Rapid online adaptation using speaker space model evolution

Dong Kook Kim a,b,*, Nam Soo Kim b,1

a Computer and Software Research Laboratory, Electronics and Telecommunications Research Institute, 161 Gajeong-dong, Yuseong-gu, Daejeon 305-350, South Korea
b School of Electrical Engineering and INMC, Seoul National University, Kwanak P.O. Box 34, Seoul 151-742, South Korea

Received 4 July 2003; received in revised form 14 October 2003; accepted 12 January 2004

Abstract

This paper presents a new approach to online adaptation of continuous density hidden Markov model (CDHMM) with a small amount of adaptation data based on speaker space model (SSM) evolution. The SSM which characterizes the a priori knowledge of the training speakers is effectively described in terms of the latent variable models such as the factor analysis or probabilistic principal component analysis. The SSM provides various sources of information such as the correlation information, the prior density, and the prior knowledge of the CDHMM parameters that are very useful for rapid online adaptation. We design the SSM evolution based on the quasi-Bayes estimation technique which incrementally updates the hyperparameters of the SSM and the CDHMM parameters simultaneously. In a series of speaker adaptation experiments on the continuous digit and large vocabulary recognition tasks, we demonstrate that the proposed approach not only achieves a good performance for a small amount of adaptation data but also maintains a good asymptotic convergence property as the data size increases.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Speaker space model; Prior evolution; Latent variable model; Quasi-Bayes estimate; Online adaptation; Rapid speaker adaptation

1. Introduction

Automatic speech recognition (ASR) systems have been used for various applications in real environments. It is well known, however, that the performance of ASR systems often degrades drastically when there is a mismatch between the training and test conditions (Lee, 1997). Many adaptation techniques have been proposed to try to reduce such mismatch (Woodland, 2001). Since there is only a very limited amount of adaptation data available in many applications, efficient use of the adaptation data has been an important research direction in the speaker adaptation study (Kuhn et al., 2001; Lee and Huo, 2000). Generally, two desirable properties of a good adaptation algorithm are required as follows (Huo and Ma, 2001; Kuhn et al., 2001):

Rapid: The adaptation algorithm should be effective even with a small amount of adaptation data, and approach asymptotically to the
matched-condition performance as the data size grows.

**Online:** The adaptation algorithm should continuously adapt to the changing environments without a requirement to store the previously used adaptation data. One of the important advantages of online (incremental, sequential) adaptation is its capability to update the model parameters with less computation and reduced storage requirement. The goal of this paper is to develop a rapid online adaptation technique which incrementally updates the model parameters with a small amount of adaptation data.

Many model-based adaptation approaches adjust the continuous density hidden Markov model (CDHMM) parameters so as to better fit to the acoustic characteristics of a new speaker (Woodland, 2001; Lee and Huo, 2000; Kuhn et al., 2001). The speaker space methods such as eigenvoice and cluster adaptive training (CAT) which perform speaker adaptation by constructing a new speaker model as a linear combination of canonical speaker models were introduced for rapid adaptation (Kuhn et al., 2000; Gales, 2000). The parameters associated with the canonical speaker models can be viewed as a set of basis vectors of the speaker space, which accounts for the a priori knowledge of the training speakers. Given a speaker space, a new speaker model can be found by estimating an appropriate point within this speaker space. Significant performance improvement has been achieved with a small amount of adaptation data. Recently, we proposed a rapid speaker adaptation technique called the speaker space model (SSM) approach (Kim and Kim, 2000, 2001, 2002, 2004) based on the latent variable models (LVMs) such as the factor analysis (FA) (Rubin and Thayer, 1982) or probabilistic principal component analysis (PPCA) (Tipping and Bishop, 1999). The SSM finds the speaker space based on the expectation maximization (EM) algorithm (Dempster et al., 1977) from a set of well-trained speaker-dependent (SD) models. The proposed approach modifies the original eigenvoice technique such that it can be naturally combined with the maximum a posteriori (MAP) adaptation framework. The SSM technique simply interpolates the maximum likelihood (ML) estimate with the prior speaker model estimate obtained within the speaker space (Kim and Kim, 2000). Similarly, a method to use the eigenvoice as the prior model for the MAP adaptation and anisotropic MAP defined by eigenvoice have been suggested (Kuhn et al., 2000; Botterweck, 2001). Also, eigenspace-based maximum likelihood linear regression (MLLR) and maximum a posteriori linear regression (MAPLR) techniques were developed by analyzing the transformation matrices associated with the training speakers using the conventional principal component analysis (PCA) (Jolliffe, 1986) and PPCA, respectively (Chen et al., 2000; Chen and Wang, 2001). Similar to the SSM, the transformation space model (TSM) which characterizes the a priori information of the training speakers can be defined as well by taking the transformation parameters, instead of the CDHMM parameters, as the representative vectors for each training speaker (Gales, 2000; Chen and Wang, 2001; Kim et al., 2003).

