the hierarchical prior with the UBM prior in the MAP framework. We name this adaptation technique multiprior MAP (MMAP). The benefit of MMAP is shown here by giving the results of experiments conducted on NIST SRE 2006 10sec4w-10sec4w tasks.

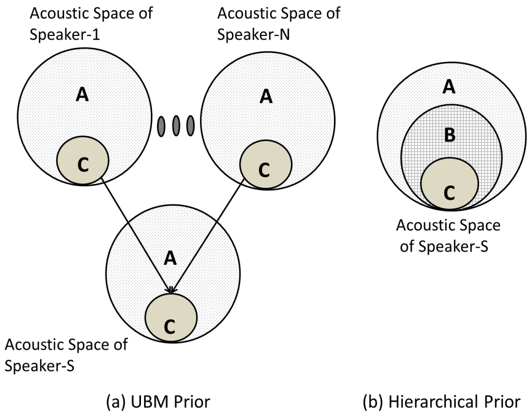
Acoustic Space of Speaker-1
Acoustic Space of Speaker-N
A
C
A
C
A
C
Acoustic Space of Speaker-S
A
B
C
Acoustic Space of Speaker-S
(a) UBM Prior
(b) Hierarchical Prior

Fig. 1 Comparison between UBM prior and hierarchical prior.

2 MAP Adaptation

Gauvain et al. [2] proposed using the MAP estimation framework for speaker adaptation in speech recognition. Let λ be a random vector in a parameter space Λ. Then the MAP estimate of λ is obtained as the mode of its posteror p.d.f. denoted as $g(\lambda|X)$,

$$\begin{aligned}\lambda_{\mathrm{MAP}} &= \arg\max_{\lambda\in\Lambda} g(\lambda|X),\\ &= \arg\max_{\lambda\in\Lambda} f(X|\lambda)g(\lambda), \qquad (1)\end{aligned}$$

where $g(\lambda)$ is the prior p.d.f. of λ and X is the set of T feature vectors,$\{x_1, x_2, x_3, ..., x_T\}$, extracted from speech. Reynolds et al. [1] proposed using the MAP adaptation technique in speaker recognition. In GMM-based speaker verification, the adaptation process is typically applied to the mean vector, μ, of a mixture component, but the variance is kept fixed.

In [2] and [1], the mean vectors of a well-trained world model or UBM were used for the priors of the mean vectors of the speaker specific models. Assuming fixed variance, the conjugate prior for the mean, μ, can be written as:

$$g_{\mathrm{u}}(\mu) = \mathcal{N}(\mu_u, \sigma_u), \qquad (2)$$

where μ_u is the mean of the UBM and σ_u is the standard deviation of the prior. Assuming that

* 話者同定のための複数事前分布を用いたＭＡＰ適応
サンギータビスワス、ヨハンロディーン、篠田浩一、古井貞熙 (東工大)

$\sigma_u = \sigma/\sqrt{\tau}$, where σ is the standard deviation of the UBM, the relevance MAP estimate of the mean vector is:

$$\hat{\mu} = \frac{N\tilde{\mu} + \tau\mu_u}{N + \tau}, \tag{3}$$

where $\tilde{\mu}$ is the ML estimate of mean vector μ. The parameter τ, called relevance factor, regulates the amount of prior knowledge used. N is the total occupation count which is calculated as:

$$N = \sum\nolimits_{t=1}^{T} \sum\nolimits_{k=1}^{K} \gamma_k(x_t), \tag{4}$$

where K is the number of Gaussian components of the GMM, $\gamma_k(x_t) = f(x_t|\lambda_k)$ is the occupation probability of feature vector, x_t being at Gaussian k, with constraints $\gamma_k(x_t) > 0$ and $\sum_{k=1}^{K} \gamma_k(x_t) = 1$.

There are many other ways to define the priors used in MAP adaptation technique. For that reason, Lucey et al. [8] referred to the above method as relevance MAP adaptation. In this paper, we refer to the choice of priors in the relevance MAP adaptation as UBM priors.

