Segmental Eigenvoice for Rapid Speaker Adaptation

Yu Tsao, Shang-Ming Lee, Fu-Chiang Chou, and Lin-Shan Lee

Graduate Institute of Communication Engineering,
National Taiwan University, Taipei, Taiwan, R.O.C.
E-mail:lsl@iis.sinica.edu.tw

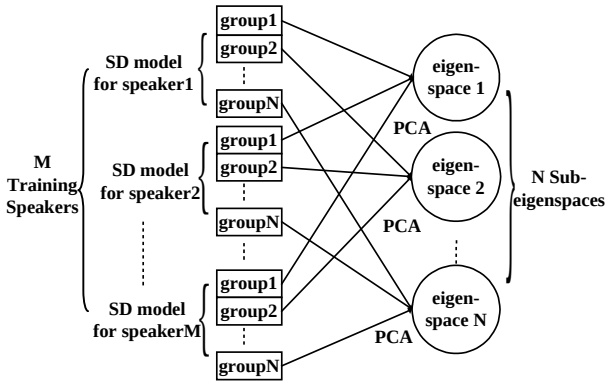
Abstract

This paper presents a new approach to improve the conventional eigenvoice technique. In the conventional eigenvoice, an eigenspace is established by introducing *a priori* knowledge of training speakers via PCA. The adaptation data is then used to determine a group of coefficients with respect to the eigenspace and build the SD model for the testing speaker. In the proposed approach, the eigenspace in the conventional eigenvoice is segmented into N sub-eigenspaces. Each sub-eigenspace is established by those components in the training speaker SD models with similar properties to each other. With the adaptation data, N groups of coefficients corresponding to the N sub-eigenspaces can be determined to build SD model for the new testing speaker. Here, both mixture-based and feature-based segmentation of eigenspace were tested, and improved results compared to the conventional eigenvoice were obtained in both cases. Even better results were obtained when these approaches were properly combined.

1. Introduction

Eigenvoice has been shown to possess the capabilities for rapid speaker adaptation with only limited quantities of adaptation data [1][2]; and there have been substantial research efforts made in this area [3][4]. The basic idea of eigenvoice is that principal component analysis (PCA) [5] was applied on the vector space constructed by the many parameters of the speaker dependent (SD) models for a group of training speakers, such that only those dimensions (eigenvectors) carrying the most data variations are extracted and used to establish the eigenspace. In the adaptation process, the adaptation data of a new speaker is used to determine a group of coefficients corresponding to the eigenvectors to indicate the coordinates of the new speaker in the eigenspace, with which the SD model for the new speaker can be built. In this paper, two new approaches to improve the conventional eigenvoice techniques are proposed. The basic concept of the approaches is shown in Figure1. In both approaches, we try to segment the eigenspace of the conventional eigenvoice technique into N sub-eigenspaces. The parameters of the SD models of each training speaker are somehow classified into N groups, each with some common properties. Each sub-eigenspace is then established with one such group of parameters obtained from all the training speakers. With the adaptation data for a new testing speaker, we obtain N groups of coefficients, each for one of the sub-eigenspace, to be used to build the SD model for the new testing speaker. By segmenting the eigenspace into N smaller sub-eigenspace, the construction of the SD model for the new speaker can be more precise. The first approach is a mixture-based segmentation of the eigenspace, i.e., all the mixtures in the training speaker SD models are classified into N clusters by the relative similarity in acoustic-phonetic properties among them. The second approach, on the other hand, is a feature-based segmentation of the eigenspace, i.e., all the components of the mean vectors for the mixtures in the training speaker SD models are segmented into N different groups of feature parameters (energy, MFCC, ΔMFCC…etc), and therefore the eigenspace is segmented into N sub-eigenspaces in this way. It will be shown in the experiments below that not only these two approaches can both provide better adaptation performance than the conventional eigenvoice technique, but also they can be properly combined (segmentation in both mixtures and features) to offer even better performance.

