ON-LINE HIERARCHICAL TRANSFORMATION OF HIDDEN MARKOV MODELS FOR SPEAKER ADAPTATION

Jen-Tzung Chien

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
jtchien@mail.ncku.edu.tw

ABSTRACT

This paper presents a novel framework of on-line hierarchical transformation of hidden Markov models (HMM's) for speaker adaptation. Our aim is to incrementally transform (or adapt) all the HMM parameters to a new speaker even though part of HMM units are unseen in adaptation data. The transformation paradigm is formulated according to the approximate Bayesian estimate, which the prior statistics and the transformation parameters are incrementally updated for each consecutive adaptation data. Using this formulation, the updated prior statistics and the current block of data are sufficient for on-line transformation. Further, we establish a hierarchical tree of HMM's and use it to dynamically control the transformation sharing for each HMM unit. In the speaker adaptation experiments, we demonstrate the superiority of proposed on-line transformation to other method.

1. INTRODUCTION

It is no doubt that speaker adaptation technique is a practical approach to improve the speaker-independent (SI) speech recognition system for an enrolled speaker by using some adaptation data. Generally, the adaptation techniques can be employed in three strategies; (1) batch adaptation, (2) self adaptation, and (3) on-line adaptation. Batch adaptation is an off-line adaptation where the models are adapted by using batch data. Self adaptation executes the adaptation on testing data itself at runtime and in an unsupervised manner. It is able to trace the changing variabilities during recognition. However, owing to the insufficient observations and unreliable transcription, the resulting performance is constrained. Besides, on-line adaptation is a tradeoff strategy between batch adaptation and self adaptation. It is aiming at performing adaptation incrementally only when a block of data is observed. This block of data is then thrown away after completing the adaptation. Consequently, the merit of on-line adaptation is to continuously update the speech models without waiting long history of batch data. Its flexible characteristics have been attracted many studies focusing on this issue [3][5-6].

In the literature, there are two categories of adaptation algorithms. One is the transformation-based adaptation, where clusters of HMM's are individually transformed by using some transformation parameters [7]. The other is the maximum *a posteriori* (MAP) adaptation of HMM parameters. By serving the SI HMM's as prior statistics, the HMM parameters are adapted accordingly based on the MAP estimate [4]. In case of limited adaptation data, the transformation-based adaptation can efficiently transform all the HMM parameters by using cluster-dependent transformation functions. Conversely, in case of sufficient adaptation data, the MAP adaptation can effectively merge the adaptation tokens into the SI HMM parameters. By jointly performing MAP transformation and adaptation, we can obtain better performance than separate methods for a wide range of adaptation data [1]. On the other hand, the construction of tree structure of HMM's in transformation-based adaptation can dynamically capture the goodness of transformation parameters and also benefit the adaptation performance for various amounts of adaptation data [8-9].

As explained above, we are motivated to propose the *on-line hierarchical transformation* of HMM parameters for speaker adaptation. The proposed method is based on the *approximate Bayesian* (or quasi-Bayes, QB) estimate described by Huo and Lee [6]. Using QB, the unknown parameters are estimated by maximizing the approximate posterior pdf, which is a product of likelihood function of current block data and a prior density given the updated parameter statistics (or hyperparameters). The hyperparameters are obtained from previous observed data. By specifying the prior density as conjugate prior family, we may generate a *reproducible prior/posterior* pair and then formulate a recursive MAP estimate for on-line adaptation. In [6], the QB learning of continuous-density HMM (CDHMM) parameters was derived for on-line speaker adaptation. Their algorithm relied on the speaker providing at least one example of each vocabulary in adaptation data. Such method may not be feasible to the adaptation with increasing vocabulary size and limited adaptation data. In this paper, we present a transformation-based on-line adaptation approach, where the overall HMM parameters are incrementally transformed. We build a hierarchical tree of HMM parameters such that each HMM unit can search its most likely transformation parameters from leaf node to root node. For each HMM unit, we extract the node containing adaptation tokens and use its parameters for on-line transformation. Experiments demonstrate that proposed method performs well for various numbers of adaptation data and lengths of adaptation interval.

