ON-LINE BAYES ADAPTATION OF SCHMM PARAMETERS FOR SPEECH RECOGNITION

Qiang Huo[†‡] and Chorkin Chan[†]

[†]Department of Computer Science, The University of Hong Kong, Hong Kong

[‡]Department of Electronic Engineering & Information Science, University of Science and Technology of China, China

ABSTRACT

In this paper, on-line adaptation of semi-continuous (or tied mixture) hidden Markov model (SCHMM) is studied. A theoretical formulation of the segmental quasi-Bayes learning of the mixture coefficients in SCHMM for speech recognition is presented. The practical issues related to the use of this algorithm for on-line speaker adaptation are addressed. A pragamatic on-line adaptation approach to combine the long-term adaptation of the mixture coefficients and the short-term adaptation of the mean vectors of the Gaussian mixture components are also proposed. The viability of these techniques are confirmed in a series of comparative experiments using a 26-word English alphabet vocabulary.

1. INTRODUCTION

In many speech recognition systems, there usually exists a performance gap between the recognition accuracies on training and on testing data. One major reason lies in the possible mismatch between the underlying acoustic characteristics associated with the training and testing conditions. To bridge this performance gap, one possible solution is to design a speech recognition system that are robust to the above types of acoustic mismatch, and this has been a long standing objective of many researchers over the past 20 years. Another way to reduce the possible acoustic mismatch between the training and testing conditions is to adopt the so called *adaptive learning* approach. The scenario is like this: starting from a pre-trained (e.g., speaker-independent) speech recognition system, for a new user (or a group of users) to use the system for a specific task, a small number of adaptation data is collected from the user, and these data are used to construct a speaker adaptive system for the speaker in the particular environment for that specific application. By doing so, the mismatch between training and testing can generally be reduced. The most fascinating adaptation scheme with great practical value is the so called *on-line* (or *incremental, sequential*) adaptation and this scheme makes the recognition system continuously adapted to the new adaptation data without the requirement of the storage of previous training data. It is this kind of approach that this paper focuses on.

Recently, Bayesian adaptive learning (specifically, Maximum *a posteriori* (MAP) estimation) of Hidden Markov Model (HMM) parameters has been proposed and adopted in a number of speech recognition applications [6, 1, 2, 3, 4, 5]. It was shown that, for HMM-based speech recognition applications, the MAP framework provides an effective way for combining adaptation data and the prior knowledge, and then creating a set of adaptive HMMs to cope with the new acoustic conditions in the test data. The prior knowledge, which is embodied in a set of seed HMMs as well as in the assumed distributions of the model parameters being adapted, is made use of to mitigate the effect of adaptation data shortage to improve the system robustness. This approach works in a *block* (or *batch*) adaptation mode using a history of all the adaptation data. It finds applications such as fast adaptation to new speaker(s) and/or new speaking environments where only a small amount of adaptation data is needed. In order to consider the long-term and short-term adaptations simultaneously, an ideal adaptation approach should work in an *incremental* adaptation mode. A related work is conducted by Matsuoka and Lee in [7]. They used the segmental MAP algorithm to conduct the so called on-line adaptation. Due to its missing mechanism of updating the hyperparameters of the prior distribution incrementally, all the previously seen adaptation data need to be stored. A full-scale *on-line* adaptation approach should be able to update both the hyperparameters of the prior distribution and the HMM parameters themselves simultaneously upon the presentation of the latest adaptation data. An *on-line* adaptation approach of this nature for semi-continuous HMMs (SCHMMs, also called tied-mixture HMMs) is presented in this paper. It is based on the segmental quasi-Bayes estimation algorithm for the mixture coefficients of SCHMM recently developed in [5].

2. SEGMENTAL MAP ESTIMATE

Consider an N-state SCHMM with parameter vector $\lambda = (\pi, A, \theta)$, where π is the initial state distribution, A is the state transition matrix, and θ is the parameter vector composed of mixture parameters $\theta_i = \{\omega_{ik}, m_k, r_k\}_{k=1,2,\cdots,K}$ for each state i with the state observation probability density function (PDF) being a mixture of a *common* set of Gaussian PDFs shared by all the HMM states. For state i, its observation PDF has the form of

$$p_i(x_t|\theta_i) = \sum_{k=1}^{K} \omega_{ik} f_k(x_t) = \sum_{k=1}^{K} \omega_{ik} N(x_t|m_k, r_k), \qquad (1)$$

where $N(x|m_k, r_k)$ is the k-th normal mixand, with m_k being the D-dimensional mean vector and r_k being the $D \times D$ precision (inverse covariance) matrix. Each state

0-7803-2431-5/95 $4.00 © 1995 IEEE

observation density differs from another by its corresponding mixture coefficients, ω_{ik}, which satisfy the constraint $\sum_{k=1}^{K}\omega_{ik}=1$.