Many online adaptation algorithms have been studied under a Bayesian inference framework to make the ASR system capable of continuously adapting to a new environment (Lee and Huo, 2000). Huo and Lee (1997) applied the quasi-Bayes (QB) learning algorithm to incrementally update both the CDHMM parameters and the hyperparameters through a prior evolution procedure. They extended the QB estimate to cope with the correlated CDHMM parameters in which all CDHMM mean vectors are correlated with a joint prior distribution (Huo and Ma, 2001). Chien (1999) presented an online transformation-based QB adaptation algorithm by adopting a simple transformation and assuming a specific prior probability density function (pdf) for the transformation matrices. In addition, he proposed the QB linear regression (QBLR) (Chien, 2002) algorithm for online linear regression adaptation of the CDHMM and showed that the QBLR is a generalized framework with MLLR (Leggetter and Woodland, 1995) and MAPLR (Chou, 1999) as special cases. More
recently, we proposed the rapid online adaptation approaches called the SSM evolution and TSM evolution (Kim and Kim, 2002; Kim et al., 2003).

In this paper, we propose a new approach to online adaptation of the CDHMM mean parameters based on SSM evolution. Due to the flexibility of online adaptation and the effectiveness of speaker space approach for rapid adaptation, we are motivated to develop the SSM evolution using both the SSM and the QB learning for rapid online adaptation of HMM parameters. Even though the basic idea of the proposed approach is similar to that of our previous work (Kim and Kim, 2002), a more comprehensive description of the algorithm with extensive experimental results is given here.

We establish SSM evolution based on the QB estimation technique that incrementally updates the hyperparameters of the SSM and the CDHMM mean parameters simultaneously. Experimental results show that the proposed approach achieves a rapid online adaptation performance for a small amount of adaptation data, and it maintains the good asymptotic convergence property as the data increases. Furthermore, online adaptation using the SSM evolution attains almost the same level of recognition performance as the batch SSM adaptation.

This paper is organized as follows. In Section 2, we introduce the SSM based on the LVMs such as the FA or PPCA. In Section 3, we present an online adaptation approach based on SSM evolution. QB learning of the SSM and its evolution are addressed in this section. The experimental results on online speaker adaptation using SSM evolution are given in Section 4. Finally, in Section 5, we summarize our work.

2. Speaker space model

Consider an $N$-state CDHMM with $K$ mixture components, $\mathcal{X} = \{ \lambda_j \} = \{ w_{jk}, \mu_{jk}, \Sigma_{jk} \}$, $j = 1, \ldots, N$, $k = 1, \ldots, K$. The state observation pdf of an observation vector $x$ is defined to be a mixture of multivariate Gaussians

$$p(x|\lambda_j) = \sum_{k=1}^{K} w_{jk} N(x|\mu_{jk}, \Sigma_{jk})$$  \hspace{1cm} (1)$$

where $w_{jk}$ is the weight for the mixture component $k$ in state $j$ with $\sum_{k=1}^{K} w_{jk} = 1$, and $N(x|\mu_{jk}, \Sigma_{jk})$ is the $k$th normal distribution given by

$$N(x|\mu_{jk}, \Sigma_{jk}) \propto |\Sigma_{jk}|^{-1/2} \exp \left( -\frac{1}{2}(x - \mu_{jk})^T \Sigma_{jk}^{-1} (x - \mu_{jk}) \right)$$  \hspace{1cm} (2)$$

with $\mu_{jk}$ and $\Sigma_{jk}$ being the $d$-dimensional mean vector and $d \times d$ covariance matrix, respectively. Here, $\propto$ denotes proportionality and $|\Sigma|$ means the determinant of the matrix $\Sigma$.

Let $\{ \mu_1, \ldots, \mu_R \}$ be a set of $R$ well-trained SD models, which can be obtained from the given training database. Here, $\mu_j = [\mu_{j1}, \ldots, \mu_{jK}]^T$ is a supervector of dimension $D$ that corresponds to all the Gaussian mean vectors from the $r$th speaker model. Specifically, $\mu_{jk}$ represents a mean vector of the $k$th Gaussian in the $j$th state of the $r$th speaker CDHMM. We assume that the set of SD models, $\{ \mu_1, \ldots, \mu_R \}$ are generated by a LVM such as FA with parameters $\phi = \{ W, \mu, A \}$ such that

