Optimization of Temporal Filters in the Modulation Frequency Domain for Constructing Robust Features in Speech Recognition

Jeih-weih Hung

Department of Electrical Engineering, National Chi Nan University
Nantou Hsien, Taiwan
jwhung@ncnu.edu.tw

10.21437/Interspeech.2007-111

Abstract

In this paper, we derive new data-driven temporal filters that employ the statistics of the modulation spectra of the speech features. The new temporal filtering approaches are based on the constrained version of Principal Component Analysis (C-PCA) and Maximum Class Distance (C-MCD), respectively. It is shown that the proposed C-PCA and C-MCD temporal filters can effectively improve the speech recognition accuracy in various noise corrupted environments. In experiments conducted on Test Set A of the Aurora-2 noisy digits database, these new temporal filters, together with cepstral mean and variance normalization (CMVN), provides average relative error reduction rates of over 40% and 27%, when compared with the baseline MFCC processing and CMVN alone, respectively.

Index Terms: temporal filters, modulation frequency, principal component analysis, maximum class distance

1. Introduction

The performance of a speech recognition system is often degraded due to the mismatch between the training and testing environments. One category of approaches, including Cepstral Mean Subtraction (CMS), Cepstral Mean and Variance Normalization (CMVN) [2], and Relative Spectral (RASTA) [3], etc., attempt to filter the time trajectories of speech features so as to alleviate the harmful effects of various distortions. In contrast to these conventional temporal filtering techniques, where the filter form is fixed, the data-driven temporal filters can be tuned in order to be suitable for the speech feature characteristics or the environment, and they are often obtained according to a specific optimization criterion. Linear Discriminant Analysis (LDA) has been widely applied [4,5] as the optimization process to yield these temporal filters. Besides LDA, Principal Component Analysis (PCA) [5] and Minimum Classification Error (MCE) [5], have also been applied in the optimization process to obtain temporal filters. All of them have shown excellent performance in enhancing the robustness of speech features and improving the speech recognition accuracy.

A common characteristic of the above data-driven temporal filtering approaches is that they are often obtained according to the characteristics of the features in the temporal domain. However, we are often concerned with the frequency response of these filters and how they influence the modulation spectra of the original feature trajectories. From this point of view, it seems more natural to obtain the temporal filters directly based on how the features behave in the modulation frequency domain, rather than in the temporal domain.

Following this direction, in our previous work [6] we have developed the temporal filters with the characteristics of the modulation spectra of speech features. The temporal filters were derived according to the optimization technique of constrained Linear Discriminant Analysis (C-LDA). It was shown that the resulting temporal filters significantly improved the noise robustness of the MFCC features. In this paper, however, we develop the temporal filters in the modulation frequency domain with two other optimization techniques, constrained Principal Component Analysis (C-PCA) and constrained Maximum Class Distance (C-MCD). After the desired magnitude-squared response of the filter is found through C-PCA or C-MCD, the corresponding finite-impulse-response (FIR) filter coefficients can be obtained approximately via some FIR filter design algorithms like the Parks-McClellan algorithm [7]. The obtained filter coefficients are symmetric to ensure a linear-phase response.

With the Aurora-2 database as the experimental platform, we develop the proposed temporal filters and investigate their corresponding performance on the recognition task. It will be shown that the C-PCA filters are low-pass, while most of the C-MCD filters ban-pass filters. In most cases, C-MCD filters outperform C-PCA ones, although all of them significantly enhance the recognition performance of the original MFCC features. Experimental results also show that when the proposed filtering approaches are integrated with cepstral mean and variance normalization (CMVN), extra improvement in recognition accuracy can be obtained.

The remainder of this paper is organized into 5 sections. In section 2, the formulation used to derive the data-driven temporal filters in the modulation frequency domain is presented, and in section 3, the techniques of C-PCA and C-MCD to optimize temporal filters are described. The experimental environment is given in section 4. Section 5 contains the magnitude response of the obtained temporal filters and their corresponding recognition performance. Finally, concluding remarks are made in section 6.

2. Temporal filter design in the modulation frequency domain for time trajectories of feature parameters

Assume a finite-impulse-response (FIR) filter $h_m(n)$ with length L is applied to a specific time trajectory $\{x_m(n)\}$ of an ordered sequence of feature vectors $\{\mathbf{x}(n)\}$, where n is the time index and m is the feature index. Then the output samples $\{y_m(n)\}$ are the linear convolution of the time trajectory $\{x_m(n)\}$ with the impulse response $\{h_m(n)\}$ of the FIR filter. That is,

$$y_m(n) = \sum_{u=0}^{L-1} h_m(u)\, x_m(n-u)\,. \qquad (1)$$

Since the filter $h_m(n)$ is to be designed via its frequency response, the sequences, $\{x_m(n)\}$ and $\{h_m(n)\}$, which are assumed to be real sequences here, are transformed into the modulation frequency domain. At first, the filter input $\{x_m(n)\}$ is processed by a running window of length L, to obtain a set of L -length segments

$$\tilde{\mathbf{x}}_m(n) = \begin{bmatrix} x_m(n-L+1) & \cdots & x_m(n-1) & x_m(n) \end{bmatrix} . \quad (2)$$