For an observation sequence $\mathbf{x}=(x_1,x_2,\cdots,x_T)$, let $\mathbf{s}=(s_1,s_2,\cdots,s_T)$ be the unobserved associated state sequence. By maximizing the joint posterior density of the parameters λ and state sequence $\mathbf{s}$, $p(\lambda,\mathbf{s}|\mathbf{x})$, one has

$$\tilde{\lambda}=\operatorname*{argmax}_{\lambda}\max_{\mathbf{s}}p(\lambda,\mathbf{s}|\mathbf{x})=\operatorname*{argmax}_{\lambda}\max_{\mathbf{s}}p(\mathbf{x},\mathbf{s}|\lambda)g(\lambda), \tag{2}$$

where $g(\lambda)$ is the prior density for parameter λ and $\tilde{\lambda}$ is called the segmental MAP estimate of λ [6]. It can be shown that by starting with any estimate $\lambda^{(l)}$, alternate maximization over $\mathbf{s}$ and λ gives a sequence of estimates with non-decreasing values of $p(\lambda,\mathbf{s}|\mathbf{x})$, i.e. $p(\lambda^{(l+1)},\mathbf{s}^{(l+1)}|\mathbf{x})\geq p(\lambda^{(l)},\mathbf{s}^{(l)}|\mathbf{x})$ with

$$\mathbf{s}^{(l)}=\operatorname*{argmax}_{\mathbf{s}}p(\mathbf{x},\mathbf{s}|\lambda^{(l)}), \tag{3}$$

$$\lambda^{(l+1)}=\operatorname*{argmax}_{\lambda}p(\mathbf{x},\mathbf{s}^{(l)}|\lambda)g(\lambda). \tag{4}$$

The most likely state sequence $\mathbf{s}^{(l)}$ is decoded by means of the Viterbi algorithm. The maximization over λ in equation (4) is usually accomplished with an EM algorithm which itself is an iterative algorithm and very time consuming [6, 1, 2, 3, 4]. We have proposed previously [5] and summarize in this paper an approximate but efficient solution which is called the quasi-Bayes method which estimates the mixture coefficients alone.

3. SEGMENTAL QUASI-BAYES ESTIMATE

By applying the Viterbi algorithm to the training data, sets of observations (e.g., x_1, x_2, $\cdots$, x_T) associated with each HMM state can be identified. Given the sequence of observations, the updating formula for $\{\omega_{ik}\}$ corresponding to the maximization in equation (4) can be derived by solving the following quasi-Bayes estimation problem for a general finite mixture distribution.

Conditional on $\omega_i=(\omega_{i1},\ \omega_{i2},\ \cdots,\ \omega_{iK})$ and density functions f_1, f_2, $\cdots$, f_K, each x_n is assumed independently observed with the PDF as shown in equation (1). Assuming that the prior density for ω_i has the form of a Dirichlet density

$$g(\omega_i)=D(\omega_i|\nu_{i1}^{(0)},\cdots,\nu_{iK}^{(0)})\propto\prod_{k=1}^{K}\omega_{ik}^{\nu_{ik}^{(0)}-1}, \tag{5}$$

where $\nu_{ik}^{(0)}>0$, $k=1,\cdots,K$, and " $\propto$ " denotes proportionality. After observing x_1, the posterior density of ω_i becomes

$$p(\omega_i|x_1)=\sum_{k=1}^{K}p_{ik}(x_1)D(\omega_i|\nu_{i1}^{(0)}+\delta_{k1},\cdots,\nu_{iK}^{(0)}+\delta_{kK}), \tag{6}$$

where

$$p_{ik}(x_1)=\frac{f_k(x_1)\nu_{ik}^{(0)}}{\sum_{m=1}^{K}f_m(x_1)\nu_{im}^{(0)}} \tag{7}$$

and δ_{ij} is the Kronecker delta function $\delta_{ij}=\delta(i-j)$. Many well-known approximate Bayesian learning procedures to solve this problem arise from approximating the RHS of equation (6) by

