SPEAKER ADAPTATION BY HIERARCHICAL EIGENVOICE

Yoshifumi Onishi and Ken-ichi Iso

Multimedia Research Laboratories, NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki, 216-8555 Japan
y-onishi@bp.jp.nec.com, k-iso@bx.jp.nec.com

ABSTRACT

We propose a novel speaker adaptation method, Hierarchical EigenVoice (HEV). This method extends the EigenVoice [1] through clustering the Gaussian components of HMMs into a hierarchical tree structure. It enables to autonomously control a number of adaptation parameters (model complexity) depending on the amount of adaptation utterances from a new speaker. The experimental results of Japanese large vocabulary continuous speech recognition confirmed the significant performance increase in all range of the adaptation utterance amounts compared with the conventional speaker adaptation methods.

1. INTRODUCTION

For Speaker-Independent (SI) speech recognition systems, a large amount of speech data collected from many speakers is used to train SI HMMs. And speaker adaptation techniques are used to compensate the mismatch between the SI HMMs and a new speaker. Maximum a posteriori (MAP) adaptation [2] and maximum likelihood linear regression (MLLR) adaptation [3] are popular model-based adaptation techniques. However, these adaptation techniques do not utilize a priori knowledge of speaker variations obtainable from the SI training database mentioned above.

Eigenvoice (EV) speaker adaptation technique [1] enables to exploit that knowledge for rapid speaker adaptation. It represents the HMMs adaptation parameters as a superposition of a few principal eigenvectors extracted from the SI training database. Its implication is that the adapted HMMs for a new speaker are represented as a single point in the eigenspace spanned by the eigenvectors. Since the number of eigenvectors for adaptation is fixed and limited in EV, the adaptation performance will saturate immediately as more adaptation data becomes available. Additionally, it is computationally expensive to apply EV to large vocabulary speech recognition (LVCSR) because the HMMs have so many Gaussian components that the dimensions of the eigenvectors become too large.

Autonomous Model Complexity Control (AMCC) speaker adaptation technique [4][5] introduces a tree structure in the acoustic space to adjust the degree of parameter sharing depending on the amount of available adaptation data. This method covers a wide range of adaptation data amount, however, it does not utilize a priori knowledge of speaker variations.

In this paper we propose a novel adaptation method, Hierarchical EigenVoice (HEV) method, which unites the EV and AMCC to improve speaker adaptation performance over all range of adaptation data amounts. In the next section, the original Eigenvoice method and a tree structure for continuous density mixture Gaussian HMMs are briefly introduced, and our proposed method is explained. Sec. 3 evaluates the algorithms on a Japanese LVCSR task.

2. SPEAKER ADAPTATION BY HIERARCHICAL EIGENVOICE (HEV)

2.1. EigenVoice(EV)

In EV, a priori knowledge of speaker variations is extracted by principal component analysis (PCA) from many speaker dependent (SD) HMMs. Although we only discuss on adaptation of the Gaussian mean vectors of HMMs in the following, its extension to other model parameters must be straightforward. First, we prepare well trained speaker independent (SI) HMMs and many SD HMMs. For instance, using the SI HMMs as an initial set of models, each SD HMMs can be trained with each speaker's utterances in the SI training database. Here, it is assumed that the SD HMMs and SI HMMs have the same topology and the same number of Gaussian mixtures. We denote $\boldsymbol{x}_i$ as the difference between Gaussian mean vectors in the SD HMMs and the SI HMMs, where $1 < i < M$ and M is the total number of Gaussian components in the HMMs. Secondly, we perform PCA as follows. Supervectors of the speaker p, and their correlation matrix are computed as,

$$\boldsymbol{X}_p^{\mathrm{t}} = [\boldsymbol{x}_1^{\mathrm{t}}\, \boldsymbol{x}_2^{\mathrm{t}} \cdots \boldsymbol{x}_M^{\mathrm{t}}]_p, \tag{1}$$

$$\hat{S} = \frac{1}{N}\sum_{p=1}^{N}(\boldsymbol{X}_p - \boldsymbol{\mu})(\boldsymbol{X}_p - \boldsymbol{\mu})^{\mathrm{t}}, \tag{2}$$

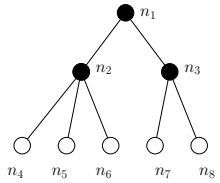

Fig. 1. Example of a tree structure of Gaussian components.

where the superscript "t" means the transpose of vectors and matrices, N is the number of speakers and $\boldsymbol{\mu} = \frac{1}{N}\sum_p \boldsymbol{X}_p$. Diagonalizing the correlation matrix (2), eigenvalues, $\lambda_1 > \lambda_2 > \cdots > \lambda_D$, and their eigenvectors, $\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_D$, are obtained, where D is the number of principal components.

