論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Voting Approach in SMAP Adaptation for Speaker Verification
著者(和文)	ﾋﾞｽﾜｽ ｻﾝｷﾞｰﾀ, 篠田 浩一, 古井 貞熙
Authors(English)	Sangeeta Biswas, Marc Ferras, Koichi Shinoda, Sadaoki Furui
出典(和文)	日本音響学会2011年春季講演論文集, Vol. , No. 2-5-2, pp. 45-48
Citation(English)	, Vol. , No. 2-5-2, pp. 45-48
発行日 / Pub. date	2011, 3

Voting Approach in SMAP Adaptation for Speaker Verification *

◎Sangeeta Biswas, Marc Ferras, Koichi Shinoda and Sadaoki Furui
(Tokyo Institute of Technology)

1 Introduction

Practical application of automatic speaker verification demands high verification accuracy using very short speech even in the text-independent case. However, it is hard for a speaker verification system to find clear speaker-specific characteristics from very short speech when users are not bound to say the same text all the times. For 10 seconds short speech, Vogt et al. [9] proposed using speaker subspace MAP adaptation into factor analysis (FA) modeling. Fauve et al. [2] proposed a well-tuned speech detection front-end for improved frame selection followed by eigenvoice modeling. Kenny et al. [5] extended joint factor analysis (JFA) to model within session-variability over a shorter time span.

We try to handle 10 seconds short speech by structural modeling of human voice characteristics using structural maximum-a-posteriori (SMAP) adaptation technique. The SMAP adaptation technique was proposed by Shinoda et al. [8] in speech recognition. In speaker verification, Liu et al. [6] and Xiang et al. [10] successfully used it for speech segments of about 2 minutes long or shorter.

In SMAP adaptation, a tree structure is used to model the acoustic space of all the speakers. However, during our work on speaker verification, we notice that one particular tree structure is not always optimal for modeling the acoustic space of every speaker. In this paper, we propose a voting approach as a way to combine decisions of multiple systems with different tree structures. We expect that this approach is more robust than SMAP adaptation with a single tree structure.

2 Speaker Modeling

In text-independent speaker verification, a GMM-SVM system proposed by Campbel et al. [1], is accepted as one of the state-of-art systems. This system associates robustness of the GMM-UBM system proposed by Reynolds et al. [7] with discrminative power of the SVM system. In this system, at first a speaker-inpendent Gaussian mixture model (GMM) is trained using hours of speech by hundreds of speakers. This GMM is called a universal background model (UBM). After training the UBM, adaptation methods are used to make a speaker-dependent GMM from UBM using a small amount of speech data for the target speaker. For adaptation, the most popular method is the relevance MAP proposed by Gauvain et al. [3]. After making the GMM, a supervector is made by stacking the mean vectors of the GMM. Supervectors for a set of background speakers, used as negative data in the support vector machine (SVM) classifier are obtained in the same way. Then the supervectors are used as inputs to a SVM with linear kernel to train a GMM-SVM system for the target speaker. For each test speech segment x, score is calculated as follows:

$$\mathcal{S}(x) = wx + b, \tag{1}$$

where b is a *constant* and w is calculated as follows:

$$w = \sum_{i=1}^{N} \alpha_i t_i \hat{x}_i, \tag{2}$$

where N is the number of supportvectors, $\hat{x}_i$ is the i-th supportvector, t_i is the class ID {1,-1} of $\hat{x}_i$, α_i is the Lagrange multiplier, $\alpha_i > 0$, and $\sum_{i=1}^{N} \alpha_i t_i = 0$.

3 SMAP Adaptation

In the SMAP-based method, at first, 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 of the leaf nodes corresponds to a Gaussian component in the UBM as shown in Fig. 1. After building the tree, a speaker-dependent model is obtained by using each non-leaf node as prior information for its child nodes.

