行政院國家科學委員會專題研究計畫　成果報告

子計畫二:無線通訊環境下國語中文之聲學及語言處理基礎技術之研究(3/3)

計畫類別： 整合型計畫
計畫編號： NSC91-2219-E-002-036-
執行期間： 91 年 08 月 01 日至 92 年 07 月 31 日
執行單位： 國立臺灣大學資訊工程學系暨研究所

計畫主持人： 李琳山

報告類型： 完整報告

處理方式： 本計畫可公開查詢

中　華　民　國 92 年 8 月 28 日

行政院國家科學委員會專題研究計畫成果報告

無線通訊環境下以語音擷取網路資訊相關技術之研究 (3/3)

J 子計畫二：無線通訊環境下國語中文之聲學及語言處理基礎技術之研究

計畫編號：NSC 91-2219-E-002-036

執行期限：91 年 8 月 1 日至 92 年 7 月 31 日

主持人：李琳山　　國立台灣大學資訊工程學系

E-mail: lslee@gate.sinica.edu.tw

ABSTRACT

It was previously proposed to use the Principal Component Analysis (PCA) to derive the data-driven temporal filters for obtaining robust features in speech recognition, in which the first principal components are taken as the filter coefficients [1,2]. In this report, a multi-eigenvector approach is proposed instead, in which the first M eigenvectors obtained in PCA are weighted by their corresponding eigenvalues and summed to be used as the filter coefficients. Experimental results showed that the multi-eigenvector filters offer significant recognition performance as compared to the previously proposed PCA-derived filters under all different conditions tested with the AURORA2 database, especially when the training and testing environments are highly mismatched.

摘要

前人提出以主成分分析產生與語料相關的時域濾波器，以獲得較具強健性的語音辨識特徵參數；這種方法即是把第一個特性向量作為濾波器係數。.在本報告中，我們提出了多特性向量法，也就是將由主成份分析所得到的前 M 個特性向量乘上對應的特性值後相加來作為濾波器係數。在 AURORA 2 語料的各種條件下，實驗結果皆顯示多特性向量濾波器比起前人提出的主成份分析濾波器有著顯著的進步，尤其是在訓練及測試環境嚴重不匹配時。

1. INTRODUCTION

Real applications of speech recognition strongly demand the recognition performance to be robust with respect to environmental changes. However, the recognition accuracy of almost all existing recognition systems drops dramatically when there is a mismatch between the training and testing conditions. Substantial research e have been made in this area. One ca of such approaches, among many othe focused on finding a set of rela robust feature representation for signals, so that it is less sensit various environmental distortion Cepstral Mean Subtraction (CMS) Cepstral Normalization (CN) [4 Relative Spectral (RASTA) [5] tech are typical examples of this categor of them can be considered as perf filtering on the time trajectories speech features in order to al harmful effects of various disto and corruptions. Such approache been proved to be able to impro robustness of the recognition perfo significantly, and the subject o report is also along this directio

The filters used by CMS, CN and are independent of the recognition Although they are very effective, th no guarantee that these solutio optimal for a specif ic recognition application environment. Ther filtering coefficients optimized specific recognition task or appli environment via data-driven appr based on some optimization c become highly desirable. The criter Linear Discriminant Analysis (LD been widely applied [6,7] i approaches in the optimization pr In recent works, the criteria o Principal Component Analysis (PCA) and Minimum Classification Error (M were also used as the optimi criteria to derive the data temporal filters. It was shown [2 all these data-driven temporal fi

although derived from different criteria, can offer reasonable recognition performance improvements for recognition tasks with mismatched conditions.

In this report, PCA is again used as the optimization criterion for deriving temporal filters as well. However, different from the previous works [1,2], here we take into consideration multi-eigenvectors rather than the first principal component only as used previously [1,2], and the resulted temporal filters are the properly weighted linear combination of these multi-eigenvectors. It is shown that these new temporal filters can offer significant improvements over the previously obtained PCA-derived filters based on the first principal components only.