(a) *Training process of the proposed approach*



(b) *Adaptation process of the proposed approach*

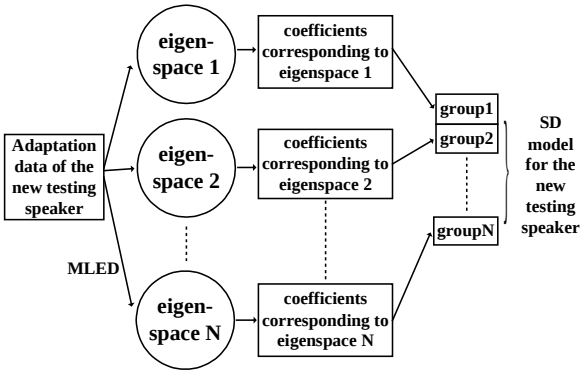


Figure 1:(a) *training process* (b) *adaptation process of the proposed approach.*

2. Conventional Eigenvoice Approach

The eigenvoice approach has two steps - training and adaptation steps. In the training step, an eigenspace is established by introducing *a priori* knowledge of a large number training speakers and PCA is performed to reduce the total number of dimensions. In the adaptation step, the model for a new testing speaker is located in this lower-dimensional eigenspace. Because the number of parameters is significantly reduced, we can adapt the testing speaker with relatively smaller quantities of adaptation data.

2.1. Training step

M well-trained SD models for M training speakers were obtained. Each SD model is used to construct a supervector with some large dimension D. We can then evaluate the covariance matrix for these supervectors.

PCA is applied to this covariance matrix to obtain M eigenvectors and M eigenvalues. K out of the M eigenvectors with relatively larger corresponding eigenvalues are selected to construct a K-dimensional eigenspace. These K eigenvectors capture most of the information of the training data. K is significantly smaller than M.

2.2. Adaptation Step

The new testing speaker is represented by a single point in the K-dimensional eigenspace by finding a set of coefficients $\{ \omega(j), j = 1,2...K \}$ corresponding to the eigenvectors $\{ e(j), j = 1,2...K \}$ of this eigenspace by maximum likelihood eigen-decomposition approach (MLED). Then we can build the SD model for the new testing speaker:

$$\vec{x} = \sum_{j=1}^{K} \omega(j) e(j);$$

Where $\vec{x}$ is the supervector for the parameters of the SD model for the new test speaker.

3. The Proposed Approach:

3.1. Mixture-based Segmental Eigenvoice

In this approach, all the mixtures in the SD models are classified into N clusters based on the acoustic-phonetic characteristics. In the initial experiments to be presented below, this classification is based on the Bhattacharyya distance between mixtures, although other distance measures may also be used. Then, all the components in the supervector belonging to a specific cluster are used to establish a "sub-supervector", with which a sub-eigenspace for the specific cluster is constructed. In the adaptation step, the position of a new testing speaker corresponding to the mixtures belonging to a specific class is determined in each sub-eigenspace. In other words, each new testing speaker has a total of N sets of eigenvoice-coefficients corresponding to the N sub-eigenspaces. The SD model for the new testing speaker is constructed based on these N sets of coefficients and these N sub-eigenspaces.

3.2. Feature-based Segmental Eigenvoice

In this approach, it is assumed that the feature parameters can be divided into several groups, such as those based on the energy components, the MFCC components, the Δ MFCC components ...etc, and the correlation among different groups of feature parameters are in fact limited. Therefore, in the training step, we segment all the feature vectors into N parts. In the initial experiments to be presented below, this is done for three parts, the energy components part, the MFCC components part and the Δ MFCC components part. All the components in each part are tied together to form a "sub-supervector" individually. After performing PCA, N sub-eigenspaces are obtained. In the initial experiments below, there are the sub-eigenspace established with energy components, sub-eigenspace established with MFCC components, and sub-eigenspace established with Δ MFCC components. In the adaptation step, N sets of coefficients corresponding to the N sub-eigenspaces are obtained and each set of coefficients is used to determine the corresponding part of feature parameters in the SD model for the new testing speaker.