2. ON-LINE TRANSFORMATION

In the continuous-density HMM framework, we are given a set of parameters $\lambda = \{\omega_{ik}, \mu_{ik}, r_{ik}\}$, where ω_{ik}, μ_{ik} and r_{ik} are

This work has been partially supported by National Science Council, Taiwan, R.O.C.

the mixture gain, mean vector, and precision (or inverse covariance) matrix, i.e. $r_{ik} = \Sigma_{ik}^{-1}$, of the *k*th mixture component from the *i*th state. Let the HMM parameters λ be grouped into *C* clusters. Our goal is to transform the clusters of HMM parameters to a new environment through some transformation functions $G_\eta(\cdot)$, $\eta = \{\eta_c\}$, $c = 1,\cdots,C$. Let $\chi^n = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ be *n* i.i.d. and successively observed adaptation samples/blocks, which are used to estimate the transformation parameters η. The *a posteriori* density of η satisfies the following *recursive* relation [10]

$$p(\eta\big|\chi^n) = \frac{p(\mathbf{X}_n\big|\eta)\cdot p(\eta\big|\chi^{n-1})}{\int p(\mathbf{X}_n\big|\eta)\cdot p(\eta\big|\chi^{n-1})d\eta}\,. \tag{1}$$

To overcome the computational difficulties in (1), an QB estimate of $\eta^{(n)}$ after observing the current sample $\mathbf{X}_n$ is approximated by [6]

$$\begin{aligned}\eta^{(n)} &= \arg\max_\eta\, p(\eta\big|\chi^n) = \arg\max_\eta\, p(\mathbf{X}_n\big|\eta)\cdot p(\eta\big|\chi^{n-1}) \\ &\cong \arg\max_\eta\, p(\mathbf{X}_n\big|\eta)\cdot g(\eta\big|\varphi^{(n-1)})\,,\end{aligned} \tag{2}$$

where $g(\eta \mid \varphi^{(n-1)})$ is the closest tractable prior density for posterior density $p(\eta \mid \chi^{n-1})$ and $\varphi^{(n-1)}$ is the updated hyperparameters after observing previous blocks χ^{n-1}. Using QB estimate with an initial hyperparameters $\varphi^{(0)}$, we can estimate the transformation parameters $\eta^{(1)}$ by applying $\mathbf{X}_1$ in (2). Then, the hyperparameters $\varphi^{(1)}$ are updated and stored for the estimation of next parameters $\eta^{(2)}$. Accordingly, a recursive formulation for parameter sequence $\eta_1, \eta_2, \cdots, \eta_n$ is established. Because QB estimate in (2) is an incomplete data problem, we use the EM algorithm to iteratively improve the approximate posterior likelihood of current estimate $\eta^{(n)}$ and derive the new estimate $\hat{\eta}^{(n)}$ in an optimal manner [2]. Applying the EM algorithm, we perform the following two steps.

E-step: Calculate the auxiliary function

$$\begin{aligned}R(\hat{\eta}^{(n)} \mid \eta^{(n)}) = E\{\log p(\mathbf{X}_n, \mathbf{s}_n, \mathbf{l}_n \mid \hat{\eta}^{(n)}) + \\ \log g(\hat{\eta}^{(n)} \mid \varphi^{(n-1)}) \mid \mathbf{X}_n, \eta^{(n)}\}\end{aligned}\,, \tag{3}$$

where $\mathbf{s}_n = \{s_t^{(n)}\}$ is the state sequence, $\mathbf{l}_n = \{l_t^{(n)}\}$ is the mixture component sequence, and $(\mathbf{X}_n, \mathbf{s}_n, \mathbf{l}_n)$ is our choice of complete data.

M-step: Find the new estimate

$$\hat{\eta}^{(n)} = \arg\max_{\hat{\eta}^{(n)}} R(\hat{\eta}^{(n)} \mid \eta^{(n)})\,. \tag{4}$$

The iterative EM steps guarantee that the approximate posterior density never decreases.