* 話者照合のための SMAP 適応化における投票法

The formulation of SMAP adaptation is similar to that of the relevance MAP, except that it uses hierarchical priors and uses normalized pdfs in the formulation. For the adaptation data $X = \{x_1, x_2, ..., x_T\}$, the SMAP estimate of the mean vector is:

$$\hat{\mu}_m^{(p)} = \mu_m^{(p)} + \sum\nolimits_m^{1/2} \hat{\nu}^{(p)}, \tag{3}$$

where $\mu_m^{(p)}$ is the unadapted mean vector for Gaussian m of node p and $\hat{\nu}^{(p)}$ is the hierarchical prior which is calculated as follows:

$$\hat{\nu}^{(p)} = \frac{N_p \tilde{\nu}^{(p)} + \tau \hat{\nu}^{(p-1)}}{N_p + \tau}, \tag{4}$$

where $N_p = \sum_{t=1}^{T} \sum_{m=1}^{M^{(p)}} \gamma_{mt}^{(p)}$ is the average number of frames assigned to node pdf p and τ is the MAP relevance factor that weights the priors at the parent node $p-1$. $\gamma_{mt}^{(p)}$ is the occupation probability for Gaussian m at tree node p and time t. $\tilde{\nu}$ is the ML estimation of the mean vector of normalized pdf of node p which is estimated as follows:

$$\tilde{\nu}^{(p)} = \frac{\sum_{t=1}^{T} \sum_{m=1}^{M^{(p)}} \gamma_{mt}^{(p)} y_{mt}^{(p)}}{\sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{mt}^{(p)}}, \tag{5}$$

where $y_{mt}^{(p)}$ is computed from the adaptation data as follows:

$$y_{mt}^{(p)} = \mathbf{\Sigma}_m^{-1/2}(x_t - \mu_m^{(p)}), \tag{6}$$

When a sufficient amount of training data is not available for a Gaussian component, it is not shifted in relevance MAP. In SMAP adaptation, in such case, it takes prior information from its parent Gaussian. Accordingly, every Gaussian component is shifted from its position in UBM. Fig. 1 shows a schematic example, where {*a*, *b*, *c*} get prior information from *h*, {*d*, ..., *g*} from *i*, and {*h*, *i*} from *j*.

4 Voting Approach

Different speakers have different acoustic spaces depending on factors such as their language, accents or pronunciation particularities. It is therefore reasonable to think that the optimal tree structure differs from speaker to speaker. In other words, some tree structures may be adapted more efficiently to some speakers than others. Our preliminary experiments indicated that decisions involving certain speakers are slightly sensitive to the chosen tree

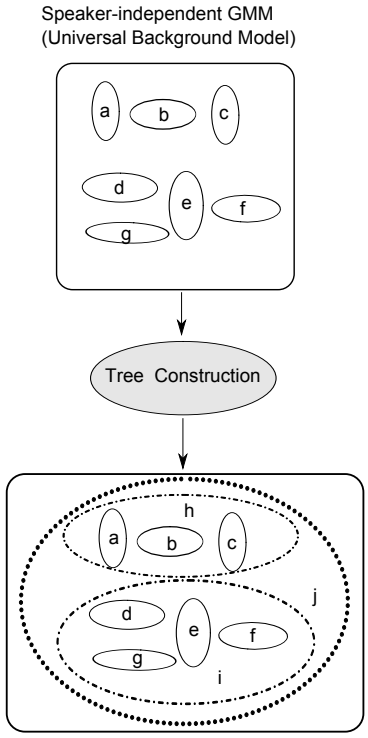


Fig. 1 Example of a tree structure of Gaussian components in SMAP. Each of *a*, *b*, ..., *g* is Gaussian component of UBM. *h*, *i* and *j* are parent Gaussians of {*a*, *b*, *c*}, {*d*, ..., *g*} and {*a*, ..., *g*}, respectively.

structure. In this paper, we propose a simple voting approach to combine decisions of multiple systems with different tree structures as a way to mitigate this problem. To proceed, we construct a set of K SMAP adapted systems with different tree structures and:

1. For each trial x, ask $yes(Y)/no(N)$ vote to each of the K systems

$$\mathcal{V}(x) = \begin{cases} Y & \text{if} \quad \mathcal{S}(x) \geq \theta_k, \\ N & \text{if} \quad \mathcal{S}(x) < \theta_k, \end{cases} \tag{7}$$

where $\mathcal{S}(x)$ is the score of trial x and θ_k is the speaker independent threshold of system k.