The remainder of this report has 4 sections. The approach to obtain the multi-eigenvector temporal filters is first described in section 2. The experimental setup and the experimental results are then presented and discussed in sections 3 and 4. Section 5 finally gives the concluding remarks.

2. TEMPORAL FILTER DESIGN USING PRINCIPAL COMPONENT ANALYSIS (PCA)

Given an ordered sequence of K-dimensional feature vectors $\boldsymbol{x}(n)$ as shown in Figure 1(a), with time index $n=1, \ldots\ldots, N$, and feature index $k=1, 2, \ldots K$,

$$\boldsymbol{x}(n)=[x(n,1)\ \ x(n,2),\ \ \ldots,\ \ x(n,k),\ \ \ldots,\ \ x(n,K)]^T,\ \ n=1,\ldots,N \qquad (1)$$

then the k'th time trajectory of $\boldsymbol{x}(n)$ is the sequence $[x(1,k)\ \ x(2,k) \ldots\ldots x(N,k)]$, denoted as $y_k(n)$, where $y_k(n)=x(n,k)$. Now we'd like to design an L-sample FIR- $W_k(z)$ to be performed on the k't trajectory $y_k(n)$. First, an L- rectangular window is shifted alc k'th time trajectory to obta sequences of L-dimensional vectors $n=1 \ldots N-L+1$,

$$\boldsymbol{z}_k(n)=[\ y_k(n)\ \ y_k(n+1)\ \ y_k(n+2)\ \ y_k(n+L-1)]^T . \qquad (2)$$

So $\boldsymbol{z}_k(n)$ is the windowed vector of th time trajectory started at the time n, on which the L-sample FIR fil applied, as depicted in Figure 1(b

2.1 Previously proposed PCA-d temporal filters [1]

In the previous work [1], the L-dimensional vectors, $\boldsymbol{z}_k(n)$, are as the samples of a random vector hence the mean vector and the cova matrix of $\boldsymbol{z}_k$ can be calculated,

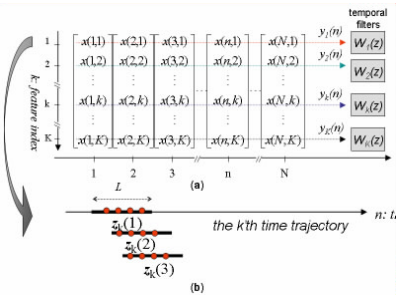

Figure 1. The representation of th trajectories of feature sequence the time index n and the feature i and (b) the windowed vector $\boldsymbol{z}_k(n)$ along a time trajectory

$$\mu_{z_k} = \frac{1}{N-L+1}\sum_{n=1}^{N-L+1} y_k(n)$$

$$\sigma^2_{z_k} = \frac{1}{N-L+1}\sum_{n=1}^{N-L+1}(y_k(n)-\mu_{z_k})(y_k(n)-\mu_{z_k})$$

Following the procedure of PCA, the components of the first eigenvector $\ddot{o}_k$ corresponding to the largest eigenvalue of the covariance matrix Σ_{z_k} is then taken as the coefficients of the L-sample filter, which maps the L-dimensional random vector z_k into a one-dimensional random variable with maximum variance. Such process is carried out for each time trajectory, thus yielding a separate FIR filter for each time trajectory.

2.2 The multi-eigenvector temporal filtering approach

Based on the PCA theory [8], assume $\ddot{o}_{i,k}$, $i=1,2,...L$ are the L distinct normalized eigenvectors of the covariance matrix Σ_{z_k} with corresponding decreasing eigenvalues $\ddot{e}_{i,k}$, $i=1,2,...L$, i.e., $\ddot{e}_{1,k} \geq \ddot{e}_{2,k} \geq ... \geq \ddot{e}_{L,k}$, and $y_{i,k}$, $i=1,2,...L$ are the random variables representing the projections of the random vector z_k on $\ddot{o}_{i,k}$. It can then be shown that the variance of $y_{i,k}$ is equal to $\ddot{e}_{i,k}$. As a result, with the previously proposed PCA-derived temporal filters, the filter coefficients used are the components of the eigenvector $\ddot{o}_{1,k}$, therefore the filter output is the random variable $y_{1,k}$ with variance being the largest eigenvalue $\ddot{e}_{1,k}$. Therefore $y_{1,k}$ can be viewed as the most "expressive" 1-dimensional representation of z_k, and this is apparently why the recog nition can be improved.

