UNSUPERVISED ADAPTATION USING STRUCTURAL BAYES APPROACH

Koichi Shinoda and Chin-Hui Lee

Bell Laboratories, Lucent Technologies
600 Mountain Ave. Murray Hill, NJ07974-0636
shinoda@hum.cl.nec.co.jp, chl@research.bell-labs.com

ABSTRACT

It is well-known that the performance of recognition systems is often largely degraded when there is a mismatch between the training and testing environment. It is desirable to compensate for the mismatch when the system is in operation without any supervised learning. Recently, a structural maximum a posteriori (SMAP) adaptation approach, in which a hierarchical structure in the parameter space is assumed, was proposed. In this paper, this SMAP method is applied to unsupervised adaptation. A novel normalization technique is also introduced as a front end for the adaptation process. The recognition results showed that the proposed method was effective even when only one utterance from a new speaker was used for adaptation. Furthermore, an effective way to combine the supervised adaptation and the unsupervised adaptation was investigated to reduce the need for a large amount of supervised learning data.

1. INTRODUCTION

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 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 adopted to compensate the degradation.

Bayesian adaptation approach(e.g. [1, 2]) has been successfully applied to speaker adaptation. In this approach, prior distributions are assumed for the parameters in HMMs and the *maximum a posteriori* (MAP) estimates for the parameters are calculated instead of the conventional maximum likelihood (ML) estimates. Since this approach requires less amount of data than 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, the improvement by this adaptation is rather small because the number of parameters to be estimated is usually large. For the other mismatches, the transformation based approach (e.g. [3, 4, 5]) has been extensively studied. It has been used 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 when the amount of data is large. This is partly because the number of free parameters is too small.

Recently, Shinoda and Lee proposed a *structural maximum a posteriori* (SMAP)[6] approach, in which hierarchical priors are introduced to combine these two approaches. In SMAP, 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 (child) 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, SMAP is expected to be more robust against the change in the amount of data than the conventional approaches. Since the MAP estimates are calculated and it is well known that the MAP estimate is asymptotically equivalent to the ML estimate[2], its recognition performance converges to that of speaker-dependent HMMs when the amount of data becomes large. It was reported that SMAP achieved better recognition rates than the conventional methods for both small and large amounts of adaptation data.

In real use, it is desirable to adopt unsupervised adaptation, in which no supervising information is required. In this paper, we focus on this unsupervised adaptation scenario. We propose a normalization technique which compensates for the mismatch and can be used as a front-end for SMAP approach. The experimental results showed that the proposed method was effective even when only one utterance was used for adaptation. Furthermore, we investigate effective ways to combine fast supervised adaptation and on-line unsupervised adaptation to achieve a sufficient recognition accuracy for real use.

2. SMAP ADAPTATION

In this paper, we focus on the adaptation of the parameters of Gaussian pdfs in continuous-density(CD) HMMs. Let

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

$\mathbf{x} = (x_1, \ldots, x_T)$ denote a given set of T observation vectors for adaptation (adaptation data). Let g_m be a normal density function for mixture component m, $N(\mathbf{x}|\mu_m, \Sigma_m)$, 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 in the CDHMMs.

2.1. Mismatch PDF

At the first step in our method, each sample vector x_t is normalized into a vector y_{mt} for each mixture component m as follows:

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

Obviously, the pdf for $\mathbf{y}_m = y_{m1}, \ldots, y_{mT}$ is the standard normal distribution $N(\mathbf{y}|0, I)$ when there is no mismatch between the training data and the adaptation data. However, when there is a mismatch between them, the pdf for $\mathbf{y}$ is different from $N(\mathbf{y}|0, I)$ for the adaptation data, and assumed to be $N(\mathbf{y}|\nu, \eta)$, where $\nu \neq 0$ and $\eta \neq I$. It can be said this pdf for $\mathbf{y}$ better represents the difference of the acoustic characteristics between the training data and the adaptation data, rather than the characteristics of the adaptation data. Therefore, we call this pdf *mismatch pdf*. We assume that the mismatch can be modeled by simpler models than that for speech recognition. In other words, we assume that the number of the mismatch pdfs required to model the acoustic difference is smaller than that of the mixture components of HMMs. In our method, to have a smaller number of the pdfs, the whole set of mixture components, G, is divided into several subsets $G_1, \ldots, G_P$, where P is the number of subsets, and one common mismatch pdf $h_p = N(\mathbf{y}|\nu_p, \eta_p)$ is assigned to all the mixture components in subset G_p.

