International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

DATA-DRIVEN TEMPORAL FILTERS BASED ON MAXIMUM MUTUAL INFORMATION FOR ROBUST FEATURES IN SPEECH RECOGNITION

Yung-Sheng Huang and Jeih-weih Hung
Dept of Electrical Engineering, National Chi-Nan University
Taipei
e-mail: s1323512@ncnu.edu.tw, jwhung @ncnu.edu.tw

Abstract

Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) and Minimum Classification Error (MCE) have been used to derive data-driven temporal filters in order to improve the robustness of speech features for speech recognition. In this paper, the criterion of Maximum Mutual Information (MMI) is proposed for constructing the temporal filters, and detailed comparative analysis among these various approaches are presented and discussed. Experimental results show that the MMI-derived temporal filters significantly improve the recognition performance of the original MFCC features as LDA/PCA/MCE-derived filters do. Also, while the MMI-derived filters are combined with the conventional temporal filters, Cepstral Mean and Variance Normalization (CMVN), the recognition performance can be further improved.

1. Introduction

When there is a mismatch between the acoustic conditions of training and application environments for a speech recognition system, the performance of the system is very often seriously degraded. Various sources give rise to this mismatch, such as additive noise and channel distortion. The robustness of speech recognition techniques with respect to these different mismatched acoustic conditions thus becomes very important. The Cepstral Mean Subtraction (CMS) [1] , Cepstral Mean and Variance Normalization (CMVN) [2] and the Relative Spectral (RASTA) [3] techniques are typical examples in performing filtering on the time trajectories of speech features in order to alleviate harmful effects of various distortions and corruptions. These approaches have been widely proved to be able to effectively improve the performance of speech recognition systems without changing the core training/recognition processes.

It is known that the temporal filters developed by CMS, CMVN and RASTA are independent of the recognition tasks. Although they are very effective, there is no guarantee that these solutions are optimal to any recognition task or environment. It is therefore desirable to obtain optimal sets of filtering coefficients for a specific recognition task or an application environment, which are often obtained in a data-driven process according to some optimization criterion. Linear Discriminant Analysis (LDA) has been widely applied [4,5] as the optimization criterion to yield the time-trajectory temporal filters, which have been found to give better recognition performance than the conventional RASTA filters.

In recent works, Principal Component Analysis (PCA) [6,7] and Minimum Classification Error (MCE) [7] were also used as optimization criteria to derive the data-driven temporal filters just like LDA. It was shown [7] that all these data-driven temporal filters, although derived from different criteria, can offer obvious recognition performance improvements for recognition tasks with mismatched conditions.

In this paper, we proposed that the criterion of Maximum Mutual Information (MMI) [8] can be also applied in the optimization process to obtain temporal filters similar to those obtained using LDA, PCA, and MCE. In addition, comparative performance analysis among the above four different optimization processes in terms of the features obtained accordingly is presented. It is found that the MMI-derived temporal filters have similar frequency response shapes as the LDA/PCA /MCE-derived ones. Experimental results also showed that the MMI-derived filter can significantly improve recognition performance as compared with the original MFCC features, just as LDA/PCA/MCE-derived filters can. Also, it was shown that further improved recognition accuracy can be obtained if the MMI-derived filters are integrated with the approach of CMVN.

The remainder of this paper contains 4 sections. The approach to obtain the data-driven temporal filters using the criterion of MMI is introduced in section 2. The experimental setup, the frequency response shapes of the resulting MMI-derived temporal filters, as well as the experimental results are then presented and discussed in sections 3 and 4. Finally, a brief conclusion is given in section 5.

2. Temporal Filter Design Based on the Maximum Mutual Information

An ordered sequence of K-dimensional feature vectors $\{\mathbf{x}(n), n = 1, 2, \cdots, N\}$, where n is the time index, is illustrated in Figure 1. Each vector $\mathbf{x}(n)$ is represented as

$$\mathbf{x}(n) = [x(n,1), x(n,2), \cdots, x(n,k), \cdots, x(n,K)]^T , \qquad (1)$$

where $x(n,k)$ is the k-th component of the feature vector $\mathbf{x}(n)$ at time n. Therefore, the k-th time trajectory of

