PAPER

Improved Reference Speaker Weighting Using Aspect Model

Seong-Jun HAHM[†a)], ***Student Member***, **Yuichi OHKAWA**[††], **Masashi ITO**[†], **Motoyuki SUZUKI**[†††], **Akinori ITO**[†], ***and*** **Shozo MAKINO**[†], ***Members***

SUMMARY We propose an improved reference speaker weighting (RSW) and speaker cluster weighting (SCW) approach that uses an aspect model. The concept of the approach is that the adapted model is a linear combination of a few latent reference models obtained from a set of reference speakers. The aspect model has specific latent-space characteristics that differ from orthogonal basis vectors of eigenvoice. The aspect model is a "mixture-of-mixture" model. We first calculate a small number of latent reference models as mixtures of distributions of the reference speaker's models, and then the latent reference models are mixed to obtain the adapted distribution. The mixture weights are calculated based on the expectation maximization (EM) algorithm. We use the obtained mixture weights for interpolating mean parameters of the distributions. Both training and adaptation are performed based on likelihood maximization with respect to the training and adaptation data, respectively. We conduct a continuous speech recognition experiment using a Korean database (KAIST-TRADE). The results are compared to those of a conventional MAP, MLLR, RSW, eigenvoice and SCW. Absolute word accuracy improvement of 2.06 point was achieved using the proposed method, even though we use only 0.3 s of adaptation data.

key words: *speaker adaptation, aspect model, reference speaker weighting, latent reference model*

1. Introduction

Speaker adaptation is an attractive field for commercialization of automatic speech recognition (ASR) systems. Concomitant with the development and improvement of the hidden Markov models (HMMs) approaches [1], [2], speech recognition systems have been shown to be functional for large vocabulary, continuous speech, and speaker-independent (SI) tasks. However, despite the high quality of SI systems, there remains a considerable gap in performance between these systems and their speaker-dependent (SD) counterparts. The difference in a system's error rate between SI and SD systems can be greater than 50% [3]. This gap arises from the wide variation that can be present in any speech waveform. This variation can result from changes in an individual speaker, environment, a microphone and a channel of the recording device. As described in this paper, we specifically examine the variability from different speakers.

In actual ASR systems, users want fast responses for the input utterances. Use of an SD system is an ideal way of recognizing specific speaker's utterances. However, in general, it is difficult to gather a sufficient amount of training utterances for a specific speaker for training an SD system. For example, in an automated call center or information desk, a certain user can disappear after using the system just one time. The conversation is usually short and the system has to start working from the first conversation or utterance. In this case, we cannot train the SD model for the user because we do not know the information of the user in advance. Therefore, speaker adaptation is a realistic way of obtaining a speech recognizer suitable to a specific user.

Model-based adaptation methods such as the speaker-clustering based methods [4], Bayesian-based maximum a posteriori (MAP) adaptation [5], and the transformation-based maximum likelihood linear regression (MLLR) adaptation [6] have been popular for many years. In such approaches, when the amount of adaptation data is small, reasonable performance cannot be obtained. Other approaches are necessary to reduce the number of adaptation parameters and to obtain reasonable performance for small amounts of adaptation data.

Reference speaker weighting (RSW) [3] was proposed to overcome such problems. Eigenvoices [7] were also proposed by extending the idea of the RSW. Both approaches are based on the reference speaker model. They differ only in the ways in which the reference vectors are computed. Both methods also assume that a new speaker model can be produced through a linear combination of the reference speakers' models. Eigenvoices employs eigen (principal component) analysis [8] to identify a set of orthogonal basis vectors. Other extended approaches have been proposed based on eigenspace such as eigen-MLLR [9], eigenspace mapping [10], and kernel eigenvoice [11], [12]. In eigenspace-based approach, the selection of eigenvectors is not based on likelihood of the training or adaptation utterances. Although the reference-speaker-based methods are designed to be effective for small adaptation data, those methods are not effective enough when the amount of adaptation data is extremely small (e.g. less than 1 s).

In this work, we assume that the adaptation is performed using a dedicated word like "Hello" in a supervised fashion. After adaptation using this short word, the adapted model is used. For actual ASR systems, adaptation must be

Manuscript received September 24, 2009.
Manuscript revised February 22, 2010.

[†]The authors are with the Graduate School of Engineering, Tohoku University, Sendai-shi, 980–8579 Japan.
[††]The author is with the Graduate School of Educational Informatics, Tohoku University, Sendai-shi, 980–8576 Japan.
[†††]The author is with Institute of Technology and Science, The University of Tokushima, Tokushima-shi, 770–8506 Japan.
a) E-mail: branden65@makino.ecei.tohoku.ac.jp

DOI: 10.1587/transinf.E93.D.1927

fast with very few adaptation data. To realize rapid adaptation, efficient approximation of inherent speaker-specific characteristics is needed using extremely small number of adaptation data. As described in this paper, we propose a Bayesian adaptation method, which exploits an aspect model: a "mixture-of-mixture" model. An aspect model obtained from a set of reference speakers is used for performance improvement of RSW. In the proposed framework, small number of "latent reference model" are trained first, which are mixtures of the distributions of the reference speakers. When adaptation data are given, the latent reference models are mixed so that the likelihood for the adaptation data is maximized. The mixture weights are determined based on the EM algorithm. Finally, the distributions of the mixture model are merged into a single distribution using the determined weights.

The organization of the paper is as follows. In Sect. 2, we review related works for speaker adaptation approaches. In Sect. 3, we describe an overview of the aspect model and discuss the potential of the techniques using the aspect model. In Sect. 4, we will present the experimental results obtained using MAP, MLLR, eigenvoice, RSW, SCW and the proposed method, and conclude the paper in Sect. 5.