The parameters for the mismatch pdfs can be estimated using the EM algorithm. It is assumed that the transition probabilities and the weight coefficients are neglected. The auxiliary function Q for the HMM parameters is given by:

$$Q(\hat{\theta}, \theta) = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{mt}(\theta) \log N(x_t|\hat{\mu}_m, \hat{\Sigma}_m), \tag{2}$$

where $\theta = \{\mu_m, \Sigma_m; m = 1, \ldots, M\}$ is the current HMM parameter set and $\hat{\theta}$ is the new HMM parameter set to be estimated. And γ_{mt} is the posterior probability of using mixture component m at t. The relation between the original pdf and the mismatch pdf is as follows when the mixture component m belongs to subset G_p:

$$\begin{aligned} N(x_t|\hat{\mu}_m, \hat{\Sigma}_m) &= \frac{N(y_{pt}|\nu_p, \eta_p)}{|J_m|}, \\ &= \frac{N(y_{pt}|\nu_p, \eta_p)}{|\Sigma_m^{1/2}|}, \end{aligned} \tag{3}$$

where $J_m = \Sigma_m^{1/2}$ is the Jacobian matrix for the normalization in Eq.(1). Using this relation, the auxiliary function is modified as follows:

$$Q(\hat{\theta}, \theta) = \sum_{t=1}^{T} \sum_{p=1}^{P} \sum_{m_p=1}^{M_p} \gamma_{m_p t}(\theta) \log \frac{N(y_{m_p t}|\nu_p, \eta_p)}{\Sigma_m^{1/2}}, \tag{4}$$

where m_p denotes each mixture component in subset G_p and M_p is the number of mixture components in subset G_p. By differentiating this equations, the ML estimates of the parameters are estimated as follows:

$$\hat{\nu}_p = \frac{\sum_{t=1}^{T} \sum_{m_p=1}^{M_p} \gamma_{m_p t} y_{m_p t}}{\sum_{t=1}^{T} \sum_{m_p=1}^{M_p} \gamma_{m_p t}}, \tag{5}$$

$$\hat{\eta}_p = \frac{\sum_{t=1}^{T} \sum_{m_p=1}^{M_p} \gamma_{m_p t} (y_{m_p t} - \hat{\nu}_p)(y_{m_p t} - \hat{\nu}_p)^t}{\sum_{t=1}^{T} \sum_{m_p=1}^{M_p} \gamma_{m_p t}}, \tag{6}$$

where $(y_{m_p t} - \hat{\nu}_p)^t$ is a transition of $(y_{m_p t} - \hat{\nu}_p)$. Using the mismatch pdf parameters, new HMM parameters, $\hat{\mu}_m$ and $\hat{\Sigma}_m$, are calculated as follows,

$$\hat{\mu}_m = \mu_m + \Sigma_m^{1/2} \hat{\nu}_p, \tag{7}$$

$$\hat{\Sigma}_m = \hat{\eta}_p \Sigma_m. \tag{8}$$

Let us compare this method with the Stochastic Matching (SM) [4]. As can be seen from Eqs.(7) and (8), $\Sigma_m^{1/2}\nu_p$ corresponds to the bias in SM, and η_p corresponds to the scaling factor in SM when the diagonal covariance is used for η_p. In our method, the bias for each mixture component changes according to the variance; when the variance is large, the bias is also large. Besides, both parameters ν_p and η_p can be simultaneously calculated in our method, while an iterative process is required in SM.