$\{\mathbf{x}(n)\}$ is the sequence $[x(1,k), x(2,k), \cdots, x(N,k)]$, denoted here as $\{x_k(n), n = 1,2,\cdots,N\}$, where

$$x_k(n) = x(n,k),\ n = 1,2,\cdots,N,\ k = 1,2,\cdots,K \qquad (2)$$

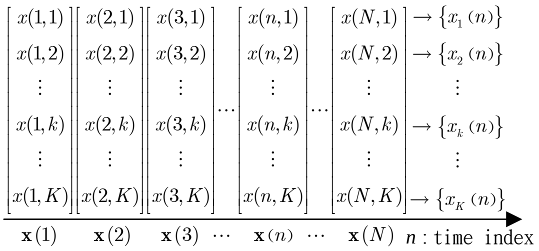

x(1,1) x(2,1) x(3,1) ⋯ x(n,1) ⋯ x(N,1) → {x_1(n)}
x(1,2) x(2,2) x(3,2) x(n,2) x(N,2) → {x_2(n)}
x(1,k) x(2,k) x(3,k) x(n,k) x(N,k) → {x_k(n)}
x(1,K) x(2,K) x(3,K) x(n,K) x(N,K) → {x_K(n)}
x(1) x(2) x(3) ⋯ x(n) ⋯ x(N)
n : time index

Figure 1. The representation of the time trajectories of feature parameters.

Now we would like to design an *L*-point FIR filter which is performed on the time trajectory $\{x_k(n)\}$. First, an *L*-point rectangular window is shifted along the sequence $\{x_k(n)\}$ to obtain the sequences of *L*-dimensional vectors $\{\mathbf{z}_k(n), n = 1,2,\cdots,N-L+1\}$, where

$$\mathbf{z}_k(n) = [x_k(n)\ \ x_k(n+1)\ \ \cdots\ \ x_k(n+L-1)]^T, \qquad (3)$$

So $\mathbf{z}_k(n)$ is the windowed vector of $\{x_k(n)\}$ started at the time index *n*, on which the *L*-point FIR filter is applied, as depicted in Figure 2. Below, the *L*-point FIR filter is designed using the statistics of $\mathbf{z}_k(n)$ with the criterion of Maximum Mutual Information (MMI).

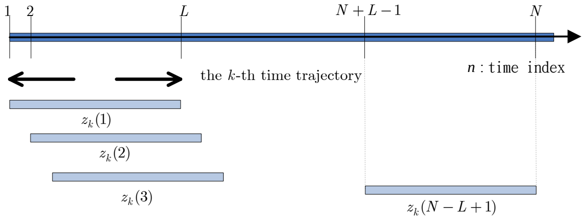


Figure 2. The windowed segments **z**$_k$(*n*) to be used in the temporal filter design.

In the general formulation of MMI analysis, a Mutual Information function $I_j(\bullet)$ is defined for a certain class *j*, an observation feature *X* that belongs to this class *j*, and a model set Λ={λ_i, *i*=1,2,…,*J*}, where λ_i is the model representing class *i*. $I_j(\bullet)$ is often defined as

$$I_j(X,\Lambda) = \log P(X|\lambda_j) - h(\log P(X|\lambda_i),\ i = 1,2,\cdots,J), \qquad (4)$$

where $h(\bullet)$ is a function defining how the class-conditioned likelihoods $\log P(X|\lambda_i)$ for the competing models λ_i are counted in the Mutual Information function. Now, the Mutual Information function in equation (4) for the *k-th* time trajectory is defined as

$$\begin{aligned} I_j\left(X_k^{(j)}(n), \Lambda_k\right) &= \log P\left(X_k^{(j)}(n) = H_k^T \mathbf{z}_k^{(j)}(n) \middle| \lambda_{j,k}\right) \\ &- \log\left\{\frac{1}{J}\sum_{m=1}^{J} P\left(X_k^{(j)}(n) = H_k^T \mathbf{z}_k^{(j)}(n) \middle| \lambda_{m,k}\right)\right\}, \end{aligned} \qquad (5)$$

where *H*$_k$ is the weight vector representing the temporal filter coefficients. Then the sum of the mutual information function for the *k*-th time trajectory can be written as

$$R_{k,MMI} = \sum_{j=1}^{J}\sum_{n=1}^{N_j} I_j\left(H_k^T \mathbf{z}_k^{(j)}(n), \Lambda_k\right), \qquad (6)$$

