interspeech 2008 incorporating SST 08

Improvement of Eigenvoice-Based Speaker Adaptation by Parameter Space Clustering

Shutaro Tanji[1], *Koichi Shinoda*[1], *Sadaoki Furui*[1], *and Antonio Ortega*[2]

[1]Department of Computer Science, Tokyo Institute of Technology
[2]Department of Electrical Engineering, University of Southern California

tanji@ks.cs.titech.ac.jp, {shinoda, furui}@cs.titech.ac.jp, antonio.ortega@sipi.usc.edu

10.21437/Interspeech.2008-373

Abstract

The segmental eigenvoice method has been proposed to provide rapid speaker adaptation with limited amounts of adaptation data. In this method, the speaker-vector space is clustered to several subspaces and PCA is applied to each of the resulting subspaces. In this paper, we propose two new techniques to improve the performance of this segmental eigenvoice approach. First, we propose a soft-clustering method in which each element in a speaker vector can be assigned to more than one cluster. Second, those elements far apart from any of the clusters are removed. Our experiments using the JNAS and S-JNAS databases show that the proposed method outperforms both the original eigenvoice and the segmental eigenvoice methods, e.g., 3.3% average improvement when only 10 utterances are used for adaptation.

Index Terms: principal component analysis, speaker adaptation, eigenvoice, parameter space clustering

1. Introduction

Many effective speaker adaptation techniques have been proposed, including such popular approaches as maximum a posterior (MAP) estimation [1], maximum likelihood linear regression (MLLR) [2], and eigenvoices [3]. It is well-known that eigenvoices are the most effective among them when the amount of adaptation data available is extremely small, for example, a few utterances. In the eigenvoice method, mean vectors of all the states in all the HMMs from one speaker form a supervector, which we call a *speaker vector*. Then, principal component analysis (PCA) is applied to a set of speaker vectors from many speakers. In this paper, we focus on improving this eigenvoice-based technique.

Due to the high dimensional nature of speaker vectors, covariance matrix estimates needed for PCA may be unreliable. Thus, researchers have focused on improving the performance of eigenvoice techniques by using more robust estimates of covariance. For example, in the segmental eigenvoices approach [4], speaker vectors are first clustered structurally into several clusters; a cluster is constructed for each feature stream, or for each dimension in feature vectors, or for each state in HMMs, etc. PCA is then applied to each segment independently (this is equivalent to forcing the covariance matrix to be block diagonal, with each block corresponding to one segment). This method, however, has three problems which can lead to performance degradation. First, since it mainly uses *structural* clustering rather than *correlation-based* clustering, the correlation among the elements in the speaker vector might not be well preserved in the resulting clusters. Second, while some elements may have strong correlation to more than one cluster, each element should be assigned to only one cluster. Third, while there exist outlier elements which have no correlation to any other elements, they must be assigned to one of the clusters.

In this paper, we describe two new techniques to improve the performance of this segmental eigenvoice approach. First, we propose a soft-clustering method in which each element in the speaker vector can be assigned to more than one cluster. Second, those elements far apart from any of the clusters are removed before applying PCA.

It has been proved that elements *closer* to each other in a speaker vector tend to have larger correlations. For example, Shinoda *et al.* used the symmetrized Kullback-Leibler divergence as a distance measure for clustering the elements in structural MAP adaptation and proved its effectiveness [5]. Here, instead of using structural clustering, we propose using this distance measure in order to cluster the elements of a speaker vector.

This paper is organized as follows. Section 2 briefly summarizes the eigenvoice and the segmental eigenvoice methods. Section 3 proposes the soft-clustering and the outlier removal. Section 4 describes the experimental results in a large vocabulary recognition task, including the comparison between the segmental eigenvoice method and the proposed method. Section 5 is the conclusion of this paper.

2. Eigenvoice

Eigenvoice-based adaptation consists of two steps. In the training step, we obtain a set of eigenvectors, which we call eigenvoices, applying PCA to data obtained from a large number of training speakers. In the adaptation step, the model for a new speaker is constructed on-line by projecting the speaker's data to the subspace spanned by the eigenvoices.