3. TRANSFORMATION FORMULATION

Before the derivation of on-line transformation (also referred as OLT), the definitions of transformation function and prior density should be addressed. In this study, the HMM parameters are transformed by

$$\hat{\lambda} = G_{\eta^{(n)}}(\lambda) = \{\omega_{ik}, \mu_{ik} + \mu_c^{(n)}, \theta_c^{(n)} r_{ik}\}\,, \tag{5}$$

where $\mu_c^{(n)}$ is a bias vector and $\theta_c^{(n)}$ is a scaling matrix. Herein, the HMM unit with indices *i* and *k* is attributed to the *c*th cluster membership Ω_c. On the other hand, we constraint the prior density in *conjugate family* due to mathematical attractiveness. The joint prior density of transformation parameters $\eta_c^{(n)} = (\mu_c^{(n)}, \theta_c^{(n)})$ of membership Ω_c is defined as a normal-Wishart density of the form [4]

$$\begin{aligned}g(\mu_c^{(n)}, \theta_c^{(n)}) = g(\eta_c^{(n)} \mid \varphi_c^{(n-1)}) \propto \left|\theta_c^{(n)}\right|^{(\alpha_c^{(n-1)}-d)/2} \\ \times \exp\left[-\frac{1}{2}(\mu_c^{(n)} - m_c^{(n-1)})^t \theta_c^{(n)} \tau_c^{(n-1)} (\mu_c^{(n)} - m_c^{(n-1)})\right] \\ \times \exp\left[-\frac{1}{2}\mathrm{tr}(u_c^{(n-1)}\theta_c^{(n)})\right],\end{aligned} \tag{6}$$

where $\varphi_c^{(n-1)} = (\tau_c^{(n-1)}, m_c^{(n-1)}, \alpha_c^{(n-1)}, u_c^{(n-1)})$ are the hyperparameters of prior density determined from previous successive data. Under this definition, the posterior density of complete data (i.e. $\mathrm{K}\cdot\exp\{R(\hat{\eta}_c^{(n)} \mid \eta_c^{(n)})\}$), can be also expressed in a form of normal-Wishart density $g(\hat{\eta}_c^{(n)} \mid \hat{\varphi}_c)$ with the new hyperparameters $\hat{\varphi}_c = (\hat{\tau}_c, \hat{m}_c, \hat{\alpha}_c, \hat{u}_c)$ given as follows:

$$\hat{\tau}_c = \tau_c^{(n-1)} + \sum_{i,k\in\Omega_c} c_{ik} r_{ik}\,, \tag{7}$$

$$\hat{m}_c = \left(\tau_c^{(n-1)} + \sum_{i,k\in\Omega_c} c_{ik} r_{ik}\right)^{-1} \cdot \left(\tau_c^{(n-1)} m_c^{(n-1)} + \sum_{i,k\in\Omega_c} c_{ik} r_{ik} \overline{\mathbf{b}}_c\right), \tag{8}$$

$$\hat{\alpha}_c = \alpha_c^{(n-1)} + \sum_{i,k\in\Omega_c} c_{ik}\,, \tag{9}$$

$$\begin{aligned}\hat{u}_c = u_c^{(n-1)} + \sum_{i,k\in\Omega_c} S_{ik} r_{ik} + \left(\tau_c^{(n-1)} + \sum_{i,k\in\Omega_c} c_{ik} r_{ik}\right)^{-1} \cdot \\ \left(\tau_c^{(n-1)} \sum_{i,k\in\Omega_c} c_{ik} r_{ik}\right) \cdot (\overline{\mathbf{b}}_c - m_c^{(n-1)})(\overline{\mathbf{b}}_c - m_c^{(n-1)})^t\end{aligned}\,, \tag{10}$$

where $\xi_t(i,k) = \Pr(s_t^{(n)} = i, l_t^{(n)} = k \big| \mathbf{X}_n, \eta_c^{(n)})$ is the posterior probability of being in state *i* and mixture component *k* given that the current parameters $\eta_c^{(n)}$ generate $\mathbf{X}_n = \{\mathbf{x}_t^{(n)}\}$ and