3 Structural MAP Adaptation

The structural MAP (SMAP) adaptation was first proposed by Shinoda et al. [3] for speech recognition. Liu et al. [4], Xiang et al. [5] and Ferras et al. [6] successfully applied it to speaker verification, where the duration of the speech segments was around two minutes . Recently we proposed to use it for GMM-SVM(support vector machine)-based speaker verification using 10-second speech segments [7].

SMAP adaptation consists of two steps. In the first step, a tree is obtained by clustering the Gaussian components of the UBM. The root node of the tree represents the whole acoustic space and each of the non-leaf nodes has a Gaussian component that summarizes its child node distributions. Each leaf node corresponds to a Gaussian component in the UBM.

In the second step, a speaker-dependent model is obtained by using the distribution of each non-leaf node for the hierarchical prior of the parameters of its child nodes. The hierarchical prior, $g_h(\mu)$, for a node is:

$$g_h(\mu) = \mathcal{N}(\mu_h, \sigma_h), \tag{5}$$

where μ_h is the mean and σ_h is the standard deviation of hierarchical prior. Let node o be the parent node of node p which is the parent node of node q. Then for node q, μ_h is estimated as:

$${\mu_h}^{(q)} = \mu_u^{(q)} + {\boldsymbol{\Sigma}^{(q)}}^{1/2}\hat{\nu}^{(p)}, \tag{6}$$

where $\mu_u^{(q)}$ is equal to the UBM prior of the node q and $\hat{\nu}^{(p)}$ is the hierarchical shift for node q which is calculated as:

$$\hat{\nu}^{(p)} = \frac{N^{(p)}\tilde{\nu}^{(p)} + \eta\hat{\nu}^{(o)}}{N^{(p)} + \eta}, \tag{7}$$

where $N^{(p)}$ is the total occupation count of frames assigned to node p. η is the relevance factor that weights the shifting value at the parent node o. $\tilde{\nu}^{(p)}$ is the ML estimate of the mean vector of normalized p.d.f. of node p which is estimated as follows:

$$\tilde{\nu}^{(p)} = \frac{\sum_{t=1}^{T} \sum_{k=1}^{K^{(p)}} \gamma_k^{(p)}(x_t) y_{kt}^{(p)}}{\sum_{t=1}^{T} \sum_{k=1}^{K^{(p)}} \gamma_k^{(p)}(x_t)}, \tag{8}$$

where $K^{(p)}$ is the number of siblings of node q and $y_{kt}^{(p)}$ is computed from the adaptation data as follows:

$$y_{kt}^{(p)} = {\boldsymbol{\Sigma}_k^{(p)}}^{-1/2}(x_t - \mu_k^{(p)}). \tag{9}$$

The SMAP estimate of the mean vector is:

$$\bar{\mu}^{(q)} = \mu_u^{(q)} + {\boldsymbol{\Sigma}^{(q)}}^{1/2}\hat{\nu}^{(q)}. \tag{10}$$

4 Multiprior MAP Adaptation

The main difference between the UBM prior and the hierarchical prior is that the UBM prior comes from the average characteristics of many other speakers whereas the hierarchical prior comes from the global characteristics of the same speaker. The prior for a small subset of the acoustic space is based on the estimated parameters of a larger subset of the acoustic space of the same speaker. The combination of these two priors, $g_{\rm u}(\mu)$ and $g_{\rm h}(\mu)$, could improve the performance of speaker verification.

Naturally, there are many ways to combine priors. An intuitive way would be to use a mixture of the priors, i.e., a weighted sum of the two Gaussian priors,

$$g_c(\mu) = (1 - w)g_{\rm u}(\mu) + wg_{\rm h}(\mu) \tag{11}$$

where w controls the weights of the two priors,

$$0 \leq w \leq 1 \tag{12}$$

Unfortunately, its corresponding posterior distribution will also have two Gaussian components as