$$\mu_j = Wv + \bar{\mu} + \epsilon$$  \hspace{1cm} (3)$$

where $\bar{\mu}$ is the mean of the supervectors, $W = [w_1, \ldots, w_R]$ is a matrix that represents the subspace of the observation data spanned by the $P$ column vectors of $W$, $v = [v_1, \ldots, v_R]^T$ is a latent variable of dimension $P$ and $\epsilon$ is a Gaussian random noise independent of $v$. Traditionally, the latent variable $v$ is defined to be an independent Gaussian of unit variance, $p(v) \sim N(0, I_P)$ with $I_P$ being the $P \times P$ identity matrix. In the FA model, the noise is distributed according to a Gaussian with a diagonal covariance matrix $A$ such that $p(\epsilon) \sim N(0, A)$. On the other hand, the PPCA defines the noise covariance matrix to be isotropic, i.e., $A = \sigma^2 I_D$, where $I_D$ is the $D \times D$ identity matrix. Based on the above assumption, it is possible to derive the conditional distribution of $\mu$ given $v$ as

$$p(\mu|v) = (2\pi)^{-D/2} |A|^{-1/2} \exp \left\{ -\frac{1}{2} (\mu - Wv - \bar{\mu})^T A^{-1} (\mu - Wv - \bar{\mu}) \right\}$$  \hspace{1cm} (4)$$

and construct the prior pdf of $\mu$ such that
latent variables

according to the ML criterion. Since, however, the

470


¼

the following criterion:

g

f

new parameter values

Let

iteratively updates the parameter values is applied.
the EM algorithm (Dempster et al., 1977) which

latent variable sequence

corresponds to the largest eigenvalues are se-

clected by means of the PCA technique, and then

adapted model to be a linear com-

bination of the speaker basis vectors and the

weight associated to each element of the basis is

estimated according to the ML criterion resulting in

a great reduction of the free parameters to be

estimated from the adaptation data. Unlike the
eigenvoice approach, the SSM finds the speaker

space based on the iterative EM algorithm. In the
FA, the subspace defined by the columns of
W

generally does not correspond to the principal
subspace of the training speakers. On the other
hand in the case of PPCA, Tipping and Bishop
(1999) showed that at the global maximum of the
likelihood of the observation data \{x_i\}, the
columns of W span the principal subspace of the
training speakers.

Given the latent variable \(v\), the speaker model \(\mu\)
doing not appear as a single point in the speaker
space but a random variable with the pdf \(p(\mu|v)\)
due to the noise process. Estimating the latent
variable \(v\) generates the a priori pdf of the speaker
model instead of producing directly the adapted
model. This prior pdf which represents the esti-
mated prior knowledge for a specific speaker
enables us to employ the MAP-based speaker
adaptation scheme (Kim and Kim, 2000, 2004).
The MAP adaptation method which incorporates
the SSM is illustrated in Fig. 1 (Kim and Kim,
2004). Given the SSM and speaker model \(\mu^{(i)}\) at
the \(i\)th step, the latent variable is estimated
according to \(v^{(i)} = E[v|\mu^{(i)}]\) and \(p(\mu|v^{(i)})\)
is used as the prior pdf for the MAP criterion to
derive \(\mu^{(i+1)}\). This iteration continues until
convergence, and the final speaker model \(\hat{\mu}\)
represents the adapted

\[ g(\mu|\phi) = (2\pi)^{-D/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mu - \mu)^T \Sigma^{-1} (\mu - \mu) \right\} \]

with the covariance matrix \(\Sigma = WW^T + A\).

Given an observation vector sequence \(Y = \{\mu_1, \ldots, \mu_N\}\), the LVM enable us to estimate the
latent variable sequence \(V = \{v_1, \ldots, v_N\}\) and find
the optimal model parameters \(\hat{\phi} = \{\hat{W}, \hat{\mu}, A\}\)
according to the ML criterion. Since, however, the
latent variables \(\{v_i\}\) are considered to be hidden,
the EM algorithm (Dempster et al., 1977) which
iteratively updates the parameter values is applied.
Let \(\phi\) be the current parameter values. Then, the
new parameter values \(\hat{\phi}\) are obtained according to
the following criterion:

\[ \hat{\phi} = \arg \max E[\log p(Y, V|\hat{\phi})|Y, \phi] \]

The E-step involves estimating the posterior dis-
tribution of latent variable \(v\) using the current
parameter values and the M-step maximizes the
auxiliary function based on the posterior distribu-
tion (Rubin and Thayer, 1982; Tipping and
Bishop, 1999).

Section 3 defines the SSM which characterizes
the a priori knowledge of the training speaker
associated with the CDHMM mean vectors. It
describes the prior information of the speaker
variability by analyzing the mean supervectors
related to the training speakers. It is noted that the
SSM provides various sources of information such
as the correlation information among different
speech units, the prior distribution, and the prior
knowledge associated with CDHMM mean
parameters, which are very useful for rapid online
speaker adaptation. The columns of the matrix,
\(\{w_1, \ldots, w_P\}\) represent a set of basis vectors which
we call the speaker basis.

