Telephone Speech Recognition with Data-Driven Selective Temporal Filtering based on Principal Component Analysis

Sungyun Jung*, Jongmok Son* , and Keunsung Bae*
* School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 702-701, Korea
Tel : +81-053-940-8627 Fax : +81-053-950-5505 E-mail: yunij@mir.knu.ac.kr

Abstract: The performance of a speech recognition system is generally degraded in telephone environment because of distortions caused by background noise and various channel characteristics. In this paper, data-driven temporal filters are investigated to improve the performance of a specific recognition task such as telephone speech. Three different temporal filtering methods are presented with recognition results for Korean connected-digit telephone speech. Filter coefficients are derived from the cepstral domain feature vectors using the principal component analysis.

Selective temporal filter, Principal component analysis, Telephone speech recognition, Feature extraction

1. INTRODUCTION

Generally, the performance of a speech recognition system is degraded in telephone environment because of distortions caused by background noise and various channel characteristics. The cepstral mean normalization(CMN)[1] method has been typically used to reduce the time-varying channel effect and to improve the recognition performance in telephone speech. The CMN can be thought as a filter that is independent of recognition tasks to alleviate the harmful effects of channel distortions. For a specific recognition task such as telephone speech recognition, it might be more effective to get filters whose coefficients are driven from the time trajectories of speech features in training database.

Recently, J. Hung et al.[2,3] has proposed data-driven temporal filters for robust speech recognition in mismatch environment between training and testing database. So we applied this technique to a specific task, Korean connected- digit telephone speech recognition, and proposed a selective temporal filtering method with new set of filter coefficients. The principal component analysis was used to derive the data-driven filter coefficients.

This paper is organized as follows. In section 2, mel-frequency cepstral coefficient(MFCC) feature parameters based on the direct weighted filter bank analysis(DWFBA) and data-driven temporal filters are explained. Experimental condition and results are shown with our discussions in section 3. Finally, conclusion is given in section 4.

2. FEATURE EXTRACTION AND TEMPORAL FILTERING

2.1. DWFBA-based MFCC feature parameters

The DWFBA emphasizes the spectral band that has large energy in the log filter bank [4]. The block diagram for DWFBA-based MFCC feature extraction is shown in Fig. 1. The log filter bank energies are multiplied by a set of weighting factors prior to discrete cosine transform(DCT) to get the MFCC. Since the band energy that contains more signal information tends to be large, the DWFBA results in putting emphasis on signal components in log spectral domain, and, it has shown to provide better recognition performance than conventional MFCC [4]. First, a weighting factor, w_i, is computed depending on the log energy in each band by eq. (1). It is directly proportional to the log energy of each critical band. Then, each weighted filter bank log energy is obtained by eq. (2).

$$w_i = \frac{\log(e_i + 1.0)}{\sum_{j=1}^{Q} \log(e_j + 1.0)} \qquad (1)$$

$$dw\log_i = w_i \log(e_i + 1.0), \qquad 1 \le i \le 19 \qquad (2)$$

where e_i represents the i th band energy, and $dw\log_i$ is i th band energy recomputed in the log domain. Finally, MFCC based on DWFBA, dwc_m, is obtained by taking DCT with the weighted log spectral energy as given in eq. (3) .

$$dwc_m = \sum_{i=1}^{Q} w_i \log(e_i + 1.0) \cos(m(\frac{2i-1}{2})\frac{\pi}{Q}), \qquad (3)$$
$$1 \le m \le 13$$

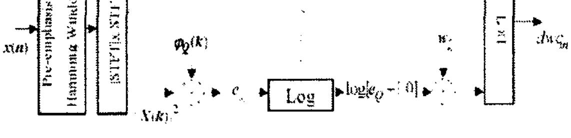

Fig. 1. Block diagram for DWFBA-based MFCC extraction

In [3], the MFCCs are first processed by cepstral mean normalization before temporal filtering in order to handle the slowly-varying channel bias problem as well as the syllabic-rate (about 4Hz) information in speech signals. The CMN is widely used as a channel compensation technique. However, to get the effective compensation of channel distortion with CMN, speech signals having a few seconds of length is needed to get appropriate cepstral mean [1].

