論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Acoustic Model Adaptation using Structural Bayes Approach
著者(和文)	篠田 浩一
Authors(English)	Koichi Shinoda, Chin-Hui Lee
出典(和文)	日本音響学会講演論文集, , , pp. 47-48
Citation(English)	, , , pp. 47-48
発行日 / Pub. date	1998, 9

2—1—5

Acoustic Model Adaptation using Structural Bayes Approach *

○ Koichi Shinoda(NEC Corporation) and Chin-Hui Lee(Lucent Technologies)

1. Introduction

Maximum a posteriori(MAP) estimation has been successfully applied to speaker adaptation for speech recognition systems using Hidden Markov Models[1, 2]. It has a good asymptotic nature that its performance converges to that of maximum-likelihood(ML) estimation when the amount of data is sufficiently large. In this paper, we propose a structural maximum a posteriori(SMAP) approach[3, 4] to enhance the performance of MAP when the amount of data is extremely small.

2. SMAP Adaptation using Hierarchical Priors

In this paper, we focus on the adaptation of the parameters of Gaussian pdfs in continuous-density(CD) HMMs. Let g_m be a normal density function for mixture component m, $N(\boldsymbol{x}|\mu_m, \Sigma_m)$, where μ_m is a mean vector and Σ_m is a covariance matrix, and let $G = \{g_m; m = 1, \ldots, M\}$ be the whole set of mixture components in CDHMMs, where M is the sum of the number of mixture components in all the states of all the speech units.

Let $\boldsymbol{x} = (x_1, \ldots, x_T)$ denote a given set of T observation vectors for adaptation (adaptation data). At the first step, each sample vector x_t is transformed into a vector y_{mt} for each mixture component m as follows:

$$y_{mt} = \Sigma_m^{-1/2}(x_t - \mu_m), \qquad t = 1, \ldots, T,\ m = 1, \ldots, M. \tag{1}$$

The pdf for $\boldsymbol{y}_m = y_{m1}, \ldots, y_{mT}$ is assumed to be $N(\boldsymbol{y}|\nu, \eta)$, where $\nu \neq 0$ and $\eta \neq I$. We call this pdf *mismatch pdf*.

The optimal number of mismatch pdfs is likely to change according to the amount of data available. Therefore, a method that can utilize both global and local structures is preferable to achieve good performance with any amount of data. For this purpose, we introduce hierarchical Bayes analysis (see [5] and references therein). At first, we consider a *tree structure* for the set G be given as shown in Fig.1, where K is the number of layers. Each node in the K-th layer (leaf node) corresponds to one mixture component of CDHMMs. The root node (the first layer) corresponds to the whole set of the mixture components, G. Each intermediate node corresponds to a subset of G, each of whose elements corresponds to one of its subordinate leaf nodes. At each node N_k in the tree, a mismatch pdf, which is shared among the mixture components in the corresponding subset G_k, is assigned. The ML estimates of the mismatch pdf parameters, $\tilde{\nu}_k$ and $\tilde{\eta}_k$, are calculated using the adaptation data as:

$$\tilde{\nu}_k = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M_k}\gamma_{mt}^{(k)} y_{mt}^{(k)}}{\sum_{t=1}^{T}\sum_{m=1}^{M_k}\gamma_{mt}^{(k)}}, \tag{2}$$

$$\tilde{\eta}_k = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M_k}\gamma_{mt}^{(k)}(y_{mt}^{(k)} - \tilde{\nu}_k)(y_{mt}^{(k)} - \tilde{\nu}_k)^t}{\sum_{t=1}^{T}\sum_{m=1}^{M_k}\gamma_{mt}^{(k)}} \tag{3}$$

where M_k is the number of mixture component in G_k, $\gamma_{mt}^{(k)}$ is the posterior probability of using the m-th mixture component in G_k, $g_m^{(k)}$, at time t, and $(y_{mt}^{(k)} - \tilde{\nu}_k)^t$ is a transpose of $(y_{mt}^{(k)} - \tilde{\nu}_k)$. In the tree structure, each mixture component corresponds to one node sequence from the root to a leaf. From now on, we focus on estimation of the parameter set $\theta_m = (\mu_m, \Sigma_m)$ for one mixture component m in G, and omit the suffix for mixture components. The same procedure can be applied to the other mixture components in CDHMMs.