2. Review of Related Works

SD models usually perform better than SI models. Speaker adaptation refers to the set of techniques that are used to modify a SI model to approximate SD models. In the following, important adaptation methods, MAP, MLLR, RSW, and eigenvoices are explained briefly.

2.1 Maximum a Posteriori Estimation

In most speech recognition systems using HMMs, the model parameters such as means and variances are estimated using maximum likelihood estimation (MLE). The MAP reduces the amount of training data by combining the original distributions and the distributions calculated from the adaptation data. The formula for MAP adaptation of mean parameters is the following:

$$\boldsymbol{\mu}_{new} = \frac{N_{adp}\boldsymbol{\mu}_{adp} + \tau\boldsymbol{\mu}_{ori}}{N_{adp} + \tau}, \tag{1}$$

where $\boldsymbol{\mu}_{new}$ is the updated mean, $\boldsymbol{\mu}_{adp}$ is the mean of the adaptation data, $\boldsymbol{\mu}_{ori}$ is the original mean, N_{adp} is the number of available adaptation data and τ is a control variable determined empirically. In Eq. (1), one can see if $\tau \to 0$ the updated mean is only dependent on the adaptation data (which is equivalent to the MLE). If $\tau \to \infty$, the updated mean keeps the original mean. The MAP method can be regarded as finding the optimal combination of existing data and the adaptation data [5].

2.2 Maximum Likelihood Linear Regression

MLLR adjusts model parameters using a transformation that is shared globally or across different units within a class. Global mean vector scaling, rotation, and translation are

$$\boldsymbol{\mu}_{new} = \mathbf{W}\boldsymbol{\mu}_{ori} + B, \tag{2}$$

where $\mathbf{W}$ is a regression matrix, and B stands for a bias term. A detailed explanation of the method can be found in [6].

2.3 Reference Speaker Weighting

The fundamental idea of the RSW is that the model parameters of a speaker adapted model can be constructed from a weighted combination of model parameters from a set of individual reference speakers [3]. Letting the reference speaker be r, the speaker vector for r is given as

$$\mathbf{m}_r = \begin{bmatrix} \boldsymbol{\mu}_{1,r} \\ \vdots \\ \boldsymbol{\mu}_{P,r} \end{bmatrix}, \tag{3}$$

where P is the number of distributions of the phonetic models. The entire set of reference speaker vectors can be represented by the matrix $\mathbf{M}$ which is defined as

$$\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_K], \tag{4}$$

where K is the number of reference speakers. The value of the updated mean vector $\mathbf{m}_{sa}$ can be constrained to be a weighted average of the speaker vectors contained in $\mathbf{M}$. This can be expressed as

$$\mathbf{m}_{sa} = \mathbf{M}\mathbf{w}. \tag{5}$$

Here $\mathbf{w}$ is a weighting vector that allows a new speaker vector to be created via a weighted summation of the reference speaker vectors in $\mathbf{M}$. The optimum weighting vector $\hat{\mathbf{w}}$ can be obtained using MLE.

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{X}|\mathbf{M}, \mathbf{w}), \tag{6}$$

where $\mathbf{X}$ signifies the adaptation data.

2.4 Eigenvoices

Eigenvoices extends the idea of the RSW. The goal is to learn uncorrelated features of the speaker space. The set of eigenvectors, $\mathbf{E}$, is obtained after applying eigen (principal components) analysis on matrix $\mathbf{M}$ in Eq. (4), as

$$\mathbf{E} = \{\mathbf{e}_0, \mathbf{e}_1, \cdots, \mathbf{e}_K\}, \tag{7}$$

where K is the number of reference speakers. The new speaker vector, the set of updated means, is combination of top N eigenvectors:

$$\mathbf{m}_{sa} = \mathbf{e}_0 + w_1\mathbf{e}_1 + \cdots + w_N\mathbf{e}_N. \tag{8}$$

The adaptation procedure for eigenvoices closely resembles RSW. The value of $\hat{\mathbf{w}}$ is calculated using maximum likelihood eigen-decomposition (MLED) [7].

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{X}|\mathbf{E}, \mathbf{w}). \tag{9}$$

2.5 Speaker Cluster Weighting

Speaker cluster weighting (SCW) [3] is an interpolation method which calculates the output probability as a weighted-sum of output probabilities of models, each of which is trained by a cluster of speakers. The weighting factors are determined on the fly to match the current speaker. Let $p(x|\Phi_l)$ represent the acoustic model trained from the l-th speaker cluster Φ_l. When we have L different clusters, then the final SCW model is a weighted combination of the L different models as represented by:

$$p_{scw}(x|\Phi, \mathbf{w}) = \sum_{l=1}^{L} w_l p(x|\Phi_l), \tag{10}$$

where

$$\mathbf{w} = \{w_1, \ldots, w_L\}. \tag{11}$$

The weights are determined by the maximum likelihood criterion, which is easily performed by the EM algorithm.

3. Speaker Adaptation Using an Aspect Model

3.1 Proposed Approach Using Aspect Model

Generally speaking, a large number of adaptation parameters can be a problem in speaker adaptation systems. The fundamental idea of the proposed method is that the adapted model is a linear combination of a few latent reference models obtained from a set of reference speakers. It therefore drastically reduces the number of free parameters to be estimated from the adaptation data. As explained, the basic strategy of the proposed method is similar to that of eigenvoice. The difference is that the proposed method is not a decomposition of a mean vector but a decomposition of a target distribution into mixtures of distributions of the latent reference models. A target distribution is calculated in the following three steps:

1. Calculate a small number of latent reference models as a mixture of Gaussian distribution of the reference speakers using the mixture weights λ_k.
2. Calculate the target distribution as a mixture of the aspect models using the mixture weight ξ_z.
3. Merge the Gaussian components of the target distribution into single Gaussian distribution.

