Structural MAP Speaker Adaptation Using Hierarchical Priors

Koichi Shinoda[1] **and Chin-Hui Lee**
Multimedia Communications Research Laboratory
Bell Laboratories, Lucent Technologies
Murray Hill, NJ

Abstract - Most adaptation methods for speech recognition using hidden Markov models fall into two categories; one is the Bayesian approach, where prior distributions for the model parameters are assumed, and the other is the transformation based approach, where a predetermined simple transformation form is employed to modify the model parameters. It is known that the former is better when the amount of data for adaptation is large, while the latter is better when the amount of data is small. In this paper, we propose a new approach, *structural maximum a posteriori* (SMAP) approach, in which hierarchical priors are introduced to combine the two approaches above. The experimental results showed SMAP achieved better recognition accuracy than the two approaches for both small and large amounts of adaptation data.

1 Introduction

Recently, speech recognition using hidden Markov models(HMMs) has been successfully applied to various applications. However, it has been reported that the performance of recognition system is often largely degraded when the testing conditions, including speakers, microphones, channels, and noise levels, are different from those with which training data are collected. Conventionally, these differences have been considered separately, and accordingly, different approaches have been taken to compensate the degradation. However, since it is difficult to distinguish the influence of one factor from the other, one method that can be applied to all the factors is preferable. There have been two major adaptation approaches, the Bayesian approach and the transformation based approach. But neither of these suits this purpose as is explained in the following.

In the Bayesian adaptation approach(e.g. [1, 2]), prior distributions are assumed for the parameters in HMMs and the maximum a posteriori(MAP) estimates for the parameters are calculated instead of the maximum likelihood(ML) estimates. Since this approach requires less amount of data than

[1] This work has been carried out while on leave from NEC Corporation, 4-1-1 Miyazaki, Miyamae-ku, Kawasaki, 216 Japan

0-7803-3698-4/97/$10.00 © 1997 IEEE.

ML estimation when the priors are appropriately chosen, it has been widely used for compensating the difference in speaker characteristics. When the amount of data is extremely small, however, improvement by this adaptation is rather small, because the number of parameters to be estimated is usually large.

On the other hand, transformation based adaptation (e.g. [3, 5, 4]) is mainly used when the amount of data is small. It has been successfully applied to compensate the difference due to microphones, channels, and noise levels. In this approach, a simple transformation, such as a shift, or an affine transformation, is defined in the acoustic feature space or the HMM parameter space and its parameters are estimated using the adaptation data. However, the recognition performance does not improve as much compared with the improvement obtained with MAP adaptation when the amount of data is large. This is partly because the number of free parameters is too small. For this problem, it has been proved effective to divide the acoustic space into a number of subspaces and estimate the transformation parameters in each subspace. In real use, however, it is rather impractical to optimize of the number of subspace for various amounts of data. Shinoda et al.[6, 7] proposed the autonomous control of the number of subspaces according to the amount of data.

From the explanation above, it is clear that neither of the two approaches can be used to deal with all differences in various conditions. Chien et al.[8] reported that the combination of these approaches performed well, while the number of subspaces were still need to be optimized in their method. Here we propose a structural maximum a posteriori (SMAP) approach in which hierarchical priors are employed. In this method, a hierarchical structure in the parameter space is assumed and the transformation parameters for each level in the structure are estimated. The parameters in one level are used as the priors for its immediate subordinate levels. The resulting transformation parameter, corresponding to each HMM parameter, is a combination of the transformation parameters at all levels, in which the weight for each level autonomously changes according to the amount of adaptation data used. Accordingly, this method is more robust against the change in the amount of data than the conventional approaches. Since MAP estimates are calculated and it is well known that the MAP estimate is asymptotically equivalent to the ML estimate, its recognition performance converges to that of speaker-dependent HMMs when the amount of data becomes large.

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. It is assumed that each Gaussian pdf has a diagonal covariance. Let $G = \{N(\mu_m, \sigma_m^2); m = 1, \ldots, M\}$ be the whole set of the mixture components in CDHMMs, where M is the sum of the number of mixture components in all the states of the CDHMMs. In our method, a bias Δ_m for each mean vector μ_m and a scaling factor η_m for each variance σ_m^2 are estimated and the pdf parameters are updated by:

$$\begin{aligned} \hat{\mu}_m &= \mu_m + \Delta_m, \quad m = 1, \ldots, M, & (1) \\ \hat{\sigma}_m^2 &= \eta_m \sigma_m^2 \quad m = 1, \ldots, M. & (2) \end{aligned}$$

The adaptation process is now described in the following.

