Segmental Eigenvoice With Delicate Eigenspace for Improved Speaker Adaptation

Yu Tsao, Shang-Ming Lee, and Lin-Shan Lee

***Abstract*—Eigenvoice techniques have been proposed to provide rapid speaker adaptation with very limited adaptation data, but the performance may be saturated when more adaptation data become available. This is because in these techniques an eigenspace with reduced dimensionality is established by properly utilizing the *a priori* knowledge from the large quantity of training data. The reduced dimensionality of the eigenspace requires less adaptation data to estimate the model parameters for the new speaker, but also makes it less easy to obtain more precise models with more adaptation data. In this paper, a new segmental eigenvoice approach is proposed, in which the eigenspace can be further segmented into N subeigenspaces by properly classifying the model parameters into N clusters. These N subeigenspaces can help to construct a more delicate eigenspace and more precise models when more adaptation data are available. It will be shown that there can be at least mixture-based, model-based and feature-based segmental eigenvoice approaches. Not only improved performance can be obtained, but these different approaches can be properly integrated to offer better performance. Two further approaches leading to improved segmental eigenvoice techniques with even better performance are also proposed. The experiments were performed with both a large vocabulary and a small vocabulary recognition tasks.**

***Index Terms*—Eigenvector approach, principal component analysis, speaker adaptation.**

I. Introduction

THE mismatch in acoustic characteristics between speech signals produced by the training speakers and those by the testing speakers has been causing serious performance degradation for automatic speech recognition (ASR) systems. Substantial efforts have been made to try to reduce such mismatch, which are usually referred to as speaker adaptation techniques. In these techniques, a limited quantity of adaptation data produced by the new speaker is used to generate a set of speaker dependent (SD) models for the new speaker by adapting the parameters of a set of speaker independent (SI) models toward the directions of the acoustic characteristics of the new speaker. Efficient use of the limited quantity of the adaptation data to obtain the highest achievable recognition accuracy has been an important direction in speaker adaptation.

Many different speaker adaptation techniques have been developed and shown to be successful. Maximum a posterior (MAP) approach [4] and maximum likelihood linear regression (MLLR) [5], [6] approach are good examples. MAP approach adapts the model parameters based on the MAP criterion. It has been shown to offer probably the best performance if enough adaptation data are available. When the available data are not adequate, however, special adaptation processes are necessary to provide very good performance [7], [8]. MLLR approach adapts the model parameters using a set of linear regression functions with parameters estimated by maximum likelihood criteria. It was shown to be able to achieve very good performance with smaller quantity of training data. But the performance turned out to be saturated when more adaptation data become available. Some improved MLLR approaches have been proposed to deal with this problem [9]–[13]. Eigenvoice techniques, on the other hand, have been shown to possess the distinct feature of very rapid speaker adaptation. Significant performance improvements have been obtained with only very limited quantity of adaptation data [1]–[3], [14], [15].

The basic idea of eigenvoice approach is to apply the principal component analysis (PCA) [16] 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 (or the eigenvectors) carrying the largest data variations are extracted and used to establish the eigenspace with reduced dimensionality. The acoustic properties of each speaker is projected onto this eigenspace as a vector. In the adaptation process, the adaptation data of a new speaker are used to determine this projection in the eigenspace representing the new speaker by estimating its component on each dimension of the eigenvectors, with which the SD models for the new speaker can be constructed. Because the PCA process has chosen the eigenvectors carrying the largest data variations, and the dimensionality of the eigenspace is low, only small numbers of parameters are needed, and, therefore, good performance can be achieved with very limited quantity of adaptation data. However, the low dimensionality of the eigenspace or the small number of parameters to be estimated in the eigenvoice approach may also imply some inherent limitations. The eigenspace representation of the speaker characteristics can't be made more delicate, more exquisite or more precise when more data become available. As a result, in many cases the performance of eigenvoice approaches may be saturated quickly as the adaptation data increase. Some new approaches have been developed toward this direction recently [17].

In this paper, a new approach to improve the eigenvoice technique by "segmenting" the eigenspace into N subspaces (called subeigenspaces below) is proposed, with a goal to develop a more delicate or precise eigenspace when more adaptation

Manuscript received May 3, 2003; revised January 9, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jerome R. Bellegarda.

The authors are with the Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan, R.O.C. (e-mail: jtsao@speech.ee.ntu.edu.tw; sammy@speech.ee.ntu.edu.tw; lslee@gate.sinica.edu.tw).

Digital Object Identifier 10.1109/TSA.2005.845819

data become available. In this new approach, the parameters of the SD models for the training speakers are somehow classified into N clusters. Each of such cluster of parameters collected from the SD models for all the training speakers are then used to construct a subeigenspace individually. In this way the eigenspace of eigenvoice techniques is "segmented" into N subeigenspaces, and it becomes possible to make the eigenspace more delicate and precise when more adaptation data become available. This new approach is referred to as the segmental eigenvoice [18] here in this paper. As will be clearer later on in this paper, there can be at least mixture-based, model-based, and feature-based approaches to classify the parameters and develop the subeigenspaces. Experimental results show that the recognition performance can actually be significantly improved by this segmental eigenvoice approach, and furthermore, the different mixture-based, model-based and feature-based segmental eigenvoice approaches can be properly integrated to offer even better performance.

The rest of this paper is organized as follows. Section II briefly summarizes the eigenvoice approach, and Section III presents the concept of the segmental eigenvoice. Section IV describes the experimental results for a large vocabulary recognition task, including those for baseline experiments and various segmental eigenvoice approaches, and those for some improved approaches for segmental eigenvoice. Section V then presents an extra set of experiments to verify the superiority of the segmental eigenvoice approaches for a small vocabulary digit recognition task. Section VI is the conclusion.

II. Brief Summary of the Eigenvoice Approach

The eigenvoice approach is summarized very briefly here for illustration purposes. In the training step, an eigenspace is established off-line by analyzing the *a priori* knowledge from a large number of training speakers with PCA performed to reduce the dimensionality. In the adaptation step, the model for a new speaker is constructed on-line according to the projection of the vector for the new speaker on the lower-dimensional eigenspace. Because the number of parameters to be estimated is significantly reduced, the adaptation can be achieved with relatively smaller quantity of adaptation data.

A. Training Step for Eigenvoice Approach

M sets of well-trained SD models for M training speakers, $\{s = 1, 2 \ldots \mathrm{M}\}$, were first obtained. The parameters in the SD models for a speaker s are used to construct a supervector $\mathbf{X}_s$ with a large dimension d, where d is the total number of distinct parameters for the SD models for a specific speaker. $\mathbf{X}_s$ is considered as a sample of a random vector $\mathbf{X}$ with dimension d. This random vector $\mathbf{X}$ describes the acoustic characteristics across different speakers. PCA is then performed on the ensemble of the supervectors $\mathbf{X}_s$, $s = 1, 2 \ldots \mathrm{M}$, collected from all the M training speakers, by finding the eigenvectors and eigenvalues of the covariance matrix $\mathbf{G}$ for the random vector $\mathbf{X}$, where $\mathbf{G}$ is estimated with the ensemble of the supervectors $\mathbf{X}_s$, $s = 1, 2 \ldots \mathrm{M}$. k eigenvectors, $\{\boldsymbol{e_j}, j = 1, 2 \ldots \mathrm{k}\}$, with the largest corresponding eigenvalues are finally selected to construct a k-dimensional eigenspace $\mathbf{S}$, where k is a significantly smaller number. These k eigenvectors and, therefore, the k-dimensional eigenspace carry most of the speaker information for the M training speakers.

B. Adaptation Step for Eigenvoice Approach

A new speaker is represented by a vector projected onto the k-dimensional eigenspace

$$\nu = \sum_{j=1}^{\mathrm{k}} w_j \mathrm{e}_j \tag{1}$$

where the coefficients w_j, $j = 1, 2 \ldots \mathrm{k}$, are estimated by the maximum likelihood eigen-decomposition (MLED) approach, which is achieved with the expectation–maximization (EM) algorithm [19] via an auxiliary function [1]–[3]. The models of the new speaker are then constructed from the obtained vector ν.

III. Concept of the Segmental Eigenvoice Approach

As mentioned before, the segmental eigenvoice approach "segments" the eigenspace of the original eigenvoice approach into N subeigenspaces. Everything else is very similar to the original eigenvoice approach. The formulation of the whole concept is given below and illustrated in Figs. 1–3.