The probability distribution function for the sample x is as follows. First, we consider adaptation of a distribution of a specific state. In this case, the probability distribution is expressed as

$$p(x|\Xi, \Lambda) = \sum_z \xi_z \sum_k \lambda_{k,z} \psi_k(x) \tag{12}$$

where

$$\Xi = \{\xi_1, \ldots, \xi_Z\}, \tag{13}$$

$$\Lambda = \{\lambda_{1,1}, \ldots, \lambda_{K,Z}\}. \tag{14}$$

The variable, $\lambda_{k,z}$, is the first-level weighting, ξ_z is the second-level weighting of the z-th latent reference model, and function $\psi_k(x)$ is a Gaussian distribution function, $\mathcal{N}(x; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Furthermore, k is a specific speaker, x is a feature vector, and z means a latent class. In an actual HMM, a distribution depends on a state s. Therefore, Eq. (12) is rewritten as follows.

$$p(x|\Xi_s, \Lambda_s) = \sum_z \xi_{z,s} \sum_k \lambda_{k,z,s} \psi_{k,s}(x), \tag{15}$$

where

$$\Xi_s = \{\xi_{1,s}, \ldots, \xi_{Z,s}\}. \tag{16}$$

$$\Lambda_s = \{\lambda_{1,1,s}, \ldots, \lambda_{K,Z,s}\}. \tag{17}$$

Here, $\lambda_{k,z,s}$ is a state-dependent weight from the speaker k to the latent reference z at the state s, ξ_z is the z-th state-dependent weight at the state s. If we regard the second-level weights ξ_z as being independent of the state, Eq. (15) becomes as follows.

$$p(x|\Xi, \Lambda_s) = \sum_z \xi_z \sum_k \lambda_{k,z,s} \psi_{k,s}(x). \tag{18}$$

In this case, $\lambda_{k,z,s}$ is a state-dependent weight, whereas ξ_z is independent of states. By sharing ξ_z among all states, we can adjust the models of all phonemes using very small number of parameters.

Figure 1 presents structure of the latent reference model. A latent reference model is a mixture model of the SD models. The mixture weights $\lambda_{k,z,s}$ are trained from the training samples using the EM algorithm. On adaptation, only the mixture weights of the aspect models ξ_z are estimated from the adaptation samples. As these weights are shared across all states, the number of adaptation parameters to be estimated is small, enabling adaptation using extremely small number of adaptation data.

Figure 2 shows a block diagram of the speaker adaptation system using an aspect model. In the training phase, SI and SD models are trained, respectively, using the training data. Using the SD model and training data, the aspect model is computed using the method explained in Sect. 3.2. In the adaptation phase, the original aspect model is adjusted using the adaptation data. Here the second-level weightings of each latent reference model, ξ_z, are the unit for adaptation instead of the SD model set. The adaptation is performed by linear combination using mean parameters from

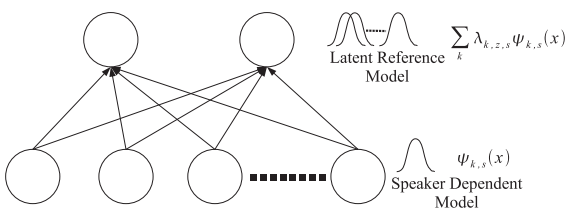


Fig. 1 Structure of latent reference model.

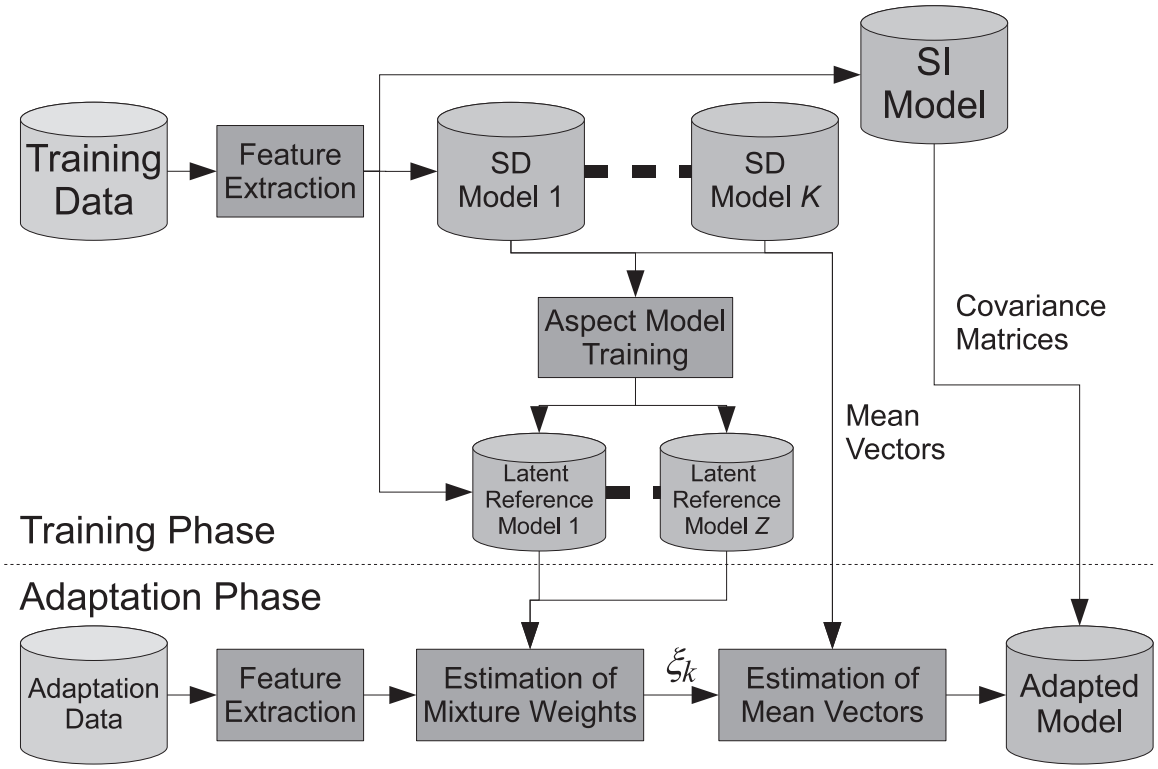


Fig. 2 Speaker adaptation system using an aspect model.