4. Initial Experiments

4.1. Experimental Setup

In the initial experiments, a Mandarin dictation speech database recorded at Taipei produced by 100 male speakers was used in training. Each speaker produced 200 utterances or sentences. The average length of the utterances or sentences is roughly 3 seconds or 11 syllables. Because the focus here is rapid adaptation, we limited the adaptation data for each new speaker to be less than 40 utterances or sentences, which corresponds to less than 2 minutes. A total of 4 testing speakers different from the 100 training speaker were tested, in which from 2 to 40 adaptation utterances or sentences for each testing speaker were used. The results presented below are the average of the 4 testing speakers. The speech signal was sampled at 16 kHz, and parameterized into 1 dimension of energy component, 1 dimension of delta energy component, 14 dimensions of MFCC components and 14 dimensions of $\triangle$ MFCC components. CMS was performed on a per-utterance/sentence basis to remove the channel effect of the features. The SI model was trained by the 100 male speakers. Because of the monosyllabic structure of Chinese Language, we adopted the Right-Context-Dependent (RCD) Initial –Final model format, which includes 112 RCD Initial models along with 38 Context-Independent (CI) Final models. Here, Initial is the initial consonant of a Mandarin syllable, while Final is the vowel (diphthong) part of the syllable plus an optional medial and an optional nasal ending. These models were then used as the acoustic units for SI model training. Each Initial model has 3 states, whereas each Final has 4 states and the silence model has 1 state. Each state of Initial or Final models has mixture numbers ranging from 1 to 4, and the state of the silence model has 8 mixtures. Each SD model of the training speakers was then further adapted by the 200 utterances or sentences with MAP adaptation. In the experiments for conventional eigenvoice, all the mixtures of the SD models for the 100 training speakers were used to obtain 100 57780-dimensional supervectors. After performing PCA, we had an

eigenspace of dimension K, where K is the number of eigenvectors to be chosen. In the first proposed approach of mixture-based segmental eigenvoice, three, six and fifteen clusters were tested. Take the three-cluster case as an example; all the mixtures in each training speaker's SD model were classified into three clusters according to the Bhattacharyya distance between them. Other distance measures can also be used, but not tested yet. For the second proposed approach of feature-based segmental eigenvoice, the feature parameters were divided into three groups, which are the two energy components, the 14 MFCC components and the 14 △MFCC components. Other partitions are also possible, but not tested yet.

4.2. Testing Results

The first four rows (1), (2), (3), (4) of Table 1, show the syllable recognition error rate performance of the conventional eigenvoice and the first proposed approach of mixture-based segmental eigenvoice with N=3, 6, 15 cluster. We can tell from Table 1 that 3-cluster mixture-based segmental eigenvoice performed better than the 6-cluster and 15-cluster cases, and almost always better than the conventional eigenvoice (except for the 2 utterances case), when the available adaptation data is less than 40 utterances or sentences. When the available adaptation data is too small, the performance of the mixture-based eigenvoice may become worse than the conventional eigenvoice (2 utterances or sentences for 3 clusters and less than 12 utterances or sentences for 6 clusters, for example). This is probably because the amount of available adaptation data in each cluster is too small; the accuracy of coefficient estimation will inevitably be degraded. In the case of more adaptation data are available, the 6-cluster, or 15-cluster cases may offer better performance than 3 clusters, although this is not tested yet.

The fifth row (5) of Table 1 lists the syllable error rates for the second proposed approach of feature-based segmental eigenvoice. The achieved syllable error rate is almost in parallel to and always better than the conventional eigenvoice by 0.5% to 1% no matter the quantities of adaptation data are small or not, except for the case of 2 utterances/sentences only, apparently due to too limited adaptation data. These results again verified the concept here, i.e., estimating parameters of energy components, MFCC components and ΔMFCC components by segmenting the conventional eigenspace into several sub-eigenspace is better than using a single eigenspace. Since both mixture-based and feature-based approaches have shown better performance than the conventional eigenvoice, we tried to combine these two approaches to see if the improvements achieved are additive. First, all the mixtures in each training speaker's SD model are classified into three clusters. Then, each mean vector for mixtures in a certain cluster is segmented into energy components, MFCC components and ΔMFCC components and tied together individually. So each training speaker's SD model was used to form 9 sub-supervectors. After performing PCA, we got 9 sub-eigenspaces. In the testing phase, for each testing speaker the adaptation data were used to determine 9 sets of coefficients corresponding to these 9 sub-eigenspaces, and the SD model for the testing speaker was then built. The last row (6) of Table 1 are the results of adaptation performance for the combined approach. As can be easily seen, the combined approach performed better than either one of the individual approaches.

