ASPECT-MODEL-BASED REFERENCE SPEAKER WEIGHTING

Seongjun Hahm\textsuperscript{1}, Yuichi Ohkawa\textsuperscript{2}, Masashi Ito\textsuperscript{1}, Motoyuki Suzuki\textsuperscript{3}, Akinori Ito\textsuperscript{1} and Shozo Makino\textsuperscript{1}

\textsuperscript{1}Graduate School of Engineering, Tohoku University, Japan
\textsuperscript{2}Graduate School of Educational Informatics, Tohoku University, Japan
\textsuperscript{3}Institute of Technology and Science, The University of Tokushima, Japan

\{branden65, itojin, aito, makino\}@makino.ecei.tohoku.ac.jp, kuri@ei.tohoku.ac.jp, moto@m.ieice.org

ABSTRACT

We propose an aspect-model-based reference speaker weighting. The main idea of the approach is that the adapted model is a linear combination of a set of reference speakers like reference speaker weighting (RSW) and eigenvoices. The aspect model is the mixture model of speaker-dependent (SD) models. In this paper, aspect model weighting (AMW) is proposed for finding an optimal weighting of a set of reference speakers unlike RSW and the aspect model which is a kind of cluster models is trained based on likelihood maximization with respect to the training data. The number of adaptation parameters can also be reduced using aspect model approach. For evaluation, we carried out an isolated word recognition experiment on Korean database (KLE452). The results were compared to those of conventional MAP, MLLR, RSW, and eigenvoices. Even though we use only 0.5s of adaptation data, 27.24\% relative error rate reduction in comparison with speaker-independent (SI) baseline performance was achieved.

Index Terms— Speech Recognition, Speaker Adaptation, Reference Speaker Weighting, Aspect Model Weighting

1. INTRODUCTION

Speech is a key factor for human-machine communication. Speech recognition system has been developed for the human’s demand; free human and machine communication. With the development and improvement of the hidden Markov models (HMMs) approach \cite{1}, speech recognition systems have been shown to be functional for large vocabulary continuous speech recognition (LVCSR) with speaker-independent (SI) task. However, despite the high quality of LVCSR systems, there remains a considerable gap in performance between SI and speaker-dependent (SD) systems \cite{2}. This gap arises from the wide variability that can be present in any speech waveform. As described in this paper, we focus on the variability from different speakers. Use of an SD system is a perfect way for 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. Therefore, speaker adaptation is a realistic way of obtaining a speech recognizer suitable for a specific user.

There are three kinds of major approaches for speaker adaptation systems. Model-based adaptation methods such as the speaker-clustering based methods \cite{3}, Bayesian-based maximum a posteriori (MAP) adaptation \cite{4}, and the transformation-based maximum likelihood linear regression (MLLR) adaptation \cite{5}. In such approaches, a sufficient amount of adaptation data is required for the reasonable performance. Other approaches are necessary to obtain reasonable performance for small amount of adaptation data. Reducing the number of adaptation parameters is also one of the important issues for fast speaker adaptation. Reference speaker weighting (RSW) \cite{2} was proposed to overcome such problems. Eigenvoices \cite{6} were also proposed by extending the idea of the RSW. Both approaches are based on the reference speaker model. Both methods also assume that a new speaker model can be produced through a linear combination of the reference vectors. Eigenvoices employs eigen (principal component) analysis to find a set of orthogonal basis vectors. 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 1s).

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 improving performance of RSW. In the proposed framework, small number of “aspect models” are trained first, which are mixtures of the distributions of the reference speakers. When adaptation data are given, the aspect models are mixed so that the likelihood for the adaptation data is maximized. The mixture weights are determined using aspect model weighting (AMW) based on EM algorithm. Finally, the distributions of the mixture model are merged into single distributions using the determined weights.