In the eigenvoice approach, a set of eigenvectors
corresponding to the largest eigenvalues are
selected by means of the PCA technique, and then
applied as the speaker basis. These basis vectors
are orthonormal and represent the most important
components (i.e., principal components) of varia-
tion among the training speakers. The eigenvoice
constrains the adapted model to be a linear com-

Fig. 1. Speaker space model and batch adaptation by SSM
with \(D = 3\) and \(P = 2\).
speaker model. The mean adaptation based on the SSM approach is performed as follows:

$$\hat{\mu}_{j,k,\text{SSM}} = \alpha_{jk} \hat{\mu}_{j,k,\text{ML}} + (1 - \alpha_{jk}) \hat{\mu}_{j,k,\text{SS}}$$

(7)

with a weight vector $\alpha_{jk}$. This tells us that the SSM-based adaptation solution provides a unified framework which simply interpolates the ML estimate of the adaptation data, $\mu_{\text{ML}}$, with the prior estimate within the speaker space, $\mu_{\text{SS}}$. As the amount of adaptation data increases, the adapted mean vector converges to the ML solution while it depends more on the prior estimate, $\mu_{\text{SS}}$ for a small amount of adaptation data (Kim and Kim, 2000, 2004).

We can consider the eigenvoice followed by MAP and anisotropic MAP approaches to be special cases of SSM in which the latent variable do not have valid distributions. However there is an important difference between the FA/PPCA and eigenvoice with MAP or anisotropic MAP approach. There is a probability density model underlying FA/PPCA, but no such model is found in eigenvoice with MAP or anisotropic MAP approach. This probability model generates the a priori/posteriori pair for the HMM parameters, such that it is naturally incorporated into the QB estimate for online adaptation.

3. Online adaptation based on SSM evolution

3.1. QB learning for SSM

We briefly review the basic concept and formulation of QB learning (Huo and Lee, 1997, 1998). Let $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ be $n$ independent sets of observation data which are incrementally collected to update the CDHMM parameters, $\lambda$. A recursive expression for the a posteriori pdf of $\lambda$ is given by

$$p(\lambda|\mathcal{X}_n) = \frac{p(\mathcal{X}_n|\lambda) \cdot p(\lambda|\mathcal{X}_{n-1})}{\int p(\mathcal{X}_n|\lambda) \cdot p(\lambda|\mathcal{X}_{n-1})d\lambda}.$$  

(8)

This provides a basis for making a recursive Bayesian estimate of the given parameter $\lambda$. However, direct application of this type of recursive Bayesian estimation technique to the CDHMM has been found very difficult. To alleviate this problem, an approach called the QB learning was proposed by Huo and Lee (1997).

The QB procedure, at each step of the recursive Bayes learning, approximates the true posterior density $p(\lambda|\mathcal{X}_{n-1})$ by the closest tractable parameter density $g(\lambda|\phi^{(n-1)})$ under the criterion that both densities should have the same mode. Here, $\phi^{(n-1)}$ denotes the updated hyperparameters after observing $\mathcal{X}_{n-1}$.

Let us assume that at time instant $n$, we are given a set of observation vectors $\mathcal{X}_n = \{x_1^{(n)}, \ldots, x_T^{(n)}\}$ and the approximate prior pdf $g(\lambda|\phi^{(n-1)})$. Since we assume that the means vectors are generated through a model given by (3), which has a hidden variable $v$ with the hyperparameter $\phi_{v}^{(n-1)}$, the complete-data likelihood for $\lambda$ can be easily defined. Let $(\mathcal{X}_n, S_n, L_n)$ denote the complete-data for $\mathcal{X}_n$ where $S_n = \{S_t^{(n)}\}$ represents the state sequence and $L_n = \{l_t^{(n)}\}$ is the mixture component sequence. We can derive an approximate MAP estimate $\lambda^{(n)}$ of $\lambda$ by repeating the following EM steps:

E-step: Compute the auxiliary function

$$R(\lambda|\hat{\lambda}_{m-1}) = E[\log p(\mathcal{X}_n, S_n, L_n|\lambda)$$

$$+ \rho \log g(\lambda, v|\phi^{(n-1)})|\mathcal{X}_n, \hat{\lambda}_{m-1}].$$

(9)

where $0 < \rho \leq 1$ is a forgetting factor to reduce the effect of past observations $\mathcal{X}_{n-1}$ relative to the new data $\mathcal{X}_n$.

M-step: Find the parameter value which maximizes the auxiliary function such that

$$\hat{\lambda}_m = \arg \max_{\lambda} R(\lambda|\hat{\lambda}_{m-1})$$

(10)

where $m = 1, \ldots, M$ and $M$ denotes the total number of iterations performed with $\lambda_0 = \lambda^{(n-1)}$. Thus, iteratively applying the EM steps of (9) and (10) guarantees that the approximate posterior density never decreases. At the last EM iteration, the set of hyperparameters $\phi^{(n)}$ is computed to satisfy

$$g(\lambda|\phi^{(n)}) \propto \exp \left\{ R(\lambda|\lambda_{M-1}) \right\}.$$ 

(11)

Finally the QB estimated CDHMM parameters $\hat{\lambda}^{(n)}$ are updated by taking the mode of $g(\lambda|\phi^{(n)})$.  

