Simultaneous Estimation of Weights of Eigenvoices and Bias Compensation Vector for Rapid Speaker Adaptation

Hwa Jeon Song and Hyung Soon Kim

Department of Electronics Engineering, Pusan National University, Korea
{Hwajeon, kimhs}@pusan.ac.kr

Abstract

Eigenvoice based speaker adaptation method is known to be very effective tool for rapid speaker adaptation. Stochastic matching approach is also known as a powerful method to reduce the mismatch between training and test environments. In this paper, we simultaneously applied two methods for speaker adaptation and environment compensation space based on the eigenvoice adaptation framework. In experiments for vocabulary-independent word recognition task with supervised mode adaptation, the proposed method shows higher performance improvement than conventional eigenvoice adaptation method with a small adaptation data. We obtained 19~30% relative improvement with only single adaptation utterance and obtained 37% relative improvement with 50 adaptation utterances by proposed method.

1. Introduction

When there is a mismatch between the training and testing environments, automatic speech recognition (ASR) systems have a degradation in performance. Some sources of mismatch include variation of speaking environments and speakers and so on. This mismatch can be generally compensated in the feature space or model space.

Change of speaking environment caused by additive noises and channel distortions is one of primary factors to degrade the performance of ASR system. Various methods to remove these mismatches are mostly carried out in the feature space. Cepstrum mean normalization (CMN) is very simple and powerful tool to remove the channel distortion and spectral subtraction (SS) to enhance speech signal with additive noise. These methods are not to use any phonetic sequence information of utterances. Contrary to CMN or SS, stochastic matching (SM) approach [6] to differently compensate with each state according to phonetic information of utterance is presented. If exact phonetic information of utterance is known in advance, SM approach is more effective than CMN or SS. SM approach can be also carried out in the model space.

Compensation for variability among speakers is generally realized in model space and speaker adaptation is the popular compensation method in model space. Adaptation techniques are also used for environment compensation. Typical adaptation methods include maximum a posteriori (MAP) adaptation [1], maximum likelihood linear regression (MLLR) adaptation [2] and speaker clustering based adaptation. Eigenvoice method, one of speaker clustering based adaptation, is known to be advantageous in fast speaker adaptation because the number of estimated parameters is smaller than other standard model-based adaptation algorithms [3][4].

We recently presented that eigenvoice adaptation scheme after compensating speaker independent (SI) model with bias vector is very useful to improve the performance of rapid speaker adaptation [7]. In eigenvoice adaptation scheme, the mean supervector of speaker dependent (SD) models constructed from the training DB roughly describes the environment to collect training DB while eigenvoices represents the most important component of variation between SD models. If there is a mismatch between training and test environments, not only speaker adaptation but also bias compensation or environment compensation is more effective to improve the ASR performance.

In this paper, we propose the simultaneous speaker and environment adaptation based on the eigenvoice adaptation framework to improve the recognition performance as an extension of previous work [7].

In section II eigenvoice speaker adaptation is briefly described. In section III we propose our method of the simultaneous estimation of weight of eigenvoice and bias vector. Section IV shows experimental results of conventional methods and the proposed method. Finally, section V provides the conclusion.

2. Eigenvoice Speaker Adaptation

Eigenvoices are basis vectors that are orthogonal to each other and guaranteed to represent the most important components of variation between the reference speakers from training DB. Eigenvoice adaptation represents a new speaker with the weighted sum of $K$ eigenvoices as

$$
\hat{\mu} = \mu(0) + \sum_{j=1}^{K} w(j) \mu(j)
$$

where $\mu(0)$ is the mean supervector of SD models and $w(k)$ is the weight of $\mu(k)$, eigenvoices $k$.

$$
\hat{\mu} = [\hat{\mu}_1^T \hat{\mu}_2^T \cdots \hat{\mu}_m^T]^T
$$

is adapted model of supervector form with $L$ dimension and subvector $\hat{\mu}_m^T$ is the mean vector of mixture Gaussian $m$ in state $s$.

$$
\mu(j) = [\mu_1^{(j)} \mu_2^{(j)} \cdots \mu_m^{(j)}]^T \quad j = 1, \cdots, K
$$
is eigenvoice \( j \) and \( e_{m}^{(i)}(j) \) represents the subvector of eigenvoice \( j \) corresponding to \( \mu_{m}^{(i)} \) of SI model. The dimension of subvector \( \mu_{m}^{(i)} \) and \( e_{m}^{(i)}(j) \) is equal to that of observation vector.