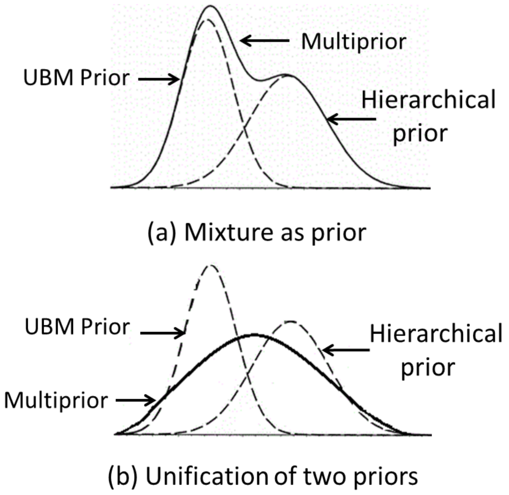

Fig. 2 Ways of combining priors.

shown in Fig. 2(a). Since the MAP approach considers only the global maximum ignoring all other local maxima of the posterior distribution, this form of distribution may not be suitable for MAP adaption. Also, there is no closed expression for the mode of the posterior distribution.

Therefore, instead of using a GMM of two priors, we estimate a new Gaussian prior with a mean equal to a weighted sum of the means of the two Gaussian priors in Eq. (2) and (5) with weights, w, fulfilling (12),

$$g_c(\mu) = \mathcal{N}(\mu_c, \sigma_c) \tag{13}$$

where μ_c for a child node q is:

$$\begin{aligned}\mu_c{}^{(q)} &= (1-w)\mu_u^{(q)} + w(\mu_u^{(q)} + \mathbf{\Sigma}^{(q)^{1/2}}\hat{\nu}^{(p)}) \\ &= \mu_u^{(q)} + w\mathbf{\Sigma}^{(q)^{1/2}}\hat{\nu}^{(p)}.\end{aligned} \tag{14}$$

This can be seen as approximating the two Gaussian priors with one Gaussian as shown in Fig. 2(b). Alternatively this distribution would arise if one considered one of the priors to be a prior for the other prior. The standard deviation, σ_c, may depend on difference of the means of the two priors as well as their variances but perhaps mostly on the properties of the approximation itself. As for MAP adaptation we assume $\sigma_c^{(q)} = \sigma^{(q)}/\sqrt{\zeta}$ where $\sigma^{(q)}$ is the standard deviation of the Gaussian to be adapted. This gives the multiprior MAP (MMAP) estimate:

$$\ddot{\mu}^{(q)} = \frac{N^{(q)}\tilde{\mu}^{(q)} + \zeta(\mu_u^{(q)} + w\mathbf{\Sigma}^{(q)^{1/2}}\hat{\nu}^{(p)})}{N^{(q)} + \zeta}, \tag{15}$$

where ζ is the relevance factor for the multiprior. The only difference from SMAP, Eq. (10), is the weight w. Setting $w = 1$ gives us SMAP and setting $w = 0$ gives us relevance MAP. Our proposed method allows us to choose how much of the shift we want to use.

While MMAP can be used for all nodes, in this study we use MMAP only for the leaf nodes. The hierarchical shifts, $\hat{\nu}^{(p)}$ of the parent nodes are estimated according to the standard SMAP procedure. Three parameters, η, ζ and w, need to be optimized.

5 Experiment

We compared the MMAP adaptation with ML estimation, relevance MAP and SMAP adaptation for text-independent speaker verification.

5.1 Experimental Condition

Performance of our speaker verification system was evaluated on the 10sec4w-10sec4w task of the 2006 NIST SRE [9]. In this task, the length of each training and test segment is approximately 10 seconds. Speaker specific models are trained using only one segment. The training set consists of 731 files for 731 speakers among which 316 are males and 415 are females. The test set consists of 2,971 true trials and 30,584 false trials. As development data, we used the NIST SRE 2005 training database. It has 3,143 true trials and 32,001 false trials for 627 speakers among which 265 are males and 362 are females.

In our evaluation, we used a GMM-UBM system proposed by Reynolds et al. [1]. We trained two gender-dependent UBMs using 4806 speech segments from the NIST SRE 2004 training database. Each speech segment was 2 minutes long on average. Our UBM had 512 Gaussian components.

Regarding feature extraction, we first removed the non-speech part from the speech segments using the information in the transcript files. We broke each segment into frames of 30 ms, with a frame rate of 100 frames/sec. We pre-emphasized each frame with a pre-emphasis factor of 0.97 and applied a Hamming window. We computed 15 perceptual linear prediction (PLP) coefficients, augmented with energy and first-order derivatives, resulting in 32 features per frame. Cepstral mean subtraction was applied to remove static channel effects.