The organization of the paper is as following. In Section 2, we review a RSW adaptation approach. In Section 3, we describe an overview of the aspect model and formulate and discuss the potential of the techniques using the aspect model. In Section 4, we will present the experimental results obtained using MAP, MLLR, RSW, eigenvoices, and the proposed method, and conclude the paper in Section 5.

2. REFERENCE SPEAKER WEIGHTING

The fundamental idea of the reference speaker weighting 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 \cite{2}. Letting the speaker vector for reference speaker be defined as $\mathbf{r}$, then $\mathbf{m}_r$ is given as

$$\mathbf{m}_r = [\mu_{1,r}, \mu_{2,r}, \cdots, \mu_{P,r}]^T,$$

(1)

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

$$\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_R],$$

(2)
where $R$ is the number of reference speakers. The value of $m_{ma}$ can be constrained to be a weighted average of the speaker vectors contained in $M$. This can be expressed as

$$m_{ma} = Mw.$$  

Here $w$ is a weighting vector that allows a new speaker vector to be created via a weighted summation of the reference speaker vectors in $M$. The weighting vector $w$ is obtainable using MLE.

$$w = \arg \max_w p(X|M, w),$$  

where $X$ signifies the adaptation data.

### 3. SPEAKER ADAPTATION USING AN ASPECT MODEL

#### 3.1. Proposed approach using aspect model

Generally numerous 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 set of reference speakers. The proposed approach can drastically reduce the number of free parameters to be estimated from adaptation data using a few aspect models. The basic strategy of the proposed method is similar to that of eigenvoice. The differences are 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 reference speakers. A target distribution is calculated in the following three steps:

1. Calculate a small number of aspect models as a mixture of Gaussian distributions of the reference speakers using the mixture weights, $\lambda_{k,z}$.
2. Calculate the target distribution as a mixture of the aspect models using the mixture weight, $\xi_{z,j}$.
3. Merge the Gaussian components of the target distribution into single Gaussian distribution.

#### 3.2. Derivation of EM update formulae for the aspect model

**Training**

The basic idea of training the aspect models is that we train a few aspect models so that mixture of the aspect models can approximate each of the reference speaker models. As the number of the aspect models is smaller than that of the reference speaker models, we can expect that the trained aspect 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. In the training phase, aspect models are trained based on the likelihood of training utterances. The purpose is the approximation of the output layer with respect to speaker layer using latent layer. The training procedure for the aspect model is shown in Fig. 1. In that figure, $\lambda_{k,z}$ is the first-level weighting, $\xi_{z,j}$ is the second-level weighting of the $z$-th aspect model. Furthermore, $k$ is a specific speaker, $x^j_i$ is the $i$-th feature vector labeled as the $j$-th speaker, and $z$ stands for a latent class. First, let us think about estimating a distribution for a specific state. Definitions of symbols are as follows.

- $u_{i,z}$: Data for speaker of the $i$-th sample such that
  $$u_{i,z} = \begin{cases} 1 & \text{if the speaker of the } i \text{-th sample is } z \\ 0 & \text{otherwise} \end{cases}$$

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

where the function $\psi_k(x)$ is a Gaussian distribution function, $\mathcal{N}(x; \mu_k, \Sigma_k)$. We define the probability distribution function for the speaker $j$ and sample $x$ as

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

where

$$\Xi_j = \{\xi_{1,j}, \ldots, \xi_{z,j}\}.$$  

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

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

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

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

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

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

From

$$\frac{dQ}{d\lambda_{k,z}} = \sum_i \sum_j u_{i,z} a_{i,z} b_{i,k} \frac{\partial \log \lambda_{k,z}}{\partial \lambda_{k,z}} - d_z = 0,$$

$$\frac{dQ}{d\xi_{z,j}} = \sum_i \sum_k u_{i,z} a_{i,z} b_{i,k} \frac{\partial \log \xi_{z,j}}{\partial \xi_{z,j}} - c_j = 0,$$