The weights \( w(k) \) can be estimated by maximum likelihood eigen-decomposition (MLED) [3][4] using adaptation data of a new speaker. Matrix-vector form equation to estimate the weights in MLED is

\[
y = Aw
\]

where

\[
y = [y_1 \quad \cdots \quad y_K]^T
\]

with

\[
y_j = \sum_{s} \sum_{m} \gamma_{m}^{(t)}(i) e_{m}^{(s)}(j)^T C_{m}^{(-1)} o_t, \quad j = 1, \ldots, K.
\]

The elements of \( A \) are

\[
a_{ij} = \sum_{s} \sum_{m} \gamma_{m}^{(t)}(i) e_{m}^{(s)}(i)^T C_{m}^{(-1)} e_{m}^{(s)}(j), \quad i, j = 1, \ldots, K
\]

and

\[
w = [w(1) \quad \cdots \quad w(K)]^T
\]

is weight vector to be estimated. \( \gamma_{m}^{(t)}(i) \) is the occupation probability of mixture Gaussian \( m \) in state \( s \) at time \( t \). And \( o_t \) is a feature vector normalized by subtracting the corresponding component of mean supervector \( e(0) \) at time \( t \).

3. Simultaneous Speaker and Environment Adaptation based on Eigenvoice method

For rapid speaker adaptation, it is favorable that the number of estimated parameters is small and estimation method is simple and powerful. Otherwise the stability of adapted model cannot be guaranteed. For example, MLLR adaptation method usually yields performance degradation with very small amount of adaptation data [4][7]. Eigenvoice adaptation method is more advantageous than other adaptation method when the amount of adaptation data is very small.

However, the model adapted by eigenvoices has poor representation of location of a new speaker in speaker space. Fig. 1 is an example of speaker space with 3-dimensions and eigenspace with 2-dimensions. If \( i \) is base unit vectors of each axis of speaker space. Eigenspace consisted of \( e(1) \) and \( e(2) \) that the origin is \( e(0) \) is also implicated in the information of training environment. If test environment is far away from training environment, adapted model in eigenspace dose not depict the new speaker well. Therefore, the procedure with bias compensation followed by eigenvoice adaptation [7] is more reasonable than that only using eigenvoice adaptation. SM approach is also effective for robust speech recognition by compensating the environmental mismatch. Both methods, namely, eigenvoice method and SM approach are based on ML estimation.

\[\begin{align*}
\mu_{m}^{(i)} &= \mu_{m}^{(i)}(0) + \sum_{j=1}^{d} b(d) \hat{a}(d) \\
\hat{a}(d) &= \frac{w(j) e(j)^{(i)}(s) + \sum_{d=1}^{D} b(d) \hat{a}(d)}{\sum_{j=1}^{d} w(j) e(j)^{(i)}(s) + \sum_{d=1}^{D} b(d) \hat{a}(d)}
\end{align*}\]
and $\delta(\cdot)$ is delta function. In (11) new $D$ eigenvectors concerning environment compensation are combined with $K$ eigenvoices for speaker adaptation. That is, the total number of eigenvectors used in MLED is $K+D$ and the weights for eigenvoices and the bias compensation vector are simultaneously estimated by MLED.

After substituting (1) with (11), matrix-vector form to estimate the weights by MLED is given by

$$
\begin{bmatrix}
y_{ev} \\
y_b
\end{bmatrix} =
\begin{bmatrix}
A_{ev} & C \\
D & A_b
\end{bmatrix}
\begin{bmatrix}
w_{ev} \\
w_b
\end{bmatrix}
$$

(13)

where $y_{ev}$ and $A_{ev}$ are the same vector and matrix with $y$ and $A$ in (4) when using eigenvoices adaptation method only, respectively. $y_b$ and $A_b$ are the vector and matrix for bias compensation using (9) in MLLR framework, $w_{ev}$ and $w_b$ are the weight vector of eigenvoices and bias compensation vector, respectively. $A_b$ is $K\times K$ symmetric matrix and $A_{ev}$ $D\times D$ matrix. And $C=DA_{ev}$. Especially, it should be noted that $A_b$ is diagonal matrix.