SD model set and weighting parameters from adjusted aspect model set. The variance parameters of SI model are used for adapted model with no change.

3.2 Derivation of EM Update Formulae for the Aspect Model (Training)

Figure 3 shows the training scheme for the aspect model. In that figure, x_i^j is the i-th feature vector labeled as the j-th speaker. The basic idea of training of the aspect models is that we train a few latent reference models so that mixture of the latent reference models can approximate each of the speaker models. As the number of the latent reference models is smaller than that of the speaker models, we can expect that the trained latent reference models are a kind of "basis" distributions that express any distribution. The optimization of "basis" distributions is based on the maximum likelihood criterion, which is the advantage of the proposed method over the eigenvoice.

First, let us think about estimating a distribution for a specific state. Definitions of symbols are as follows.

- $u_{i,j}$ Data for speaker of the i-th sample such that

$$u_{i,j} = \begin{cases} 1 & \text{if the speaker of the } i\text{-th sample is } j \\ 0 & \textit{otherwise} \end{cases}$$

- v_i A speaker of the i-th sample (i.e. $u_{i,v_i} = 1$)
- x_i The i-th sample
- $\psi_k(x)$ A pdf trained for speaker k

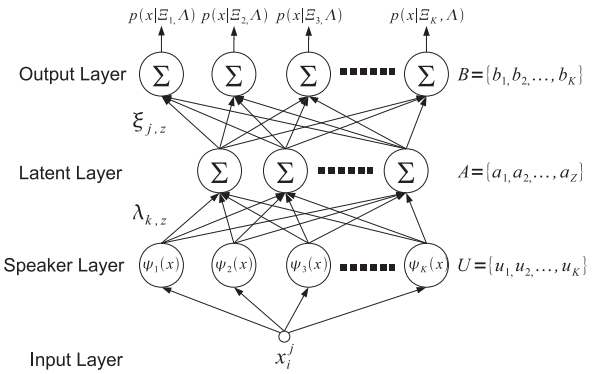


Fig. 3 Training procedure for the aspect model.

We define the probability distribution function for the speaker j and sample x as

$$p(x|\Xi_j,\Lambda) = \sum_{z=1}^{Z} \xi_{j,z} \sum_{k=1}^{K} \lambda_{k,z} \psi_k(x), \tag{19}$$

where

$$\Xi_j = \left\{\xi_{1,1},\ldots,\xi_{j,Z}\right\}. \tag{20}$$

The complete data are assumed as $\Gamma = \{a_{i,z}, b_{i,k}\}$, where

$$a_{i,z} = \begin{cases} 1 & \text{if the } z\text{-th latent layer is} \\ & \text{selected at the } i\text{-th sample} \\ 0 & \textit{otherwise} \end{cases},$$

$$b_{i,k} = \begin{cases} 1 & \text{if the } k\text{-th output layer is} \\ & \text{selected at the } i\text{-th sample} \\ 0 & \textit{otherwise}. \end{cases}$$

Letting $U = \{v_1, \ldots, v_N\}$, the probability of the samples and the complete data are

$$p(X, \Gamma|U, \theta) = \prod_i \prod_j \prod_z \prod_k \left(\xi_{j,z}\lambda_{k,z}\psi_k(x_i)\right)^{u_{i,j}a_{i,z}b_{i,k}}, \tag{21}$$

$$\log p(X, \Gamma|U, \theta) = \sum_i \sum_j \sum_z \sum_k u_{i,j}a_{i,z}b_{i,k}\left(\log \xi_{j,z} + \log \lambda_{k,z} + \log \psi_k(x_i)\right). \tag{22}$$

The expectations of $a_{i,z}$ and $b_{i,k}$ are calculated next.

$$\alpha_{i,z} = E\left[a_{i,z}\right] = \frac{\xi_{v_i,z}\sum_k \lambda_{k,z}\psi_k(x_i)}{\sum_z \xi_{v_i,z}\sum_k \lambda_{k,z}\psi_k(x_i)}, \tag{23}$$

$$\beta_{i,k} = E\left[b_{i,k}\right] = \frac{\sum_z \xi_{v_i,z}\lambda_{k,z}\psi_k(x_i)}{\sum_k \sum_z \xi_{v_i,z}\lambda_{k,z}\psi_k(x_i)}, \tag{24}$$

$$E\left[\log p(X, \Gamma|U, \theta)\right] = \sum_i \sum_j \sum_z \sum_k u_{i,j}\alpha_{i,z}\beta_{i,k}\left(\log \xi_{j,z} + \log \lambda_{k,z} + \log \psi_k(x_i)\right). \tag{25}$$

Let the Q-function be

$$Q = E\left[\log p(X, \Gamma|U, \theta)\right] + \sum_{j=1}^{K} c_j\left(1 - \sum_z \xi_{j,z}\right) + \sum_{z=1}^{Z} d_z\left(1 - \sum_k \lambda_{k,z}\right). \tag{26}$$

From

$$\frac{dQ}{d\lambda_{k,z}} = \sum_i \sum_j \frac{u_{i,j}\alpha_{i,z}\beta_{i,k}}{\lambda_{k,z}} - d_z = 0, \qquad \frac{dQ}{d\xi_{j,z}} = \sum_i \sum_k \frac{u_{i,j}\alpha_{i,z}\beta_{i,k}}{\xi_{j,z}} - c_j = 0, \tag{27}$$

the optimal $\lambda_{k,z}$ and $\xi_{j,z}$ can be found as

$$\lambda_{k,z} = \frac{\sum_i \alpha_{i,z}\beta_{i,k}}{\sum_k \sum_i \alpha_{i,z}\beta_{i,k}}, \tag{28}$$

$$\xi_{j,z} = \frac{\sum_{i:v_i=j} \alpha_{i,z}}{\sum_z \sum_{i:v_i=j} \alpha_{i,z}}. \tag{29}$$