Then, by padding $h_m(n)$ and each of $\tilde{\mathbf{x}}_m(n)$ with $K-L$ zeros, where $K \geq 2L$ and K is an even number, according to the Parseval's theorem [7] the right-hand side of equation (1) can be re-written as,

$$y_m(n) = \frac{1}{K}\sum_{k=0}^{K-1} H_m(k) X_m^*(n,k) = \sum_{k=0}^{K-1} H_m^*(k) X_m(n,k), \quad (3)$$

where $H_m(k)$ and $X_m(n,k)$ are the K-point DFTs of zero-padded versions of $h_m(n)$ and $\tilde{\mathbf{x}}_m(n)$, respectively. Note that there is no time index n in $H_m(k)$, because the temporal filter is assumed to be invariable with time. For the sake of compact notation, we omit the subscript m in the following discussions.

Next, the instantaneous energy of the temporal filter output at time n is

$$|y(n)|^2 = \left| \frac{1}{K}\sum_{k=0}^{K-1} H(k) X^*(n,k) \right|^2 , \quad (4)$$

and by assuming

$$\mathbf{H} = \begin{bmatrix} |H(0)|^2 & |H(1)|^2 & \cdots & |H(K/2)|^2 \end{bmatrix}^T , \quad (5)$$

and

$$\mathbf{X}(n) = \begin{bmatrix} |X(n,0)|^2 & |X(n,1)|^2 & \cdots & |X(n,K/2)|^2 \end{bmatrix}^T , \quad (6)$$

it can be shown that

$$|y(n)|^2 \leq \frac{2(K+2)}{K^2}\sum_{k=0}^{K/2} |H(k)|^2 |X(n,k)|^2 = \frac{2(K+2)}{K^2}\mathbf{H}^T\mathbf{X}(n) . \quad (7)$$

Thus $\mathbf{H}$ and $\mathbf{X}(n)$ are the vector forms of the magnitude-squared response for the filter and the squared magnitude spectrum for the filter input, respectively. If we define the instantaneous modulation spectral energy of the filter output as

$$\mathcal{E}_Y(n) = \sum_{k=0}^{K/2} |H(k)|^2 |X(n,k)|^2 = \mathbf{H}^T\mathbf{X}(n), \quad (8)$$

then equation (7) can be re-written as

$$|y(n)|^2 \leq \frac{2(K+2)}{K^2}\mathcal{E}_Y(n) . \quad (9)$$

Note that the range of the summation in equations (7) and (8) is within $[0, K/2]$ since both $\{|H(k)|^2\}$ and $\{|X(n,k)|^2\}$ are symmetric with respect to $k = K/2$. From equation (9), it is observed that the instantaneous energy of the filter output, $|y(n)|^2$, is bounded by a scaled version of $\mathcal{E}_Y(n)$. As a result, $\mathcal{E}_Y(n)$ can be approximately used to characterize the behavior of $|y(n)|^2$. In the subsequent discussions we will use $\mathcal{E}_Y(n)$ rather than $|y(n)|^2$ since it is more directly related to the magnitude-squared response for the filter, $\mathbf{H}$.

Let $\mathcal{E}_Y(n)$ and $\mathbf{X}(n)$ be the samples of a random variable $\mathcal{E}_Y$ and a random vector $\mathbf{X}$, respectively, then

$$\mathcal{E}_Y = \mathbf{H}^T\mathbf{X} . \quad (10)$$

Now the optimal $\mathbf{H}$ is found to maximize a specific objective function of $\mathcal{E}_Y$, which is related to the statistics of $\mathbf{X}$. The objective function is determined by the chosen optimization technique. In this paper, we apply two optimization techniques: constrained Principal Component Analysis (C-PCA) and constrained Maximum Class Distance (C-MCD), which is described in the next section. In order to obtain the statistics of $\mathbf{X}$, we first collect all the segments $\tilde{\mathbf{x}}_m(n)$, as in equation (2), for a specific time trajectory in the training database, and then calculate the squared magnitude spectrum, $\mathbf{X}(n)$, of each of them. These $\mathbf{X}(n)$ can be regarded as the samples of $\mathbf{X}$, with which the statistics of $\mathbf{X}$ are obtained.