Let the node sequence from the root to the leaf correspond to mixture component m as $\{N_1, \ldots, N_k, \ldots, N_K\}$, where N_1 is the root node and N_K is the leaf node directly attached to mixture component m. We denote $\lambda_k = (\nu_k, \eta_k)$ as the pdf parameters for node N_k. In our approach, a set of priors $\{\lambda_0, \lambda_1, \ldots, \lambda_k, \ldots, \lambda_{K-1}\}$ is used as *hierarchical priors* for λ_K, where $\lambda_0 = (0, I)$. The pdf with λ_0 is assumed to be the prior for the parameter set λ_1 of the root node, and the pdf for node N_k, which has the parameter set λ_k is assumed to be the prior for the parameter set λ_{k+1} in its immediate subordinate node, N_{k+1}. The MAP estimates for each node N_k are calculated as follows:

$$\nu_0 = 0 \tag{4}$$

$$\eta_0 = I \tag{5}$$

$$\nu_k = \frac{\Gamma_k\tilde{\nu}_k + \tau_k\nu_{k-1}}{\Gamma_k + \tau_k}, \quad k = 1, \ldots, K, \tag{6}$$

$$\eta_k = \frac{\Gamma_k\tilde{\eta}_k + \xi_k\eta_{k-1} + \frac{\tau_k\Gamma_k}{\tau_k+\Gamma_k}\Lambda_k}{\xi + \Gamma_k}, \quad k = 1, \ldots, K, \tag{7}$$

$$\Lambda_k = (\tilde{\nu}_k - \nu_{k-1})(\tilde{\nu}_k - \nu_{k-1})^t \tag{8}$$

$$\Gamma_k = \sum_{t=1}^{T}\sum_{m\in G_k}\gamma_{mt}, \tag{9}$$

where $(\tilde{\nu}_k, \tilde{\eta}_k)$ are the ML estimates for (ν_k, η_k). The parameters $\tau_k > 0$, $\xi_k > 1$ are the control parameters. By successively applying Eqs.(6) and (7) from

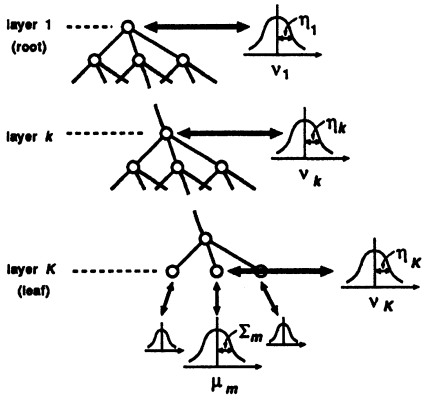

Figure 1: Tree Structure for CDHMMs

* 構造的ベイズ手法による音響モデルの適応化, 篠田浩一 (NEC), チン・フイ・リー (ルーセントテクノロジ)

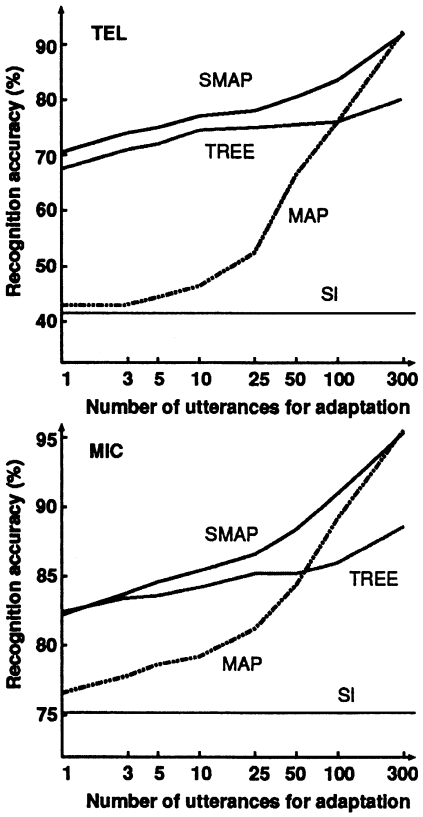

Figure 2: Recognition rates of SMAP adaptation

the root node to the leaf node, the mean ν_K and the variance η_K for the leaf node N_K are obtained. These ν_K and η_K are used to update the parameters of the corresponding mixture component:

$$\hat{\mu}_m = \mu_m + (\Sigma_m)^{1/2}\tilde{\nu}_K, \quad (10)$$

$$\hat{\Sigma}_m = \tilde{\eta}_K \Sigma_m. \quad (11)$$

We call this estimation process SMAP method.