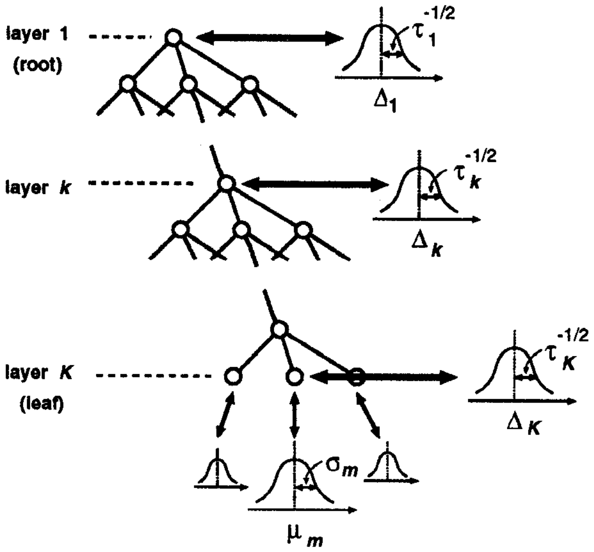

Figure 1: Tree Structure for Gaussian pdfs in CDHMMs

Let a *tree structure* for the set G be given as shown in Fig.1, where K is the number of layers. Each node in K-th layer (leaf node) corresponds to one mixture component of CDHMMs. The root node corresponds the whole set of the Gaussian pdfs, G. Each intermediate node corresponds to a subset of G, each of whose elements corresponds to one of its subordinate leaf nodes. There are many ways to construct a tree structure(e.g. [6, 7]).

At each node in the tree, a prior for a bias Δ and a prior for a factor η, which are shared among the Gaussian pdfs in the corresponding subset, are assigned. It is assumed that the prior pdf for each bias is a Gaussian pdf in which a diagonal covariance is used and the prior pdf for each scaling factor is a beta distribution.

At first, the ML estimates of the bias and the scaling factor for each node are calculated using the adaptation data. Let $Y = \{y_1, \ldots, y_T\}$ be a set of the data for adaptation and let Δ_k be the bias at node k, where the corresponding subset of G has M^k pdfs, $N(\mu_1, \sigma_1^2), \ldots, N(\mu_{M^k}, \sigma_{M^k}^2)$. Then, the ML estimate for the bias, $\tilde{\Delta}_k$, and the scaling factor, $\tilde{\eta}_k$, are estimated using the EM-algorithm [5](in the following, the suffix for the data vector dimension is omitted):

$$\tilde{\Delta}_k = \frac{1}{\Gamma_k}\sum_{t=1}^{T}\sum_{m^k=1}^{M^k}\gamma_t(m^k)\frac{y_t-\mu_{m^k}}{\eta_k\sigma^2_{m^k}}, \tag{3}$$

$$\tilde{\eta}_k^{-1} = \frac{1}{\Lambda_k}\sum_{t=1}^{T}\sum_{m^k=1}^{M^k}\gamma_t(m^k)\frac{(y_t-\mu_{m^k}-\Delta_k)^2}{\sigma^2_{m^k}}, \tag{4}$$

$$\Gamma_k = \sum_{t=1}^{T}\sum_{m^k=1}^{M^k}\frac{\gamma_t(m^k)}{\eta_k\sigma^2_{m^k}}, \quad \Lambda_k = \sum_{t=1}^{T}\sum_{m^k=1}^{M^k}\gamma_t(m^k), \tag{5}$$

where, $\gamma_t(m^k)$ is the posteriori probability of being in the state which contains the mixture component m^k at time t and using the mixture component m^k.