2.1. Training step

In the training step, using the speaker-independent (SI) HMM for N speakers as the initial model, we make a speaker-dependent (SD) HMM for each of the N speakers. From each SD model, we extract a supervector which we call a speaker vector, which consists of the means of the Gaussian mixture in all the HMM states for that speaker. Speaker vector $\boldsymbol{x}_p$ of

Accepted after peer review of full paper

speaker p is

$$\boldsymbol{x}_p = (\boldsymbol{\mu}_{p,1}^t, \cdots, \boldsymbol{\mu}_{p,m}^t, \cdots, \boldsymbol{\mu}_{p,M}^t)^t, \quad (1)$$

where M is the number of Gaussian mixture components over all HMM states, $\boldsymbol{\mu}_{p,m}$ is the mean vector of m-th mixture component of speaker p, and t indicates transposition. Then the covariance matrix $\boldsymbol{C}$ corresponding to the N speaker vectors used for training is

$$C = \frac{1}{N}\sum_{p=1}^{N}(\boldsymbol{x}_p - \bar{\boldsymbol{x}})(\boldsymbol{x}_p - \bar{\boldsymbol{x}})^t, \quad (2)$$

where $\bar{\boldsymbol{x}}$ is the average of speaker vector. We apply PCA to this covariance matrix C and select a number of eigenvectors with the large eigenvalues.

2.2. Adaptation step

In the adaptation step, a model for a new speaker is constructed using a small amount of data from the speaker. It is assumed that the speaker vector $\hat{\boldsymbol{x}}$ of a new speaker is well represented by a linear combination of a small number of eigenvectors (eigenvoices) $(\boldsymbol{e}_k, k = 1, \cdots, K)$

$$\hat{\boldsymbol{x}} = \sum_{k=1}^{K}\omega_k \boldsymbol{e}_k + \bar{\boldsymbol{x}}, \quad (3)$$

where ω_k is the weight coefficient for the k-th eigenvector. Maximum likelihood estimation is used to estimate the ω_k for each new speaker, using data available for adaptation. This estimation process is called maximum likelihood eigen decomposition (MLED) [3].

2.3. Segmental eigenvoice

Recently, the segmental eigenvoice method [4] was proposed as an improvement to the eigenvoice approach. In this method, the training step is modified so that the speaker vector is first segmented into several "sub-vectors". These sub-vectors are generated by grouping together components of the original feature vector that has some common structural characteristics (e.g., in terms of phonology or feature type). In the adaptation step, the weight coefficients are estimated by MLED independently for each sub-vector or cluster and the estimated speaker vector for the new speaker is simply obtained by concatenating the speaker sub-vectors (one per cluster). This method suffers from performance degradation due to the three problems mentioned in Section 1.

3. Proposed method

We propose two new techniques to improve the performance of the segmental eigenvoice approach. Here, we assume that the mean of each mixture component in Gaussian-mixture HMMs, $\boldsymbol{\mu}_{p,m}$ in Eq.(1), forms one data point whose dimension is the same as the input feature vectors, and carry out clustering to all the data points. It should be noted that our method is more general in the sense that we can use any definitions of a data point other than a mixture component.

In the following, we first explain a soft-clustering method in which each element in speaker vector can be assigned to more than one cluster. Then, we describe how to remove those elements far apart from any of the clusters.