3.2. Derivation of SSM evolution

In this paper, we consider only the case of prior evolution for the mean vectors of CDHMM in which mixture weights and covariance matrices are fixed during adaptation. The selection of the prior distribution is one of the important issues for the QB estimate. The prior density of the mean parameters should be selected such that it belongs to the conjugate distribution family where the prior distribution and the pooled posterior distribution have the same pdf form. Here the major contribution of this study is that the generation of reproducible prior/posterior pair for the CDHMM mean parameters is based on the SSM defined in (3). According to the SSM, we define the prior density of the mean parameters to be a normal distribution in (5). Under the specification of the SSM for $\mu$ in (3), the auxiliary function in the expectation step (9) can be rewritten as

$$R(\hat{\lambda}_{m-1}) = \sum_{S_n} \sum_{L_n} p(S_n, L_n | \mathcal{X}_n, \hat{\lambda}_{m-1})$$

$$+ \log p(\mathcal{X}_n, S_n, L_n | \lambda)$$

$$+ \rho E[\log p(\lambda | v, \phi^{(n-1)})] p(v) | \hat{\lambda}_{m-1}].$$

(12)

Based upon (2) and (4), it is not difficult to derive

$$R(\hat{\lambda}_{m-1}) = \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_{i}(j, k)$$

$$\left[ - \frac{1}{2} (x_{i}^{(n)} - \mu_{jk})^T \Sigma_{jk}^{-1} (x_{i}^{(n)} - \mu_{jk}) \right]$$

$$+ \rho E \left[ - \frac{1}{2} (\mu - W^{(n-1)} v)$$

$$- \hat{\mu}^{(n-1)})^T \Sigma_{jk}^{-1} (\mu - W^{(n-1)} v)$$

$$- \hat{\mu}^{(n-1)}) | \hat{\lambda}_{m-1} \right].$$

(13)

where $\gamma_{i}(j, k) = P(s_{i}^{(n)} = j, l_{i}^{(n)} = k | \mathcal{X}_n, \hat{\lambda}_{m-1})$ is the posterior probability of being in state $j$ and mixture component $k$ at time $t$ given the observation sequence $\mathcal{X}_n$. It is shown that the exponential of the expectation function in (9) multiplied by a normalized constant $C$, i.e., $C \cdot \exp(R(\hat{\lambda}_{m-1}))$ can be expressed in the normal distribution form $\mathcal{N}(\hat{m}, \hat{\Phi})$ with mean $\hat{m}$ and covariance matrix $\hat{\Phi}$ as follows (Zavaliagkos, 1995):

$$\hat{m} = (\rho A^{1,(n-1)} + CS^{-1})^{-1}$$

$$\cdot (\rho A^{1,(n-1)} W^{(n-1)} \tilde{v} + \mu^{(n-1)} + CS^{-1} \mu_{ML})$$

(14)

$$\hat{\Phi} = (\rho A^{1,(n-1)} + CS^{-1})^{-1}$$

(15)

with $S = \text{diag}(\Sigma_{11}, \ldots, \Sigma_{NK})$ and $C = \text{diag}(c_{11}, \ldots, c_{NK})$ in which $c_{jk} = \sum_{i} \gamma_{i}(j, k)$ is the count of the $k$th Gaussian in the $j$th state. Furthermore, $\tilde{v} = E[v | \hat{\lambda}_{M-1}]$

$$= (I + W^{T}(n-1) A^{1,(n-1)} W^{(n-1)})^{-1}$$

$$\times W^{T}(n-1) A^{1,(n-1)} (\hat{\mu}_{M-1} - \hat{\mu}^{(n-1)})$$

and $\mu_{ML} = [(\mu_{11})^{T}, \ldots, (\mu_{NK})^{T}]^{T}$ where $[(\mu_{jk})_{ML} = (\sum_{i} \gamma_{i}(j, k)x_{i}^{(n)}/c_{jk}.$

It is obvious that the approximated posterior density of complete data belongs to the same normal distribution family as $g(\mu | \phi^{(n)})$ in (5) with the updated hyperparameters $\hat{m}$ and $\hat{\Phi}$. Because the speaker space $W$ specifies the a prior knowledge of the training speakers, it is not supposed to change after observing the adaptation data. Thus we assume that the hyperparameter $W$ does not evolve (i.e., fixed during prior evolution) and the prior evolution for $A$ is approximated by $\hat{\Phi}$. As a result, the prior evolution of the SSM can be described in terms of the hyperparameters $\phi^{(n)}$

$$\tilde{\mu}^{(n)} = \hat{m}$$

(17)

$$A^{(n)} = \hat{\Phi}$$

(18)

$$W^{(n)} = W^{(n-1)}.$$  

(19)

