Performance Improvement of Rapid Speaker Adaptation Based on Eigenvoice and Bias Compensation

Jong Se Park, Hwa Jeon Song and Hyung Soon Kim

Department of Electronics Eng., Pusan National University, Pusan, Korea

{next,hwajeon,kimhs}@pusan.ac.kr

Abstract

In this paper, we propose the bias compensation methods and the eigenvoice method using the mean of dimensional eigenvoice to improve the performance of rapid speaker adaptation based on eigenvoice. Experimental results for vocabulary-independent word recognition task shows the proposed method yields improvements for a small adaptation data. We obtained 22~30% relative improvement by the bias compensation methods, and obtained 41% relative improvement by the eigenvoice method using the mean of dimensional eigenvoice with only single adaptation word.

1. Introduction

Speaker adaptation contributing to better representation of speaker characteristics in a speech model is a very useful tool to improve speech recognition performance. Typical adaptation methods include Maximum A Posteriori (MAP)[1], Maximum Likelihood Linear Regression (MLLR)[2] and speaker clustering. Among them, eigenvoice method, one of speaker clustering methods, is known to be advantageous in fast speaker adaptation because the number of estimated parameters is small [3].

However, eigenvoice method hardly shows additional improvement even with increased amount of adaptation data. To deal with this problem, several modified methods have developed such as eigenvoice method followed by MAP[4] and segmental eigenvoice adaptation method[5].

In this paper, we propose two approaches to improve the recognition performance with very small data as well as increased amount of adaptation data. The first approach obtains the adaptation model based on the combination of the bias compensation model and the eigenvoice adaptation model. The second approach is based on the eigenvoice adaptation method and the mean of dimensional eigenvoice models built from other speakers' utterances.

Section 2 describes eigenvoice speaker adaptation and dimensional eigenvoice speaker adaptation. In section 3 and 4 we propose various methods based on eigenvoice speaker adaptation. Section 5 shows experimental results of conventional methods and the proposed method. Finally, section 6 provides the conclusion.

2. Eigenvoice

Eigenvoices are basis vectors in eigen space effectively representing the distribution of deviation of a large number of training speakers. Eigenvoice adaptation represents a new speaker with weighted sum of K eigenvoices as

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

where $e(0)$ is the mean of T speaker dependent models, and $w(k)$ is the weight of $e(k)$, k-th eigenvoice. The weights can be estimated by Maximum Likelihood Eigen-Decomposition (MLED)[3][4] using adaptation data of a new speaker.

2.1. Dimensional eigenvoice approach

Eigenvoice method hardly shows additional improvement even with increased amount of adaptation data. The additional improvement can be achieved by estimating the weights of eigenvoices in each feature vector dimension in MLED step. In this paper this is called as 'dimensional eigenvoice model'.

3. Eigenvoice adaptation method using bias compensation

We estimated bias component between adaptation data and model and used it to reflect speaker characteristics in the adaptation model. With adaptation data, we obtained the bias compensated model as shown in Fig. 1.

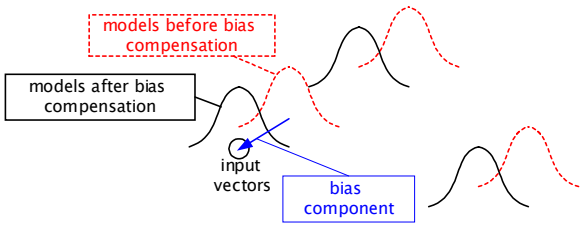

Figure 1: Compensation for the bias components

Four methods are suggested according to the type of bias compensation. The block diagram of the four methods is shown in Fig. 2.