$$p(\omega_i|x_1)\approx D(\omega_i|\nu_{i1}^{(0)}+\Delta_{11},\cdots,\nu_{iK}^{(0)}+\Delta_{1K}), \tag{8}$$

where the Δ_{ij}'s take values according to a specified method. In our proposed quasi-Bayes procedure, it is suggested that Δ_{1k} be replaced by $p_{ik}(x_1)$ shown in equation (7), and therefore

$$p(\omega_i|x_1)\approx D(\omega_i|\nu_{i1}^{(1)},\cdots,\nu_{iK}^{(1)}), \tag{9}$$

where $\nu_{ik}^{(1)}=\nu_{ik}^{(0)}+p_{ik}(x_1)$. Then, subsequent updating takes place entirely within the Dirichlet family of distributions, *viz.*, $p(\omega_i|x_1,x_2,\cdots,x_n)$ is Dirichlet with parameters

$$\nu_{ik}^{(n)}=\nu_{ik}^{(n-1)}+p_{ik}(x_n), \tag{10}$$

where $\nu_{ik}^{(n-1)}$ are parameters of $p(\omega_i|x_1,x_2,\cdots,x_{n-1})$, and

$$p_{ik}(x_n)=\frac{f_k(x_n)\nu_{ik}^{(n-1)}}{\sum_{m=1}^{K}f_m(x_n)\nu_{im}^{(n-1)}}. \tag{11}$$

The (quasi-) posterior mean for ω_{ik}, after observing x_1, x_2, $\cdots$, x_n is given by

$$\hat{\omega}_{ik}^{(n)}=\frac{\nu_{ik}^{(n)}}{\sum_{m=1}^{K}\nu_{im}^{(n)}}, \tag{12}$$

and the mode of the approximate posterior density is

$$\tilde{\omega}_{ik}^{(n)}=\frac{\nu_{ik}^{(n)}-1}{\sum_{m=1}^{K}(\nu_{im}^{(n)}-1)}. \tag{13}$$

Both equations (12) and (13) can serve as the updating formula for the mixture coefficients in the segmental quasi-Bayes learning for SCHMMs. Equation (10) is used as the updating formula of the hyperparameters.

Note that apart from its computational efficiency, another advantage of the segmental quasi-Bayes method over the segmental MAP one is due to its ***sequential*** nature in updating both the hyperparameters of the prior distribution and the SCHMM parameters. This makes the so-called ***on-line adaptation*** of the mixture coefficients very natural under the framework of the quasi-Bayes method. We will compare the so-called ***batch adaptation*** scheme and the ***on-line adaptation*** scheme in the next Section using a series of speaker adaptation experiments to substantiate the viability of the proposed techniques.

4. SPEAKER ADAPTATION EXPERIMENTS

4.1. Experimental Setup

To study the practical issues related to the use of the segmental quasi-Bayes algorithm in estimating SCHMM parameters for a speaker adaptation application, the 26-letters of the English alphabet are chosen to form the vocabulary for all experiments. Two severely mismatched databases are used for evaluating the adaptation algorithms. For speaker

Table 1: Performance comparison (% correct) of several segmental adaptation schemes for the mixture coefficients of SCHMMs only (SI recognition rate: 47.8%)

Tokens	SEG-ML	SEG-MAP	QB-BL	QB-OL
1	56.3	61.5	62.0	62.0
2	62.5	65.1	65.0	65.4
3	65.9	67.2	66.8	67.4
4	67.0	68.0	67.8	68.2
5	68.4	69.1	69.0	69.1
6	68.3	68.9	69.0	69.3
7	68.6	69.4	69.3	69.5
8	69.4	69.7	70.0	69.9
9	70.3	70.5	70.1	70.1
10	70.7	70.8	70.4	70.6

independent (SI) training and prior density estimation, the OGI ISOLET database produced by 150 speakers (75 females and 75 males) is used. Each speaker utters each of the letters twice. For speaker dependent (SD) or adaptive (SA) training and testing, the TI46 isolated word corpus produced by 12 speakers (8 females and 4 males) is used. Each person utters each of the letters 26 times, 10 of them used for SD/SA training and the remaining 16 tokens for testing. Readers are referred to [3, 4, 5] for further details.