After training $\lambda_{k,z}$ and $\xi_{j,z}$, only $\lambda_{k,z}$ are saved for calculation of the aspect models. $\xi_{j,z}$ are not used for the adaptation. We use the average of $\xi_{j,z}$ over j as initial values of the adaptation. Only ξ_z is adjusted for adaptation. The second-level weightings of the aspect models, ξ_z, are shared globally over the entire phoneme models. Following ξ_z is used for the initial value.

$$\xi_z = \frac{1}{K}\sum_{j=1}^{K} \xi_{j,z}. \tag{30}$$

When using the aspect model for HMM, we need to apply the above-mentioned method to distributions in the many states. In this case, we use state-dependent $\lambda_{k,z}$ (i.e. $\lambda_{k,z,s}$) and state-independent $\xi_{j,z}$. To estimate these parameters, $\alpha_{i,z}$ and $\beta_{i,k}$ are also changed into state-dependent (i.e. $\alpha_{i,z,s}$ and $\beta_{i,k,s}$). Note that only averages of $\xi_{j,z}$ over j are used for the adaptation process.

3.3 Adaptation Using Aspect Model

For adaptation of the aspect model, EM algorithm is applied for estimating $\bar{\xi}_z$, which is the updated ξ_z. When the adaptation data $y_1, y_2, \ldots, y_n$ are given, $\bar{\xi}_z$ is calculated as

$$\bar{\xi}_z^{(n+1)} = \frac{\sum_s \sum_i \bar{\xi}_z^{(n)} \sum_k \lambda_{k,z,s}\psi_{k,s}(y_i^{(s)})}{\sum_z \sum_s \sum_i \bar{\xi}_z^{(n)} \sum_k \lambda_{k,z,s}\psi_{k,s}(y_i^{(s)})}. \tag{31}$$

where n is the number of iterations, and

$$\psi_{k,s}(y_i^{(s)}) = \begin{cases} \psi_{k,s}(y_i) & \text{if } y_i\text{belongs to state } s \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

After estimating $\bar{\xi}_z$, we obtain a mixture model adapted to the data as

$$p(x|\bar{\Xi}, \Lambda_s) = \sum_{z=1}^{Z} \bar{\xi}_z \sum_{k=1}^{K} \lambda_{z,k,s}\psi_{k,s}(x). \tag{33}$$

This distribution can be viewed as a mixture of $\psi_k(x)$ as

$$p(x|\bar{\Xi}, \Lambda_s) = \sum_{k=1}^{K} w_{k,s}\psi_{k,s}(x) \tag{34}$$

where

$$w_{k,s} = \sum_{z=1}^{Z} \bar{\xi}_z\lambda_{k,z,s}. \tag{35}$$

We can use this mixture model directly; however, when simply using this mixture distribution, number of mixture components becomes large when large number of reference speakers is used. Therefore, we merge the distributions using the weight $w_{k,s}$. Here, the means for a new speaker are a linear combination of reference speaker models.

$$\boldsymbol{\mu}_{new}^{(s)} = \sum_{k=1}^{K} w_{k,s}\boldsymbol{\mu}_k^{(s)} \tag{36}$$

where $\boldsymbol{\mu}_{new}^{(s)}$ signifies the updated mean of the distribution of the state s and $\boldsymbol{\mu}_k^{(s)}$ denotes the mean of a specific speaker, which is the mean vector of the distribution $\psi_{k,s}$. The covariance matrix of the adapted model, $\boldsymbol{\Sigma}_{new}^{(s)}$, comes from the SI model.

3.4 Relationship between the Aspect Model and Other Adaptation Methods

As explained, the proposed method finally calculates a mean vector of an adapted distribution as a weighted sum of those

of the distributions of the reference speakers, and thus this method can be viewed as a variant of the reference speaker weighting. The original RSW tries to estimate all w_k using the adaptation data, which is difficult when number of reference speaker is large or amount of adaptation data is small, even if state-independent weights are employed. The eigenvoice method decomposes w_k as follows:

$$\boldsymbol{\mu}_{new}^{(s)} = \sum_{i=1}^{N} w_i' \sum_{j=1}^{K} v_{i,j}^{(s)} \boldsymbol{\mu}_k^{(s)}. \tag{37}$$

where $v_{i,j}^{(s)}$ is a weight for calculating j-th eigenvector. This formula has similar form as the proposed method. The difference is that the decomposition is performed based on the least mean squares criterion, while the decomposition in the proposed method is based on the maximum likelihood criterion.

As shown in Fig. 1, the aspect model can be viewed as a combination of RSW and SCW. In fact, RSW is an interpolation of models from "reference speakers," and SCW is the cluster's mixture models. From this point of view, $\lambda_{k,z}$ can be thought of as weighting of reference speakers and ξ_z can be regarded as weighting of speaker clusters if we assume that $\lambda_{k,z}$ is the speaker cluster. In our approach, RSW is performed in training phase using training utterances across the different aspect model. Then SCW is performed using the adjusted aspect model set in the adaptation phase.