The results for the conventional eigenvoice, the 3-cluster mixture-based approach, the feature-base approach and the combined approach in rows (1), (2), (5), (6) are also plotted in Figure 2.

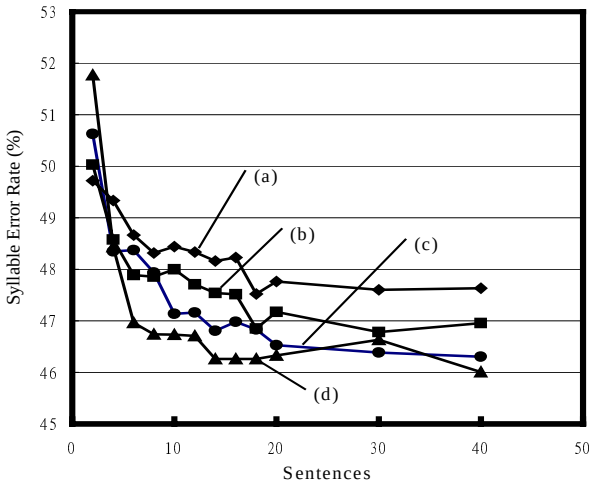


Figure 2:*Adaptation performance of mixture-based segmental eigenvoice, feature-based segmental eigenvoice and combination of the two methods.* (a) *Conventional eigenvoice* (b) *Feature-based Segmental Eigenvoice* (c) *Mixture-based Segmental Eigenvoice* (d) *Combined mixture/feature-based Segmental Eigenvoice. SI syllable error rate is 50.94%*

5. Further Discussions

Some additional discussions are made here in this section. First, to determine the dimension of eigenspace is a critical problem. If choosing a too large dimension of eigenspace when there are not enough adaptation data, the accuracy of the eigenspace coefficients will be degraded. If there are enough adaptation data, but a too small dimension of eigenspace is chosen, the opportunity to build the best SD model for the testing speaker may be lost. Here in the experiments mentioned above, we performed a preliminary experiment for each case to deal with this problem. In each of the above experiments, tests with respect to one reference testing speaker out of the 4 testing speakers were performed for dimension varying from 5 to 50, and the dimension achieving the best performance was then used for all the 4 testing speakers in each of the experiments reported here.

On the other hand, it is surely possible to segment the conventional eigenvoice using some other basis, or into some more detailed sub-eigenspaces. One possibility is based on the phones. The first two rows (1), (2) of Table 2 are the results of another approach of phone-based segmental eigenvoice by dividing the phones into 3 or 6 clusters simply based on linguistic knowledge. Comparing to row (1) of Table 1, the improvements with respect to the conventional eigenvoice is obvious. Comparing to rows (2), (3) of Table 1, it seems that mixture-based approach is better than phone-based when the adaptation data is limited. This is probably because some of the phones actually have only very little adaptation data to build the SD model accurately for the testing speaker. This phenomenon then disappeared when more adaptation data is available and the performance of phone-based segmental eigenvoice became comparable to mixture-based approach.

We also tried to combine the feature-based approach and the phone-based approach with 3 clusters of phones. The results are listed in row (3) of Table 2. It can be found that when the adaptation data is limited, the combination of mixture-based and feature-based approaches in row (6) of Table 1 gave better performance. Nevertheless, the combination of feature-based and phone-based approaches may provide even better performance when there is more data. However, when more data are needed, the advantage of rapid adaptation for eigenvoice may be lost, and better performance can in fact be achieved by Maximum Likelihood Linear Regression (MLLR) approach [6], for which the results with full matrices are listed in the last row (4) of Table 2.