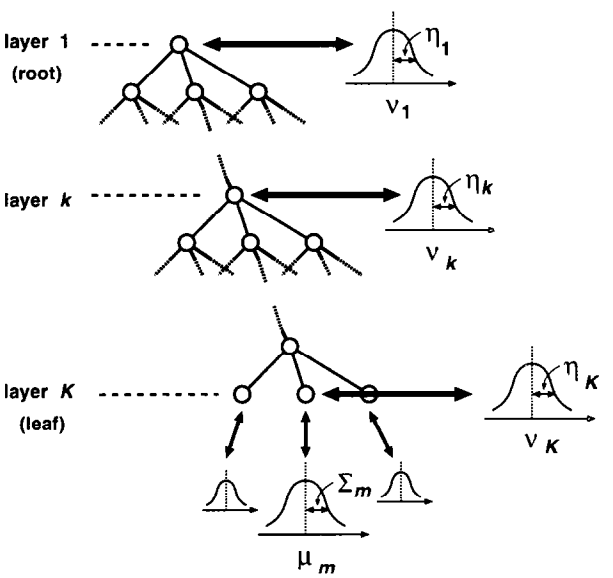

Figure 1: Tree Structure for Gaussian pdfs in CDHMMs

2.2. SMAP Adaptation Using Hierarchical Priors

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 the K-th layer (leaf node) corresponds to one mixture component of CDHMMs. The root node corresponds 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 in the tree, a mismatch pdf for $\mathbf{y}$, which is shared among the mixture components in the corresponding subset of G, is assigned. Let $N(\nu_k, \eta_k)$ be the pdf for node k, whose corresponding subset is denoted as $\{m_k; m_k = 1, \ldots, M_k\}$. First, the ML estimates of the pdf parameters, $\tilde{\nu}_k$ and $\tilde{\eta}_k$, is calculated using the adaptation data by using Eqs.(5) and (6).

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

$$\nu_k = \frac{\Gamma_k \tilde{\nu}_k + \tau_k \nu_{k-1}}{\Gamma_k + \tau_k}, \quad k = 1, \ldots, K, \tag{9}$$

$$\eta_k = \frac{\xi \eta_{k-1} + \Gamma_k \tilde{\eta}_k + \frac{\tau_k \Gamma_k}{\tau_k + \Gamma_k}(\tilde{\nu}_k - \nu_{k-1})(\tilde{\nu}_k - \nu_{k-1})^t}{\xi + \Gamma_k}, \quad k = 1, \ldots, K, \tag{10}$$

$$\Gamma_k = \sum_{t=1}^{T} \sum_{m_k=1}^{M_k} \gamma_{m_k t}, \quad k = 1, \ldots, K, \tag{11}$$

where $\tau > 0$, $\xi > 1$. The prior distribution for the root node is assumed to be the standard normal pdf, i.e., $\nu_0 = 0$, $\eta_0 = I$. By successively applying Eqs.(9) and (10) from 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 corresponding mixture components as in Eqs.(7) and (8). We call this estimation process as SMAP method.

Eq.(9) can be rewritten for the leaf node as follows:

$$\nu_K = \sum_{j=1}^{K} w_j^K \tilde{\nu}_j, \tag{12}$$

where,

$$w_j^K = \frac{\Gamma_j}{\Gamma_j + \tau_j} \prod_{i=j+1}^{K} \frac{\tau_i}{\Gamma_i + \tau_i}. \tag{13}$$

The mean 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. At 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.

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 structure for the set of mixture components, G. In this study, the Kullback divergence between the output pdfs of the mixture components is used as the distance measure between the mixture components[7]. The k-means clustering algorithm is used for clustering the Gaussian pdfs.

Although this SMAP approach is not the first to propose tree-based adaptation (e.g.[8]), we believe the proposed method is theoretically well-defined in terms of both the Bayesian framework and the tree construction principle. It demonstrates these two properties well as will be clear in the experimental result section.