where $\Lambda_k = \{\lambda_{j,k}, j = 1,2,\cdots,J\}$ is the set of models $\lambda_{j,k}$ for the *j-th* class and the *k-th* time trajectory, N_j is the total number of $\mathbf{z}_k$ that are labeled as the *j-th* class. For simplicity, the model $\lambda_{j,k}$ for the transformed value $X_k^{(j)}(n) = H_k^T \mathbf{z}_k^{(j)}(n)$ in equation (6) is assumed as a one-dimensional (single-variate) Gaussian distribution,

$$\lambda_{j,k} = N\left(H_k^T \boldsymbol{\mu}_k^{(j)}, H_k^T \boldsymbol{\Sigma}_k^{(j)} H_k\right), \qquad (7)$$

where $\boldsymbol{\mu}_k^{(j)}$ and $\boldsymbol{\Sigma}_k^{(j)}$ are the mean and covariance matrix of the vectors **z**$_k$(*n*) labeled as belonging to the class *j*.

In order to maximize the mutual information in equation (6) by choosing the weight vector *H*$_k$, the gradient-descent algorithm is used. Taking derivative of equation (6) with respect to *H*$_k$, we have

$$\frac{\partial R_{k,MMI}}{\partial H_k} = \sum_{j=1}^{J}\sum_{n=1}^{N_j} \frac{\partial I_j\left(H_k^T \mathbf{z}_k^{(j)}(n), \Lambda_k\right)}{\partial H_k} \qquad (8)$$

$$= \sum_{j=1}^{J}\sum_{n=1}^{N_j} \frac{\sum_{m=1}^{J} P\left(H_k^T \mathbf{z}_k^{(j)}(n) \middle| \lambda_{m,k}\right)\left[\frac{\partial \log P\left(H_k^T \mathbf{z}_k^{(j)}(n) \middle| \lambda_{j,k}\right)}{\partial H_k} - \frac{\partial \log P\left(H_k^T \mathbf{z}_k^{(j)}(n) \middle| \lambda_{m,k}\right)}{\partial H_k}\right]}{\sum_{m=1}^{J} P\left(H_k^T \mathbf{z}_k^{(j)}(n) \middle| \lambda_{m,k}\right)}$$

where

$$P\left(H_k^T \mathbf{z}_k^{(j)}(n) \middle| \lambda_{m,k}\right) = N\left(H_k^T \mathbf{z}_k^{(j)}(n); H_k^T \mu_k^{(m)}, H_k^T \Sigma_k^{(m)} H_k\right)\ 1 \le m \le J$$

Then, a better estimate of the temporal filter *H*$_k$ for the (*t*+1)-th iteration, $H_k^{(t+1)}$, based on its estimate obtained from the *t*-th iteration $H_k^{(t)}$,is obtained as below.

$$\bar{H}_k^{(t+1)} = H_k^{(t)} + \varepsilon_t \frac{\partial R_{k,MMI}}{\partial H_k}\bigg|_{H_k = H_k^{(t)}}, \qquad (9)$$

where ε_t is the learning rate at the *t*-th iteration, and

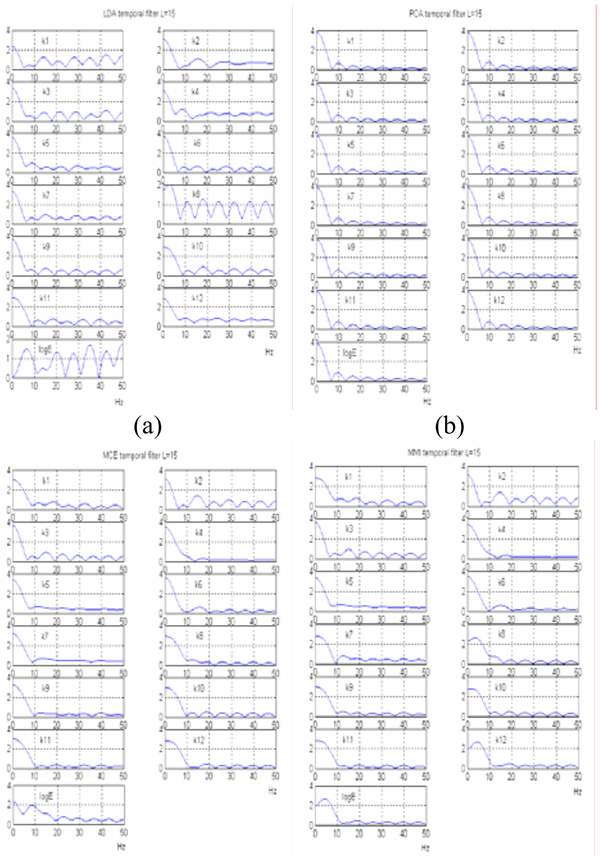
$$H_k^{(t+1)} = \bar{H}_k^{(t+1)} \Big/ \left|\bar{H}_k^{(t+1)}\right|, \qquad (10)$$