A. Training Step for Segmental Eigenvoice Approach

M sets of well-trained SD models for M training speakers, $\{s = 1, 2 \ldots \mathrm{M}\}$, were obtained, each with d distinct parameters as in the original eigenvoice approach mentioned above. These d parameters in the SD models for each speaker s are first classified in some way into N clusters, $\{c = 1, 2 \ldots \mathrm{N}\}$, where each cluster c includes d_c parameters

$$\sum_{c=1}^{\mathrm{N}} \mathrm{d}_c = \mathrm{d}. \tag{2}$$

The detailed processes of this classification of d parameters into N clusters will be discussed below. It can be at least mixture-based, model-based, or feature-based. The d_c parameters in the cluster c for the SD models of speaker s are then used to construct a subsupervector $\mathbf{X}_{c,s}$ with dimension d_c. In other words, the supervector $\mathbf{X}_s$ obtained previously for each training speaker s in the original eigenvoice approach is now "segmented" into N subsupervectors $\mathbf{X}_{c,s}$, $c = 1, 2 \ldots \mathrm{N}$. These N subsupervectors $\mathbf{X}_{c,s}$, $c = 1, 2 \ldots \mathrm{N}$, are then taken, respectively, as samples of N random subvectors $\mathbf{X}_c$, $c = 1, 2 \ldots \mathrm{N}$, each for a cluster c and describing a subset of the acoustic characteristics across different speakers. Again, the random vector $\mathbf{X}$ in the original eigenvoice approach mentioned previously describing the acoustic characteristics across different speakers is "segmented" here into N random subvectors $\mathbf{X}_c$, $c = 1, 2 \ldots \mathrm{N}$. PCA can now be performed, respectively, on each of these N random subvectors mentioned above, $\mathbf{X}_c$, $c = 1, 2 \ldots \mathrm{N}$, to find the eigenvectors and eigenvalues for the corresponding covariance matrices $\mathbf{G}_c$, $c = 1, 2 \ldots \mathrm{N}$, which are estimated from the ensembles of the subsupervector $\mathbf{X}_{c,s}$, $s = 1, 2 \ldots \mathrm{M}$, $c = 1, 2 \ldots \mathrm{N}$, collected from all the M training speakers. For the random subvector $\mathbf{X}_c$ for the cluster c of parameters, k_c eigenvectors with the largest corresponding eigenvalues, $\{\boldsymbol{e}_{\boldsymbol{j}}^{(c)}, j = 1, 2 \ldots \mathrm{k}_c\}$, are selected to construct

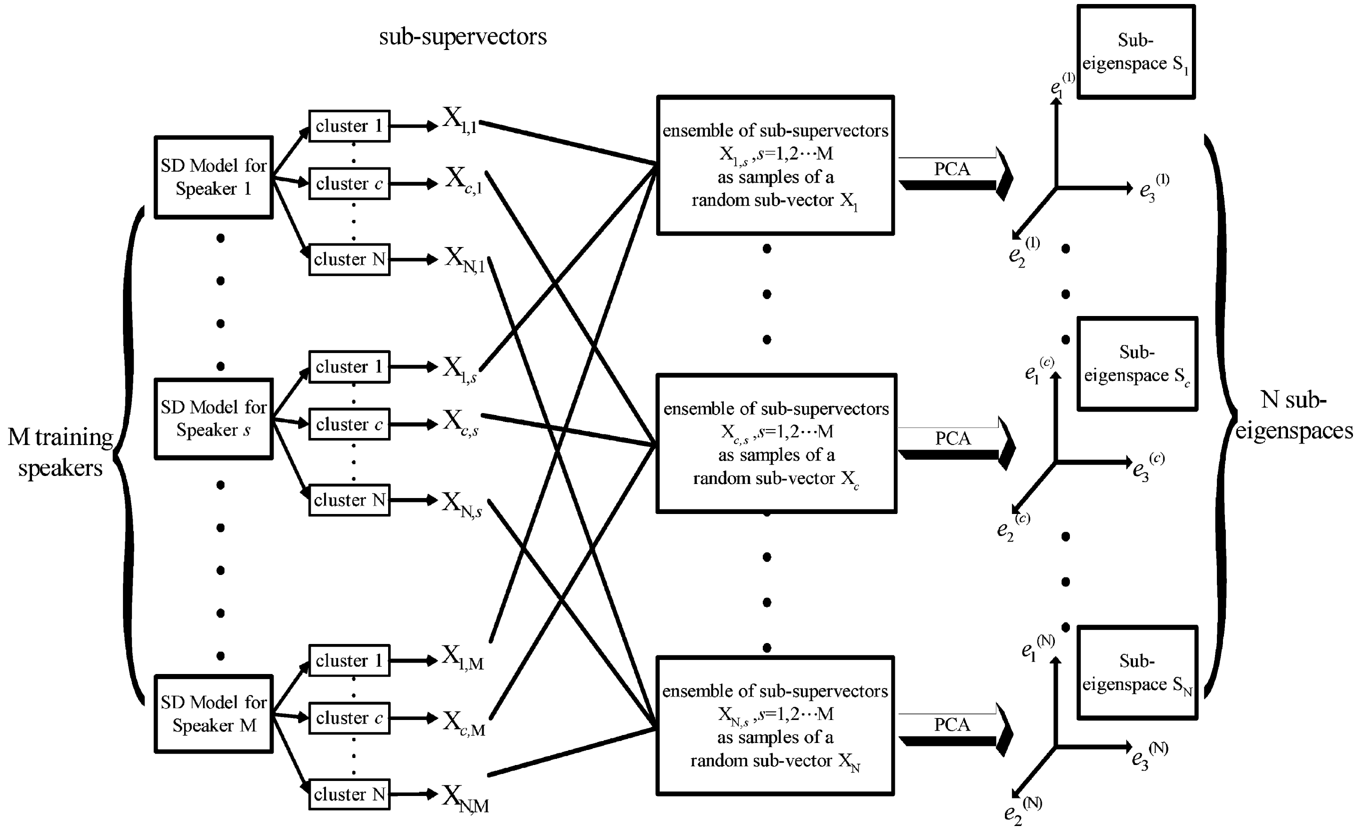


Fig. 1. Training step of the segmental eigenvoice approach.

a k_c-dimensional subeigenspace $\mathbf{S}_c$, and k_c is a significantly smaller number. As a result, the eigenspace $\mathbf{S}$ in the original eigenvoice approach mentioned previously is "segmented" into N subeigenspaces $\mathbf{S}_c$, $c = 1,2\ldots N$. The whole training process mentioned above is illustrated in Fig. 1.

B. Adaptation Step for Segmental Eigenvoice Approach

In the adaptation step, N sets of coefficients, $\{w_j^{(c)}, j = 1,2\ldots k_c\}$, $c = 1,2\ldots N$, corresponding to the N subeigenspaces $\mathbf{S}_c$ are estimated for the new speaker by the maximum likelihood eigen-decomposition (MLED) approach via the EM algorithm. Each new speaker is then represented by the N subvectors

$$\nu_c = \sum_{j=1}^{k_c} w_j^{(c)} e_j{}^{(c)}, \quad c = 1,2\ldots N \tag{3}$$

in the N subeigenspaces. The models of the new speaker are then constructed from a vector ν in the entire k-dimensional eigenspace, which is formed by concatenating these N subvectors ν_c, $c = 1,2\ldots N$

$$\nu = \{\nu_1, \ldots \nu_c, \ldots v_N\}. \tag{4}$$

The whole adaptation process mentioned above is illustrated in Fig. 2. If we consider each of the above subvector v_c , $c = 1,2\ldots N$, to be a vector $\bar{v}_c$ in the entire k-dimensional eigenspace but with all other components being zero, i.e.

$$\begin{aligned}\bar{\nu}_1 &= \{\nu_1, 0, \ldots 0, \ldots 0\}\\ \bar{\nu}_c &= \{0, \ldots 0, \nu_c, 0, \ldots 0\}\\ \bar{\nu}_N &= \{0, \ldots\ 0, \ldots\ 0, \nu_N\}.\end{aligned}$$

then the above (4) can also be represented as a summation

$$\nu = \sum_{c=1}^{N} \bar{\nu}_c.$$

Such a relation can then be further illustrated by the vector spaces in Fig. 3.

C. Mixture/Model/Feature-Based Segmental Eigenvoice

There can be at least several different ways to classify the d model parameters into N clusters as mentioned above. The most natural approach may be mixture-based, i.e., those parameters for mixtures with closer acoustic properties are classified together in a cluster. In this approach, all the Gaussian mixtures in the SD models, and, therefore, their parameters, are classified into N clusters based on some distance measure. There can be many different ways of evaluating the distance between two Gaussian mixtures, and Bhattacharyya distance [20] and the Divergence parameter [21] are two example measures used in the experiments presented below. Bhattacharyya distance is defined as

$$D_B = \frac{1}{8}(\mu_1 - \mu_2)^T \left(\frac{1}{2}(\Sigma_1 + \Sigma_2)\right)^{-1} (\mu_1 - \mu_2) + \frac{1}{2}\ln\frac{\left|\frac{(\Sigma_1+\Sigma_2)}{2}\right|}{|\Sigma_1|^{1/2}|\Sigma_2|^{1/2}} \tag{5}$$

while the Divergence parameter is defined as

$$D_D = \sum_i \left(\frac{\sigma_2(i)^2 + \Delta_{12}(i)^2}{\sigma_1(i)^2} + \frac{\sigma_1(i)^2 + \Delta_{12}(i)^2}{\sigma_2(i)^2}\right) \tag{6}$$