Next, the MAP estimates[1, 2] for these parameters are calculated using a hierarchical Bayes analysis to be explained as follows. For the estimation at each node, the pdf for the bias and the scaling factor at its parent node are used as the prior distribution. Let $N_1, \ldots, N_k, \ldots, N_K$ be a sequence of nodes from the root node to the leaf node, where N_1 is the root node and N_K is the leaf node. Each node N_{k-1} is the parent node for node N_k. Then, the MAP estimates at each node are calculated as follows:

$$\Delta_1 = \frac{\Gamma_1\tilde{\Delta}_1}{\Gamma_1+\tau_1}, \tag{6}$$

$$\Delta_k = \frac{\Gamma_k\tilde{\Delta}_k+\tau_k\Delta_{k-1}}{\Gamma_k+\tau_k}, k = 2, \ldots, K, \tag{7}$$

$$\eta_1^{-1} = \frac{\Lambda_1\tilde{\eta}_1^{-1}}{\Lambda_1+\xi_1}, \tag{8}$$

$$\eta_k^{-1} = \frac{\Lambda_k\tilde{\eta}_k^{-1}+\xi_k\eta_{k-1}^{-1}}{\Lambda_k+\xi_k}, k = 2, \ldots, K, \tag{9}$$

where τ_k is the precision of the prior distribution for Δ_k, and ξ_k is the hyper-parameter for η_k. It must be noted that Δ_k and η_k are dependent on each other, and thus should be calculated iteratively.

Finally, by successively applying Eq.(8) from the root node to the leaf nodes, the bias Δ_K and the scaling factor η_K for the leaf node N_K can be expressed as follows:

$$\Delta_K = \sum_{j=1}^{K} w_j^K\tilde{\Delta}_j, \quad \eta_K^{-1} = \sum_{j=1}^{K} v_j^K\tilde{\eta}_j^{-1}, \tag{10}$$

where,

$$w_j^K = \frac{\Gamma_j}{\Gamma_j+\tau_j}\prod_{i=j+1}^{K}\frac{\tau_i}{\Gamma_i+\tau_i}, \tag{11}$$

$$v_j^K = \frac{\Lambda_j}{\Lambda_j + \xi_j} \prod_{i=j+1}^{K} \frac{\xi_i}{\Lambda_i + \xi_i}. \quad (12)$$

These Δ_K and η_K is used to update the corresponding Gaussian pdf in the CDHMMs.

We call this estimation process as the SMAP method. Eqs.(11) and (12) indicate that the parameters estimated by SMAP can be interpreted as the weighted sum of the ML estimates at the different layers of the tree. The weight has the following characteristics:

1. In node N_j, as data amount becomes larger, Γ_j becomes larger, and thus, w_j^K becomes larger.

2. The weight w_j^K for an ancestor node N_j decays exponentially as j becomes smaller, i.e., the node approaches to the root node.

These are preferable characteristics for adaptation. When the amount of data is small, the ML estimates in the upper layers are mainly responsible for the resulting pdf. On the other hand, when the amount of data is large, the ML estimates in the lower layers are dominant. This control is done autonomously.

The prior knowledge about the embedded structure in the acoustic space should be used for the construction of the tree. In this study, the Kullback divergence between the output pdfs of the mixture components is used as the distance measure between the mixture components. The k-means clustering algorithm was used for clustering the Gaussian pdfs[9].

Although this SMAP approach is not the first to propose tree-based adaptation (e.g.[10]), we believe the proposed method is theoretically well-defined in terms of both the Bayesian framework and the tree construction principle. It demonstrated these two properties well as will be clear in the experimental result section.

3 Experiments

3.1 Experimental Conditions

We experimented with the 991-word DARPA resource management (RM) task[11]. 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 data consisted of 300 utterance for adaptation from each speaker (A,B,C,D,E) in each of the two channels (MIC and TEL). For testing, we collected 75 utterances from each speaker for each of the two channels.

The speech was first down-sampled from 16 kHz to 8 kHz. For each frame a 39-dimensional feature vector[12] was extracted based on a tenth order LPC analysis, whose components are 12 cepstral coefficients plus a normalized log energy and their first and second time derivatives. For recognition, we used 1769 context dependent units[12]. For all our experiments, we used the RM word pair grammar, which gives a perplexity of about 60. Speaker-independent models were trained using the NIST/RM SI-109 training set consisting of 3990 utterances from 109 native American talkers (31 females and 78 males), each providing 30 or 40 utterances. A diagonal covariance was used for each mixture Gaussian component.

In the experiments, only the mean vector, μ, was modified and the parameter τ in Eq.(8) were fixed. The scaling factor η was fixed to one.