3.1. Method 1 : weighted sum of bias compensation model and eigenvoice adaptation model

Method 1 is shown in Fig. 2(a). The final adaptation model $\hat{\mu}$ by method 1 is represented by weighted sum of bias compensation model and eigenvoice adaptation model as

$$\hat{\mu} = (1-\alpha)\hat{\mu}_{EV} + \alpha\,[\,\mu_{SI} + \hat{b}\,], \quad 0 \le \alpha \le 1 \tag{2}$$

where $\hat{\mu}_{EV}$ is a eigenvoice adaptation model using adaptation data and $(\mu_{SI}+\hat{b})$ is a bias compensation model based on speaker independent(SI) model. Bias component, $\hat{b}$, is estimated by maximum likelihood stochastic matching method [6] as

$$\hat{b}=\frac{\sum_{t=1}^{T}\sum_{s}\sum_{m}\gamma_m^{(s)}(t)(\mathbf{o}_t-\mu_m^{(s)})}{\sum_{t=1}^{T}\sum_{s}\sum_{m}\gamma_m^{(s)}(t)} \tag{3}$$

where T is total the number of frames of adaptation data and $\mu_m^{(s)}$ is the mean vector of m-th mixture in state s for observation data $\mathbf{o}_t$ at time t. $\gamma_m^{(s)}(t)$ is the occupation probability of m-th mixture in state s at time t. Therefore, estimated bias component represents the mean of deviation between adaptation data and SI model.

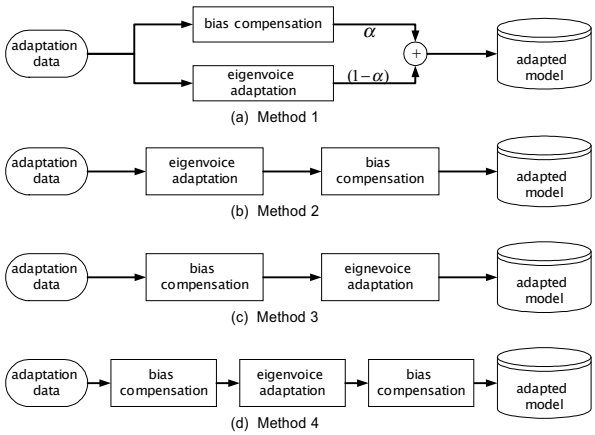

Figure 2:Bias compensation methods

In (2), α =0 means conventional eigenvoice method and α =1 means only bias compensation of SI model. For the performance improvement, we estimated α optimally representing the speaker characteristics with a given amount of adaptation data.

To estimate α , we first used the ratio of distances between the observation and the two models: the bias compensation model and the eigenvoice adaptation model as

$$r_t=\frac{|\mathbf{o}_t-\hat{\mu}_{EV}|}{|\mathbf{o}_t-(\mu_{SI}+\hat{b})|} \tag{4}$$

If input feature vector is close to the eigenvoice adaptation model, r_t increases. On the other hand, if it is close to the bias compensation model, r_t decreases. In (5), we make a logarithmic transform so that x_t is positive if r_t > 1 and otherwise x_t is negative.

$$x_t=\log(r_t) \tag{5}$$

To constrain dynamic range of α between 0 and 1, sigmoid function is adopted as (6).

$$\alpha=\frac{\sum_{t=1}^{T}\sum_{s}\sum_{m}\gamma_m^{(s)}(t)\left(\frac{1}{1+e^{-x_t}}\right)}{\sum_{t=1}^{T}\sum_{s}\sum_{m}\gamma_m^{(s)}(t)}=\frac{\sum_{t=1}^{T}\sum_{s}\sum_{m}\gamma_m^{(s)}(t)\left(\frac{1}{1+r_t^{-1}}\right)}{\sum_{t=1}^{T}\sum_{s}\sum_{m}\gamma_m^{(s)}(t)} \tag{6}$$