However, from the above discussion it is clear that there is still some part of information carried by z_k which was not used at all in the previously proposed PCA-derived filters, i.e., the other $y_{i,k}$'s, $i=2,...L$, which also carry one information of z_k that may be help improving the recognition perfor With these observations, th multi-eigenvector temporal proposed in this report are

$$\bar{w}_k = \sum_{i=1}^{M} a_{i,k} \ddot{o}_{i,k} \quad , \quad w_k = \frac{\bar{w}_k}{|\bar{w}_k|} = \frac{1}{\sqrt{\sum_{i=1}^{M} a_{i,k}^2}} \bar{w}_k \quad ,$$

where w_k is the new coefficients L-sample filter for time trajectory summation is over the first M eigenv with larger corresponding eigenv and $1 < M \leq L$. Note that the length temporal filter vector w_k is normali unity, so as to be consistent wi eigenvector used in the pre proposed PCA-derived temporal fi Also, the first M eigenvector weighted by the corresponding eigen in Eq. (5), therefore the outputs new temporal filter can be viewed samples of a new random variable

$$v_k = w_k^T z_k = \frac{1}{\sqrt{\sum_{i=1}^{M} a_{i,k}^2}} \sum_{i=1}^{M} a_{i,k} y_{i,k} \quad .$$

In other words, in addition to $y_{1,k}$ a previously, with the p multi-eigenvector filters the s "next expressive" but "ortho representations of z_k are also in here, while weighted by corresponding variance values parameter M, or the number of eigenv used here, can be determined empiri In the experiments presented belo chosen to be 3. Such a choice can a verified by the fact that $\ddot{e}_{1,k}$, $\ddot{e}_{2,k}$, are always significantly large other eigenvalues.

In the experiments below, it wil

be shown that the new multi-eigenvector filters obtained with the approach here are low-pass filters whose characteristics in speech signals is able to enhance the syllabic-rate information (about 4Hz) in speech signals. However, with the low-pass characteristics the slowly-varying channel bias components may also be emphasized. In order to handle this problem, in the experiments below the original Mel-Frequency Cepstral Coefficients (MFCC) are first proposed by *Cepstral Normalization* (CN) [4] in order to properly reduce the low-frequency components. Therefore multi-eigenvector temporal filters discussed here are derived from, and performed on, the normalized cepstral coefficients. In other words, the CN process was first performed on the MFCC's of the training speech database to obtain the multi-eigenvector filters. These filters are then applied to the time trajectories of the MFCC's of both the training and testing database to obtain the new feature parameters. These new parameters are finally used in model training as well as the testing experiments.

3. EXPERIMENTAL SETUP

The AURORA2 database distributed by ETSI committee is used for the experiments here. It contains several sets of noisy speech with additive noise of different characteristics and levels plus some channel effects, as representatives for real-world environments. The clean speech data in the AURORA2 database consisting of 8440 utterances of English connected digits were used to obtain the data-driven temporal filters in the experiments Each utterance was first converte 13-dimensional Mel-frequency ce coefficients (12 MFCCs + log coefficient) using the AURORA Front-end, as defined in the documentation. The resulted 8440 s of MFCC's were then processed *Cepstral Normalization* (CN) algo such that the mean and variance para were normalized to 0 and 1 respect Then for each time trajectory feature vectors, two versions of te filters were constructed, one wi previously proposed PCA-based ap with a single eigenvector [1,2], a other with the multi-eigenvector ap proposed here as described in sect The length L of the temporal filte empirically set to be 15.