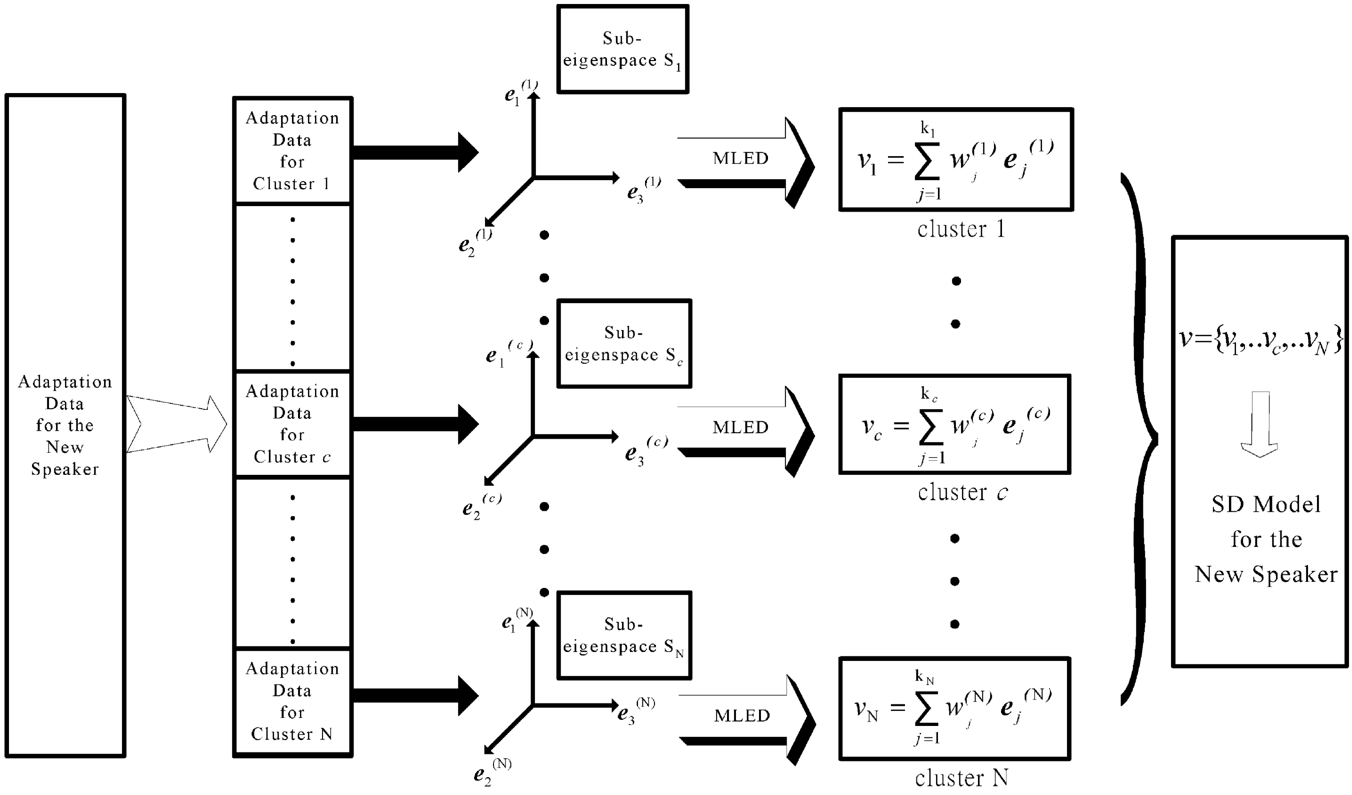

Adaptation Data for the New Speaker
Adaptation Data for Cluster 1
Adaptation Data for Cluster c
Adaptation Data for Cluster N
Sub-eigenspace S_1
Sub-eigenspace S_c
Sub-eigenspace S_N
$e_1^{(1)}$ $e_2^{(1)}$ $e_3^{(1)}$
$e_1^{(c)}$ $e_2^{(c)}$ $e_3^{(c)}$
$e_1^{(N)}$ $e_2^{(N)}$ $e_3^{(N)}$
MLED
MLED
MLED
$v_1 = \sum_{j=1}^{k_1} w_j^{(1)} e_j^{(1)}$
cluster 1
$v_c = \sum_{j=1}^{k_c} w_j^{(c)} e_j^{(c)}$
cluster c
$v_N = \sum_{j=1}^{k_N} w_j^{(N)} e_j^{(N)}$
cluster N
$v=\{v_1,..v_c,..v_N\}$
SD Model for the New Speaker

Fig. 2. Adaptation step of the segmental eigenvoice approach.

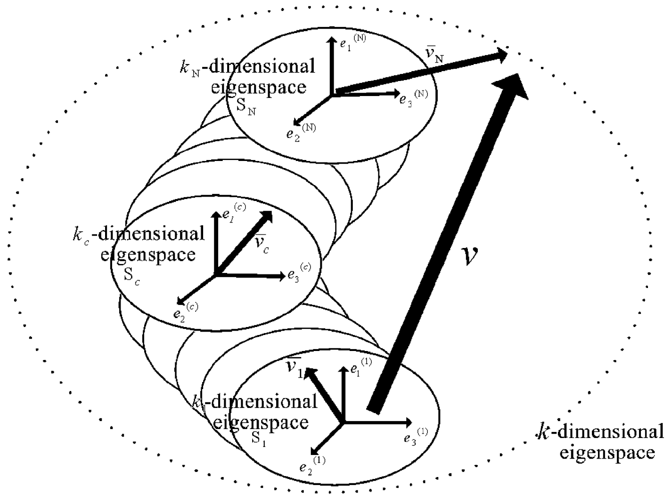

k_N-dimensional eigenspace S_N
$e_1^{(N)}$ $e_2^{(N)}$ $e_3^{(N)}$
$\bar{v}_N$
k_c-dimensional eigenspace S_c
$e_1^{(c)}$ $e_2^{(c)}$ $e_3^{(c)}$
$\bar{v}_c$
v
$\bar{v}_1$
k_1-dimensional eigenspace S_1
$e_1^{(1)}$ $e_2^{(1)}$ $e_3^{(1)}$
k-dimensional eigenspace

Fig. 3. Relationship among the k_c-dimensional subeigenspace S_c, $c = 1, 2 \ldots \mathrm{N}$, and the entire k-dimensional eigenspace.

where μ_1, μ_2 are the mean vectors of the two mixtures, Σ_1, Σ_2 are the covariance matrices of the two mixtures, $\sigma_1(i)^2$, $\sigma_2(i)^2$ are the ith diagonal components of Σ_1 and Σ_2, respectively, $\Delta_{12}(i)^2 = (\mu_1(i) - \mu_2(i))^2$, and $\mu_1(i)$, $\mu_2(i)$ are the ith components of μ_1 and μ_2, respectively.With the distance measures as summarized above, vector quantization (VQ) can be applied to classify the mixtures (and their parameters) into clusters.

Another straightforward approach is the model-based classification, i.e., those parameters for models for phone units with closer acoustic properties are classified together in a cluster. For example, the phone units can be first classified into consonants and vowels, and consonants/vowels can be further classified based on phonetic knowledge and/or data-driven approaches, etc. Another very natural classification of the parameters is feature-based, i.e., considering the type of the parameters, for example, energy parameters, MFCC parameters, first-order and second-order delta MFCC parameters, etc., and all parameters of the same type can be classified together. The effectiveness of all these different approaches of classifying the parameters will be evaluated and discussed in details below. It should be noted that the main idea for the development from the original eigenvoice to the mixture-based and model-based segmental eigenvoice approaches is very similar to the extension of MLLR from a single transformation function (for a single class) to many transformation functions (for many classes). Also, the development from the original eigenvoice to the feature-based segmental eigenvoice is very similar to the extension from the basic full-matrix MLLR to block-matrix MLLR.

IV. Experimental Results

In this section, the results for the core experiments for this research are summarized. We, therefore, first describe the experimental environment, which is a large vocabulary task, and some baseline experiments for comparison purposes in Sections IV-A and IV-B. The experimental results for the new segmental eigenvoice approaches are then given in Sections IV-C–IV-E. In Section IV-F, we will show that the segmental eigenvoice approaches can be further integrated with parameter smoothing and the Structural MAP (SMAP) techniques to produce improved performance. An extra set of experiments for a very different task with a small vocabulary will then be presented in Section V to show that the performance improvements achieved here are in fact consistent across many different tasks.

A. Experimental Environment

In the core experiments for this research, a large vocabulary Mandarin speech database for recognition purposes recorded at Taipei produced by 104 male speakers was used, in which 100 speakers used in training, and the rest 4 as testers. The recording was done in office-like environment directly through microphones, with low-level noise but negligible channel effects. Each speaker produced 200 utterances. The average length of the utterances is roughly 3 s or 11 Mandarin syllables. In the following experiments, because the focus here is rapid adaptation, we limited the adaptation data for each new speaker to be less than 40 utterances, which corresponds to less than 2 min or 440 Mandarin syllables. The adaptation performance reported below was obtained with 40 testing utterances produced by each of the four testers but not used in the adaptation processes, and the results are the average of the four testers.

The speech signal was sampled at 16 KHz, and parameterized into one energy component, one delta energy component, 14 MFCC components and 14 delta MFCC components. So the total dimension of a feature vector is 30. Cepstral Mean Subtraction (CMS) [22] was performed first on a per-utterance basis. The SI model was trained with the 100 training speakers, each with all the 200 utterances. Considering the monosyllabic structure of the Chinese language, the acoustic units used were the intra-syllabic Right-Context-Dependent (RCD) Initial/Final units [23], which include 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 (or diphthong) part of the syllable plus an optional medial and an optional nasal ending. Each Initial model has three states, each Final has four states, while a silence model with one state is also used. Each state of Initial or Final models has mixture numbers ranging from 1 to 4 depending on the quantity of available training data. The state of the silence model has eight mixtures. The recognition performance measure used in the experiments below is the free-decoded syllable error rates for continuous utterances without knowledge of lexicon or constraints of language models. So it tells directly the achievable improvements in acoustic recognition, not interfered by any other linguistic knowledge or constraints. This error rate for the SI model trained with the 100 training speakers is 36.31%.