3.2. Method 2 : bias compensation after eigenvoice adaptation

Method 2, shown in Fig. 2(b), is an approach compensating bias component after eigenvoice adaptation. First, the eigenvoice adaptation model $\hat{\mu}_{EV}$ can be established from adaptation data, and then the final adaptation model $\hat{\mu}$ is constructed by compensating bias component of $\hat{\mu}_{EV}$ as

$$\hat{\mu}=\hat{\mu}_{EV}+\hat{b} \tag{7}$$

where bias component $\hat{b}$ can be estimated by (8) using adaptation data and $\hat{\mu}_{EV}$.

$$\hat{b}=\frac{\sum_{t=1}^{T}\sum_{s}\sum_{m}\gamma_m^{(s)}(t)(\mathbf{o}_t-\hat{\mu}_m^{(s)})}{\sum_{t=1}^{T}\sum_{s}\sum_{m}\gamma_m^{(s)}(t)} \tag{8}$$

where $\hat{\mu}_m^{(s)}$ is the mean vector of m-th mixture in state s of $\hat{\mu}_{EV}$ for observation data $\mathbf{o}_t$ at time t. Difference between (3) and (8) is that $\mu_m^{(s)}$ in (3) is the mean vector of SI model and $\hat{\mu}_m^{(s)}$ in (8) is the mean vector of $\hat{\mu}_{EV}$.

3.3. Method 3 : eigenvoice adaptation after bias compensation

In Fig. 2(c), contrary to the method 2, the final eigenvoice adaptation model is constructed after bias compensation as

$$\hat{e}(0)=e(0)+\hat{b} \tag{9}$$

where bias component $\hat{b}$ can be estimated from adaptation data and $e(0)$ as

$$\hat{b}=\frac{\sum_{t=1}^{T}\sum_{s}\sum_{m}\gamma_m^{(s)}(t)(\mathbf{o}_t-e_m^{(s)}(0))}{\sum_{t=1}^{T}\sum_{s}\sum_{m}\gamma_m^{(s)}(t)} \tag{10}$$

where $e_m^{(s)}$ is m-th mixture in state s of $e(0)$ assigned to observation $\mathbf{o}_t$.

3.4. Method 4 : Method 2 + Method 3

Method 4 is a hybrid of the method 2 and the method 3 as shown in Fig. 2(d). The second bias compensation module is employed for compensation of residual bias after eigenvoice adaptation.

4. Eigenvoice adaptation method using the mean of dimensional eigenvoice models

If speech data collected from the same environment with recognition is available, additional performance improvement can be achieved with small adaptation data. In this paper, we investigated the methods using the mean of dimensional eigenvoice models obtained from the additional speakers' data.

4.1. Method A : weighted sum method of the mean of dimensional eigenvoice models and eigenvoice adaptation model

Method A obtains the final adaptation model $\hat{\mu}$ from the weighted sum of the mean of dimensional eigenvoice models and the conventional eigenvoice model.

$$\hat{\mu} = (1-\alpha)\mu_{EV} + \alpha\,\bar{\mu}_{EV\_DIM}, \quad 0 \le \alpha \le 1 \tag{11}$$

where μ_{EV} is conventional eigenvoice adaptation model and $\bar{\mu}_{EV\_DIM}$ is the mean of dimensional eigenvoice models estimated from several speakers. If α=0, it is equivalent to the conventional eigenvoice method. If α =1, it only uses $\bar{\mu}_{EV\_DIM}$. Therefore, the final adaptation model for a new speaker is somewhere between μ_{EV} and $\bar{\mu}_{EV\_DIM}$. The evaluation was conducted by varying α.

4.2. Method B : eigenvoice adaptation after adopting mean of dimensional eigenvoice models

Another method uses the mean of dimensional eigenvoice models where $e(0)$ is replaced with $\bar{\mu}_{EV\_DIM}$ as

$$\hat{e}(0) = \bar{\mu}_{EV\_DIM} \tag{12}$$

where $\bar{\mu}_{EV\_DIM}$ is the mean of dimensional eigenvoice models of several speakers. Only in MLED step, modified eigenvoice $\hat{e}(0)$ instead of $e(0)$ is used.