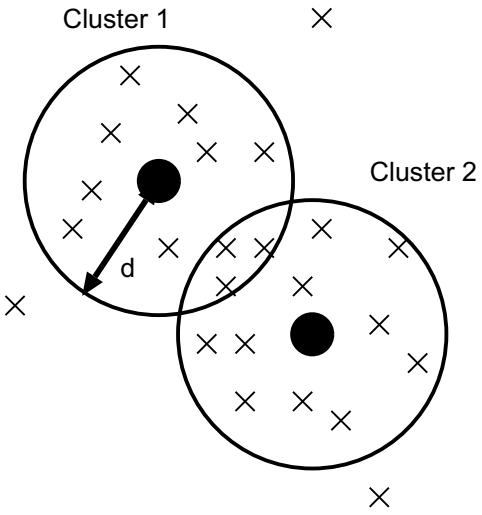

Figure 1: A soft-clustering example. A speaker vector is first segmented into elements with equal dimension and close elements to each other are clustered. Elements in the same cluster are then concatinated to form a sub-vector of the speaker vector. In this example, the dimension of each element is two and the number of clusters (i.e., sub-vectors) is two. "×" represents an element, "●" is a center of a cluster and d is a predetermined threshold for the distance between elements. Elements having smaller distance than d to two clusters are assigned to both of those two clusters. Elements having larger distance than d to any clusters are labelled as outliers.

3.1. Soft clustering

We propose a soft-clustering method that allows one element in the speaker vector to be assigned to more than one cluster. In this method, Gaussian mixture components of all the states of SI HMMs are clustered by k-means [6] clustering. As the distance measure, we use the average of symmetrized Kullback-Leibler divergence (KLD) $D(P, Q)$, which is defined as follows

$$D(P, Q) = \frac{D_{\mathrm{KL}}(P||Q) + D_{\mathrm{KL}}(Q||P)}{2}, \quad (4)$$

where P and Q are Gaussian distributions and $D_{\mathrm{KL}}(P||Q)$ is KLD from P to Q. If the distance between a mixture and a cluster is less than the predetermined threshold, the mixture is assigned to the cluster. If there is a mixture far apart from any of the clusters, it is labelled as *outlier*. An example of this soft-clustering is illustrated in Figure 1.

3.2. Modification of covariance matrix

In the next step, we modify the elements of the covariance matrix according to the result of soft clustering. First, all the outlier elements are set to zero, i.e., we ignore the corresponding crosscorrelation terms in the covariance matrix. Second, for each pair of mixtures which do not belong to the same cluster, the elements of the covariance matrix which correspond to this pair are also set to zero. Finally, we apply PCA to the resulting covariance matrix. This process is illustrated in Figure 2. It should be noted that, unlike the segmental eigenvoice method,

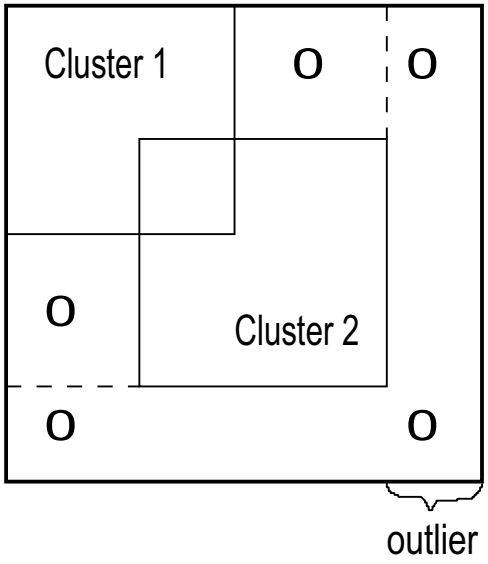

Figure 2: Covariance matrix after the modification process described in Subsection, when the number of clusters is two.

our method does not increase the number of weight coefficient to be estimated in MLED process. We expect this method to be robust when the amount of adaptation data is extremely small.

4. Experiments

4.1. Experimental Setup

We evaluated the proposed method in phoneme recognition using monophone HMMs. We compared the performance of our method with that of the speaker-independent (SI) HMMs, that of MLLR [2], and that of the segmental eigenvoice method.

We used two Japanese databases, JNAS [7] and S-JNAS [8]. JNAS consists of read speech of news paper articles. S-JNAS uses the same language resources as JNAS but consists of the speech data from elderly speakers. Each utterance in those databases correponds to one Japanese sentence whose duration is one to three seconds. JNAS consists of 266 (133 males and 133 females) speakers, each of which utters 150 sentences. S-JNAS consists of 300 (150 males and 150 females) speakers, each of which utters 200 sentenses. As training data we used 522 (222 in JNAS and 300 in S-JNAS) speakers' utterances. As adaptation data we used 20 utterances from 44 (22 males and 22 females) speakers, who were not involved in the training data. As test data we used another 50 utterances from the same speaker as those in the adaptation data.