In the experiments for eigenvoice or segmental eigenvoice to be discussed below, we need SD models for the 100 training speakers, i.e., $\mathrm{M} = 100$ as in Sections II and III. These SD models for each training speaker were adapted by the 200 utterances for the respective speaker using MAP adaptation. All the mixtures in the SD models for each speaker were used to construct a 57 780-dimensional supervector, i.e., $\mathrm{d} = 57\,780$ as in Sections II and III. After performing PCA, we had an eigenspace of dimension k, where k is a parameter to be chosen empirically.

B. Baseline Experimental Results – MLLR, MAPLR, SMAP, and the Original Eigenvoice

The block matrix MLLR (with optimal number of classes) was first taken as the reference approach for MLLR for comparison here. It is referred to as experiment (a) with performance shown in Fig. 4 as curve (a). Also compared here is a well-known improved version of MLLR approach, Maximum A Posteriori Linear Regression (MAPLR) [12]. In this approach, the transformation parameters and the model parameters are jointly estimated based on the maximum a posteriori (MAP) criterion. This is referred to here as experiment (b), with results plotted in Fig. 4 as curve (b). It can be found in Fig. 4 that this improved version of MAPLR [curve (b)] performs better than block matrix MLLR [curve (a)] in all cases, from 2 up to 40 adaptation utterances. Curve (c) in Fig. 4, on the other hand, is for another experiment (c) for the Structural MAP (SMAP) adaptation [7], in which a structural Bayes approach is used in the MAP parameter optimization. It can be clearly seen from Fig. 4 that for the SMAP adaptation [curve (c)] the syllable error rate decreases monotonically and continuously as the quantity of adaptation data increases beyond about 14 utterances, but this approach offers less stable performance for less than 14 utterances, which is the nature of SMAP adaptation. With the above comparison in Fig. 4, curve (b) for MAPLR technique is selected as the reference to be compared with for the purpose here.

The capability of the original eigenvoice approach was then obtained in an experiment (d). In this experiment, in each case (number of adaptation utterances) the number of eigenvectors, or the dimensionality k in Section II, has been optimized empirically. The result for this experiment (d) is then compared with the above reference of MAPLR in Fig. 5. The curve (b) in Fig. 5 is exactly the same curve (b) in Fig. 4 for experiment (b), for the MAPLR technique, and the curve (d) is for the experiment (d) for the original eigenvoice. It can be found that the original eigenvoice approach [curve (d)] performs significantly better than MAPLR [curve (b)] for adaptation data less than 14 utterances, but becomes similar or slightly worse for more utterances. The former is the distinct feature of the original eigenvoice approach, while the latter is the inevitable limitation of it.

C. Mixture-Based Segmental Eigenvoice

Here and below we briefly summarize the experimental results for the proposed mixture-based, model-based and feature-based segmental eigenvoice approaches in Sections IV-C–IV-E, respectively. We start in this subsection with mixture-based approach. The next experiment (e) is for mixture-based segmental eigenvoice with Bhattacharyya distance in (5) used as the distance measure in VQ clustering. The experiment was performed with number of clusters N varying from 2 to 15, and it was found that $\mathrm{N} = 3$ offered the best results. These best results for $\mathrm{N} = 3$ are listed in the second row (e) in Table I. Also listed in the first row (d) in Table I are the results for the original eigenvoice for experiment (d) mentioned above for comparison. We can tell from the first two rows (d) (e) of Table I that if the available adaptation data are too limited (two utterances, for example), the performance of the mixture-based segmental eigenvoice may become worse than the original eigenvoice, apparently because of the difficulties in accurate estimation of the necessary coefficients. However, an error rate reduction of up to 7% becomes achievable when the adaptation data were more than ten utterances, which is an obvious improvement. Note that $\mathrm{N} = 3$ was found to be a good choice only for the cases tested here. Larger number of N may possibly offer better performance than three

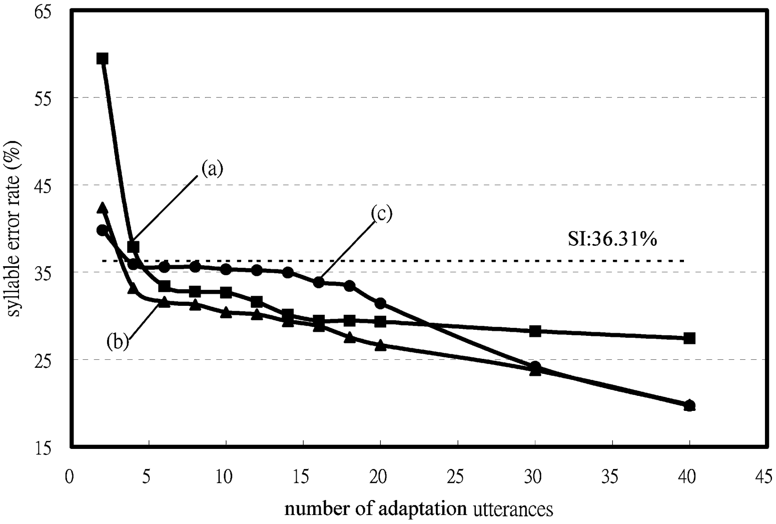


Fig. 4. Comparison of experiments (a) block-matrix MLLR, (b) MAPLR, and (c) S MAP adaptation.

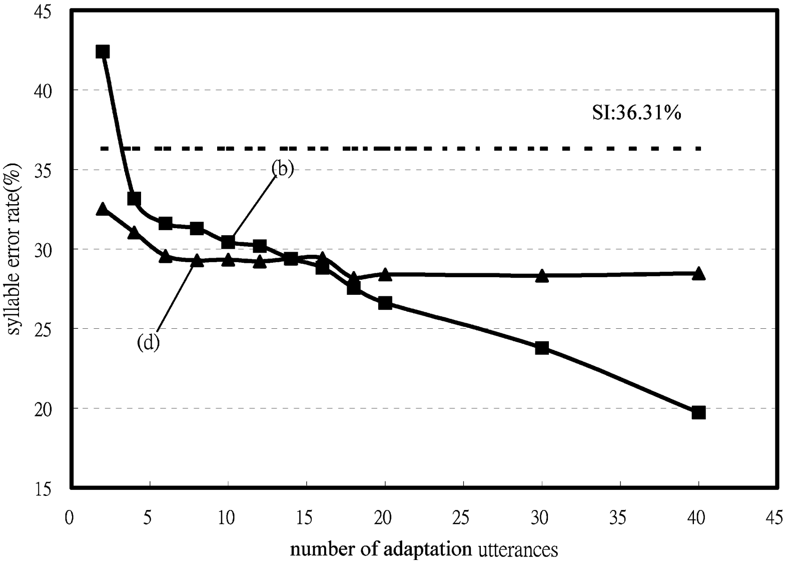


Fig. 5. Comparison of experiments (b) MAPLR technique with (d) the original eigenvoice approach.

clusters when more adaptation data become available, but that part is out of the scope of this paper.

It should be pointed out that in either the original or any kind of the segmental eigenvoice approaches to be discussed here, the proper choice of the dimensionality k of the original eigenspace, or k_c of each subeigenspace, is critical for the achievable performance. In all the cases in all the experiments here and discussed below in this paper, the dimensionalities of the subeigenspace have actually been empirically optimized. Therefore, the improvements obtained by the segmental eigenvoice approaches here are definitely not achievable by simply increasing the dimensionality of the eigenspace of the original eigenvoice approach, although the parameters describing the acoustic characteristics of the new speaker can be increased in that way too.

Another experiment (f) for mixture-based segmental eigenvoice was then performed with the Divergence parameter in (6) used as the distance measure in VQ. The results showed very similar trends to those in experiment (e), and the best performance again occurs when $\mathrm{N} = 3$ and the results are listed in the next row (f) of Table I. Comparing the results in rows (e) and (f) of Table I, it can be clearly found that lower error rates can be obtained as compared to the original eigenvoice with either the Bhattacharyya distance or the Divergence parameter as the distance measure (except for the case with too limited adaptation data), and the Bhattacharyya distance produces better performance than the Divergence parameter in most cases. Therefore, Bhattacharyya distance with 3 clusters will be taken as the representative for mixture-based segmental eigenvoice in the discussions below.