The performance measure was equal error rate (EER) and minimum detection cost (MDC). For

SMAP adaptation, we chose ten different tree structures having odd number of branches in intermediate layers. The feature extraction and GMM construction were implemented by using the hidden markov model toolkit (HTK).

5.2 Results

First we conducted experiments on the development set to optimize model parameters. For SMAP adaptation, when $\eta = 20$, we found that the lowest EER was obtained from the 21_21 tree structure based system. We changed the values of τ and η from 20 to 1 for our relevance MAP and SMAP adapted systems. The lowest EER was obtained when $\tau = 15$ for relevance MAP and $\eta = 10$ for SMAP-adapted system, respectively. For MMAP, we achieved the lowest EER by setting $(\eta, \zeta, w) = (1, 15, 0.1)$.

Table 1 shows our results on the test data.We did not achieve a big improvement in MDC by using MMAP. We achieved only 5.7% and 2.5% relative improvement in MDC over ML estimation and SMAP adaptation, respectively. MDCs were the same for MAP and MMAP adaptation. One of the reasons could be that the optimization of model parameters were only based on the improvement in EER. Since speech segments were very short, ML estimation-based GMMs did not represent speakers well. Therefore, its EER was high. The relevance MAP was better than the SMAP. One of the reasons could be that the optimization of tree structures in SMAP was not good enough. Our proposed MMAP adaptation outperformed ML estimation, relevance MAP adaptation and SMAP adaptation. In MMAP, we achieved 27.6%, 3.3%, and 11% relative improvement in EER over ML estimation, relevance MAP and the SMAP, respectively. Thus, we confirmed that our proposed method was significantly effective.

Table 1 *MDC and EER(%) for GMM-UBM systems using different types of priors in MAP adaptation on the test set.*

Adaptation Techniques	EER	MDC
ML	35.9	0.0991
MAP	26.9	**0.0934**
SMAP	29.2	0.0958
MMAP	**26.0**	**0.0934**

6 Conclusions

In this paper, we proposed a technique to combine the hierarchical prior used in the SMAP adaptation with the UBM prior used in the relevance MAP adaptation in the MAP framework. We named this adaptation technique multiprior MAP (MMAP). We compared the speaker verification performances of classical ML estimation, relevance MAP, SMAP, and MMAP adaptation techniques for short speech segments in NIST SRE 2006 10sec4w-10sec4w task. Our proposed method, MMAP achieved 27.6%, 3.3%, and 11.0% relative improvement in EER over ML estimation, relevance MAP and SMAP, respectively. Our experimental results showed that it is better to set a small value to η i.e., use a small amount of hierarchical prior in the multiprior. One of the reasons could be that the hierarchical prior was not estimated properly. This fits well with the fact that relevance MAP performs better than SMAP.

In order to obtain a better baseline we need better channel compensation techniques. In future, we therefore, would like to introduce JFA in our approach. We would also like to find a better approach to optimize the tree structure in order to improve the performance of SMAP as well as MMAP adapted systems.

Our MMAP framework can also be used in other applications such as acoustic modeling for speech recognition. We also plan to study in this direction.

References

[1] Reynolds et al., Digital Signal Processing, 2000, pp 19-41.

[2] Gauvain et al., IEEE Trans. on Speech and Audio Processing, 1994, vol. 2, no. 2, pp 291-298.

[3] Shinoda et al., IEEE Trans. on Speech and Audio Processing, 2001, vol. 9, no. 3, pp 276-287.

[4] Liu et al., Proc. ICSLP, 2002, pp 1353-1356.

[5] Xiang et al., IEEE Trans. on Speech and Audio Processing, 2003, pp 447-456.

[6] Ferras et al., Proc. ICASSP, 2011, pp 5432-5435.

[7] Biswas et al., Proc. Interspeech, 2011, pp 2377-2380.

[8] Lucey et al., ICME, 2003, pp I-69 - I-72.

[9] http://www.itl.nist.gov/iad/mig/tests/sre/2006/sre-06_evalplan-v9.pdf, 2006.