For large HMMs such as triphone HMMs used in LVCSR, the dimension of supervectors (1) and their correlation matrix becomes too large to be diagonalized (computationally expensive).

Since the larger eigenvalue means the wider speaker variation in its eigenvector direction, a few principal eigenvectors with largest eigenvalues should be chosen for rapid speaker adaptation. The supervector for a new speaker is represented as a superposition of those eigenvectors. Using only a small amount of adaptation utterances from the new speaker, the weighting coefficients of the eigenvectors can be estimated by the maximum likelihood eigen decomposition (MLED) method. Its formulation is discussed in the following sections.

2.2. Autonomous model complexity control (AMCC) by a tree structured Gaussian Components

In AMCC [4], all Gaussian components (their total number is denoted as M) in the HMMs are clustered in a hierarchical tree structure using a top-down clustering with the k-means algorithm and the KL divergence as a distance measure between Gaussians. Fig. 1 shows an example of a tree structure of Gaussian components. Leaf nodes, $n_4 \cdots n_8$, represent Gaussian components of the HMMs. The model complexity (the number of adaptation parameters) is controlled by node selection in this tree. For example, if the node n_2 is selected for adaptation, the adaptation parameters such as shift vectors for Gaussian mean vectors are shared in its child nodes, n_4, n_5 and n_6. Adaptation node selection should be determined depending on the amount of adaptation utterances. Namely, fine models for lower adaptation nodes and coarse models for upper ones are autonomously prepared.

2.3. Hierarchical PCA

As pointed out in the previous sections, EV selects a predefined number of eigenvectors for adaptation. Therefore, EV works fine for rapid speaker adaptation with small amount of adaptation data, but fails to further improve its performance when the adaptation data amount increases. To overcome this drawback, we introduce a hierarchical tree structure in a set of HMMs Gaussian components which enables to control the model complexity in EV depending on the amount of adaptation data.

First, we prepare the SI HMMs and many speakers' SD HMMs in the same manner as EV (Sec. 2.1). Second, a tree structure of Gaussian components is constructed based on the SI HMMs. Each leaf node has a difference vectors of Gaussian mean vectors between the SD and SI HMMs. Third, we perform PCA hierarchically along the tree structure from bottom (leaf nodes) to top (a root node) as follows. In Fig.1, the node n_2 has three child nodes, n_4, n_5 and n_6. Then we are able to apply EV method to these Gaussians as mentioned in Sec. 2.1, namely, we construct the supervectors (3) for each speaker, p, and the correlation matrix (4),

$$\boldsymbol{X}^{\mathrm{t}}_{p\,n_2} = [\boldsymbol{x}^{\mathrm{t}}_{n_4}\, \boldsymbol{x}^{\mathrm{t}}_{n_5}\, \boldsymbol{x}^{\mathrm{t}}_{n_6}]_p, \tag{3}$$

$$\hat{S}_{n_2} = \frac{1}{N}\sum_{p=1}^{N}(\boldsymbol{X}_{p\,n_2} - \boldsymbol{\mu}_{n_2})(\boldsymbol{X}_{p\,n_2} - \boldsymbol{\mu}_{n_2})^{\mathrm{t}}, \tag{4}$$

where supervectors consist of not all Gaussian mean vectors in the HMMs but their small subset belonging to the child nodes of n_2 (actually not mean vectors themselves but their difference vectors as previously mentioned). Then we extract D_{n_2}-dimensional eigenspace on the node n_2 spanned by principal eigenvectors, $\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{D_{n_2}}\,|_{n_2}$, which have eigenvalues, $\lambda_1, \lambda_2 \cdots, \lambda_{D_{n_2}}|_{n_2}$ respectively (in descending order), where we use notation "$\cdot|_{n_2}$" to denote the eigenspace which belongs to the node n_2. The number of selected principal eigenvectors, D_{n_2}, can be determined by the cumulative contribution ratio,