D. Model-Based Segmental Eigenvoice

The next experiment (g) is for model-based segmental eigenvoice. The Initial and Final models used here as discussed previously in Section IV-A were classified manually into $\mathrm{N} =$ 2, 3, and 6 clusters based on phonetic knowledge only. Note that there can actually be many different ways to perform the model classification, and, therefore, these three cases $\mathrm{N} = 2$,

TABLE I
SYLLABLE ERROR RATES (%) FOR THE LARGE VOCABULARY TASK FOR DIFFERENT VERSIONS OF SEGMENTAL EIGENVOICE TECHNIQUES PROPOSED IN THIS PAPER, COMPARED WITH THE ORIGINAL EIGENVOICE (N: NUMBER OF CLUSTERS)

Number of Utterances			**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**30**	**40**
(d) Original Eigenvoice			32.53	31.05	29.57	29.29	29.34	29.37	29.40	29.42	28.18	28.40	28.33	28.48
Mixture-based Segmental Eigenvoice	(e) Bhattacharyya distance	N=3	33.15	30.16	29.47	29.26	27.75	27.54	27.33	27.41	27.25	27.19	26.74	26.71
	(f) Divergence parameter	N=3	34.14	31.63	28.87	28.45	27.89	28.37	28.51	28.37	28.18	28.13	27.69	27.54
(g) Model-based Segmental Eigenvoice		N=3	34.02	30.83	30.28	29.89	28.50	28.47	28.33	27.99	28.02	27.80	27.91	27.56
(h) Hybrid Mixture/model-based Segmental Eigenvoice		N=3	34.25	30.92	28.18	28.77	27.76	27.65	27.70	27.44	27.28	27.22	27.11	27.03
(i) Feature-based Segmental Eigenvoice		N=3	32.16	29.75	28.65	28.78	29.22	28.75	28.56	28.62	27.94	28.18	27.80	27.89
(j) Integrated Mixture/feature-based Segmental Eigenvoice		N=9 or N=27	34.88 (N=9)	29.26 (N=9)	27.44 (N=9)	26.92 (N=9)	26.58 (N=9)	26.72 (N=9)	26.05 (N=9)	25.94 (N=9)	25.94 (N=9)	26.26 (N=9)	26.09 (N=27)	25.57 (N=27)

3 and 6 mentioned here are simply three examples. The results for this experiment (g) showed that again $\mathrm{N} = 3$ offered the best results for most cases, which are listed as the next row (g) in Table I. Similar to the mixture-based segmental eigenvoice, improved performance as compared to the original eigenvoice is achievable except for the cases with very limited adaptation data. But comparing the results in row (g) with those in row (e) of Table I, it can be found that in all cases the model-based segmental eigenvoice performs slightly worse than the mixture-based segmental eigenvoice approach. One possible reason for this may be that each model includes many mixtures; therefore, classification based on mixtures can be done more precisely than that based on models, and as a result more delicate subeigenspaces can be obtained with mixture-based approach. However, it is still possible that the model-based classification approach, specially the phonetic knowledge used here, may provide some better directions for mixture-based classification, because the best approach for classifying the mixtures is actually unknown, while VQ with Bhattacharyya distance as used above is simply a relatively better way. A good example for this concept is then to try to do some coarse classification first with model-based approaches by the phonetic knowledge, then some fine classification with data-driven mixture-based approaches. One such approach, referred to as hybrid mixture/model-based segmental eigenvoice here, is that all the models can be first classified into 2 clusters, one with all Initial models (i.e., all consonants) and the other with all Final models (i.e., primarily vowels). Then all the mixtures in the clusters of Initial models and Final models are further classified into $\mathrm{N_I}$ and $\mathrm{N_F}$ clusters, respectively, with Bhattacharyya distance to give a total of $\mathrm{N} = \mathrm{N_I} + \mathrm{N_F}$ clusters. The experiment for this approach with $\mathrm{N_I} = 2$, $\mathrm{N_F} = 1$ and, therefore, $\mathrm{N} = 3$ is referred to as experiment (h) here, with the results listed in the next row (h) of Table I. The results indicated that there existed cases in which this hybrid mixture/model-based approach performed slightly better than the mixture-based approach alone, and 5% to 6% error rate reduction can be achieved as compared to the original eigenvoice for this mixture/model-based segmental eigenvoice.

It should be noted that the hybrid mixture/model-based approach is also a special case of the mixture-based approach. It is, therefore, reasonable that these two approaches perform very closely in many cases. Also, it is worth mentioning that the method used in classifying the parameters and the number of clusters are apparently two important keys. Furthermore, for the several initial examples shown here, mixture-based, model-based, or hybrid mixture/model-based segmental eigenvoice approaches as in rows (e), (f), (g), and (h) in Table I, the achievable improvements seem to be not very significant, although the trends of improvements are obvious and consistent across different approaches. Moreover, a common limitation for all these segmental eigenvoice techniques mentioned here is that they may produce worse performance than the original eigenvoice when there are too limited adaptation data available. This will not be the case when we examine the feature-based segmental eigenvoice in the next Section IV-E.

E. Feature-Based Segmental Eigenvoice

In the feature-based segmental eigenvoice approach, the parameters in the supervectors are naturally classified based on the type of the parameters. The energy components can be classified into one cluster, the MFCC components into another and the delta MFCC components into the third, which gives $\mathrm{N} = 3$. All the rest processes are the same as before. This is referred to as experiment (i) here. The results are listed in the next row (i) of Table I. It can be seen from the experimental results that the feature-based segmental eigenvoice offered better performance than the original eigenvoice in all cases for adaptation data ranging from two up to 40 utterances, regardless of the quantity of the adaptation data. An important observation is that some improvements are obtainable even with only two utterances, which is not true for any other segmental eigenvoice approaches discussed previously. Note that in the feature-based approach all the feature vectors for all the adaptation data are used to estimate the sets of coefficients for each subeigenspace, while in mixture- or model-based approaches the feature vectors for the adaptation data are first divided into N clusters, and then only one cluster of the feature vectors is used to estimate the set of coefficients for one subeigenspace. This may be the reason why the performance of the feature-based segmental eigenvoice is not very sensitive to the quantity of the adaptation data.

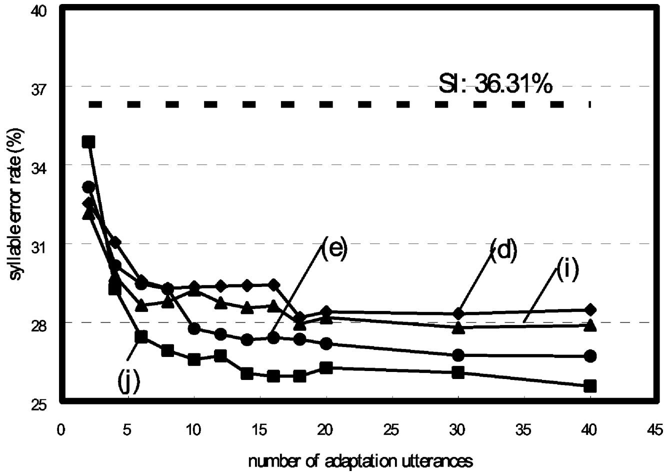


Fig. 6. Experiment (j) for the integrated mixture/feature-based approach ($N = 9$ or 27), to be compared with the individual component approaches: experiments (e) and (i) for mixture-based ($N = 3$) and feature-based ($N = 3$) approaches alone, respectively. Also shown is the original eigenvoice in experiment (d).

Because the basic concepts and processing procedures are quite similar for mixture-based and model-based segmental eigenvoice, yet very different for feature-based segmental eigenvoice, it is, therefore, reasonable to believe that the integration of these different categories of approaches may give even better performance. For example, all the mixtures or models can be first classified into N_1 clusters by mixture- or model-based approaches, and the mean vectors for the mixtures in each cluster can then be further segmented by feature-based approaches into N_2 clusters to give a total of $N = N_1 \times N_2$ clusters and so on. The initial test for this concept is for the integration of the mixture-based approach with feature-based approach, referred to as the integrated mixture/feature-based approach here. In the experiment, it was found that for this integrated approach, with two to 20 utterances of adaptation data the best results were obtained with $N_1 = 3$, $N_2 = 3$ and $N = 9$, while with 30 to 40 utterances the best results were obtained with $N_1 = 9$, $N_2 = 3$ and $N = 27$. This is reasonable, i.e., larger N for more adaptation data. This is referred to as experiment (j) here. The complete results of experiment (j) together with the number N of clusters used in each case are listed in the last row (j) of Table I. Quite significant improvements can be observed for this experiment (j) as compared to the original eigenvoice in row (d). For example, the error rate reduction is 7.20% (29.57% to 27.44%), 9.41% (29.34% to 26.58%), 11.83% (29.42% to 25.94%), and 10.22% (28.48% to 25.57%), respectively, for 6, 10, 16, and 40 utterances of adaptation data. This situation is, therefore, further illustrated in Fig. 6, in which the curve (j) for the experiment (j) for row (j) in Table I with $N = 9$ or 27 is compared with each of its individual component approaches, the curve (i) for experiment (i) in row (i) in Table I for the feature-based segmental eigenvoice alone with $N = 3$ and the curve (e) for experiment (e) in row (e) in Table I for the mixture-based segmental eigenvoice alone with $N = 3$, along with the curve (d) in row (d) in Table I for the original eigenvoice. Note here again that in Fig. 6 for 30 and 40 utterances of adaptation data, the results for curve (j) were obtained with $N_1 = 9$, $N_2 = 3$ and $N = 27$, while those for both curves (i) and (e) were with $N = 3$. These were the best results for the respective cases. In other words, when different approaches are integrated to segment the eigenspace, when more adaptation data are available, a larger value of $N = N_1 \times N_2$ may give better results, although the component number, N_1 and N_2, may not be the best number when only a single individual component approach was used. Fig. 6 clearly verified the previous statement, i.e., the integration of two different categories of approaches gives better results than any of its component approaches. Therefore, the improvements obtained in the feature-based and mixture-based segmental eigenvoice approaches are apparently additive. Also note that here in the experiment (j), $N = 9$ or 27. This is significantly larger than $N = 3$, with which the best results were obtained with previous approaches. So we have 9 and even 27 subeigenspaces here and better results were obtained. This verifies that as long as good approaches can be found, more delicate subeigenspaces may be constructed to provide better performance.