Equation (10) is used here to normalize the norm of the vector representing the temporal filter to unity, in order to let it be consistent with the eigenvectors used in LDA, PCA. The final temporal filter *H*$_k$ is obtained when the iteration process converges.

3. Experimental Setup

The speech database for the initial experiments included 8000 Mandarin digit strings produced by 50 male and 50 female speakers, taken from the database NUM-100A provided by the Association for Computational Linguistics and Chinese Language Processing at Taipei. The speech signal was recorded in normal laboratory environment at 8 kHz sampling rate and encoded with 16-bit linear PCM. The 8000 digit strings included 1000 each for 2, 3, 4, 5, 6 and 7-digit strings respectively plus 2000 single digit utterances. Among the 8000 Mandarin digital strings, 7520 were used in training, while the other 480 in testing. A 25ms Hamming window shifted with 10ms steps and a pre-emphasis factor of 0.97 were used to evaluate 13 mel-frequency cepstral coefficients (MFCCs, c1~c12 plus log-energy). The LDA/PCA/MCE/MMI-derived temporal filters were then obtained using these 13-dimensional MFCC vectors of the 7520 training digital strings. The Length *L* of the FIR filter was preliminarily set to be 15.

Figure 3 show respectively the frequency responses of the 13 LDA-, PCA-, MCE- and MMI-derived FIR filters. From these figures, several phenomena can be observed as follows.



(a) (b)

(c) (d)

Figure 3. The frequency response shapes of (a) the 13 LDA-derived temporal filters (b) the 13 PCA-derived temporal filters (c) the13 MCE-derived temporal filters(d) the13 MMI-derived temporal filters

1. All these data-driven filters do not attenuate the low modulation frequency components as CMS or RASTA does. In other words, They are all low-pass filters. (with the log-energy component of the LDA-derived filters in Figure 3 being the only one exception).

2. The widths of the main-lobes for filters derived with different criteria are of similar order. For LDA-derived filters, the main-lobe widths are roughly between 7-9 Hz. For PCA-derived ones, they are roughly 7 Hz. For these MCE- and MMI-derived ones, the cutoff frequencies of the main-lobes are not very clear, whose range is roughly between 6-20Hz.

3. For PCA-, MCE-, and MMI-derived filters, the magnitudes of the side-lobes are relatively low compared with that of the main-lobe. This is not true for some of LDA-derived filters.

4. Experimental Results

In the training process, the LDA-, PCA-, MCE-, and MMI-derived FIR filters were first respectively applied on the time trajectories of the MFCC feature vectors for the training database. The resulting 13-dimensional new features plus their delta and delta-delta features were the components of the finally used 39-dimensional feature vectors. With these new feature vectors, the HMMs for each digit and silence with 10 states and 4 mixtures per state were trained. Similarly, three conventional temporal filtering approaches, CMS, CMVN and RASTA, were also applied to the same original MFCC feature vectors and their respective HMMs were trained for recognition.

For the recognition experiments, the 480 clean speech testing digit strings were manually added with eight kinds of noise at different SNR levels (ranging from 5dB to 20dB) to produce noise corrupted speech data. The eight kinds of noise are divided into two sets, Set A (subway, babble, car, exhibition) and Set B (restaurant, street, airport, station) noise, all taken from the AURORA2 database.

4.1 Recognition Results

Table 1 lists the digit recognition accuracy averaged over four SNR conditions, 20dB, 15dB, 10dB and 5dB for four types of noise in Set A and four types of noise in Set B, respectively, for various temporal filtering techniques. In addition, the word error rate (WER) improvements compared with the baseline experiments (plain MFCC) were also listed in Table 1. From Table 1, several observations can be found as follows.