Since the utterance of connected 4-digit is not long enough, in this paper, we used the modified real-time cepstral normalization(MRTCN)[1] method which estimates the cepstral mean and its variance every utterance. Each cepstral vector is then normalized with estimated mean and variance as given in eq. (4).

$$\overline{C}_{t,i} = \frac{C_{t,i} - \overline{X}_t}{\sqrt{\overline{\sigma}_t^2}} \qquad (4)$$

where $C_{t,i}$ is the MFCC vector of the i th frame in the t th 4-digit speech signal. The estimated cepstral mean vector at the t th speech signal, $\overline{X}_t$, and its estimated variance, $\overline{\sigma}_t^2$, are updated using eqs. (5) and (6), respectively, and α is the forgetting factor and set to be 0.125 empirically.

$$\overline{X}_t = \alpha X_t + (1-\alpha)\overline{X}_{t-1} \qquad (5)$$

$$\overline{\sigma}_t^2 = \alpha\sigma_t^2 + (1-\alpha)\overline{\sigma}_{t-1}^2 \qquad (6)$$

where X_t and α_t^2 is the local mean and variance vector of the t th speech signal, respectively.

2.2. Data-driven temporal filtering: SETF and METF

Temporal filters are designed using the principal component analysis in the following procedure[3]. Let's assume that an ordered sequence of K -dimensional feature vectors are given as shown in Fig. 2, with time index n and feature index k . The k th time trajectory of x(n) is then the sequence of [x(1,k), x(2,k),...,x(n,k),...,x(N,k)], denoted as $y_k(n)$. Data-driven temporal filter is the L -sample FIR filter to be performed on the k th time trajectory $y_k(n)$.

Firstly, an L -sample rectangular window is shifted along the k th time trajectory to obtain the sequences of L -dimensional vectors $z_k(n)$. The L -dimensional vectors, $z_k(n)$, are viewed as the samples of a random vector z_k, and hence the mean vector and covariance matrix of z_k can be calculated from eqs. (7) and (8).

$$\mu_{z_k} = \frac{1}{N-L+1}\sum_{n=1}^{N-L+1} z_k(n) \qquad (7)$$

$$\Sigma_{z_k} = \frac{1}{N-L+1}\sum_{n=1}^{N-L+1}(z_k(n)-\mu_{z_k})(z_k(n)-\mu_{z_k})^T \qquad (8)$$

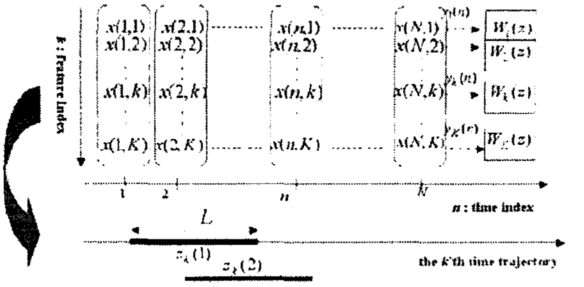
k : feature index
x(1,1) x(2,1) x(n,1) x(N,1)
x(1,2) x(2,2) x(n,2) x(N,2)
x(1,k) x(2,k) x(n,k) x(N,k)
x(1,K) x(2,K) x(n,K) x(N,K)
y1(n)
yk(n)
yK(n)
W1(z)
W2(z)
Wk(z)
WK(z)
1
2
n
N
n : time index
L
zk(1)
zk(2)
the k'th time trajectory

Fig. 2. Representation of the time trajectories of feature sequences

Following the procedure of principal component analysis in eq. (9), then, the components of the first eigenvector $\phi_{1,k}$ having the largest eigenvalue $\lambda_{1,k}$ of the covariance matrix Σ_{z_k} are taken as the coefficients of the L -sample filter. Here it is referred to as the single-eigenvector temporal filter(SETF).

$$\Sigma_{z_k}\phi_{i,k} = \lambda_{i,k}\phi_{i,k} \qquad (9)$$

Based on the principal component analysis theory, assume $\phi_{i,k}$ are L distinct normalized eigenvectors of the covariance matrix Σ_{z_k} with corresponding decreasing eigenvalues $\lambda_{i,k}, i = 1,2,3,...L$. The filter coefficients used in SETF are the components of the eigenvector $\phi_{1,k}$ corresponding to the largest eigenvalue, and the filter output can be viewed as the most expressive 1-dimensional representation of z_k. But, the filter output of other eigenvectors, which also carry some information of z_k , may be useful to improve the recognition performance. Under these assumptions, the filter coefficients with multiple eigenvectors can be obtained from eq. (10). It is referred to as a multi-eigenvector temporal filter(METF) in [3].