F. Further Approaches for Improved Segmental Eigenvoice

Here we presented two additional approaches, which can be used to further improve the performance of the segmental eigenvoice proposed here. The first is to use the parameter smoothing concept, while the second is to integrate with the Structural MAP (SMAP) approach. They are, respectively, discussed in Sections IV-F1 and IV-F2 below.

1) Improved Segmental Eigenvoice With Parameter Smoothing: In Section IV-C it has been pointed out that for all the cases in all the experiments discussed in this paper, the dimensionalities of the eigenspace or subeigenspaces have actually been empirically optimized. In fact, there is still an extra issue not mentioned. When performing the segmental eigenvoice experiments as described above, in each case (number of adaptation utterances) the dimensionality of all different subeigenspaces were always set to be equal for simplicity, although such an equal-dimensionality assumption for all different subeigenspaces is apparently not necessarily optimal. This in fact creates some problem worth further investigating as discussed here in this subsection. When classifying the adaptation data into clusters, it very often happened that more adaptation data were distributed to some clusters, while some other clusters had only very small quantity of adaptation data. If all the subeigenspaces are given the same dimensionality, some sets of coefficients may be poorly estimated and the SD model parameters thus obtained may become imprecise.

Here we propose an additional approach to handle the above problem due to insufficient adaptation data in some subeigenspaces. This approach adopts the concepts of parameter smoothing, taking into consideration the reliability of the coefficient estimation based on the sufficiency of the adaptation data. This parameter smoothing is expressed as in (7) below, and is depicted in Fig. 7

$$\hat{\mu} = \frac{\Gamma}{\Gamma + \sum_{t=1}^{T} \gamma_c(t)} \mu^{SI} + \frac{\sum_{t=1}^{T} \gamma_c(t)}{\Gamma + \sum_{t=1}^{T} \gamma_c(t)} \mu^{\text{eigen}} \qquad (7)$$

where μ^{SI}, μ^{eigen} represent the mean vectors for a certain mixture in a certain state for, respectively, the SI model and the SD model obtained from a specific segmental eigenvoice approach, $\hat{\mu}$ is the newly obtained mean vector for the desired mixture and state with the parameter smoothing approach, Γ represents the weight of the *a priori* knowledge, which is usually an empirically determined parameter, $\gamma_c(t)$ is the occupation probability for the desired mixture and state included in the cluster c for the segmental eigenvoice here, and the summation is over all adaptation data frames $\boldsymbol{t}$. The adapted mean vectors obtained in this way will remain closer to the mean vectors for SI models if the accumulated occupation probability for the adaptation data and the desired cluster c is small compared to Γ. $\hat{\mu}_2$ in Fig. 7 is an example for such cases. On the other hand, if the accumulated occupation probability is large compared to Γ, the adapted mean vector for this cluster c will be closer to that obtained by the segmental eigenvoice directly. $\hat{\mu}_1$ and $\hat{\mu}_3$ in Fig. 7 are examples for such cases. This is referred to as the improved segmental eigenvoice approach with parameter smoothing here in this paper.

All the segmental eigenvoice approaches discussed above can be further improved with this approach. Here we take the approach with the best results obtained previously, the integrated mixture/feature- based segmental eigenvoice in experiment (j) or row (j) in Table I, curve (j) in Fig. 6 above as an example. The results for the improved version with parameter smoothing for this case, referred to as experiment (k) here, is plotted as curve (k) in Fig. 8, as compared to the results for the previous experiment (j) and the original eigenvoice approach for experiment (d) (curves (j) and (d) in Fig. 8). From Fig. 8, the distinct feature of the improved segmental eigenvoice with parameter smoothing can be clearly seen, i.e., when the adaptation data were very limited (2 utterances only) the performance of the previous segmental eigenvoice (the integrated mixture/feature-based approach in the case here) can be made better than the original eigenvoice. When the adaptation data were increased, slightly better performance as compared to the previous segmental eigenvoice approach is also obtainable in some cases. This is of course due to properly taking into consideration the reliability of the parameter estimation in the segmental eigenvoice approaches. Therefore, one natural limitation of the segmental eigenvoice approaches, i.e., poor coefficient estimation when the adaptation data are very limited, can be overcome properly. In fact, some preliminary experiments performed at this stage also indicated that the value of Γ in (7) should be dynamically adjusted according to the quantity of the adaptation data. Of course, when more and more adaptation data become available, the performance of the improved segmental eigenvoice with parameter smoothing will eventually converge to the previous segmental eigenvoice approach, as suggested in (7) and Fig. 8.

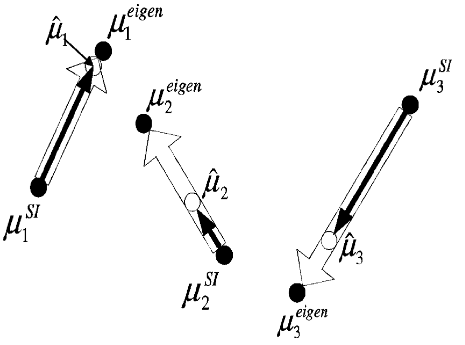


Fig. 7. Improved segmental eigenvoice approach with parameter smoothing: white arrows pointing from μ_i^{SI} to μ_i^{eigen} represent the adaptation process for the segmental eigenvoice, and black arrows pointing from μ_i^{SI} to $\hat{\mu}_i$ represent the adaptation process for the improved segmental eigenvoice with parameter smoothing.

2) Integration of the Segmental Eigenvoice With SMAP: The above Section IV-F.I indicated that the parameter smoothing approach can properly handle one end of the natural limitation of the segmental eigenvoice approaches, i.e., the imprecise parameter estimation for very limited adaptation data. But the achievable improvements for that approach diminish gradually when more adaptation data become available. Therefore, the other end of the natural limitation of the segmental eigenvoice approaches, i.e., the performance saturation when more adaptation data become available, remains to be considered. Of course this can be handled to a good extent by increasing the number of clusters or subeigenspaces, N, as discussed previously. But this may not be the only solution. Here we propose an extra approach to handle this other end of the natural limitation. We consider that the MAP adaptation approach in principle is able to offer the best performance when more adaptation data are available. Because the Structural MAP (SMAP) is an improved version of MAP in which the situation of relatively less adaptation data is also considered, SMAP should be a better approach to be used here. Therefore, the second approach for improved segmental eigenvoice presented here in this subsection is to integrate the segmental eigenvoice with SMAP, or performing the segmental eigenvoice approach first, followed by SMAP adaptation. The purpose is to handle the other end of the natural limitation of the segmental eigenvoice, i.e., the performance saturation when more adaptation data become available. This integration with SMAP can be applied to any segmental eigenvoice approach presented above (of course including the original eigenvoice as well).

In the experiment (l) to be presented here, the best approach obtained above, the integrated mixture/feature-based approach with parameter smoothing in experiment (k), was further followed by SMAP. The results are shown as the curve (l) in Fig. 9, compared with the curve (k) for the experiment (k) without SMAP, exactly the same curve (k) in Fig. 8. Also depicted is the curve (m) for another experiment (m), the original eigenvoice also followed by SMAP, as well as the curve (d) for the original eigenvoice. From Fig. 9, it is obvious that the performance of both the integrated mixture/feature-based segmental eigenvoice with parameter smoothing [curve (k)] and the original eigenvoice [curve (d)] can be significantly improved when more adaptation data (more than 14 utterances) become available [curves (l) versus (k) and curves (m) versus (d)], if followed

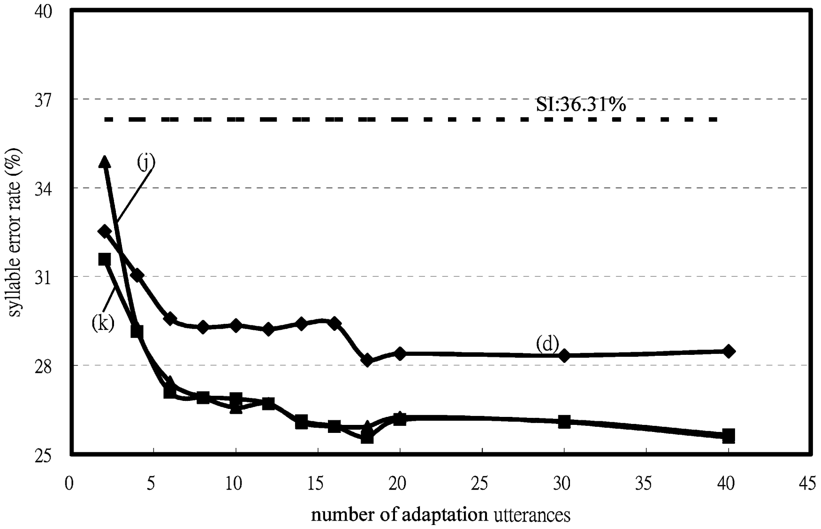


Fig. 8. Performance of the improved integrated mixture/feature-based segmental eigenvoice with parameter smoothing, experiment (k), compared with experiments (j) (without parameter smoothing) and (d) (original eigenvoice).