$$c_{ik} = \sum_t \xi_t(i,k)\,, \tag{11}$$

$$\overline{\mathbf{b}}_c = \sum_t \sum_{i,k\in\Omega_c} \xi_t(i,k)(\mathbf{x}_t^{(n)} - \mu_{ik}) \Big/ \sum_t \sum_{i,k\in\Omega_c} \xi_t(i,k)\,, \tag{12}$$

$$S_{ik} = \sum_t \xi_t(i,k)(\mathbf{x}_t^{(n)} - \mu_{ik} - \overline{\mathbf{b}}_c)(\mathbf{x}_t^{(n)} - \mu_{ik} - \overline{\mathbf{b}}_c)^t\,. \tag{13}$$

The E-step of EM algorithm is therefore completed. In the M-step, we maximize $g(\hat{\eta}_c^{(n)} \mid \hat{\varphi}_c)$ with respect to $\hat{\eta}_c^{(n)}$ and derive the new estimate of transformation parameters

$\hat{\eta}_c^{(n)} = (\hat{\mu}_c^{(n)}, \hat{\theta}_c^{(n)})$ shown below

$$\hat{\mu}_c^{(n)} = \hat{m}_c , \tag{14}$$

$$\hat{\theta}_c^{(n)-1} = (\hat{\alpha}_c - d)^{-1} \hat{u}_c . \tag{15}$$

By iteratively performing E-step and M-step for several times, we finally obtain the transformation parameters $\hat{\eta}_c^{(n)}$. Using $\hat{\eta}^{(n)} = \{\hat{\eta}_c^{(n)}\} = \{\hat{\mu}_c^{(n)}, \hat{\theta}_c^{(n)}\}$, the HMM parameters are transformed according to (5). After the transformation, the hyperparameters are refreshed by

$$\varphi^{(n)} = \{\tau_c^{(n)}, m_c^{(n)}, \alpha_c^{(n)}, u_c^{(n)}\} = \{\hat{\tau}_c, \hat{m}_c, \hat{\alpha}_c, \hat{u}_c\} . \tag{16}$$

These hyperparameters $\varphi^{(n)}$ are then kept in memory and served as the new hyperparameters for on-line estimation of next transformation parameters $\eta^{(n+1)}$ when consecutive data $\mathbf{X}_{n+1}$ are collected. As shown in above derivation, the merit of proposed method is focused on the generation of reproducible prior/posterior pair in EM algorithm so that the transformation parameters and the associated hyperparameters can be efficiently and recursively computed for OLT. Generally, this set of formulas can be easily extended in terms of the segmental QB estimate which the state sequence $\mathbf{s}_n$ and transformation parameters η are alternately maximized [6].

4. HIERARCHICAL TRANSFORMATION

In OLT, it is crucial to dynamically control the number of transformation parameters such that the recognition accuracy can be improved for limited adaptation data as well as abundant adaptation data. To achieve this goal, a hierarchical tree of HMM parameters should be established prior to the adaptation [8-9]. In this study, we built the tree by clustering HMM parameters (or pdfs) using the K-means algorithm [11]. During clustering process, the divergence measure [11] was served as the distance measure. After building the tree, the node labels of HMM units in each layer are determined. Theoretically, the HMM units connected to the same node possess similar acoustical behaviors and can be suitably transformed via the shared transformation parameters. In case of missing adaptation data, part of nodes in lower layer may miss adaptation tokens. As a result, we usually obtain the transformation parameters for most nodes in higher layer and few nodes in lower layer. To reinforce the OLT precision, the HMM parameters should be transformed using the parameters nearest to leave layer. Thus, our aim is to automatically extract the transformation parameters for each HMM unit based on a *bottom-up* search strategy. This strategy captures the transformation factors along the hierarchical path corresponding to each HMM unit. The algorithm of bottom-up search strategy is described and shown below.