$$w_k = \frac{\overline{w_k}}{|w_k|} = \frac{1}{\sqrt{\sum_{i=1}^{M}\lambda_{i,k}^2}}\overline{w_k} \qquad (10)$$

$$\overline{w_k} = \sum_{i=1}^{M}\lambda_{i,k}\phi_{i,k} \qquad (11)$$

where w_k is the new coefficients of the L -sample filter for time trajectory k , the summation is over the first M eigenvectors with larger corresponding eigenvalues.

2.3. Selective temporal filtering: SVTF

According to Kanedera [5], in clean environment most of useful information is contained in the frequency range between 1 and 16 Hz of the modulation spectrum. And, in noisy environment, the components of the modulation spectrum below 2 Hz and above 10 Hz are less important for speech intelligibility. Generally, important ranges in modulation frequency are determined depending on specific task environment. Thus we examined frequency responses of other eigenvectors, and included them in a specific task, i.e., Korean connected-digit telephone speech recognition. While the SETF and METF take only one set of filter coefficients, we used M eigenvectors as different set of filter coefficients, respectively. We call it the selective temporal filtering (SVTF). Two types of SVTF are used: SVTF01 and SVTF02. SVTF01 employs M eigenvectors but SVTF02 employs only $M-1$ eigenvectors except the first one.

3. EXPERIMENTAL RESULTS

We used Korean connected-digit telephone speech DB released by Speech Information Technology & Industry Promotion Center (SITEC)[6] in our recognition experiments. This corpus consists of 4 kinds of telephone environment, i.e., cord, cordless, cellular and personal communication service (PCS). All the data were digitized with a sampling rate of 8 kHz, and 16bits/sample quantization. This DB consists of 58292 training data from 1800 speakers and 6468 test data from 200 speakers. All the training data were used to design the temporal filters. The order of the MFCC features was 13, and the length of the temporal filters, L, was empirically set to 7. The total number of eigenvectors, M, was set to 3 in our experiments.

Fig. 3 shows the frequency response of designed temporal filters for the 13 time trajectories. A horizontal axis shows the modulation frequency which presents the components between 0 Hz and 50 Hz in log scale, and a vertical axis shows magnitude in dB. Both SETF and METF show low-pass filtering characteristics with cutoff frequencies around 6 Hz. The first filter of SVTF is same with SETF, but the second filter and the third filter shows the characteristics of band-pass filter with pass-band between 4 Hz and 17 Hz, and between 10 Hz and 24 Hz, respectively.

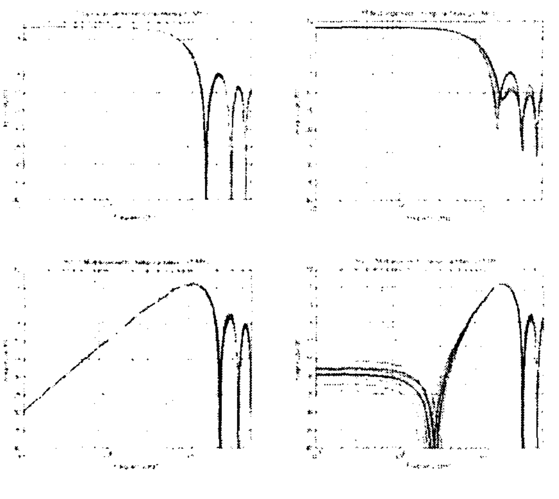

Fig. 3. The frequency response of temporal filters
(a) SETF, SVTF (1st filters), (b) METF, (c) SVTF (2nd filters), (d) SVTF (3rd filters)