$$\frac{\sum_{i=1}^{D_{n_2}} \lambda_i}{\sum_{\mathrm{all}} \lambda_i}, \tag{5}$$

comparing the predefined threshold α. On the node n_3, the eigenspace is constructed in the same manner as that of n_2. To define the eigenspace for the upper layer node, n_1, we project the supervectors of its child nodes, n_2 and n_3 to their eigenspaces. For the node n_2, a projected supervector of each speaker is obtained as follows:

$$\hat{M}_{n_2} \equiv [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{D_{n_2}}|_{n_2}], \tag{6}$$

$$\boldsymbol{x}_{p\,n_2} = \hat{M}^{\mathrm{t}}_{n_2}(\boldsymbol{X}_{p\,n_2} - \boldsymbol{\mu}_{n_2}). \tag{7}$$

Combining the projected supervectors of child nodes, $\boldsymbol{x}_{p\,n_2}$ and $\boldsymbol{x}_{p\,n_3}$, we get supervectors for their parent node n_1,

$$\boldsymbol{X}^{\mathrm{t}}_{p\,n_1} = [\boldsymbol{x}^{\mathrm{t}}_{n_2}\, \boldsymbol{x}^{\mathrm{t}}_{n_3}]_p, \tag{8}$$

and the correlation matrix S_{n_1}, just as in (4). And then, PCA of S_{n_1} extracts the eigenspace of n_1. Fig 2 illustrates

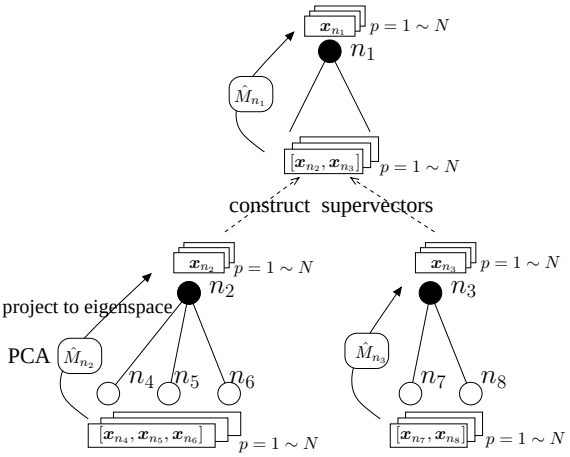

Fig. 2. An illustration of hierachical PCA.

this hierarchical EV method for the tree structure defined in Fig. 1.

Proposed method avoids the direct diagonalization of the correlation matrix between supervectors of all Gaussians in the HMMs and divides the problem into smaller diagonalizations at tree nodes. This technique should be legitimate since the low correlation directions discarded in the lower layer are unimportant in extracting the high correlation direction at the higher layer.

2.4. Model Complexity Control

Hereafter we consider speaker adaptation process based on the hierarchical eigenspaces generated in the previous sections. In this section, we focus on the node selection algorithm based on the amount of adaptation utterances. New speaker's utterances for adaptation are aligned with SI HMMs to assign each frame in the utterances to the Gaussian in the tree leaf node. Therefore some of the leaf node Gaussians seen in the adaptation utterances has new speaker's feature vectors for adaptation and unseen ones do not have any data. To achieve reliable speaker adaptation, the amount of adaptation data and the model complexity (the number of adaptation parameters) have to be balanced. In our hierarchical tree (Fig. 1), nodes in the higher layers such as n_2, n_3 and n_1 have their eigenspaces for adaptation. The model complexity can be defined as the dimensions of their eigenspaces which are the number of adaptation parameters. For node selection, the number of adaptation feature vectors per the eigenspace dimensions on the node should be compared with the predefined threshold θ. For instance, we denote the total number of feature vectors for leaf node Gaussians n_4, n_5 and n_6 as N_2 (total frame counts at n_2), and for n_7 and n_8 as N_3. If N_2/D_{n_2} exceeds threshold θ, the node n_2 is selected for the adaptation of n_4, n_5 and n_6. If not, n_1 is used for the adaptation.