4. Experimental Evaluation

4.1 Training and Evaluation Data

The rapid speaker adaptation performance of the aspect model was tested using the Korean (KAIST-TRADE) database [13] in a supervised fashion. The KAIST-TRADE database consists of 150 speakers (100 male speakers and 50 female speakers) and 14,746 sentences. Each speaker utters about 100 sentences. The speech in the database was recorded in an office environment with sampling rate of 16 kHz. Among the 100 male speakers, 90 males were used for training. For testing and adaptation, 10 male speakers not included in the training set were used. The speech data uttered by these 10 male speakers were divided into two groups for testing and adaptation. Eighty sentences (from the 1st to 80th sentence) were used for testing, and the remaining sentences (from the 81st to 100th sentence) were used for adaptation for each speaker. Each speaker utters different sentences from the sentences spoken by other speakers. The amount of adaptation speech is from 0.1 s to 20 s. We performed two sets of evaluation. For the first evaluation, the adaptation data were started at an each speaker's 81st sentence, in the ascending order (after using all samples of the 81th sentence, the next sample was taken from the 82th sentence). For the second evaluation, the adaptation was performed using the last sentence as the starting sentence, in the descending order (the 99th sentence was used after using the 100th sentence). The adapted model was tested for each speaker's speech for testing (i.e., 80 sentences from the 1st to 80th sentence). Finally, the two adaptation results were averaged. For calculating linguistic scores, a trigram language model was used for experiments. Trigram language model was trained using 13,751 sentences which were not included in testing and adaptation set (995 sentences). We used CMU-Cambridge toolkit [14] for training and Witten-Bell discounting [15] as a smoothing method.

4.2 Acoustic Modeling

A 13-dimensional Mel-frequency cepstral coefficients (MFCCs) feature vectors including frame log power were extracted from the pre-emphasized speech signal every 10 ms using a 25 ms Hamming window. The MFCCs, ΔMFCCs and ΔΔMFCCs were concatenated to form 39-dimensional feature vectors. Cepstral mean normalization was used. The SI model consists of 37 monophones. Each was modeled as a continuous density HMM which is strictly left-to-right and has three states with one Gaussian mixture density per state. As shown in Eq. (36), we also employed single-Gaussian HMMs for all of the adapted HMMs. The reason why we used the single Gaussian models was for verifying the effectiveness of the proposed method, and comparing with RSW and eigenvoices. The RSW and eigenvoices are not well defined for more than 2-mixture acoustic models, in which the order of the reference speaker vectors in matrix $\mathbf{M}$ in Eq. (4) cannot be defined appropriately.

The number of aspect model was set to 5, 10, 20, and 40. Each SD model was created using a typical EM training procedure using SI model as an initial model. We used the large vocabulary speech recognition decoder: Julius rev.4.1 [16]. The SI model has a word accuracy of 74.50% on the test data.

4.3 Effect of the Number of Aspect Models

The idea of using the aspect model is to make use of the most important latent information to reduce the number of estimation parameters. In this experiment, we investigate the effect of the different number of aspect models.

The number of reference speaker model is the same as the number of training speaker (i.e., 90). For a more detailed evaluation, the x-axis units are set not to sentences but to seconds. The performance was evaluated using word accuracy. The results for global weighting are portrayed in Fig. 4. The figure shows the following facts.

- Although the model with 40 aspect models is better than that with 5 aspect models, effects of the different number of aspect models are not very large.
- The performance shows a slightly increasing tendency across the number of adaptation data.
- The performance saturates at about 3 s in most cases.
- However, the performance rapidly reached to the best one for all cases.

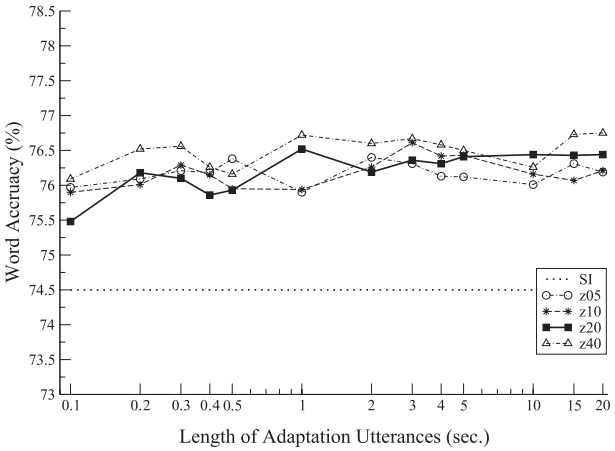


Fig. 4 Effect of the number of aspect models (Global weighting).

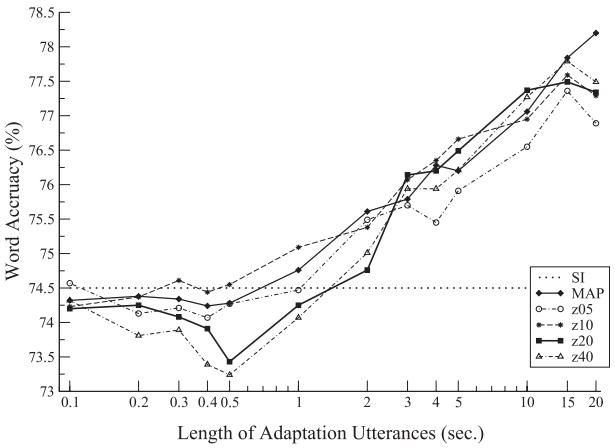


Fig. 5 Effect of the number of aspect models (State-dependent weighting).

To investigate the reason why the proposed method showed very slow improvement of performance with respect to the amount of adaptation data, we tried to apply the proposed method in the state-dependent manner. Figure 5 shows the results using state-dependent weighting, where the state-dependent weights $\xi_{z,s}$ were used. In this experiment, number of parameters to be adapted was 37(number of phonemes) $\times$ 3(number of states) $\times$ Z. The performance of this method was almost same as those of MAP even though only weightings are adjusted for adaptation. The performance was starting to increase from about 0.5 s.