6. Conclusions

The approaches in this paper are based on the concept that, if segmenting the eigenspace into some smaller sub-eigenspaces, the construction of the SD model for the new speaker can be made more precisely. Mixture-based segmental eigenvoice is based on the known acoustic characteristics of all the mixtures, while the feature-based segmental eigenvoice is based on the hypothesis of limited correlation among different groups of feature parameters. Phone-based segmental eigenvoice is also possible and is tested in our further discussion. Experimental results indicate that all these approaches, feature-based, mixture-based or phone-based segmental eigenvoice provided better performance than the conventional eigenvoice technique, and the combination of them provided even better results.

7. References

[1] R. Kuhn, et. al, "Eigenvoice for Speaker Adaptation," *proc.* ICSLP' 98, pp.1771-1774.

[2] R. Kuhn, et. al., "Fast Speaker Adaptation Using *a priori* Knowledge," *proc.* ICASSP' 99, pp.1587-1590.

[3] Kuan-ting Chen, Wen-wei Liau, Hsin-min Wang and Lin-shan Lee "Fast Speaker Adaptation Using Eigenspace-Based Maximum Likelihood Linear Regression," *proc.* ICSLP' 2000.

[4] Nick T.-C. Wang, Sammy S.-M. Lee, Frank Seide, and Lin Shan Lee"EigenMLLR Adaptation fast Speaker Adaptation Using *a priori* Knowledge by Transformations Combination Based Approach," *proc.* ICASSP' 2000.

[5] I.T. Jolliffe, "Principal Component Analysis," Springer-Verlag, 1986.

[6] C.J. Leggetter & 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.

Table 1: Syllable error rates (%) for the conventional eigenvoice, mixture-based, feature-based and the combination approaches of segmental eigenvoice. SI syllable error rate is 50.94%

Number of Utterances/sentences		2	4	6	8	10	12	14	16	18	20	30	40
Conventional Eigenvoice (1)		49.72	49.34	48.67	48.32	48.44	48.34	48.16	48.23	47.52	47.77	47.60	47.63
Mixture-based Segmental Eigenvoice	3 Clusters (2)	50.63	48.35	48.37	47.94	47.14	47.16	46.81	46.98	46.83	46.53	46.38	46.31
	6 Clusters (3)	52.95	49.65	49.47	48.95	48.49	48.49	47.96	48.18	47.61	47.91	46.14	47.04
	15 Clusters (4)	53.66	49.56	49.38	48.84	48.61	48.52	48.81	48.57	47.88	47.85	47.30	47.22
Feature-based Segmental Eigenvoice (5)		50.03	48.58	48.14	47.86	48.00	47.71	47.54	47.52	46.85	47.17	46.78	46.96
Combined mixture-based and feature-based Approach (6)		51.74	48.44	46.96	46.74	46.73	46.71	46.26	46.26	46.26	46.33	46.63	46.01

Table 2: Syllable error rates (%) for the phone-based approach, combined phone-based and feature-based approach, and MLLR (the last row) with full matrix.

Number of Utterances/sentences		2	4	6	8	10	12	14	16	18	20	30	40
Phone-based Segmental Eigenvoice	3 Clusters (1)	51.20	49.15	49.10	48.65	47.50	47.58	47.61	47.24	47.52	47.06	46.98	46.76
	6 Clusters (2)	53.28	49.83	48.63	48.37	48.55	48.42	48.85	48.29	47.79	47.06	46.98	47.03
Combined phone-based and feature-based Approach (3)		53.26	49.07	48.24	48.03	47.87	47.55	47.51	46.61	46.56	45.91	46.04	45.79
MLLR with Full Matrix (4)		90.35	82.67	73.17	69.90	66.39	63.84	58.67	51.71	50.71	50.48	47.36	45.20