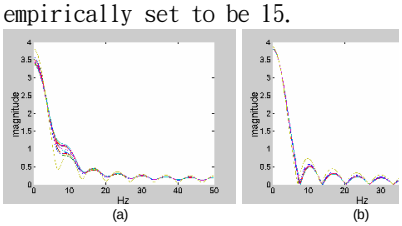

Figure 2. The frequency responses of (a) the 13 multi-eigenvector te filters and (b) the 13 previously proposed PCA-derived filters

Figure 2 shows the frequency respon the obtained temporal filters: (a the new multi-eigenvector ap proposed here in this report, and (b the previously proposed PC approach with a single eigenvector for the 13 time trajectories, traine the clean speech database of AURORA2 these figures we can have some i observations as follows.

1. Both sets of the temporal filte low-pass. They don't attenuate t

modulation frequency components, although CMS or RASTA does. This is why the *Cepstral Normalization* (CN) process is useful here.

2. The 13 previously proposed PCA-derived filters for the 13 temporal trajectories are very close, which the 13 multi-eigenvector temporal filters are obviously different around 0~15Hz.

3. The previously proposed PCA-derived filters have the shape of a main-lobe and several side-lobes; but the multi-eigenvector filters proposed here have a wider pass-band without apparent zeros.

4. For the new multi-eigenvectors filters proposed here, the modulation-frequency components around 0~4.5Hz are specially emphasized, therefore they may somehow enhance the syllabic-rate information (about 4Hz [5]) in the speech signals, which is a possible reason for the recognition performance improvements as found in the data below.

4. EXPERIMENTAL RESULTS

In the recognition experiments, there are two training modes in AURORA2: *clean speech training* and *multi-condition training*. In the multi-condition training, the acoustic models were trained with speech data under different noisy conditions, added with different types of noise at different levels and so on. For each training mode, three sets (Sets A, B and C) of utterances artificially contaminated by different types of noise (subway, babble, car, etc.) at different SNR levels (ranging from -5dB to 20dB) were tested. Since the proposed approach here only has to do with the fr feature extraction, all the follow procedures for training and recog are exactly identical to the ref experiments stated in documentation.

In the training process 13-dimensional normalized MFCC fea i.e., the MFCC features but proces CN, were used to construct bo multi-eigenvector and the prev proposed PCA-derived temporal fi and these two sets of temporal f were then applied on these 13-dimen normalized MFCC features. The re 13-dimensional new features plus delta and delta-delta features we components in the finally 39-dimensional feature vectors. these new feature vectors the HMM each digit were trained. Similarly, testing phase the clean and corrupted testing speech data wer first converted to MFCC's, processe and then individually processed above two sets of temporal optimized with the training data, t various sets of feature vecto testing.

4.1 Recognition results

Table 1 lists the recognition r respectively for the baseline expe (Baseline), i.e., with the origina features without any further proce and the experiments with the fe processed by CN only (CN), first by then by the previously p PCA-derived filters (CN+PCA), firs and then by the multi-eigenvector f

proposed here (CN+M-eigen). The results include those for two training modes, clean speech training and multi-condition training, and three testing sets, sets A, B, and C. The length of the filters, L, for both of the two latter cases are set to be 15, and the number of eigenvectors, M, for the multi-eigenvector filters proposed here in eq. (6) is set to be 3. The word accuracy listed in Table 1 is the average of the recognition rates between 0-20dB. The overall word error rate (WER) improvements in Table 1 were calculated with respect to the baseline results. Note that with the proposed multi-eigenvector filters proposed in this report applied after CN, very significant improvements over CN only or CN plus the previously proposed PCA-derived filters were obtained in both clean speech and multi-condition training modes, and the improvements in the clean training mode are specially high. This verified that the new multi-eigenvector temporal filters are particularly effective when the training and testing environmen mismatched. Retailed comparison data in Table 1 indicates th proposed multi-eigenvector f performed better than the prev proposed PCA-derived filters in case, for all the testing sets i training modes.