by the SMAP technique. Of course the improvements become insignificant when the quantity of the adaptation data become relatively small. Therefore, the parameter smoothing approach presented in the above Section IV-F1 is helpful in handling the natural limitation of the segmental eigenvoice at one end (i.e., with very limited adaptation data), while the integration with SMAP presented in this subsection is helpful in handling the natural limitation of the segmental eigenvoice at the other end (i.e., saturation with more adaptation data). Thus both ends of the natural limitation have been well taken care of. Now comparing curves (l) and (m), the achieved improvements for the segmental eigenvoice approaches presented here in this paper is quite clear. Even if the original eigenvoice followed by SMAP can also offer very good performance including the situation when more adaptation data are available [curve (m) in Fig. 9], the segmental eigenvoice followed by SMAP as proposed here is able to achieve significantly better results over the whole range of adaptation data from 2 to 40 utterances to be considered here [curves (l) versus (m) in Fig. 9]. Note that the error rate reduction achieved here [curve (l) versus (m)] is 9.90% (30.10% to 27.12%), 8.94% (29.30% to 26.68%) and 10.80% (28.99% to 25.86%), respectively, for 6, 10, and 14 adaptation utterances, which can be considered significant.

V. Extra Set of Experiments for a Small Vocabulary Task

In all the above experiments presented in Section IV, obvious improvements were achieved with the various segmental eigenvoice approaches proposed in this paper for a large vocabulary task. One may wonder whether the segmental eigenvoice approaches also work for small vocabulary tasks. Moreover, all the results presented above in Section IV are the average for four testing speakers, which is relatively small. This is why in this section an extra set of experiments were performed on a small vocabulary task, with a much larger number of testing speakers. The results show that the proposed segmental eigenvoice approaches indeed offer improved performance for a quite variety of different environments.

In this task to be reported here in this section, the testing database used in this extra set of experiments was the NUM-100A digit database provided by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) at Taipei. This database includes 8000 Mandarin digit strings produced by 50 males and 50 females. The speech signal was recorded under normal laboratory environment through microphone at 8 KHz sampling rate. This 8000 Mandarin digit strings consist of 1000 2, 3, 4, 5, 6, 7-digit strings each, and 2000 single-digit utterances. In the experiment presented below, 40 male speakers were selected as testers, and each test speaker has, respectively, twenty single-digit utterances to be used as the adaptation data, and five 7-digit strings for testing. Because the ten Mandarin digits are simply ten distinct syllables, the digit models needed for this task can directly be the ten corresponding syllable models used in the above large vocabulary experiments constructed from the RCD Initial models and the CI Final models as presented in Section IV-A. So all the ten SI digit models for testing and the 100 sets of SD digit models for the original eigenvoice or segmental eigenvoice approaches can simply be picked up from those mentioned in Section IV-A and used in the whole Section IV, and, therefore, all the training environment and experimental setup are exactly the same as in the above. The only difference is that the testing data here include only the ten digits and, therefore, the recognition process is to select one out of the ten digits. In addition, the testing data need to be interpolated from 8 KHz sampling rate to 16 KHz in order to be compatible to the training conditions. In all the experiments presented below, the number of adaptation data for each testing speaker was increased from 2 to 20 single-digit utterances. The recognition results presented below are the free-decoded digit error rates for the 7-digit strings, averaged over all the 40 testing speakers. The digit error rate for the ten SI digit models trained with the 100 training speakers (as mentioned in Section IV-A) is 8.69%.

The results for the experiments are listed in Table II. SMAP was taken as the baseline, together with the original eigenvoice as well as the segmental eigenvoice approaches both followed

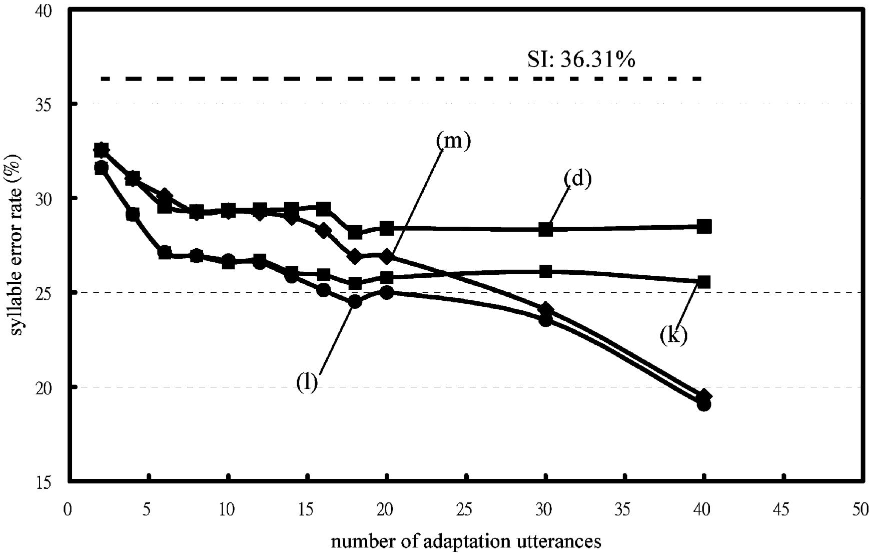


Fig. 9. Performance of the integrated mixture/feature-based segmental eigenvoice with parameter smoothing followed by SMAP [experiment (l)] and the original eigenvoice followed by SMAP [experiments (m)], also shown are experiments (k) and (d) without SMAP.

by SMAP, respectively, referred to as experiments (n) (o) (p) here. The results are listed in Table II as the first three rows (n) (o) (p). In the case of original eigenvoice or segmental eigenvoice, the dimensionality of the eigenspaces or subeigenspaces has been empirically optimized. Note that here not only the vocabulary was very small, only ten digits, but the adaptation data were very limited either. For example, in the 2- or 4- utterances cases we have only two or four syllables as the adaptation data. It can be found from row (n) that SMAP can provide some error rate reduction, though to a relatively limited extent, even with 2 single digits (two syllables only) of adaptation data, as compared to the SI result of 8.69%. The error rate was then monotonically decreased as the adaptation data were increased all the way up to 20 single digits, although with some minor exception cases. The original eigenvoice followed by SMAP in row (o) of Table II, on the other hand, was able to perform similarly. In this case some error rate reduction was achieved starting with six single digits of adaptation data, and the error rate was then reduced monotonically with increased adaptation data. But the error rates in row (o) were worse if only two or four single digits of adaptation data were available, and, by comparing rows (o) to (n), it can be found that in this task the original eigenvoice followed by SMAP was not able to do better than SMAP alone in almost all cases of adaptation data, probably because the adaptation data were too limited to estimate precise parameters. The segmental eigenvoice followed by SMAP, on the other hand, listed in row (p), performed differently. Note that here the number of clusters for the segmental eigenvoice approaches has been optimized in each case. For 2-6 adaptation digits feature-based alone with three clusters ($N_1 = 1$, $N_2 = 3$) gave the best results, while for 8-20 adaptation digits mixture-based alone with 2 clusters ($N_1 = 2$, $N_2 = 1$) gave the best results. These conditions are also shown in row (p) of Table II. The nice feature of row (p) is that the segmental eigenvoice followed by SMAP was able to offer improvements with respect to the baseline of SMAP alone in all cases right starting with two single digits of adaptation data, all the way to 20 single digits. In particular, the error rate reduction was 12.05% (8.55% to 7.52%), 9.42% (8.17% to 7.40%), and 14.69% (7.83% to 6.68%) for 4, 6, and 8 single digits (single syllables) of adaptation data, for example, which are reasonably significant. This verified the capabilities of the segmental eigenvoice approaches proposed here. The improvements become slightly less for ten single digits of adaptation data or more, which is natural. Because in this task only the ten digits are to be distinguished, the performance becomes better when more adaptation data become available for all adaptation techniques.