2.5. Parameters Estimation in Eigenspace

In the original EV, the maximum likelihood eigen decomposition (MLED) method is used to estimate adaptation parameters, $\boldsymbol{z}$, which are the weighting coefficients of the eigenvectors. These parameters maximize the likelihood p for the adaptation data $\boldsymbol{O}$ from a new speaker as follows:

$$\boldsymbol{z}^{\mathrm{opt}} = \underset{\boldsymbol{z}}{\mathrm{argmax}}\ p(\boldsymbol{O}|\boldsymbol{Y} = \hat{M}\boldsymbol{z}), \tag{9}$$

where Y is a supervector of all Gaussian mean vectors, and $\hat{M}$ is a linear transformation matrix from eigenspace to the supervector space. The $\hat{M}$ consists of eigenvectors as column vectors.

In our HEV, the MLED method can be also used at each selected node. For example, at the node n_2 in Fig.1, it is apparent that (9) can be used for $\boldsymbol{Y}_{n_2} = [\boldsymbol{y}_{n_4}, \boldsymbol{y}_{n_5}, \boldsymbol{y}_{n_6}]$, a sub-component vector of $\boldsymbol{Y}$, and $\boldsymbol{z}_{n_2}$, a point in the eigenspace of node n_2. When the node n_1 is selected for adaptation, (9) can be applied with $\tilde{M}_{n_1}$ and $\boldsymbol{z}_{n_1}$ using the following relation,

$$\tilde{M}_{n_1} \equiv \begin{bmatrix} \hat{M}_{n_2} & 0 \\ 0 & \hat{M}_{n_3} \end{bmatrix} \hat{M}_{n_1}, \tag{10}$$

$$\boldsymbol{Y} = \tilde{M}_{n_1}\boldsymbol{z}_{n_1}. \tag{11}$$

Equation (10) shows the advantage of our method over the original EV. For LVCSR with triphone HMMs, the dimension of matrix $\tilde{M}_{n_1}$ may become extremely large. In HEV, it can be replaced by a multiplication of block diagonal matrices with smaller dimensions. With this manner, matrices that transform eigenspaces of any node to supervectors of Gaussian components can be generated to perform the MLED optimization process.

To further improve the speaker adaptation performance with less amount of data, we introduced a prior probability density g of adaptation parameter vector $\boldsymbol{z}$ and extended MLED to MAPED (maximum a posteriori eigen decomposition) [6],

$$\boldsymbol{z}^{\mathrm{opt}} = \underset{\boldsymbol{z}}{\mathrm{argmax}}\ p(\boldsymbol{O}|\boldsymbol{Y} = \hat{M}\boldsymbol{z})g(\boldsymbol{z}). \tag{12}$$

Since the eigenvalues measure the extent of the speaker variation of z, we use a multivariate Gaussian (13) as a prior density.

$$g(\boldsymbol{z}) = N(\boldsymbol{0}, \boldsymbol{\lambda I}), \tag{13}$$

$$\boldsymbol{\lambda}^{\mathrm{t}} = [\lambda_1, \lambda_2, \cdots, \lambda_D], \tag{14}$$

where $\boldsymbol{I}$ is the $D \times D$ identity matrix and D is dimension of eigenspace on the selected node.

3. EXPERIMENTS

3.1. Experimental Conditions

The proposed method was evaluated on a large-vocabulary continuous speech recognition for Japanese. Speech was

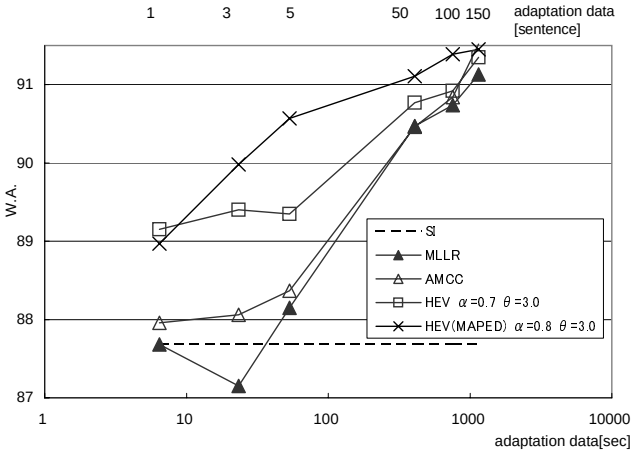

Fig. 3. Word accuracy rate for SI, HEV, AMCC and MLLR.