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

$$\lambda_{k,z} = \sum_i a_{i,z} b_{i,k} \frac{\sigma_{i,z} \beta_{i,k}}{\sum_k a_{i,z} \beta_{i,k}},$$

$$\xi_{z,j} = \sum_i u_{i,z} b_{i,k} \frac{\sigma_{i,z} \alpha_{i,z}}{\sum_j u_{i,z} b_{i,k} \sigma_{i,z} \alpha_{i,z}},$$

where $\alpha_{i,z}$ and $\beta_{i,k}$ are the expectations of $a_{i,z}$ and $b_{i,k}$. After training $\lambda_{k,z}$ and $\xi_{z,j}$, only $\lambda_{k,z}$ are saved for calculation of the aspect
models. \( \xi_{z,j} \) are not used for the adaptation. We use the average of \( \xi_{z,j} \) over \( j \) as initial values of the adaptation. Only \( \xi_z \) is adjusted for adaptation. The second-level weightings of the aspect models, \( \xi_z \), are shared globally over the entire phoneme models.

When using the aspect models 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 \( \xi_z,j \). To estimate these parameters, \( \alpha_{k,x} \) and \( \beta_{k,y} \) are also changed into state-dependent (i.e. \( \alpha_{k,x,s} \) and \( \beta_{k,y,s} \)). Note that only averages of \( \xi_{z,j,s} \), over \( j \) and \( s \) are used for the adaptation process.

\[
\xi_z = \frac{1}{K} \sum_{j=1}^{K} \xi_{z,j}.
\]  

(11)

3.3. Aspect Model Weighting (adaptation)

For adaptation of the aspect models, AMW is performed using available adaptation data. For AMW, EM algorithm is applied for estimating \( \xi_z \), which is the updated \( \xi_z \). When the adaptation data \( y_1, y_2, \ldots, y_N \) are given, \( \xi_z \) is calculated as

\[
\xi_z^{(n+1)} = \frac{\sum_{s=1}^{S} \sum_{k=1}^{K} \bar{\xi}_z^{(n)} \lambda_{k,z,s} \psi_{k,s}(y_s^{(n)})}{\sum_{s=1}^{S} \sum_{k=1}^{K} \bar{\xi}_z^{(n)} \lambda_{k,z,s} \psi_{k,s}(y_s^{(n)})}.
\]  

(12)

where \( n \) means the number of iteration, \( s \) stands for a state, and \( y_s^{(n)} \) represents \( s \)-th adaptation data belongs to state \( s \).

After estimating \( \xi_z \), we obtain a mixture model adapted to the data as

\[
p(x|\bar{\Xi}, \Lambda) = \sum_{z=1}^{Z} \sum_{k=1}^{K} \xi_z^{(n)} \lambda_{k,z,s} \psi_{k,s}(x).
\]  

(13)

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 mean for new speaker are linear combinations of reference speaker models, as

\[
\mu_{(s)}^{(k)} = \sum_{k=1}^{K} w_{k,s} \mu_{(s)}^{(k)}, \quad \text{where} \quad w_{k,s} = \sum_{z=1}^{Z} \bar{\xi}_z \lambda_{k,z,s},
\]  

(14)

where \( \mu_{(s)}^{(k)} \) signifies the updated mean of the distribution of the state \( s \) and \( \mu_{(s)}^{(k)} \) stands for the mean of the state \( s \) of the \( k \)-th reference speaker model. The covariance matrix of the adapted model, \( \Sigma_{(s)}^{(k)} \), comes from the SI model.

Fig. 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 models are computed using the method explained in section 3.2. In the adaptation phase, the original aspect models are adjusted using the adaptation data. Here the second-level weightings of each aspect model, \( \xi_z \), are the unit for adaptation instead of the SD model set. The adaptation is performed by a linear combination using mean parameters from the 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.

4. EXPERIMENTAL EVALUATION

The rapid speaker adaptation performance of the aspect model was tested using the Korean (KLE452) database [7] in a supervised fashion.