In the next set of experiments (q) and (r), parameter smoothing as presented in Section IV-F1 was applied in addition, i.e., we have the original eigenvoice with parameter smoothing and followed by SMAP [experiment (q)], and the segmental eigenvoice with parameter smoothing and followed by SMAP [experiment (r)], with results listed in the last two rows (q) and (r) in Table II. Again here the number of clusters for the segmental eigenvoice approaches has been optimized in each case. For 2-6 adaptation digits feature-based alone with 3 clusters ($N_1 = 1$, $N_2 = 3$) gave the best results, while for 8-20 adaptation digits mixture-based alone with 2 clusters ($N_1 = 2$, $N_2 = 1$) gave the best results. These conditions are also shown in row (r) of Table II. By comparing rows (q) with (o), we see that for the original eigenvoice followed by SMAP, the additional parameter smoothing did offer some improvements for 2-14 single digits of adaptation data, although the improvements are kind of limited, and in all the cases (2-20 single digits of adaptation data) the results in row (q) are actually not too much different from the baseline of SMAP alone in row (n). For the segmental eigenvoice f1ollowed by SMAP, on the other hand, the additional parameter smoothing [in row (r)] was able to offer much more improvements, for example, comparing rows (r) with (p) in Table II, the error rate reduction was 13.26% (8.37% to 7.26%), 15.82% (7.52% to 6.33%), 16.35% (7.40% to 6.19%) and 13.47% (6.68% to 5.78%) for 2, 4, 6, and 8 single digits of adaptation data. In fact, row (r) includes the lowest error rate for all the cases (2 up to 20 single digits

TABLE II
DIGIT ERROR RATES (%) FOR THE SMALL VOCABULARY TASK FOR SMAP AS THE BASELINE, COMPARED WITH THE ORIGINAL EIGENVOICE AND SEGMENTAL EIGENVOICE TECHNIQUES BOTH FOLLOWED BY SMAP, WITHOUT OR WITH PARAMETER SMOOTHING. THE DIGIT ERROR RATE FOR THE SI MODEL IS 8.69%. THE BEST PERFORMANCE FOR EACH CASE (NUMBER OF ADAPTATION UTTERANCES) IS SHOWN WITH BOLD DIGITS

Number of Utterances (Single Digits)	2	4	6	8	10	12	14	16	18	20
(n) SMAP	8.49	8.55	8.17	7.83	7.01	6.25	6.14	5.71	5.55	5.47
(o) Original Eigenvoice, Followed by SMAP	9.06	8.85	8.53	7.92	7.14	6.31	6.01	5.73	5.62	5.59
(p) Segmental Eigenvoice, Followed by SMAP (N_1 for mixture-based,N_2 for feature-based, N= $N_1 \times N_2$)	8.37 (N_1=1 N_2=3)	7.52 (N_1=1 N_2=3)	7.40 (N_1=1 N_2=3)	6.68 (N_1=2 N_2=1)	6.47 (N_1=2 N_2=1)	5.95 (N_1=2 N_2=1)	5.68 (N_1=2 N_2=1)	5.53 (N_1=2 N_2=1)	5.38 (N_1=2 N_2=1)	5.31 (N_1=2 N_2=1)
(q) Original Eigenvoice with Parameter Smoothing, Followed by SMAP	8.57	8.41	8.06	7.51	6.92	6.21	5.99	5.72	5.62	5.59
(r) Segmental Eigenvoice with Parameter Smoothing, Followed by SMAP (N_1 for mixture-based,N_2 for feature-based, N= $N_1 \times N_2$)	**7.26** (N_1=1 N_2=3)	**6.33** (N_1=1 N_2=3)	**6.19** (N_1=1 N_2=3)	**5.78** (N_1=2 N_2=1)	**5.62** (N_1=2 N_2=1)	**5.36** (N_1=2 N_2=1)	**5.26** (N_1=2 N_2=1)	**5.24** (N_1=2 N_2=1)	**5.15** (N_1=2 N_2=1)	**5.12** (N_1=2 N_2=1)

of adaptation data) as compared to all other rows in Table II. The "real improvements" achieved by the segmental eigenvoice proposed in this paper, however, should probably be observed by comparing rows (r) with (q), i.e., the segmental eigenvoice as compared to the original eigenvoice, both with parameter smoothing and followed by SMAP. The error rate reduction for this case is then 15.29% (8.57% to 7.26%), 24.73% (8.41% to 6.33%), 23.20% (8.06% to 6.19%) and 23.04% (7.51% to 5.78%) for 2, 4, 6, and 8 single digits of adaptation data and so on, which should be considered significant. All these results verified the superiority of the segmental eigenvoice approaches proposed here for speaker adaptation even for small vocabulary tasks.

VI. CONCLUSION

This paper proposes the concept of segmenting the eigenspace into smaller subeigenspaces, so as to construct more delicate eigenspace and more precise SD models for new speakers for rapid speaker adaptation. The mixture-based segmental eigenvoice is primarily based on the statistical characteristics of all the model mixtures, while the model-based segmental eigenvoice may utilize the phonetic knowledge in segmenting the eigenspace. Both of these two versions of segmental eigenvoice approaches can produce better performance than the original eigenvoice, except when the adaptation data are too limited. But the actually achievable performance depends heavily on how the eigenspace is segmented, or how the mixtures or models are clustered. The feature-based segmental eigenvoice, on the other hand, has a different point of view. It is based on the hypothesis of limited correlation among different types of feature parameters. The performance improvements achievable by feature-based segmental eigenvoice are found to be less sensitive to the quantity of the adaptation data. Experimental results indicate that not only each of these approaches can offer better performance than the original eigenvoice, but the proper integration of them may provide even better results than the individual approaches. Further approaches for improved segmental eigenvoice were also developed. One uses parameter smoothing to consider the reliability of the coefficient estimation to handle the case of very limited adaptation data. The other integrates the segmental eigenvoice approaches with SMAP to handle the case when more adaptation data are available. Both approaches were shown to offer extra improvements for the segmental eigenvoice proposed in this paper.

REFERENCES

[1] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 11, pp. 695–707, Nov. 2000.

[2] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proc. Int. Conf. Speech Language Processing*, vol. 5, Sydney, Australia, 1998, pp. 1771–1774.

[3] R. Kuhn, P. Nguyen, J.-C. Junqua, R. Boman, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Fast speaker adaptation using *a priori* knowledge," in *Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Phoenix, AZ, Mar. 1999, pp. 749–752.

[4] 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. 4, pp. 291–298, Apr. 1994.

[5] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.

[6] M. Gales and P. Woodland, "Mean and variance adaptation within the MLLR framework," *Comput. Speech Lang.*, vol. 10, pp. 250–264, Oct. 1996.

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

[8] L. He, J. Wu, D. Fang, and W. Wu, "Speaker adaptation based on combination of MAP estimation and weighted neighbor regression," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Istanbul, Turkey, Jun. 2000, pp. 981–984.

[9] J. Ishii and M. Tonomura, "Speaker normalization and adaptation based on linear transformation," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing* , vol. 2, Munich, Germany, Apr. 1997, pp. 1055–1058.

[10] J.-T. Chien, C.-H. Lee, and H.-C. Wang, "A hybrid algorithm for speaker adaptation using MAP transformation and adaptation," *IEEE Signal Process. Lett.*, vol. 4, no. 6, pp. 167–169, Jun. 1997.

[11] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 7, pp. 294–300, Jul. 1996.

[12] O. Siohan, C. Chesta, and C.-H. Lee, "Joint maximum *a posteriori* adaptation of transformation and HMM parameters," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 417–428, May 2001.

[13] A. Fischer and V. Stahl, "Database and online adaptation for improved speech recognition in car environments," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Phoenix, AZ, Mar. 1999, pp. 445–448.

[14] K.-T. Chen, W.-W. Liau, H.-M. Wang, and L.-S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Proc. Int. Conf. Speech Language Processing*, vol. 3, Beijing, China, 2000, pp. 742–745.

[15] D. K. Kim and N. S. Kim, "Rapid speaker adaptation using probabilistic principal component analysis," *IEEE Signal Process. Lett.*, vol. 8, no. 6, pp. 180–183, Jun. 2001.

[16] I. T. Jolliffe, *Principal Component Analysis*. Berlin, Germany: Springer-Verlag, 1986.

[17] Y. Onishi and K.-I Iso, "Speaker adaptation by hierarchical eigenvoice," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Hong Kong, China, Apr. 2003, pp. 576–579.

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

[19] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.

[20] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

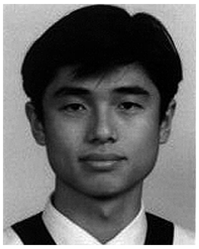
[21] T. Watanabe, K. Shinoda, K. Takagi, and E. Yamada, "Speech recognition using tree-structure probability density function," in *Proc. Int. Conf. Speech Language Processing*, Yokohama, Japan, 1994, pp. 223–226.

[22] M. G. Rahim, B.-H. Juang, W. Chou, and E. Buhrke, "Signal conditioning techniques for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 3, no. 4, pp. 107–109, Apr. 1996.

[23] B. Chen, H.-M. Wang, and L.-S. Lee, "Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 3, Istanbul, Turkey, Jun. 2000, pp. 1771–1774.

Yu Tsao received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1999 and 2001, respectively. He is currently pursuing the Ph.D. degree at the Center for Signal and Image Processing (CSIP), Georgia Institute of Technology, Atlanta.

His current research is primarily focused on the detection-based speech recognition.



Shang-Ming Lee received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1999 and 2001, respectively.

His research interests are primarily on multimedia processing and speech recognition.

Lin-Shan Lee received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

He has been a Professor of electrical engineering and computer science at the National Taiwan University, Taipei, R.O.C., since 1982 and holds a joint appointment as a Research Fellow of Academia Sinica, Taipei. His research interests include digital communications and spoken language processing. He developed several of the earliest versions of Chinese spoken language processing systems in the world including text-to-speech system, natural language analyzer, dictation systems, and voice information retrieval system.

Dr. Lee was Guest Editor of a Special Issue on Intelligent Signal Processing in Communications of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATION in December 1994 and January 1995. He was Vice President for International Affairs (1996–1997) and the Awards Committee chair (1998–1999) of the IEEE Communications Society. He has been a member of Permanent Council of International Conference on Spoken Language Processing (ICSLP), was the convenor of COCOSDA (International Coordinating Committee of Speech Databases and Assessment, 2000–2001), and is currently a member of the Board of International Speech Communication Association (ISCA).