For each HMM unit λ_{ik}, we search the transformation parameters from leaf layer to root layer and perform the following steps. First, the cluster label of λ_{ik} in a layer is extracted. Then, we check if there exist the transformation parameters for this label. If exist, we use the associated parameters $\eta_c^{(n)}$ for OLT, i.e. $G_{\eta_c^{(n)}}(\lambda_{ik})$. Otherwise, we further check if the hyperparameters of this label $\varphi_c^{(n-1)} = (\tau_c^{(n-1)}, m_c^{(n-1)}, \alpha_c^{(n-1)}, u_c^{(n-1)})$ exist. If exist, we transform the mean vector μ_{ik} by adding the bias term $m_c^{(n-1)}$ and the covariance matrix Σ_{ik} by multiplying the scalar term $(\alpha_c^{(n-1)} - d)^{-1} u_c^{(n-1)}$ as indicated in (14-15). Once the HMM unit λ_{ik} is transformed, we skip to process the next HMM unit. Finally, this algorithm is ended until all the HMM units are transformed.

Bottom-up search algorithm for OLT

```
1. for each HMM unit $\lambda_{ik}$
2.     for tree depth from leaf layer to root layer
3.         Extract cluster label of $\lambda_{ik}$ in that depth
4.         if its transformation parameters $\eta_c^{(n)}$ exist
5.             Perform on-line transformation $G_{\eta_c^{(n)}}(\lambda_{ik})$
6.             go to step 1
7.         else if hyperparameters of that label $\varphi_c^{(n-1)}$ exist
8.             Perform on-line transformation $G_{\varphi_c^{(n-1)}}(\lambda_{ik})$
9.             go to step 1
10.        end
11.    end
12.end
```

5. EXPERIMENTS

The experiments conducted in this paper are aimed at the recognition of Mandarin speech. Mandarin is a syllabic and tonal language. Without considering the tonal information, the overall number of Mandarin syllable is 408. Generally, each Mandarin syllable can be divided into an initial (consonant) part and a final (vowel) part. When the syllable only has final part, a null initial exists practically. In this study, we employed the context-dependent subsyllable modeling for constructing the HMM units of Mandarin speech. Cumulatively, there were 93 context-dependent (CD) initials, 38 context-independent (CI) finals and 33 null initials included in the experiments. We arranged the CD initials, CI finals and null initials by three, four and two HMM states, respectively. Hence, 498 HMM states (279 for CD initials, 152 for CI finals, 66 for null initials and 1 for background silence) were setup for covering all phonetic units of 408 Mandarin syllables. Herein, two speech corpora were collected and provided by Telecommunication Laboratories, Chunghwa Telecom, Taiwan. The first one consisted of 5045 phonetically-balanced Mandarin words uttered by 51 males and 50 females. It was recorded in an office room. We applied this database to generate the SI HMM parameters and estimate the initial hyperparameters for OLT. The speech frame was characterized by a feature vector comprised of 12-order LPC-derived cepstral coefficients, 12-order delta cepstral coefficients, 1 delta log energy and 1 delta delta log energy. Besides, the second database consisted of four repetitions of 408 isolated Mandarin syllables spoken by a single female speaker. This database was collected in a soundproof room. We used three repetitions for testing and the remaining one for adaptation. Only supervised adaptation was investigated. Our recognition task is to recognize 408 Mandarin syllables, which is known to be a highly confusable vocabulary. Without adaptation, the baseline result using SI speech models had a top five recognition rates 73.8%. In the following, we examine proposed OLT through two sets of experiments.

First, we compare the recognition results of OLT with various update intervals in Fig. 1. In this case, the total number of adaptation data is fixed at N=150. The update intervals of I=5, 10, 15, 30, 50, 75 and 150 are considered in the comparison. Notably, the case of I=150 corresponds to perform the batch adaptation. We can see that the top five recognition rates are increased from 85.7% of I=5 to 89.7% of I=100. This is because that longer interval of speech data contains larger knowledge of training tokens and phonetic units. The goodness of estimated transformation parameters could be guaranteed. However, long interval of data collection is less practical due to higher costs of computation and memory. Therefore, it is a tradeoff between update interval and recognition result in OLT. On the other hand, we demonstrate the superiority of proposed OLT over Huo's on-line adaptation (also referred as OLA) [6] which QB estimate was applied for estimating CDHMM parameters. Herein, the update interval are set to be I=15. As shown in Fig.2, the recognition performance of OLT is significantly better than that of OLA. The improvement is especially obvious for small N. For examples, the top five recognition rate of OLT at N=30 is 83.2%, which is excellent compared with 75.5% of OLA. The main reason is that the proposed OLT is capable of hierarchically transforming overall HMM units even though most of sounds are unheard in adaptation data. Conversely, the OLA only adjusts the HMM units appearing in adaptation data. From these prompting results, we conclude that the proposed on-line hierarchical transformation is an effective approach to incremental adaptation in large scale's HMM-based speech recognition.