5. Experiments and Results

5.1. Experimental condition

Korean phonetically optimized words (POW) DB[7] is used for constructing SI model and SD models in training session. Only a part of the POW - 40 males - is used in this experiments.

The speech data was sampled at 16kHz and segmented into 20ms frame at every 10ms. We used 36 dimensional feature parameters (12 MFCCs, its deltas 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 varied from 1 to 4. We tied the states using tree based clustering (TBC) method.

We first trained SI model from the POW DB. A set of 40 SD models is constructed by MAP adaptation with SI model. To obtain eigenvoices, we applied principal component analysis (PCA) to 40 SD models. We used 30 eigenvoices for adaptation.

For adaptation and evaluation, another DB set, gathered in an environment different from the training DB, is needed. We used Korean phonetically balanced words (PBW) DB for adaptation and evaluation. We performed supervised adaptation experiments for 10 male speakers. We used 50 words for adaptation and 400 words for the evaluation.

5.2. Results

In Fig. 3, the performance of various conventional speaker adaptation methods, MAP, MLLR and eigenvoice, are compared with the baseline system which uses SI model constructed from the POW DB.

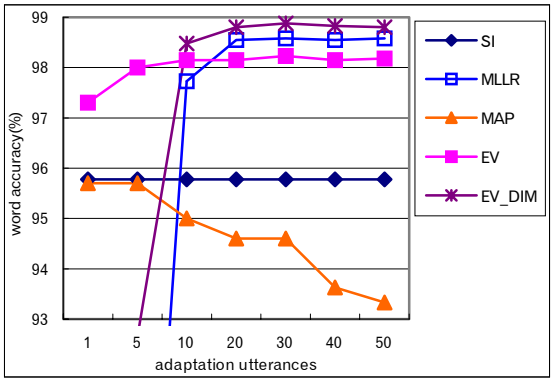

Figure 3: Performance of several adaptation methods (1-mixture)

MAP method is outperformed by the baseline system because a relatively large number of adaptation data is needed for MAP method to work effectively. When the number of adaptation utterances is less than 10, the performance improvement for MLLR is not guaranteed.

As mentioned previously, eigenvoice adaptation shows performance improvement with small data while additional performance improvement rarely is achieved with increased amount of data.

Dimensional eigenvoice method which estimates the weights of eigenvoices in each feature vector dimension (EV_DIM in Fig. 3) shows higher performance than other methods with sufficient adaptation data because it is able to represent more detailed speaker characteristics than the conventional eigenvoice method. However when the number of adaptation utterances is less than 10, its performance is not guaranteed as much as MLLR because the number of estimated parameters is larger than conventional eigenvoice method.

Figs. 4 and 5 show the result of bias compensation methods proposed in this paper. In Fig. 4, one mixture per state is used while two in Fig. 5.

The bias compensated eigenvoice method shows higher performance improvement than conventional eigenvoice method (EV in Figs. 4 and 5). Method 4 shows the best

performance because it capitalizes on the strong points of the method 2 and method 3.

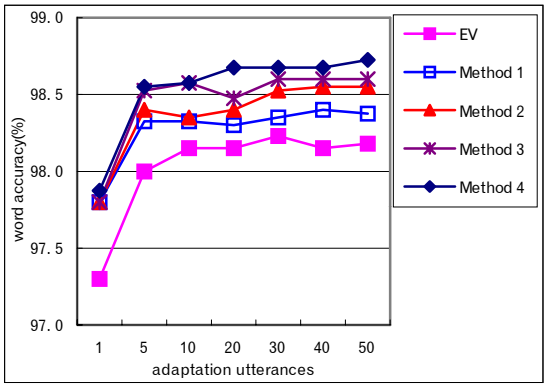

Figure 4:Results of bias compensation methods (1-mixture)