Matrices $C$ and $D$ indicate the correlation between eigenvoices vectors and basis unit vectors in speaker space. If eigenvoices and basis unit vectors for bias compensation is uncorrelated, $C$ and $D$ become zero matrices. Therefore, weights of eigenvoices $w_{ev}$ and bias vector $w_b$ are equal to those estimated individually through ML method. However, because they have correlation to each other, the weights of eigenvoices and bias compensation vectors are affected by correlation between eigenvoices and basis unit vectors.

### 4. Experiments and Results

#### 4.1. Experimental Setup [7]

Our task domain is vocabulary-independent isolated word recognition with supervised mode adaptation. First, Korean phonetically optimized words (POW) DB [8] is used for constructing SI model and SD models in training session. Only a part of the POW DB (40 males) is used in this experiment. The speech data was sampled at 16 kHz and segmented into 20 ms frame at every 10 ms. We used 36-dimension observation vectors (12 MFCCs, its delta and double deltas).

Our baseline system used triphones with continuous mixture density HMM. Each HMM has three states and the number of mixtures per state is 1 or 2. We tied states using tree based clustering method. Total number of tied states is 4050. A set of 40 SD models is constructed by MLLR followed by MAP adaptation from SI model. To obtain eigenvoices, we applied principal component analysis (PCA) to 40 SD models. We used 5 to 30 eigenvoices for adaptation. The dimension of supervector $L$ is 145800 and 291600 when the number of mixture per state is one. A global transformation matrix is used for MLLR. Dimensional eigenvoice method [7] means that the dimension of subvector in (3) is one. That is, $K$ weights of eigenvoices are estimated each observation vector dimension. The number of parameters to be estimated in dimensional eigenvoices method is $D\times K$.

In case of MAP adaptation method, as the amount of adaptation data increases, performance degrades. An additional experiment using much amount of adaptation data showed the performance improvement in MAP.

When amount of adaptation utterances is less than 10, the performance of system adapted by MLLR and dimensional eigenvoices method shows severe degradation as explained in section III. Eigenvoice adaptation method shows performance improvement with small data for rapid speaker adaptation. However, additional performance improvement rarely is achieved with increased amount of data.

#### 4.2. Experimental Results

First, performance of baseline with CMN and MLLR with only bias term for environment compensation is shown in Fig. 2. Because a mismatch between training (POW DB) and test environment (PBW DB) is more affected by channel distortion than additive noise, CMN is effective to improve the performance. It shows that SM approach (MLLR(only bias) in Fig. 2) in supervised mode is more effective than CMN.

Fig. 2 also shows the performances of various conventional speaker adaptation methods, MAP, MLLR, eigenvoices and dimensional eigenvoices method [7], when the number of mixture per state is one. A global transformation matrix is used for MLLR. Dimensional eigenvoices method [7] means that the dimension of subvector in (3) is one. That is, $K$ weights of eigenvoices are estimated each observation vector dimension. The number of parameters to be estimated in dimensional eigenvoices method is $D \times K$.

A result of the method proposed in this paper when the number of mixture per state is one is shown in Fig. 3. When the number of adaptation utterance is only one, the performance using 5 eigenvoices is the best. And the word error rate is reduced by 30%. In case of using 30 eigenvoices, the performance is higher than conventional eigenvoices method when using not only small amount of adaptation data but also increased amount of data. The word error rate is reduced by 37% when the number of adaptation utterances is

![Figure 2: Performance compensation of several adaptation methods (1-mixture per state)](image-url)
50. That is, simultaneous application with speaker adaptation by eigenvoice method and environment compensation by bias vector proposed in this paper is more effective to improve performance than conventional standard adaptation method for rapid speaker adaptation. Fig. 4 shows the result of proposed method when using two mixtures per state. In Fig. 4, we also obtain similar result to previous experiments of Fig. 3.

5. Conclusion

In this paper, we simultaneously applied two methods for speaker adaptation and environment compensation in model space based on the eigenvoice-adaptation framework. In experiments for vocabulary-independent word recognition task with supervised mode adaptation, the proposed method shows higher performance improvement than conventional eigenvoice adaptation method for a small adaptation data. We obtained 19~30% relative improvement with only single adaptation word and obtained 37% relative improvement with 50 adaptation words by proposed method.

6. Reference


The proposed method in this paper improved the performance and is simpler than the previous method[7] in the viewpoint of implementation. In addition, as the number of adaptation data is increased, the performance of proposed method stands comparison with dimensional EV method.