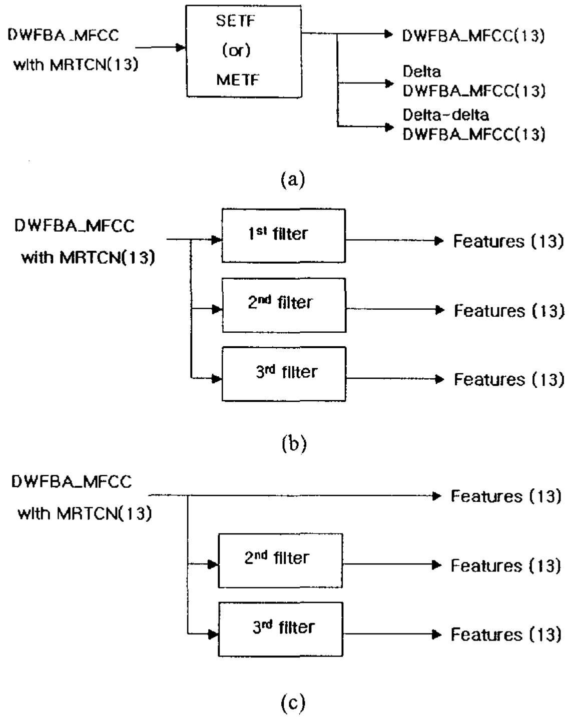

Fig. 4. Extraction of feature parameters using different temporal filters. (a) SETF, METF, (b) SVTF01, (c) SVTF02

The HTK system is used as a recognizer in the recognition experiments of connected-digit telephone speech[7]. Triphone models were used as basic units of recognition. Each unit was modeled with a five-states continuous mixture HMM having a simple left-to-right structure without skipping. 9 Gaussian mixtures with a diagonal covariance matrix were used for each state. Finite state network was also constructed for connected 4-digit strings. A 13th-order static cepstral coefficient vector was derived every 10ms with a hamming windowed analysis frame length of 20ms.

Fig. 4 shows how the feature parameters are constructed using the temporal filters for recognition. In SETF and METF, final 39-dimensional features were constructed from 13-dimensional DWFBA based MFCC with their delta and delta-delta features. In SVTF, however, delta and delta-delta features were replaced by the temporal filtered outputs of the original 13-dimensional features. The difference between SVTF01 and SVTF02 is that while the former uses 3 temporal filters the latter removes the first temporal filter to keep the original 13-dimensional features.

Table 1 shows the recognition results for different feature parameters with temporal filtering. The METF showed better performance than SETF, but the improvement was not as remarkable as shown in [3]. The proposed temporal filtering SVTF01 and SVTF02 showed slightly better performance than SETF and METF, respectively, but the improvement is not significant. Though temporal filtering with noisy speech was reported to achieve significant performance improvement in mismatch environment between training and testing database [2,3], it was not in our experiment with large size of Korean connected-digit telephone speech database.

However, proposed selective temporal filtering methods showed a little bit better performance than existing ones. Further studies are needed with noisy speech in additive noise environment.

Table 1 Recognition results according to types of feature parameters

Feature parameters	Connected-digit speech recognition rate(%)	Each word recognition rate(%)
DWFBA_MFCC	88.64	96.72
DWFBA_MFCC with MRTCN	91.48	97.55
SETF	90.74	97.30
METF	91.03	97.38
SVTF01	90.99	97.43
SVTF02	91.51	97.63

4. CONCLUSIONS

In this paper, data-driven temporal filters are investigated to improve the performance of a specific recognition task such as Korean connected-digit telephone speech. Three different temporal filtering methods were presented with recognition results for Korean connected digit telephone speech. Though temporal filtering with noisy speech was reported to achieve performance improvement in mismatch environment between training and testing database, it was not significant in telephone environment. But the proposed selective temporal filtering methods showed a little bit better performance than existing ones. Further studies are needed with noisy speech in additive noise environment.

References

[1] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, Spoken language processing, Prentice Hall PTR, 2001.

[2] J.W.Hung, L.S.Lee, "Data-driven temporal filters obtained via different optimization criteria evaluated on AURORA2 database", ICSLP 2002.

[3] N.W.Wang, J.W.Hung, "Data-driven temporal filters based on multi-eigenvectors for robust features in speech recognition," Proc. ICASSP 2003.

[4] · Wei Wen Hung, Hsiao Chuan Wang, "On the use of weighted filter bank analysis for the derivation of robust MFCCs," IEEE Signal Processing Letters, vol. 8, No. 3, March 2000.

[5] N.Kanedera, H. Hermansky, T.Arai, "On properties of modulation spectrum for robust automatic speech recognition," Proc. ICASSP'98, vol.2, pp.613-616, 1998.

[6] http://www.sitec.or.kr/index.asp

[7] Steve Young, The HTK Book (HTK Version 3.1), Cambridge, 2000.