This result depicts that improved performance can be obtained using long adaptation data when a large number of parameters are used. In other words, number of parameters shown in Fig. 4 (from 5 to 40) is too small to obtain improvement using long adaptation data compared with that shown in Fig. 5 (from 555 to 4440). At the same time, we can confirm that the proposed method with global weights outperforms the method with state-dependent weights when the adaptation data is short (under 3 s), showing that the parameter reduction works effectively for extremely small adaptation data.

4.4 Comparison with Existing Adaptation Methods

Next, MAP, MLLR, RSW, eigenvoice, and SCW were used for comparison with the proposed method. In all adaptation methods, only mean parameters were updated. For MAP, the adjustment parameter τ was set to 35 which was decided empirically. The original mean (i.e., SI mean) and the mean of adaptation data are mixed using this τ. For the MLLR, the global transformation matrix was used for adaptation. Using adaptation data, global transformation matrix which has the size of 39 $\times$ (39+1) was created for adaptation. The global weighting vector was used for both RSW and eigenvoice adaptation. The optimum weighting vector for RSW was obtained using MLE. For the eigenvoice adaptation, we used 45 eigenvoices whose cumulative contributions were greater than 80%. The optimum weighting vector for eigenvoices was estimated using MLED.

On applying the SCW, we first have to cluster all speakers. There are a variety of ways in which the speaker clustered tree can be constructed. The construction can be performed using unsupervised bottom-up clustering based on an acoustic similarity measure [4], [17], unsupervised top-down clustering based on an acoustic similarity measure [18], [19], or some supervised method. In this paper, we used a top-down clustering based on Bhattacharyya distance. In the constructed tree-structure, the root node is identical to SI model and the leaf node is the same as SD model. The depth of levels is 13 and the total number of node is 179. After constructing tree-structure, each node model which is composed of 3-state left-to-right 1-mixture monophone model is trained using MLE. For comparison, an adapted model were also merged into 1-mixture HMM. Therefore, adaptation is a linear combination of node models in tree-structured clusters. Table 1 shows the number of adaptation parameters of adaptation methods used for experiments. The experimental results are presented in Fig. 6. Results show that all other methods, especially MLLR, suffer from data sparseness when the amount of adaptation data is extremely small. MLLR methods yielded worse results than SI model when the amount of adaptation data were less than 15 s. For adaptation utterances longer than about 3 s, the performance of the proposed method was not better than those of other approaches. However, it shows that the aspect models provided the best adaptation performance when the adaptation utterances were extremely short (less than 0.5 s). From 0.5 s to 2 s, the proposed method has similar performance with SCW even though we use only 40 aspect models. Using only 0.3 s of adaptation data for 20 and 40 aspect models, word accuracy improvements of 1.60 and 2.06 points were achieved, respectively, from SI model using the proposed method.

The experimental results show that aspect model can represent speaker's characteristics effectively for extremely small amount of adaptation data (less than 1 s). We could obtain the improved results by using the estimated weight-

Table 1 Comparison of the number of parameters of adaptation methods.

Adaptation Methods	Number of Adaptation Parameters
MAP	4329 (37 phonemes × 3 states × 39 dimensions)
MLLR	1560 (39 × 40)
RSW	90 weighting vectors
Eigenvoice	45 eigenvoices
SCW	179 weighting vectors
Proposed	5, 10, 20, and 40 weighting vectors

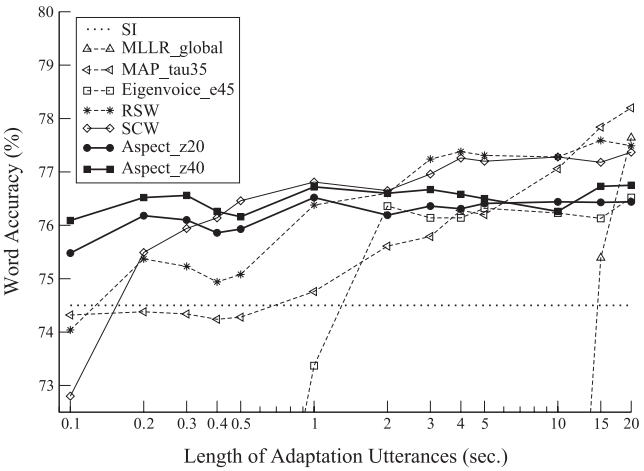


Fig. 6 Comparison with existing adaptation methods.

ing values based on EM algorithm. The small number of weights for the latent reference models can adjust a large number of parameters. The number of free parameters for adaptation is really small because only ξ_z has to be estimated in the adaptation phase.

5. Conclusions

In this paper, we proposed an improved reference speaker weighting using the aspect model based on likelihood scores derived from training utterances. The aspect model is a "mixture-of-mixture" model, which first calculates a small number of latent reference models as mixtures of distributions of the SD models. We then estimated the weights of the latent reference models using available adaptation data to obtain the adapted distribution. We used the obtained mixture weights for interpolating weights of mean parameters of the distributions. The number of free parameters to be estimated from adaptation data was reduced by the use of the aspect model. We evaluated performance of the proposed method through a speaker adaptation experiment. Even though we used only 0.3 s of adaptation data and 40 aspect models, word accuracy improvement of 2.06 point was achieved from SI model using the proposed method. Future work will involve performing experiments on many speakers and in various noisy environments.