1. The new features processed by each of the temporal filters, both data-independent and data-driven, obviously perform better than the plain MFCC.

2. Among the data-independent temporal filters, we found that CMVN performed the best. It is also better than any of the four data-driven temporal filters.

3. Among the data-driven temporal filters ,it is found that the newly proposed MMI-derived temporal filters perform as well as, and sometimes better than the LDA-, PCA-, and MCE- derived temporal filtering approaches for all the cases in Set A and Set B.

4. The MMI-derived temporal filters give higher WER improvement with Set B (nonstationary noise) than with

Set A(stationary noise). This is also the case for other temporal filters.

	Set A		Set B	
	Average	WER Imp.	Average	WER Imp.
MFCC	62.42		69.09	
RASTA	67.27	12.91%	76.31	23.36%
CMS	65.58	8.41%	74.55	17.66%
CMVN	72.93	27.97%	78.72	31.15%
LDA	64.93	6.68%	72.54	11.16%
PCA	65.39	7.91%	74.08	16.13%
MCE	66.40	10.59%	73.18	13.23%
MMI	66.63	11.21%	73.97	15.77%

Table 1. The recognition accuracy (%) averaged over the different SNR conditions, 20dB, 15dB, 10dB and 5dB, and averaged over different types of noise, together with the word error rate(WER) improvements with respect to the baseline experiments.

4.2 Combining Data-Driven Filters with Cepstral Mean and Variance Normalization (CMVN)

From the above sections, it is found that all the data-driven temporal filtering approaches discussed here, the LDA-, PCA-, MCE, and MMI -derived ones, are low-pass filters, and thus are very helpful in enhancing the low modulation frequency components of the speech information in order to improve recognition accuracy. However, it is also very likely that all these low-pass temporal filters tend to retain the most slowly-varying components (roughly 1 Hz modulation frequency or below). On the other hand, all the conventional temporal filtering approaches, whether being CMS, CMVN, or RASTA, are high-pass or band-pass filters, Especially, the CMVN is very outstanding. It performs better than all other temporal filters. This leads to the concept of integrating the CMVN and the data-driven temporal filters. The obtained experiment results are shown in Table 2.

	Set A		Set B	
	Average	WER Imp.	Average	WER Imp
MFCC	62.42		69.09	
CMVN	72.93	27.97%	78.72	31.15%
CMVN+LDA	77.02	38.85 %	82.84	44.48 %
CMVN+PCA	76.96	38.69 %	83.93	48.01 %
CMVN+MCE	77.14	39.17 %	83.59	46.91 %
CMVN+MMI	76.77	38.19 %	83.75	47.43 %

Table 2. The recognition accuracy (%) averaged over the different SNR conditions, 20dB, 15dB, 10dB and 5dB, and averaged over different types of noise, together with the word error rate (WER) improvements with respect to the baseline experiments.

Observing Table 2 and comparing it with Table 1, it can be found that the each of four data-driven temporal filters plus CMVN is always significantly better than CMVN alone, and better than the respective data-driven temporal filter alone as well. Therefore, the data-driven temporal filters are additive to conventional CMVN under noisy conditions. We believe that such improved results are due to the fact that the syllabic-rate information are enhanced by the data-driven temporal filters, while the very slowly-varying deteriorative components, which may be amplified or left due to the low-pass characteristics of the data-driven temporal filters, are reduced or eliminated by CMVN.

5. Conclusion

In this paper, we proposed a new temporal filtering approach using the criterion of Maximum Mutual Information (MMI). Significant improvements in recognition accuracy under different conditions show the effectiveness of the MMI filtering approach. The MMI temporal filtering may efficiently alleviate the mismatch caused by noise corruption. Furthermore, the newly proposed MMI-derived temporal filters outperform the previously proposed LDA-, PCA-, and MCE- derived ones under the environment of stationary noise. Also, it can be easily integrated with Cepstral Mean and Variance Normalization (CMVN) method to provide further improvements.

6. References

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

[2] 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

[3] H. Hermansky and N. Morgan, "RASTA processing of speech". IEEE Trans. Speech Audio Processing, 1994

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

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

[6] 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

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

[8] Y.L. Chow. “Maximum Mutual Information Estimation of HMM Parameters for Continuous Speech Recognition Using the N-Best Algorithm”, ICASSP1990.