2. Count each type of votes, $N_{\{\mathcal{V}(x)=Y\}}$ and $N_{\{\mathcal{V}(x)=N\}}$.

3. Take final decision $true(T)/false(F)$ about each trial as follows:

$$\mathcal{D}(x) = \begin{cases} T & \text{if} \quad N_{\{\mathcal{V}(x)=Y\}} \geq N_{\{\mathcal{V}(x)=N\}}, \\ F & \text{Otherwise} \end{cases} \tag{8}$$

5 Experimental Setup

We made a GMM-SVM system. The performance of our speaker verification system was measured by carrying out experiments on the 10sec4w-10sec4w task of the 2006 NIST SRE. In this task, the length of each training and test segment is approximately 10 seconds. There are 2971 true trials and 30584 false trials for 731 speakers among which 316 are males and 415 are females.

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 long 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 and Mel-Frequencey Cepstral Coefficients (MFCCs), augmented with energy, first and second-order derivatives, resulting 48 features per frame. Cepstral mean subtraction was applied to remove static channel effects. We trained one gender-independent UBM and two gender-dependent UBMs using 4806 speech segments from NIST SRE 2004 training database. Each speech segment was 2.5 minutes long on average. Among 4806 speech segments, 242 speech segments of male speakers and 362 speech segments of female speakers were selected as speech segments of background speakers.

We chose two groups of SMAP adapted systems. In the first group, there were eight systems using binary trees and in the second group there were 15 systems using 15 different tree structures where each node had odd number of children. The performance measure was the Equal Error Rate (EER). To calculate the EER of our proposed voting method-based system, the scores of majority group were linearly fused. The threshold θ_k of SMAP adapted system k was optimized a posteriori using the test set, based on the minimum detection cost (MDC) used in the NIST 2006 SRE [4].

6 Result

At first, we conducted an experiment on relevance MAP-adapted GMMs with 32 Gaussian components. By setting the relevance factor equal to 10, we found that the system using gender-dependent UBM was better than the system using gender-independent UBM, and PLP outperformed MFCC. We also noticed that the performance of our MAP adapted system improved, when we increased the number of Gaussian components until 512 and decreased the relevance factor to 1, and when we did not use the delta-delta coefficients. For SMAP adapted system, we used the gender-dependent UBM with 512 Gaussian Components, 32 dimensional PLP feature vector (i.e. 15 PLP + 15 ΔPLP + E + Δ E), and set the relevance factor to 1.

Fig. 2 shows the EER of our MAP and SMAP adapted systems when the length of speech segments of background speakers was 2.5 minutes on average. The general trend was that the EER decreased as the number of nodes of the trees got larger. For both groups, voting approach outperformed the best SMAP adapted system as well as relevance MAP adapted system although the performances of most of the binary tree based systems were worse than the relevance MAP adapted system. We also noticed that the large diversity of selecting tree structures improved the effectiveness of voting approach. From the MAP baseline system, we obtained an additional gain of 1.58% EER for Group-1 and 3.97% EER for Group-2. Therefore, for the further experiments, we used the tree structures of Group-2.

We noticed that, as the length of the speech segment of background speaker decreased from 2.5 minutes, the overall EER of relevance MAP and SMAP adapted systems decreased. However, it is not clear yet why relevance MAP adapted system started outperforming over SMAP adapted system for shorter background speech. The MV curve of Fig. 3 shows that by using our voting approach it is possible to get all times better performance from SMAP adapted system.