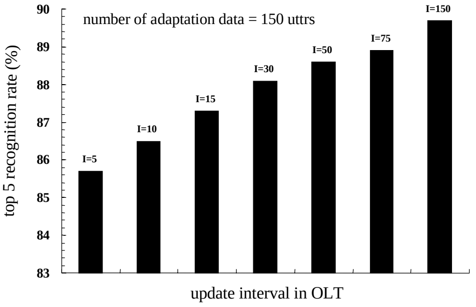

Fig. 1. Recognition result versus update interval in OLT.

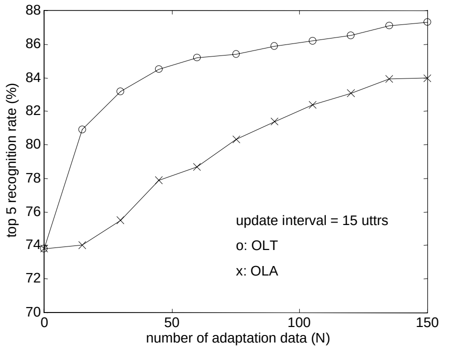

Fig. 2. Recognition comparison of OLT and OLA.

6. CONCLUSION

We have extended the framework of QB estimate to recursively learn the parameters for OLT. Our purpose is to incrementally adapt the model parameters to fit the newest variabilities without the need of storing previous adaptation data. In this study, we emphasized our contribution on the development of on-line transformation of overall HMM parameters in large-vocabulary speech recognition system even though only limited adaptation data are available. We constructed a tree structure of HMM parameters as the prior knowledge to dynamically control the transformation tying in OLT. This method is really adaptive in nature for speech recognition. In the speaker adaptation evaluation, the proposed OLT was improved asymptotically for increasing number of adaptation data. Besides, due to the capability of transforming all HMM units by using insufficient adaptation data, our OLT was significantly superior to other on-line adaptation method for various update intervals and adaptation data amounts.

7. REFERENCES

[1] J.-T. Chien, C.-H. Lee and H.-C. Wang, "A hybrid algorithm for speaker adaptation using MAP transformation and adaptation," *IEEE Signal Processing Letters*, vol. 4, no. 6, pp. 167-169, 1997.

[2] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Statist. Society (B)*, vol. 39, pp. 1-38, 1977.

[3] V. Digalakis, "On-line adaptation of hidden Markov models using incremental estimation algorithms", *Proc. EUROSPEECH*, vol. 4, pp. 1859-1862, 1997.

[4] J. L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291-298, 1994.

[5] Y. Gotoh, M. M. Hochberg, D. J. Mashao and H. F. Silverman, "Incremental MAP estimation of HMMs for efficient training and improved performance", *Proc. ICASSP*, pp. 457-460, 1995.

[6] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate", *IEEE Transactions Speech and Audio Processing*, vol. 5, no. 2, pp. 161-172, 1997.

[7] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

[8] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous control using tree structure", *Proc. EUROSPEECH*, pp. 1143-1146, 1995.

[9] K. Shinoda and C.-H. Lee, "Unsupervised adaptation using structural Bayes approach", *Proc. ICASSP*, vol. 2, pp. 793-796, 1998.

[10] J. Spragins, "A note on the iterative application of Bayes' rule", *IEEE Transactions on Information Theory*, vol. IT-11, no. 4, pp. 544-549, 1965.

[11] J.-T. Tou and R. C. Gonzalez, *Pattern Recognition Principles*, Addison-Wesley Publishing Company, 1974.