4.2. On-line Adaptation of the Mixture Coefficients

As is well-known, mixture coefficients are very important parameters in modeling speech units in SCHMM. To examine the viability and effect of the segmental quasi-Bayes algorithm presented in this paper for on-line adapting the mixture coefficients of SCHMM only, a series of comparative experiments are conducted. The first experiment is to recognize the English alphabet subset of TI46 with the SI system trained with speech tokens from OGI ISOLET. The average recognition rate is 47.8%. For simplicity, in SA/SD training, Gaussian mixture component PDFs and the transition probabilities are fixed to that of the SI system. In SA training, the hyperparameters of the prior distribution of the mixture coefficients are estimated with the *ad hoc* method discussed in [3]. The remaining experimental setups are as follows: "SEG-ML" stands for SD segmental ML (k-means) training of the mixture coefficients and "SEG-MAP" corresponds to its MAP counterpart. "QB-BL" stands for SA segmental quasi-Bayes block adaptation of the mixture coefficients, and "QB-OL" refers to its on-line adaptation counterpart. The average word recognition rates for the 12 speakers are summarized in Table 1. The rows in Table 1 correspond to the numbers of training tokens used for each SD and SA cases.

The first observation from Table 1 is that the SD recognition rate of only one training token is better than that of the SI system and this fact is a good indication of the serious mismatch between the two corpora. A second observation is that when using the same amount of training data, SA training outperforms SD training in most of the cases tested. This implies that SA training utilizes the adaptation data more effectively than SD training, especially in cases of insufficient training data. As expected, the SA performance quickly becomes equivalent to the SD performance when the number of adaptive training tokens increases. A third observation is that the recognizer performance with the segmental quasi-Bayes method is not much different from that with the segmental MAP method, and this fact also shows the viability of the quasi-Bayes approximation in maximizing the RHS of equation (4). By using the quasi-Bayes learning framework, one can update both the hyperparameters of the prior distribution and the mixture coefficients simultaneously upon the presentation of the current adaptation data. In this way, with each adaptation utterance presented, its effect upon further adaptation is accumulated into the prior distribution. Thus previous adaptation data need not be stored explicitly. The prior distribution reflects all the prior knowledge about the mixture coefficients. A true on-line adaptation of the mixture coefficients can thus be conducted. As a fourth observation, by comparing the results of "QB-OL" and that of "QB-BL", it is noticed that the on-line adaptation results are similar to the one based on the batch adaptation scheme. This confirms the effectiveness of the on-line adaptation scheme of the mixture coefficients. In the next subsection, a pragmatic procedure which combines the quasi-Bayes adaptation of mixture coefficients with an adaptation scheme of the component density will be experimentally tested and reported.

4.3. On-line Adaptation of the Mixture Coefficients and the Mean Vectors

For an SCHMM based recognizer, apart from the mixture coefficients, the adaptation of the mean vectors of the Gaussian mixture components is also very important [4]. However, the previously proposed algorithm can only be theoretically justified in the case of fixed mixture components. On the other hand, it has been shown in [4] that the mean vectors of the common Gaussian densities in SCHMM can be rapidly and effectively estimated even with a limited amount of training data by the conventional speaker dependent training. Thus a pragmatic procedure which combines the quasi-Bayes adaptation of the mixture coefficients and the adaptation of the mean vectors can be as follows:

0. Take SI trained models as initial models. The initial hyperparameters of the mixture coefficients are computed.
1. Obtain new adaptation token(s) and push it (them) into the "history data buffer".
2. Conduct segmental quasi-Bayes estimation of the mixture coefficients.
3. Fix the other parameters and SD-train the mean vectors of the mixture components with the adaptation data in the "history data buffer".
4. Repeat Steps 2 and 3 several times, and then update the hyperparameters of the prior distributions of the mixture coefficients.
5. Go to Step 1.

To examine the effects of this on-line adaptation procedure, a series of comparative experiments are conducted. Once again, for simplicity, the transition probabilities and the covariance matrices of the Gaussian mixture components are fixed to that of the SI system. In an on-line SD training of mean vectors, different block sizes of the

Table 2: Performance comparison (% correct) of several adaptation schemes for the mixture coefficients and the mean vectors of SCHMMs (SI recognition rate: 47.8%)

Tokens	SEG-ML	OL-1	OL-2	OL-3
1	63.6	66.8	66.8	66.8
2	70.4	71.5	72.4	72.4
3	74.2	72.3	73.2	73.9
4	76.2	73.7	73.9	75.0
5	77.3	73.7	74.2	75.3
6	76.1	73.6	74.9	75.6
7	77.1	75.9	76.1	76.2
8	76.9	76.0	76.6	76.0
9	77.7	76.4	76.8	77.0
10	78.5	77.1	77.2	77.4