sampled at 11kHz and analyzed by 11msec frame period to parameterize into 23 dimensional feature vectors. The SI training database comprises about 200,000 utterances (200 utterances from each of 1000 different male speakers). Triphone HMMs with diagonal covariance mixture Gaussians were trained as a set of SI HMMs. The total number of Gaussian components was about 4,000 and they were clustered into a 3-layer tree structure with a maximum of 16 children per node. Using the SI HMMs as initial HMMs, only Gaussian mean vectors were re-estimated for each speaker's SD HMMs. As mentioned in Sec. 2, eigenspaces for each node were extracted hierarchically. To decide dimensions of eigenspaces, we used thresholds for the cumulative contribution ratio. We built open 10-male speaker database where individual spoke 150 sentences for adaptation and 30 for testing.

To compare with the conventional methods, the AMCC and MLLR were examined. For AMCC, we employed the same tree structure of Gaussian components as used in HEV. For MLLR, we used a regression class tree with the maximum of 256 classes.

3.2. Results

Fig. 3 shows dictation recognition results for the test database. The vertical axis shows average word accuracy (W.A.) of 10 speakers, while the horizontal ones show the amount of adaptation utterances in second [sec] (lower axis) and in the number of sentences (upper axis). The dotted line in the figure shows the SI HMMs recognition performance (no speaker adaptation). The lines with "MLLR" and "AMCC" are the performance of the conventional methods mentioned above. The lines with "HEV" are of our proposed methods, one useing MLED for adaptation parameter estimation and the other useing MAPED. The parameter "α" is the threshold for the cumulative contribution ratio defining the dimension of node eigenspaces. "θ" is the node selection threshold for adaptation described in Sec. 2.4. It demonstrates that the proposed methods significantly outperform the conventional AMCC and MLLR. The performance increase with less amount of adaptation data reflects the effectiveness of the proposed hierarchical structures. Estimating adaptation parameters in MAP fashion (MAPED) was proved to be effective.The original EV is not shown on the figure since it was computationally too expensive to perform diagonalization of the matrix with the large dimension (which will be roughly 23*4000 in our case).

4. CONCLUSION

We proposed a new adaptation method, Hierarchical EigenVoice (HEV), that is an extension of EigenVoice (EV) with a hierarchical tree structure of Gaussians for model complexity control. The model complexity control was carried out based on the amount of adaptation data and the optimum parameters estimation formula was derived as MLED and MAPED. The results of Japanese LVCSR experiments shows that the proposed methods outperformed the conventional speaker adaptation methods, MLLR and AMCC. Using only 5 sentences for adaptation, its recognition accuracy achieved the same result as that of the conventional methods with 50 sentences.

5. REFERENCES

[1] R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski "Rapid Speaker Adaptation in Eigenvoice Space," *IEEE Trans. Speech Audio Proc.*, vol. 8, pp. 695-707, 1994.

[2] J.L. Gauvain and C.H. Lee, "Maxmum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Proc.*, vol. 2, pp. 291-298, 1994.

[3] C.J. Leggetter and P.C. Woodland, "Maxmum likelihood linear regression for speaker adaptation of continuous-density hidden markov models," *Comput. Speech Lang.*, vol. 9, pp. 171-185, 1995.

[4] K. Shinoda and T. Watanabe " Speaker adaptation with autonomous control using tree structure," in *Proc. EuroSpeech '95*, pp. 1143-1146, 1995.

[5] K. Shinoda and T. Watanabe " Speaker adaptation with autonomous model complexity control by MDL principle, " in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '96*, pp. 717-720, 1996.

[6] H. Botterweck "Anisotropic map defined by eigenvoices for large vocabulary continuous speech recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing 2001*, pp. 353-356, 2001.