3. EXPERIMENTS

We experimented with the 991-word DARPA resource management (RM) task[9]. Simultaneous recordings of five nonnative speakers (A,B,C,D,E) 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 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 utterances were first down-sampled from 16 kHz to 8 kHz. For each frame a 38-dimensional feature vector[10] was extracted based on a tenth order LPC analysis, whose components are 12 cepstral coefficients and their first and second time derivatives and the first and second time derivatives of a normalized log energy. For recognition, we used 1769 context dependent units[10]. 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. These models were then adapted using a MAP adaptation algorithm[2], with the data from the 78 male talkers, to create speaker-independent male models. These male models are used as initial models for adaptation. A diagonal covariance was used for each mixture Gaussian component. The tree structure was constructed using the speaker-independent male models. It was a binary tree with five layers. In the experiments, only the mean vector, μ, was modified and the parameter τ in Eq.(9) were fixed. The covariance matrix η was assumed to be fixed to I.

First, an on-line unsupervised adaptation method was evaluated. During unsupervised adaptation and testing, the parameters were estimated on a per-utterance basis; we first decode the word string using the initial HMMs and then estimate the parameters condition on this word string. Tables 1 and 2 shows the recognition results. By using only one utterances, the error reduction rates were 23 % for MIC and 27 % for TEL, respectively. It should be noted that the effect of the proposed method is larger for the speakers with

Table 1: Recognition rates (%) of unsupervised adaptation for MIC data

	A	B	C	D	E	Ave.
SI	74.8	51.2	74.9	85.7	89.3	75.2
SMAP	81.2	67.3	78.7	88.4	89.0	80.9

Table 2: Recognition rates (%) of unsupervised adaptation for TEL data

	A	B	C	D	E	Ave.
SI	48.8	18.6	50.8	35.6	52.6	41.3
SMAP	67.3	48.5	57.7	43.1	70.1	57.3

lower recognition rates. For example, the error reduction rates for speaker B were 33 % for MIC and 37 % for TEL.

For real use, the recognition rate in Table 1 seems still rather low. In the next experiment, we examined effective ways to combine fast supervised adaptation and on-line unsupervised adaptation to achieve a sufficient recognition accuracy for real use. This combined adaptation process was described as follows:

Step 1. Supervised adaptation using the adaptation data

Step 2. Unsupervised adaptation using the test data, in which the models obtained in Step.1 are used as the initial models

In this experiment, the supervised adaptation was carried out using only MIC adaptation data, while the unsupervised adaptation was done for both MIC and TEL test data. The recognition rates averaged over the five speakers are shown in Table 3, in which the number of utterances used in Step 1 is changed. In this table, SUP is the recognition rates only with Step 1 and S+U is the rates obtained after Step 2. Although this adaptation is only slightly effective for MIC data, its effectiveness for TEL data is clear. For example, to achieve 60% recognition accuracy, the combined method required only three utterances, while the supervised adaptation needed 100 utterances. It can be said this combined adaptation is especially effective when there exist other mismatches than the speaker differences. It should be also noted that there was no degradation in the recognition performance when MIC data were used for testing, i.e., when there was no mismatch between the adaptation data and the testing data.

4. CONCLUSION

We have presented a novel unsupervised adaptation method using the SMAP approach. Its effectiveness was confirmed by the recognition experiments. Several problems remain to be investigated. First, adaptation for the variances should be examined. Second, the way to make a tree structure that well represents the embedded structure in the acoustic space should be further studied. Third, the effect of the unsupervised adaptation for different kinds of mismatches should be evaluated.

Table 3: Recognition rates (%) of the combination of unsupervised adaptation and supervised adaptation

	MIC		TEL	
No. of Utter.	SUP	S+U	SUP	S+U
SI	75.2		41.3	
1	83.0	83.5	41.4	57.8
3	83.8	83.9	44.3	62.2
5	83.9	84.2	47.0	63.0
10	85.7	86.1	51.7	67.2
25	86.3	86.4	54.0	69.4
50	87.7	87.8	58.1	73.3
100	90.4	90.4	62.1	75.5
300	94.3	94.4	70.8	84.6

5. REFERENCES

[1] Lee,C.-H.,Lin,C.-H., Juang,B.-H., "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] 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.

[5] 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.

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

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

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

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

[10] 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.