From the right part of Table 1, w observe that in the multi-con training mode the new multi-eigen filters can successfully improv recognition performance of CN-processed speech features, but t not always true for the pre proposed PCA-derived filters.

4.2 Choice of the parameter M

We mentioned in section 2.2 th parameter M, or the number of eigenv to be included, is chosen to be 3 b the eigenvalues $\ddot{e}_{i,k}$ becomes much s for i >3. Here some

Approaches	Clean speech training				Multi-condition training		
	Set A	Set B	Set C	Overall WER improvements	Set A	Set B	Set C
Baseline	*61.34*	*55.76*	*66.14*	-	*87.82*	*86.22*	*83.76*
CN	*70.18*	*70.78*	*66.37*	*20.62*	*89.67*	*88.07*	*86.10*
CN+PCA	*78.99*	*79.69*	*77.47*	*45.31*	*87.91*	*86.90*	*86.70*
CN+M-eigen	***81.90***	***82.78***	***80.83***	***53.33***	***90.43***	***89.22***	***89.11***

Table 1. Word accuracy for the three testing sets A, B, C under clean - and multi-con training modes. The overall word error rate (WER) improvements were calculate respect to the baseline experiments.

experimental results for different values of M are presented. The results for the same recognition experiments, as in Table 1, i.e., for the two training modes and three test sets, but using different values of M (M=1, 2, 3, 4, 5, 6, 9, 12, 15) are depicted in Figure 3(a)(b). Of course here the case of M=1 is exactly the previously proposed PCA-derived filters (CN+PCA in Table 1). It can be found clearly from this figure that very sharp improvements in performance were obtainable as M was increased from 1 to 3 in all cases, but the improvements turned out to be saturated if M was further increased from 3. As mentioned above, the eigenvalues $\lambda_{4,k}$~$\lambda_{15,k}$ are very small relative to $\lambda_{1,k}$~$\lambda_{3,k}$; therefore the information projected on the first three eigenvectors $\mathbf{v}_{1,k}$~$\mathbf{v}_{3,k}$ are much more important, and actually dominate the recognition processes.

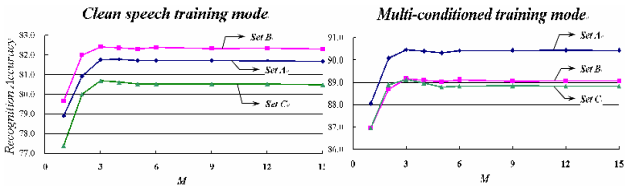

Figure 3. Recognition accuracy for multi-eigenvector filters with different M values: (a) clean speech training and (b) multi-condition training

4.3 Choice of the filter length L

When designing the temporal filters including the multi-eigenvector filters as proposed here, it is well known that if the length of the temporal filters L is smaller, the width of the pass-band of the filters will be larger. This is clear by comparing Figure 2(a) for L=15 with Figure 4(a)(b) for multi-eigenvector filters with M=3 but L=10 and 20 respectively. From these figures we can observe that the 3 dB width of the pass-band are around 6~7Hz and 4Hz for the cases of L=10 and 20 respectively, while that for L=15 is about 4.5Hz. Since the syllabic-rate of human speech is roughly around 4Hz, so the characteristics of the filters for the case L=15 may exactly emphasize the syllabic-rate information and thus the performance roughly saturates at L=15, as can be found by the recognition accuracy shown in Table 2.