For the prior evolution of the PPCA model, similar results are obtained, $\tilde{\mu}^{(n)} = \hat{m}$ and $\sigma^{2,(n)} = \text{trace}(\hat{\Phi})/D$ with $A^{(n-1)} = \sigma^{2,(n-1)} I$. After completing the SSM evolution procedure, the QB estimated CDHMM mean vectors $\tilde{\mu}^{(n)}$ are obtained by just taking the mode of the evolved prior pdf as follows:

$$\hat{\mu}^{(n)} = \hat{m}.$$  

(20)

It is noted that the updated parameter $\hat{\mu}^{(n)}$ is obtained by simply interpolating the ML estimate of
the new data with the prior speaker model estimated within the speaker space. The SSM evolution and the updated mean parameter are illustrated in Fig. 2. Given the SSM and a new block of adaptation data, the SSM evolution finds the updated mean parameter according to (20) and moves the SSM to a new position with the updated hyperparameters \( \phi^{(n)} \).

3.3. Comparison with QB adaptation with correlation

The proposed SSM evolution approach is similar to the QB adaptation of correlated CDHMM approach (Huo and Lee, 1998). In the QB adaptation approach, the CDHMM mean parameter has a joint normal distribution given by

\[
 p(\mu | \phi) \propto |U|^{-1/2} \exp \left\{ (\mu - \bar{\mu})^T U^{-1} (\mu - \bar{\mu}) \right\} \tag{21}
\]

where \( \bar{\mu} \) and \( U \) are the mean and covariance matrix of the supervector, respectively. This distribution has the mean and covariance matrix as the hyperparameters, i.e., \( \phi = (\mu, U) \). Based on the above prior distribution, the QB adaptation with correlation was developed to execute sequential adaptation where the prior statistics \( \phi^{(n)} \) are updated using QB learning as follows (Huo and Lee, 1998):

\[
 \mu_{QB}^{(n)} = (\rho U^{-1,(n-1)} + CS^{-1})^{-1} \cdot (\rho U^{-1,(n-1)} \mu^{(n-1)} + CS^{-1} \mu_{ML}) \tag{22}
\]

\[
 U_{QB}^{(n)} = (\rho U^{-1,(n-1)} + CS^{-1})^{-1} \tag{23}
\]

The QB estimate for the correlated CDHMM mean parameter in the QB adaptation is given by \( \mu_{QB}^{(n)} = \mu_{QB}^{(n)} \). However, it is difficult to directly implement the above formula due to the estimation of the huge covariance matrix \( U \). To overcome this problem, pairwise correlation information between different mean vectors are used in (Huo and Lee, 1998). Compared with the QB adaptation, the SSM evolution approach robustly estimates the hyperparameters based on the EM algorithm and updates those according to (17)–(19). Also, the SSM evolution indicates that the updated mean parameter \( \mu_{QB}^{(n)} \) in (20) is obtained by incorporating the prior speaker model estimated within the speaker space into the QB estimation with correlation given by (22).

3.4. Discussion

The SSM adaptation approaches are designed to perform batch adaptation by finding the MAP estimate of the CDHMM mean parameter \( \mu \) based on the FA and PPCA model. The expectation function found in the batch SSM adaptation is defined in the same way as (9) without the index \( n \). For that reason, the batch SSM solution of the CDHMM mean parameter is considered equivalent to (14) with the forgetting factor being set to 1. This confirms that the SSM evolution approach provides a general framework with the batch SSM approach as a special case.

4. Experiments

4.1. Database and recognition systems

To evaluate the performance of the proposed speaker adaptation algorithms, we used two kinds of speech databases provided by Korea Advanced Institute of Science and Technology (KAIST), Korea. The first task was a connected Korean digit
database, and the second task was a continuous speech database of 3000 Korean words for trade consulting. Both the two databases were recorded in a quiet environment.

Korean digits consist of 10 words with two pronunciations for zero as in English. They are highly confusable with each other and differ from the others by only one or two phonemes. Moreover, when they are spoken continuously, the articulation effects make it increasingly difficult to discriminate among them. For the first adaptation task, 939 utterances consisting of 4815 digits from 105 speakers (68 males and 37 females) constructed the training data and 3880 utterances consisting of 19891 digits from the other 35 speakers (25 males and 13 females) were used for evaluation. Each speaker contributed 30–40 sentences consisting of 3–7 digits and each sentence had an average length of 1.3 s. Each digit was modeled by a seven-state left-to-right HMM without skips and the 3 silence types were modeled by a one-state HMM. Four mixture Gaussians were used for representing the observation distribution in each state. In the recognition experiments, we drew 1–10 sentences from each target speaker for adaptation, and performed the recognition test with a loop grammar of equiprobable digits on the remaining sentences. Speech signal was sampled at 8 kHz and segmented into 30 ms frame at every 10 ms with 20 ms overlap. Each frame was parameterized by a 24-dimensional feature vector consisting of 12 mel-frequency cepstral coefficients and their first-order time derivatives. The speaker independent (SI) system with four mixture Gaussians produced 91.7% of word recognition rate.