3.2 Results

Fig 2 shows the recognition results for the two channels, MIC and TEL. For comparison, we experimented two other methods; one is MAP adaptation without tree (MAP)[2], and the other is simple bias estimation using a tree without Bayesian method (TREE)[6]. These figures show that the proposed SMAP method performed better than MAP and TREE in every data point. The recognition rates were highly improved from MAP when the amount of data were 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 were large but also when the amount of data were 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 layers were used,

4 Conclusion

The SMAP approach for adaptation has proposed. Its effectiveness was confirmed by the recognition experiments.

Several research issues remain to be investigated. First, adaptation for variances and other HMM parameters should be examined. Second, the way to make a tree structure that well represent the embedded structure in the acoustic space should be further studied. Third, unsupervised adaptation using this approach should be evaluated.

References

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

[2] Gauvain,J.-L., and Lee, C.-H., "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] Leggetter,C.J. and Woodland, P.C., “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov Models”, *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

[4] Digalakis,V.V. and Neumeyer,L.G., “ Speaker Adaptation Using Combined Transformation and Bayesian Methods,” *IEEE Trans. on Speech and Audio Processing*, vol. 4, No. 4, pp. 294-300, 1996.

[5] Sankar, A. and Lee,C-.H., “Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition”, *IEEE Trans. on Speech and Audio Processing*, vol. 4, No. 3, pp.190-202, 1996.

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

[7] Shinoda, K. and Watanabe,T., “Speaker Adaptation with Autonomous Model Complexity Control by MDL Principle”, *Proc. ICASSP-96*, pp.717-720, 1996.

[8] Chien,J.-T.,Lee,C.-H., and Wang H.-C., "Improved Baysian Learning of Hidden Markov Models for Speaker Adaptation” *ICASSP-97*, pp. 1027-1039, 1997.

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

[10] Paul, D.-B., “Extensions to Phone-State Decision-Tree Clustering: Single Tree and Tagged Clustering” *ICASSP-97*, pp. 1487-1490, 1997.

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

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

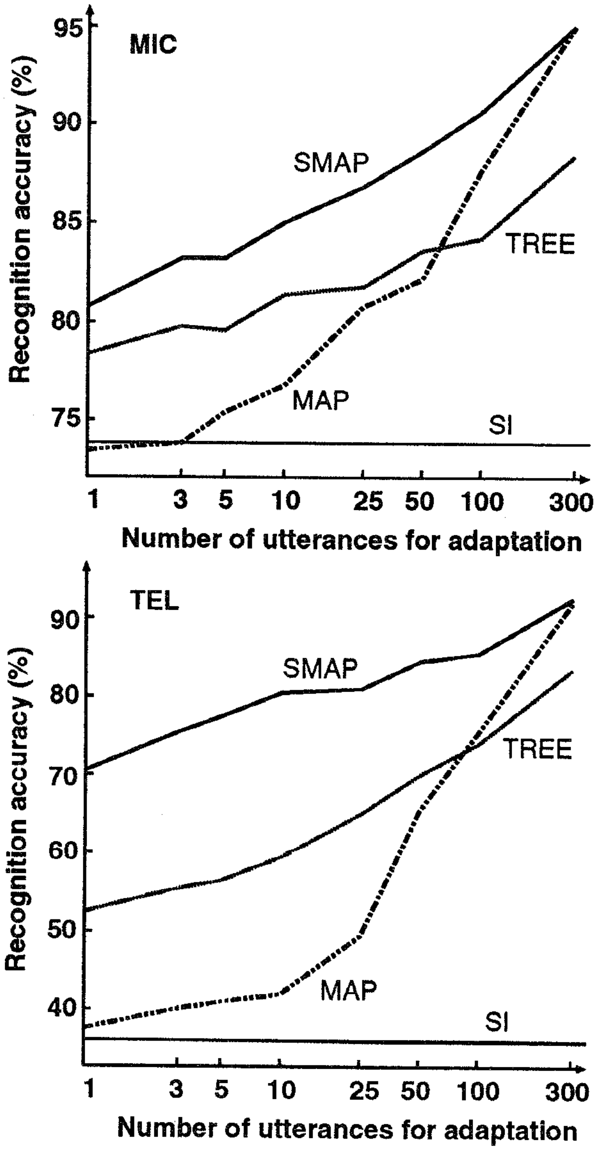

Figure 2: Recognition rates for various amounts of adaptation data