We trained SI HMM which has 43 monophones, three states for each monophone, and one mixture for each state. A feature vector consists of 12 dimensional MFCCs, 12 dimensional ΔMFCCs, and Δpower. In the segmental eigenvoice method and the proposed method, all 129 mixtures (43 phonemes $\times$ 3 states $\times$ 1 mixture) are classified to 2, 4, and 8 clusters. The dimension of a speaker vector was 3225 (43 phonemes $\times$ 3 states $\times$ 25 dimensions). In the phoneme recognition, we employed a simple grammar representing Japanese syllable structure.

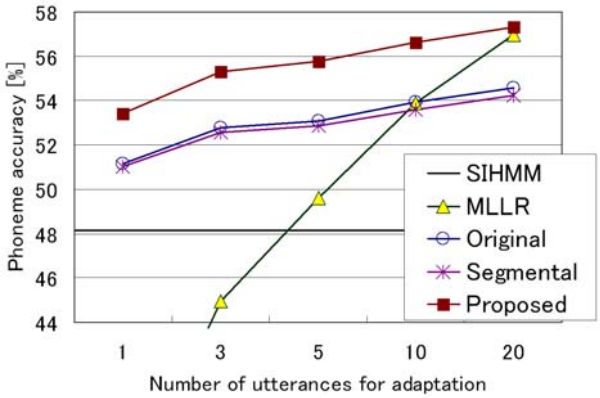

Figure 3: Comparison with the conventional methods when the number of utterances for adaptation is changed. Original is the original eigenvoice method, Segmental is the segmental eigenvoice method, and Proposed is our proposed method with soft clustering.

4.2. Comparison with the other eigenvoice methods

We first compared our method with the segmental eigenvoice method. In both methods, we used a set of 50 eigenvectors *in total* as the eigenvoice. In the segmental eigenvoice method, the symmetrized KLD was used as the distance measure for comparison with the proposed method. In both methods, the number of clusters was two. The threshold for the distance used in the proposed method was determined by our preliminary experiments using the test data. We observed that the difference of the threshold value did not influence the recognition accuracies so much.

The evaluation results with different numbers of utterances for adaptation are shown in Figure 3. The result showed that the performance of the segmental eigenvoice method was not as good as the original eigenvoice approach. This may be because the loss of the correlation information between those mixtures separated to different clusters deteriorated the performance. The performance of the MLLR method was not better than SI HMM using 1 and 3 utterances for adaptation, since the amount of data was too small. The performance of the proposed method had the best accuracy among all methods. This result confirmed the effectiveness of the proposed method.

4.3. Number of clusters

Next, we evaluated the proposed method with different number of clusters. Here, the number of eigenvectors in eigenvoice was fixed to 50 and the number of clusters was changed from 1 to 8. The threshold was optimized in the same way as in the previous experiment. The evaluation result is shown in Figure 4. The accuracies of the proposed method in all cases were better than the original eigenvoice method. When the number of clusters was four, the recognition accuracy was improved by 3.3 points using 10 utterances for adaptation.

We analyzed the relationship between the number of clusters and the phoneme accuracy. First, Cluster 1 and Cluster 2 had the same result, since the threshold in Cluster 2 was a relatively large value and one cluster was totally contained in

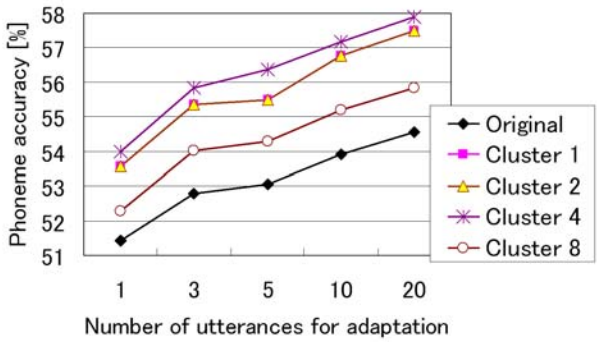

Figure 4: Recognition accuracies with different numbers of clusters. Original is the original eigenvoice method and Cluster 1, 2, 4, 8 are the results with the proposed method.