For the second adaptation task, the baseline recognizer was trained with the training data consisting of 11,789 utterances (100,408 words) spoken by 120 speakers (80 males and 40 females), and 716 utterances (6106 words) from eight test speakers (five males and three females) were used for evaluation. Each test speaker provided 98 utterances, 20 of which were taken as adaptation data and the remaining 78 as test data. Each adaptation sentence consisted of 8.4 words on average and had an average length of three seconds. Context-dependent acoustic models were built by means of the decision tree state tying algorithm (Young et al., 1994), leading to a total of about 2000 distinct clusters of states. Four Gaussians were used to characterize the observation distribution in each state. A bigram model based on word class consisting of 760 classes was applied for the language model. The baseline system with four mixture Gaussians produced 83.0% of word recognition rate.

4.2. Training of SSM

To obtain the SSM by means of the FA and PPCA, we first trained a set of SI models over the speech from all the training speakers. We obtained the SD CDHMM mean parameters by adapting the SI parameters to each training speaker with the speaker-specific data by means of the joint application of the MLLR and MAP techniques. We extracted the mean supervectors by concatenating all the CDHMM mean vectors of the trained SD models. Consequently, the mean supervector is of dimension $D = \{11 \times 7 \times 24 + 3 \times 24\} \times 4 = 7680$ and $D = \{1991 \times 24\} \times 4 = 191,136$ for the digit and large vocabulary tasks, respectively. The order in which the mixture Gaussian parameter vectors are arranged in a supervector is naturally determined due to the combined MLLR and MAP training. Before applying the PPCA and FA techniques, the supervectors were normalized by their standard deviation to prevent the variables with large absolute value from dominating the analysis (Jolliffe, 1986; Kuhn et al., 2000). We estimated the parameters for the FA and PPCA, $\{W, \mu, A\}$ with dimension $P = 20$ for each task using the EM algorithm. For the initial parameters, we used the sample mean for $\mu$, the randomly generated numbers for $W$ and $A$. The number of EM iterations was fixed to 20. We observed that the EM algorithm converged within less than 10 iterations which indicates that the PPCA and FA approaches are very effective in building the SSM. With the same speaker space dimensionality $P$, the PPCA has $(D + 2 + D \times P) + 1$ parameters, while the FA requires $(D \times 2 + D \times P)$ parameters due to the unequal diagonal components of the covariance matrix.
4.3. Evaluation of batch adaptation methods

First, we compared the recognition performance of a variety of different batch adaptation methods when only the mean parameters were adapted. We carried out supervised batch adaptation experiments by using five different methods, i.e., (1) MAP (Gauvain and Lee, 1994), (2) MAPLR (Chou, 1999), (3) eigenvoice followed by MAP (Kuhn et al., 2000) (4) SSM by PPCA and (5) SSM by FA.

For the MAP and eigenvoice with MAP adaptation, \( \tau \) was set to 20. To obtain the MAPLR prior distribution, the conventional MLLR adaptation approach was applied to each training speaker to produce the SD regression matrices. Considering the amount of adaptation data available, only a single global regression class with a block-diagonal matrix was considered for the MLLR parameter estimation. The hyperparameters of the MAPLR were determined by taking the ensemble average and covariance over the SD regression matrices. To obtain the speaker space for the eigenvoice, conventional PCA was applied to the SD mean supervectors. The eigenvectors with corresponding to the largest eigenvalues were selected for speaker space \( P = 20 \). For the eigenvoice with MAP adaptation, we estimated the position in eigenspace once and then continued to apply MAP.

Figs. 3 and 4 show the performance of the conventional MAP, MAPLR, eigenvoice with MAP and SSM adaptation techniques for the digit and large vocabulary recognition tasks, respectively. Word recognition rates are displayed against the number of adaptation sentences and SI represents the performance of the baseline system in which the original SI model was used. Here “EV+MAP(\( P = 20 \))” represents the MAP method with the eigenvoice as prior with \( P = 20 \). In addition, “SSM (PPCA, \( P = 20 \))” and “SSM (FA, \( P = 20 \))” denote the batch SSM adaptation approaches based on the PPCA and FA with \( P = 20 \), respectively.