References

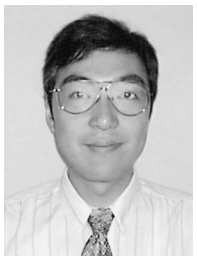
[1] L.R. Bahl, F. Jelinek, and R.L. Mercer, "A maximum likelihood approach to continuous speech recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.PAMI-5, no.2, pp.179–190, March 1983.
[2] L. Rabiner and B. Juang, Fundamentals of Speech Recognition, Prentice-Hall International, 1993.
[3] T. Hazen, "A comparison of novel techniques for rapid speaker adaptation," Speech Commun., vol.31, no.1, pp.15–33, 2000.
[4] T. Kosaka and S. Sagayama, "Tree-structured speaker clustering for fast speaker adaptation," Proc. ICASSP, pp.245–248, 1994.
[5] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observation of markov chains," IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291–298, 1994.
[6] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," Comput. Speech Lang., vol.9, no.2, pp.171–185, 1995.
[7] R. Kuhn, J.C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," IEEE Trans. Speech Audio Process., vol.8, no.6, pp.695–707, 2000.
[8] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, 1986.
[9] K.T. Chen, W.W. Liau, H.M. Wang, and L.S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," Proc. ICSLP, pp.742–745, 2000.
[10] B. Zhou and J. Hansen, "Rapid discriminative acoustic model based on eigenspace mapping for fast speaker adaptation," IEEE Trans. Speech Audio Process., vol.13, no.4, pp.554–564, July 2005.
[11] B. Mak, J.T. Kwok, and S. Ho, "Kernel eigenvoice speaker adaptation," IEEE Trans. Speech Audio Process., vol.13, no.5, pp.984–992, Sept. 2005.
[12] B. Mak and S. Ho, "Various reference speakers determination methods for embedded kernel eigenvoice speaker adaptation," Proc. ICASSP, pp.981–984, 2005.
[13] I.J. Choi, O.W. Kwon, and J.R. Park, "A Korean speech database for use in automatic translation," Proc. 11th Workshop on Speech Communication and Signal processing, pp.287–290, 1994.
[14] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-Cambridge tookit," Proc. EUROSPEECH, pp.2707–2710, 1997.
[15] I. Witten and T. Bell, "The zero frequency problem: Estimating the probabilities of novel events in adaptive text compression," IEEE Trans. Inf. Theory, vol.37, no.4, pp.1085–1094, 1991.
[16] A. Lee, T. Kawahara, and K. Shikano, "Julius — An open source real-time large vocabulary recognition engine," Proc. EUROSPEECH, pp.1691–1694, 2001.
[17] T. Kosaka, S. Matsunaga, and S. Sagayama, "Tree-structured speaker clustering for speaker-independent continuous speech recognition," Proc. ICSLP, pp.1375–1378, 1994.
[18] L. Mathan and L. Miclet, "Speaker hierarchical clustering for improving speaker-independent HMM word recognition," Proc. ICASSP, pp.149–152, 1990.
[19] S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," Proc. ICASSP, pp.286–289, 1989.

Seong-Jun Hahm was born in Seoul, Korea in 1976. He received B.S. degree from Kookmin University, Seoul, Korea in 2003 and M.S. degree from Yeungnam University, Gyeongsan, Korea in 2006, respectively. He is currently pursuing the Ph. D. degree in Tohoku University, Sendai, Japan. His research interests include rapid speaker adaptation and speaker recognition. He is a member of the Acoustical Society of Japan.

Yuichi Ohkawa was born in Osaka, Japan in 1975. He received B.E., M.E. and Ph.D. degrees from Tohoku University, Sendai, Japan in 1998, 2000 and 2006, respectively. Since 2003, he has worked with the Graduate School of Engineering, Tohoku University as a research associate. He is now an Assistant Professor of the Graduate School of Educational Informatics, Tohoku University. He has been engaged in spoken language processing and educational technology. He is a member of the Acoustical Society of Japan.

Masashi Ito was born in Fukushima, Japan in 1970. He received B.E., M.E., and Ph. D. degrees from Tohoku University, Sendai, Japan, in 1994, 1996, and 2006, respectively. From 1996 to 2004 he worked with Wako Research Center, Honda R & D Co., Ltd., Japan. He is now an Assistant Professor of the Graduate School of Engineering, Tohoku University. His interests include speech perception, processing, and recognition. He is a member of the Acoustical Society of Japan.

Motoyuki Suzuki was born in Chiba, Japan in 1970. He received B.E., M.E. and Ph. D. degrees from Tohoku University, Sendai, Japan, in 1993, 1995 and 2004, respectively. Since 1996, he has worked with the Computer Center and the Information Synergy Center, Tohoku University as a research associate. From 2006 to 2007 he worked with the Centre for Speech Technology Research, University of Edinburgh, UK, as a visiting researcher. He is now an Associate Professor of the Institute of Technology and Science, The University of Tokushima. His interests include spoken language processing, music information retrieval and pattern recognition using statistical modeling. He is a member of the Acoustical Society of Japan and the Information Processing Society of Japan.



Akinori Ito was born in Yamagata, Japan in 1963. He received B.E., M.E. and Ph. D. degrees from Tohoku University, Sendai, Japan in 1984, 1986 and 1992, respectively. Since 1992, he has worked with the Research Center for Information Sciences and Education Center for Information Processing, Tohoku University. He was with the Faculty of Engineering, Yamagata University from 1995 to 2002. From 1998 to 1999 he worked with the College of Engineering, Boston University, MA, USA as a visiting scholar. He is now an Associate Professor of the Graduate School of Engineering, Tohoku University. He has engaged in spoken language processing, statistical text processing and audio signal processing. He is a member of the Acoustical Society of Japan, the Information Processing Society of Japan and the IEEE.

Shozo Makino was born in Osaka, Japan on January 3, 1947. He received B.E., M.E. and Dr. Eng. degrees from Tohoku University, Sendai, Japan in 1969, 1971 and 1974, respectively. Since 1974, he has worked with the Research Institute of Electrical Communication, Research Center for Applied Information Sciences, Graduate School of Information Science, Computer Center and Information Synergy Center, as a Research Associate, an Associate Professor and a Professor. He is now a Professor of the Graduate School of Engineering, Tohoku University. He has been engaged in spoken language processing, CALL systems, autonomous robot systems, speech corpora, music information processing, image recognition and understanding, natural language processing, semantic web searches and digital signal processing.