"history data buffer" is examined. In the particular experimental setup here, the cases with buffer size of 1, 2, 3 token(s) per letter have been tried. The related experimental results (the average word recognition rates for 12 speakers) are summarized in Table 2. The rows in Table 2 correspond to the numbers of training tokens used for each SD and SA cases. "SEG-ML" stands for SD "segmental ML (k-means)" training of the mixture coefficients and the mean vectors. "OL-1" corresponds to on-line adaptation of the mixture coefficients and the mean vectors with the history data buffer size being 1. Similarly, "OL-2" and "OL-3" refer to respectively the cases with buffer sizes of 2 and 3.

Once again, from Table 2, it is observed that the recognizer performance with on-line adaptation outperforms that with SD training when the SD training data is insufficient (1 and 2 tokens). The SD performance improves as the number of speaker specific training tokens increases, and the on-line adaptation scheme can follow this increasing trend, although its absolute recognition rate is inferior to the SD one when relatively more SD training tokens (in particular here more than 3 tokens) are available. As for the effects of the "history data buffer" size, it is observed that the larger the buffer size, the better the on-line adaptation performance. On the other hand, larger buffer size also means more storage is required. From the practical point of view, there will be a compromise in real applications. The on-line adaptation of the mixture coefficients can be viewed as a long-term adaptation process to cope with long-term variations. All the historical knowledge is represented by the prior distributions and is updated incrementally. The effect of this long-term prior knowledge on the adaptation results can be easily controlled through some *forgetting* mechanism. This mechanism can be implemented by setting up some registers to store the most recent contributions from the adaptation data history. When it becomes time to "forget" about the long-term prior knowledge, the hyperparameters of the prior distributions can be recomputed from the stored recent contributions. The online SD training of mean vectors can be looked upon as a short-term (or fast) adaptation process to track the latest variations. This kind of on-line adaptation framework will find applications in real world adaptive speech recognition systems.

5. SUMMARY

In this paper, in order to cope with the acoustic mismatch problem between the training and testing conditions, the issues of on-line adaptation of a SÇHMM-based speech recognition system are addressed. A theoretical formulation of the segmental quasi-Bayes learning of the mixture coefficients in SCHMM for speech recognition is presented. The practical issues related to the use of this algorithm in on-line adapting the mixture coefficients of SCHMM for speaker adaptation are studied. A pragamatic on-line adaptation approach to combine the long-term adaptation of the mixture coefficients and the short-term adaptation of the mean vectors of the Gaussian mixture components are also proposed. The viability of these techniques are confirmed in a series of comparative experiments using a 26-word English alphabet vocabulary. The kind of on-line adaptation approach studied in this paper is a topic of interest both in theory and in practice. Further research is needed to develop the on-line adaptation method which can update incrementally the hyperparameters of both the mixture coefficients and the mean vectors as well as the covariance matrices.

ACKNOWLEDGEMENT

Dr. Chin-Hui Lee of AT&T Bell Laboratories is an equal partner in authoring this paper. Since he is a co-author of three papers submitted to ICASSP'95, he declines to have his name included in this paper. We must thank him for his contribution all the same.

REFERENCES

[1] J.-L. Gauvain and C.-H. Lee, "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Communication*, Vol. 11, Nos. 2-3, pp.205-213, 1992.

[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] Q. Huo, C. Chan and C.-H. Lee, "Bayesian Learning of the Parameters of Discrete and Tied Mixture HMMs for Speech Recognition," in *Proc. Eurospeech-93* (Berlin, Germany), 1993, pp.III-1567-1570.

[4] Q. Huo, C. Chan and C.-H. Lee, "Bayesian Learning of the SCHMM Parameters for Speech Recognition," in *Proc. ICASSP-94*, 1994, pp.I-221-224.

[5] Q. Huo, C. Chan and C.-H. Lee, "Segmental Quasi-Bayesian Learning of the Mixture Coefficients in SCHMM for Speech Recognition," in *Proc. ISSIPNN-94* (Hong Kong), 1994, pp.678-681.

[6] C.-H. Lee, C.-H. Lin and B.-H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans. on ASSP*, Vol. ASSP-39, No. 4, pp.806-814, 1991.

[7] T. Matsuoka and C.-H. Lee, "A Study of On-line Bayesian Adaptation for HMM-based Speech Recognition," in *Proc. Eurospeech-93*, 1993, pp.815-818.