3. Experiments

We experimented with the 991-word DARPA resource management (RM) task[6]. Simultaneous recordings of five non-native speakers were collected through two channels: 1) a close talking microphone (MIC), and 2) a telephone handset over a dial-up line (TEL). The database consist of 300 utterances for adaptation from each speaker for each of the two channels. For testing, we collected 75 utterances from each speaker for each of the two channels. For each frame a 38-dimensional feature vector[7] was extracted based on a tenth order LPC analysis. A diagonal covariance was used for each mixture Gaussian component. The speaker-independent models for adaptation were trained using the NIST/RM SI-109 training set from 76 native American male talkers, each providing 30 or 40 utterances. The tree structure was constructed using Kullback divergence between the pdfs of the mixture components[8]. It had three layers and the root node and each intermediate node had ten branches.

We compared the proposed SMAP method with two other methods to verify its effectiveness: one was the conventional MAP estimation(MAP)[2], and the other was the simple bias estimation using a tree structure without MAP estimation(TREE)[9]. In the former, no structure in the acoustic space was assumed and each parameter of HMMs was estimated separately. In the latter, one node in the tree was selected for each mixture component using a threshold for data amount, and the ML estimates for the parameters at that node were used to modify the parameter of the mixture component. Figure 2 shows the recognition results for the two channels, MIC and TEL, averaged over five speakers. This figure shows that the proposed SMAP method outperforms MAP and TREE in almost every data point. The recognition rates were highly improved from MAP when the amount of data was small, and converged to the same rates as MAP when the amount of data became large. It showed better recognition accuracy than TREE, not only when the amount of data was large but also when the amount of data was small. This is probably because the parameter estimation were more robust than that in TREE since a weighted sum of parameters in more than one layer was used.

4. Conclusions

The SMAP approach for adaptation has been proposed and its effectiveness was confirmed by the recognition experiments using the speech data from non-native speakers collected through different channels. This method has proved to be applicable to compensate the mismatch in general and was effective for any amount of data. In the future, we will further study the way to make tree structures that well represent the inner structure in the acoustic space.

References

[1]C.-H. Lee, C.-H. Lin, and B.-H. Juang, "Study on Speaker Adaptation of Continuous Density HMM parameters," *Proc. ICASSP-90*, pp. 145-148, 1990.

[2]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*, vol. 2, No. 2, pp. 291-298, 1994.

[3]K. Shinoda, and C.-H. Lee, "Structural MAP Speaker Adaptation Using Hierarchical Priors," *Proc. of IEEE Workshop on Speech Recognition and Understanding*, 1997.

[4]K. Shinoda, and C.-H. Lee, "Unsupervised adaptation using structural Bayes approach," *Proc. of ICASSP98*, pp.II793-796, 1998.

[5]J.O. Berger, "Statistical Decision Theory and Bayesian Analysis," in Springer Series in Statistics, Springer-Verlag, 1980.

[6]P. Price, W. Fisher, J. Bernstein, and D. Pallett, "A database for continuous speech recognition in a 1000-word domain," *Proc. of ICASSP-88*, pp. 651-654, 1988.

[7]C.-H. Lee, E. Giachin, L. Rabiner, R. Pieraccini, and A. Rosenberg, "Improved acoustic modeling for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 6, pp. 103-127, 1992.

[8]T. Watanabe, K. Shinoda, K. Takagi, E. Yamada, "Speech Recognition Using Tree-Structured Probability Density Function," *Proc. of ICSLP-94*, pp. 223-226, 1994.

[9]K. Shinoda and T. Watanabe, "Speaker Adaptation with Autonomous Control Using Tree Structure," *Proc. of EuroSpeech-95*, pp. 1143-1146, 1995.