4.1. KLE452 database

KLE452 database consist of 70 speakers (38 males and 32 females). The vocabulary size is 452 words and each speaker utters 452 words once. The condition for recording is an office environment. Sampling rate for speech is 16kHz. Among the 70 speakers, 60 speakers (32 males and 28 females) were used for training. For testing and adaptation, 10 speakers (6 males and 4 females) not included in the training set were used. The amount of adaptation speech is from 0.1s to 20s. For the first evaluation, the adaptation data were started at an each speaker’s 401st word. For the second evaluation, the adaptation was performed using the last word as the starting word. The adapted model was tested for each speaker’s remaining speech (i.e., 400 words). Finally, the two adaptation results were averaged.

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 10ms using a 25ms Hamming window. The MFCCs, \( \Delta \)MFCCs and \( \Delta\Delta \)MFCCs were concatenated to form 39-dimensional feature vectors. Cepstral mean normalization was not 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. The SI model has a word error rate of 24.33% on the test data. Each SD model was created by a typical EM training procedure using SI model as an initial model.

4.3. Effect of the Number of Aspect Models

In this experiment, we investigate the effect of the different number of aspect models. The results are portrayed in Fig. 3. For a more detailed evaluation, the x-axis units are set not to words but to seconds. The performance was evaluated using word error rate. The figure shows that the followings are true.

- Effects of the different number of aspect models are rapidly smoothed in 10 aspect models case.
- The performance saturates at about 2s for all cases.
- However, the performance rapidly reached to the best one for all cases.
4.4. Comparison with Existing Adaptation Methods

In fact, MAP, MLLR, RSW, and eigenvoice were used for comparison with the proposed method. In all adaptation methods, only mean parameters were updated. The experimental results are presented in Fig. 4. For the MAP, the adjustment parameter $\tau$ was set to 50 which was decided empirically. For the MLLR, the global transformation matrix was used for adaptation. The global weighting vector was used for both RSW and eigenvoice adaptation. For the eigenvoice adaptation, we used 25 eigenvoices whose cumulative contributions were greater than 80%. Results show that MAP, MLLR and eigenvoice, especially MLLR, suffer from data sparsity problem when adaptation data are extremely small. RSW and the proposed method were not degraded for the insufficient amount of adaptation data. In [8], the paper already reported the best results of maximum-likelihood RSW in comparison with eigenspace-based approach and original RSW. [8] also stated that, under the WSJ0 corpus, using all training speakers as the reference speakers provides the best adaptation performance. In our experiment, RSW not using hierarchical speaker clustering (HSC) [2] was used; i.e., the number of reference speakers was the same as the total number of training speakers. Compared to RSW, the similar error rate was achieved using only 10 aspect models.

The experimental results demonstrate that aspect models can represent speaker characteristics for extremely small amount of adaptation data (e.g. short segments of less than 1s). Using only 0.5s of adaptation data for 10 and 20 aspect models, relative error rate reductions of 26.62% and 27.24% were achieved, respectively. We obtained improved results using the estimated weighting values calculated by AMW based on EM algorithm. The small number of aspect models could adjust numerous parameters of reference speaker models. The free parameters for adaptation are small because only $\xi$ must be estimated in the adaptation phase.

5. CONCLUSIONS

In this paper, we proposed an aspect-model-based reference speaker weighting. The aspect model was trained based on likelihood of the training utterances. The aspect model is a “mixture-of-mixture” model, which first calculates a few aspect models as mixtures of distributions of the reference speakers models. We then performed AMW using available adaptation data: the aspect models were mixed to obtain the adapted distribution. We used the obtained mixture weight for interpolating weights of mean parameters of the distributions. The number of free parameters was effectively reduced using the small number of aspect model. We evaluated performance of the proposed method through a speaker adaptation experiment. Although we use only 0.5s of adaptation data and twenty aspect models, 27.24% relative improvement was achieved from SI model using the proposed method. Future work will involve performing experiments on many speakers and in various noisy environments.

6. REFERENCES