7 Conclusion

We have proposed a voting technique to optimize the tree structure for each speaker in SMAP adaptation. We tested it on a speaker verification task, namely the 10sec4w-10sec4w condition of the 2006 NIST SRE which is an inherently difficult task due to the short length of the speech segments. We showed that the voting technique is effective although relative gain is small. As a future work, we

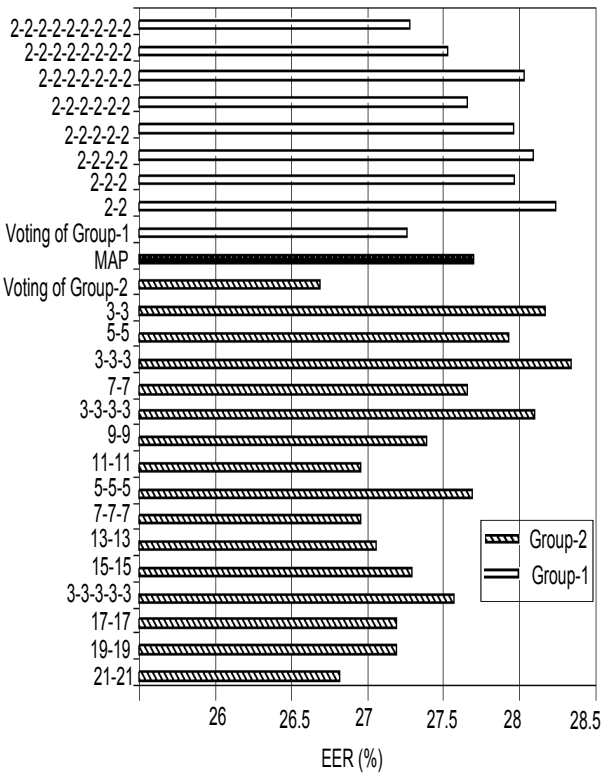


Fig. 2 EER for GMM-SVM systems using MAP and SMAP adaptation on the 10sec4w-10sec4w task of 2006 NIST SRE. The design of a tree is written as $n_1 - n_2 - ... - n_l$ where n_l represents the maximum number of child nodes belonging to each node of the l-th layer.

would investigate other score fusion strategies such as those based on neural network or logistic regression. We would also like to find out the reason of the worse performance of SMAP adapted system than MAP adapted system when the length of speech segment of background speakers decreases.

References

[1] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311, 2006.

[2] B. Fauve, N. Pearson N. Evans, J. F. Bonastre, and J. Mason. Influence of task duration in text-independent speaker verification. In *Proc. Interspeech*, pages 794–797, August 2007.

[3] J. L. Gauvain and C.-H Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. on Speech and Audio Processing*, 2:291–298, 1994.

[4] http://www.nist.gov/speech/tests/spk/.

[5] P. Kenny and N. Dehak. Factor analysis conditionning. In *Report from JHU workshop*, pages 20–42, 2008.

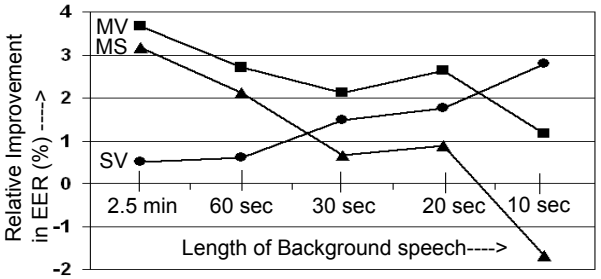
[6] M. Liu, E. Chang, and B.-Q. Dai. Hierarchical gaussian mixture model for speaker verification. In *Proc. ICSLP*, September 2002.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000.

[8] K. Shinoda and C-H. Lee. A structural bayes approach to speaker adaptation. *IEEE Trans. on Speech and Audio Processing*, 9(3):276–287, March 2001.

[9] R. Vogt, C. Lustri, and S. Sridharan. Factor analysis modelling for speaker verification with short utterances. In *Proc. IEEE Odyssey:The Speaker and Language Recognition Workshop*, January 2008.

[10] B. Xiang and T. Berger. Efficient text-independent speaker verification with structural gaussian mixture models and neural network. *IEEE Trans. on Speech and Audio Processing*, 11:447–456, September 2003.



MAP	27.7	27.4	26.9	26.8	26.0
SMAP	26.8	26.8	26.7	26.5	26.5
Voting	26.6	26.7	26.3	26.1	25.7

Fig. 3 Relative improvement in EER (%) for different length of speech segments of background speakers for three cases: (i) SV: Voting approach over SMAP, (ii) MV: Voting approach over MAP and (iii) MS: Best single tree based SMAP over MAP