From the figures it is seen that the MAP was conservative, giving performance that was close to that of the SI system initially and gradually improving as more sentences were given. The MAPLR approach did not provide a good performance for a small amount of adaptation data, but better results were shown with more adaptation data. The eigenvoice combined with MAP performed very well for a small amount of adaptation sentences, and seemed to improve the performance as the number of adaptation sentences increased. On the other hand, the SSM adaptation approaches showed nearly the same performance as the eigenvoice followed by MAP for the digit task. For the large vocabulary task, the SSM
approaches showed a better performance than the eigenvoice with MAP as the adaptation sentence increased. The SSM approaches outperformed the MAPLR for all the adaptation conditions while the performance of the eigenvoice followed by MAP showed a similar performance to that of the MAPLR with more adaptation data for the large vocabulary task. From the results, we can observe that the SSM approach possesses both the rapid and consistent adaptation properties, and as a result it can efficiently perform speaker adaptation not only for a small adaptation data size but also when a large amount of data is available.

4.4. Comparison of different online adaptation approaches

For a comparative study, we implemented two other online adaptation approaches based on the QB paradigm. One is the online adaptation method denoted by OLA where the QB estimate of the CDHMM parameters were formulated. The other is the online linear regression adaptation method called QBLR where the regression matrix was updated according to the QB scheme. The hyperparameters obtained for the MAP and MLLR approaches were used for the initial hyperparameters of the OLA and QBLR, respectively. Only a single global regression class with a block-diagonal matrix was considered for the QBLR adaptation. All the adaptation approaches were evaluated in a supervised mode with the forgetting factor \( \rho = 1 \). We did not investigate the forgetting mechanism with various factors because it did not affect the recognition performance critically in our experiments. We used only one adaptation sentence in each sequential epoch to evaluate the rapid online performance for both the recognition tasks. Totally, 10 and 20 adaptation sentences were incrementally provided for the digit and large vocabulary recognition tasks, respectively.

Figs. 5 and 6 show the performance of the OLA, QBLR and SSM evolution by PPCA and FA for the digit and large vocabulary tasks, respectively. We can observe that the recognition rates of the OLA and QBLR were similar to those of the batch adaptation using MAP and MAPLR for both of the two tasks. Also, it is shown that online adaptation based on OLA is significantly inferior to that using the QBLR and SSM evolution approaches in the large vocabulary task. This comes from the fact that the OLA only adapts the CDHMM parameters observed during the adaptation session while the QBLR and SSM evolution approaches adapt all the CDHMM parameters by means of the regression matrix and correlation information. From Figs. 5 and 6, we can find that the online adaptation approach using the SSM evolution was significantly better than using the OLA and QBLR for both of the two tasks. For
instance, when two adaptation sentences were provided in the large vocabulary task, SSM evolution based on PPCA achieved a recognition rate of 86.2% which is higher than 83.1% of OLA and 85.3% of QBLR. In all the tested conditions, SSM evolution showed better performance than the OLA and QBLR approaches. The performance improvement results from the incorporation of the SSM which provides various sources of information useful for rapid adaptation. From the results, we can see that the SSM evolution is superior to the OLA and QBLR for rapid online adaptation. Hence, it is preferable to employ the SSM evolution approach for online adaptation with little adaptation data.

4.5. Comparison of batch SSM adaptation

We examined the recognition performance of the batch adaptation and online adaptation for the SSM approaches. Two adaptation tasks were evaluated in a supervised mode with the forgetting factor set to $\rho = 1$ and the parameters were updated for each adaptation sentence. Figs. 7 and 8 respectively show the performance of the SSM batch and online adaptation techniques with various amounts of adaptation data. In the case of the digit task with one adaptation sentence, SSM evolution was equivalent to batch SSM. From Figs. 7 and 8, we can find that the SSM evolution approach achieved nearly the same adaptation performance as that of the batch approach in all the adaptation conditions. The experimental results show that the SSM batch and online adaptation approaches which were based on the PPCA and FA performed equally well.

5. Conclusions

We have presented a novel SSM evolution technique for online adaptation of the CDHMM mean parameters. The SSM based on the FA or PPCA provides various sources of information useful for rapid online adaptation such as the correlation information, the prior distribution, and the prior knowledge of the training speakers related with the CDHMM parameters. The SSM evolution is formulated such that it can incrementally update the hyperparameters of the SSM as well as the mean parameters based on the QB learning. The effectiveness of the proposed approach has been proved on the Korean connected digit and large vocabulary recognition tasks. Experimental results showed that the performance of the batch SSM adaptation approaches was better than that of the MAPLR method, while it was comparable to that of the eigenvoice with MAP technique. It was also observed that the SSM evolution outperformed the OLA and QBLR approaches in a variety of adaptation conditions. From the results of a number of online speaker adaptation experiments, we can conclude that the
proposed approach is effective in speaker adaptation both for sparse and sufficient data.

References