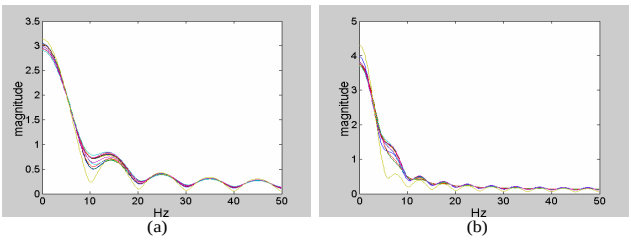

Figure 4. The frequency responses of the 13 multi-eigenvector filters as L is set to (a) 10 (b) 20

	Clean speech training		
L	**Set A**	**Set B**	**Set C**
10	*79.95*	*80.22*	*77.23*
15	*81.90*	*82.78*	*80.83*
20	*81.75*	*82.64*	*80.79*

Table 2. Recognition accuracy for multi-eigenvector filters with M=3 and different values of L with clean speech training mode

4.4 Further comparison between the two versions of temporal filters with some other metric

In this subsection, we'd like to compare the multi-eigenvector filters proposed here with the previously proposed PCA-derived filters using a metric different from recognition accuracy. The metric used is the average of the normalized distance between the corrupted features, $\hat{x}$, and the corresponding clean speech features, x,

$$d = E\left[\frac{\hat{x} - x}{x}\right], \qquad (7)$$

where the average is taken over all the testing speech. This metric is to provide an estimated measure of the robustness of the temporal -filtered MFCC features with respect to the corruption. Smaller values of d imply that the features are less influenced by the corruption. Table 3 compares this distance measure d for the three testing sets A, B, C under different SNR values. We see that the multi -eigenvector filters proposed here gives smaller averaged normalized distance in all cases. This offers another explanation why the proposed multi-eigenvector filters give better recognition accuracy.

	Approaches	**20db**	**15db**	**10db**	**5db**	**0db**	**-5db**
Set A	*CN+PCA*	*0.6826*	*0.7554*	*0.8281*	*0.9081*	*0.9932*	*1.0855*
	CN+M-eigen	*0.6211*	*0.6940*	*0.7686*	*0.8521*	*0.9473*	*1.0587*
Set B	*CN+PCA*	*0.6693*	*0.7394*	*0.8150*	*0.9053*	*1.0016*	*1.1086*
	CN+M-eigen	*0.6071*	*0.6778*	*0.7556*	*0.8493*	*0.9559*	*1.0782*
Set C	*CN+PCA*	*0.6836*	*0.7561*	*0.8305*	*0.9123*	*1.0057*	*1.1029*
	CN+M-eigen	*0.6238*	*0.7002*	*0.7806*	*0.8763*	*0.9902*	*1.1098*

Table 3. The average normalized distance between clean and corrupted speech features under various SNRs and two different temporal filter techniques

5. CONCLUSION

In this report, we proposed a multi-eigenvector approach of designing data-driven temporal filters, in which more than one eigenvectors are weighted by their corresponding eigenvalues and summed to form the filter coefficients. Very encouraging experimental results have been obtained and this approach is also shown to be particularly effective when the training and testing environments are highly mismatched.

6. REFERENCE

[1] J-W. Hung. et al “Comparative Analysis for Data-Driven Temporal Filters Obtained Via Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) In Speech Recognition”, Eurospeech 2001

[2] J-W. Hung and L-S. Lee “Data-Driven Temporal Filters Obtained Via Different Optimization Criteria Evaluated On AURORA2 Database”, ICSLP 2002

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification". IEEE Trans. Acoust. Speech Signal Process. 1981

[4] O. Viikki and K. Laurila, “Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization,” in ESCA NATO Workshop Robust Speech Recognition Unknown Communication Channels, Pont-a-Mousson, France, 1997, pp. 107–110.

[5] H. Hermansky and N. Morgan, “RASTA processing of speech”. IEEE Trans. Speech Audio Process. 2, 578-589, 1994

[6] C. Avendano, S. van Vuuren and H. Hermansky, "Data Based Filter Design for RASTA-like Channel Normalization in ASR" ICSLP 96

[7] S. van Vuuren and H. Hermansky, "Data-driven Design of RASTA-like Filters", Eurospeech 97

[8] K. Fukunaga, “Introduction to statistical pattern recognition”, E.2nd, Academic Press, 1990