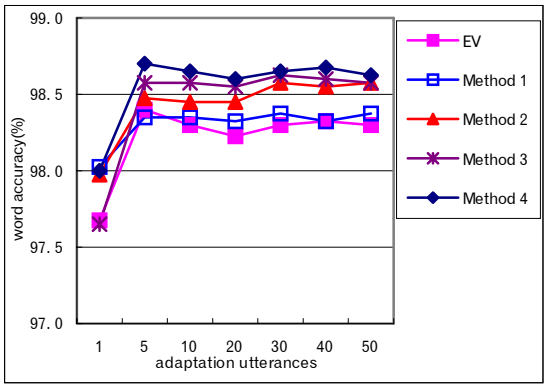

Figure 5: Results of bias compensation methods (2-mixture)

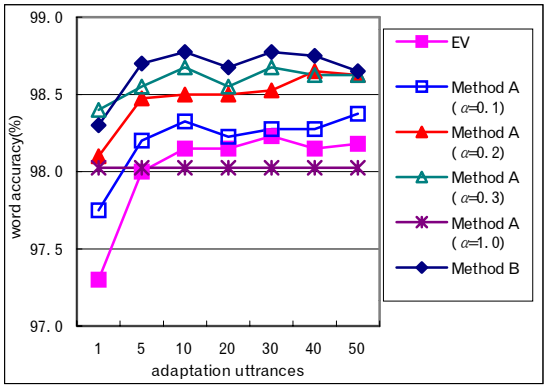

Figure 6:Results of eigenvoice adaptation using mean of dimensional eigenvoice models (1-mixture)

Fig. 6 shows the result of the experiments using method A and B described in section 4. The mean of dimensional eigenvoice models is obtained using 28 male speakers not participated in evaluation with PBW DB. Fifty isolated words are used for each speaker.

Methods A and B show good performance improvement in comparison with the conventional eigenvoice method. Especially in case of very small adaptation data, the improvement rate is high. We can say that the general performance of method B is better than method A. We obtained 41% relative improvement in error rate by using the weighted sum of eigenvoice method and the mean of dimensional eigenvoice method with only single adaptation word in method A.

6. Conclusion

In this paper, we proposed methods using the weighted sum of eigenvoice and the mean of dimensional eigenvoice method and the bias compensation method to improve the performance of eigenvoice adaptation.

We obtained 41% relative improvement by using the weighted sum of eigenvoice and the mean of dimensional eigenvoice method with only single adaptation word. It should be noted that this approach requires other speakers' utterances with the same environment as test utterances' environment. We also obtained 22~25% relative improvement by using the bias compensation method without additional information on test environment.

7. References

[1] C. H. Lee, C. H. Lin and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," IEEE Trans. Signal Processing, vol.**39**, no.4, pp.806-814, April, 1991.

[2] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.**9**, no.1, pp.171-185, Sep, 1995.

[3] R. Kuhn, P. Nguyen, J. C. Jungua, L. Goldwasser, N. Niedzielski, S. Finche, K. Field and M. Contolini, "Eigenvoices for speaker adaptation," in Proc. ICSLP, vol.**5**, pp.1771-1774, 1998.

[4] R. Kuhn, J. C. Jungua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," IEEE Trans. Speech and Audio Processing, vol.**8**, no.6, pp.695-707, Nov. 2000.

[5] Y. Tsao, S. M. Lee, F. C. Chou and L. S. Lee, "Segmental eigenvoice for rapid speaker adaptation," in Proc. Eurospeech, vol.**2**, pp.1269-1272, 2001.

[6] A. Sanker, "A maximum-likelihood approach to stochastic matching for robust speech recognition," IEEE Trans. Speech and Audio Processing, vol.**4**, no.3, pp.190-202, May, 1996.

[7] Y. Lim and Y. Lee, "Implementation of the POW (Phonetically Optimized Words) algorithm for speech database," In Proc. ICASSP, vol.**1**, pp.89-91, 1995.