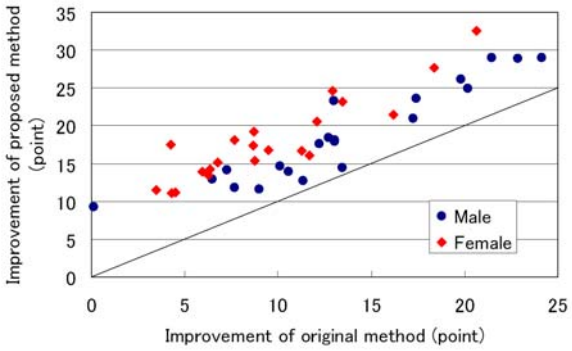

Figure 5: Improvement obtained by the original eigenvoice and the proposed method for each test speaker. Here the point means the improvement in phoneme accuracy (%) from the result by SI HMM.

another dominant cluster. Second, the difference between Original and Cluster 1 was more than that between Cluster 1 and Cluster 4. In Cluster 1, the improvement obtained by the proposed method came from the removal of outliers. This effect was much larger than the improvement obtained by increasing the number of clusters from 1 to 4. In our analysis of the clustering results, the 2nd states of four long vowels /a:/, /i:/, /e:/, /o:/, the 3rd state of /q/ (double consonant), the 1st and the 2nd states of /silB/ (silence before an utterance), the 2nd and the 3rd states of /silE/ (silence before an utterance), and the 2nd and the 3rd states of /sp/ (short pause) were categorized as outliers. In the 2nd states of long vowels, since they are the center states, the acoustic characteristic of speech tend to be steady. The feature values for Δ MFCCs and Δ power tends to come close to zero. This might be the major reason that the distance of these mixtures and each cluster becomes large. The phoneme such as /q/, /silB/, /silE/, and /sp/ mostly represent *silence*, and thus, they have little speech characteristics. Third, the performance of Cluster 8 was worse than that of Cluster 4. This may be because the mixtures which should be outliers tended to be assigned to some clusters when the number of clusters increased.

4.4. Performance for each speaker

Finally, the effect of the proposed method for each speaker was compared with the original eigenvoice. The result is shown in Figure 5, where 10 utterances were used for adaptation. The number of clusters in our method was four which gave the best performance. The result showed that the proposed method was effective for all speakers, especially for those speakers having a relatively small improvements by the original eigenvoice.

5. Conclusion

In this paper, we have proposed a soft-clustering method in which each Gaussian mixture component is assigned to more than one cluster by using the distance between mixtures. In addition, those mixtures far apart from any of the clusters are labelled as outlier. We implemented the proposed method by changing the value of the elements of the covariance matrix for speaker vectors. The result of our experiments showed that the proposed method was better than the original eigenvoice and the segmental eigenvoice methods. The phoneme accuracy was improved 3.3 points from the original eigenvoice on average using 10 utterances for adaptation. The result also indicates that the removal of the outlier mixtures was more effective than the soft clustering.

In future, we would like to apply our method to triphone HMMs. Since we can make the covariance matrix very sparse by using our techniques, it is expected that we can implement our eigenvoice method without much effort. Also, distance measure among the elements in the speaker vector should be investigated further.

6. References

[1] J. L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291-298, 1994.

[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, pp. 171-185, 1995.

[3] R. Kuhn and J.-C Junqua, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695-707, 2000.

[4] Y. Tsao, S.-M. Lee, F.-C. Chou, and L. S. Lee, "Segmental eigenvoice with delicate eigenspace for improved speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 399-411, 2005.

[5] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation", *IEEE Trans. Speech Audio Processing*, vol. 9, no. 3, pp. 276-287, 2001.

[6] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations", *Proc. 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1967.

[7] JNAS, http://www.milab.is.tsukuba.ac.jp/jnas/.

[8] S-JNAS, http://db.ciair.coe.nagoya-u.ac.jp/dbciair/koureisha_files/.