Once the optimal $\mathbf{H}$ is approximately obtained and thus the magnitude-squared response of the temporal filter is found, the corresponding impulse response $\{h[n]\}$ can be approximately obtained by a number of filter design algorithms, such as the Parks-McClellan algorithm [7]. With the help of the filter design techniques, an FIR filter with symmetric impulse response can be designed.

As we know, an excellent property of a symmetric FIR filter is that it has a linear-phase response, which implies the filter does not distort the phase of the input signal components. Consequently, one major benefit to design temporal filters in the modulation frequency domain, as stated above, is that it possesses the optimal magnitude-squared response as well as the linear phase response simultaneously.

3. Temporal filter design in the modulation frequency domain

As stated in the previous section, the optimal squared magnitude response, $\mathbf{H}$, of the temporal filter is obtained via maximizing a specific objective function. Since each component of $\mathbf{H}$ is constrained to be real and nonnegative, it gives rise to a constrained optimization problem. That is,

$$\mathbf{H}^* = \arg\max_{\mathbf{H}} J(\mathbf{H}), \text{ subject to } \mathbf{H} \geq 0 , \quad (11)$$

where $J(\mathbf{H})$ denotes the objective function, and by $\mathbf{H} \geq 0$, we mean that every component of $\mathbf{H}$ is nonnegative. The objective function is determined by the chosen optimization technique, which will be described later. First, in order to deal with the nonnegative constraint for $\mathbf{H}$, we introduce an intermediate variable vector $\bar{\mathbf{H}} = \begin{bmatrix} \bar{H}_0 & \bar{H}_1 & \cdots & \bar{H}_{K/2} \end{bmatrix}^T$, where $\bar{H}_k$ can be any real number. The relationship between $\mathbf{H}$ and $\bar{\mathbf{H}}$ is

$$H_k = \left(e^{\bar{H}_k} \Big/ \sum_{m=0}^{K/2} e^{\bar{H}_m} \right)^{\frac{1}{P}}, \; k = 0,1,\cdots,K/2 , \quad (12)$$

where P is a positive integer. Based on equation (12), the nonnegative condition for $\mathbf{H}$ in equation (11) is satisfied. As a result, the constrained optimization problem in equation (12) becomes unconstrained through the unconstrained variable vector $\bar{\mathbf{H}}$. The next step is to find the optimal $\mathbf{H}$ that maximizes $J(\mathbf{H})$ through the intermediate vector $\bar{\mathbf{H}}$. Since the closed-form solution for $\bar{\mathbf{H}}$ is often not obtainable, we use the gradient-descent algorithm to iteratively update $\bar{\mathbf{H}}$, and then $\mathbf{H}$. By arbitrarily choosing an initial guess $\bar{\mathbf{H}}^{(0)}$, the iterative procedure is as follows:

$$\bar{\mathbf{H}}^{(\theta+1)} = \bar{\mathbf{H}}^{(\theta)} + \varepsilon \frac{\partial J}{\partial \bar{\mathbf{H}}}\Big|_{\bar{\mathbf{H}}=\bar{\mathbf{H}}^{(\theta)}}, \quad (13)$$

where ε is the step size, and

$$\frac{\partial J}{\partial \bar{\mathbf{H}}} = \frac{\partial \mathbf{H}}{\partial \bar{\mathbf{H}}} \frac{\partial J}{\partial \mathbf{H}}, \quad (14)$$

where the i, j-th term of the matrix $\frac{\partial \mathbf{H}}{\partial \bar{\mathbf{H}}}$ is

$$\left(\frac{\partial \mathbf{H}}{\partial \bar{\mathbf{H}}}\right)_{ij} = \frac{\partial H_j}{\partial \bar{H}_i} = \frac{1}{P}\left(e^{\bar{H}_j} \Big/ \sum_{m=0}^{K-1} e^{\bar{H}_m}\right)^{\frac{1}{P}-1} \times \left(\left(e^{\bar{H}_j}\delta_{ij}\sum_{m=0}^{K-1} e^{\bar{H}_m} - e^{\bar{H}_i+\bar{H}_j}\right) \Big/ \left(\sum_{m=0}^{K-1} e^{\bar{H}_m}\right)^2\right), 0 \le i, j \le K \quad (15)$$

and $\frac{\partial J}{\partial \mathbf{H}}$ is determined by the chosen objective function $J(\mathbf{H})$. The above gradient-descent procedure terminates when there is no substantial difference between $\bar{\mathbf{H}}^{(\theta)}$ and $\bar{\mathbf{H}}^{(\theta+1)}$. Then, equation (12) is used to obtain each component of the final magnitude-squared response $\mathbf{H}^*$. Now, we introduce two optimization techniques, constrained Principal Component Analysis (C-PCA), and constrained Maximum Class Distance (C-MCD), to determine the objective function.

3.1 Constrained Principal Component Analysis

The squared magnitude spectra, $\mathbf{X}(n)$, of all windowed segments, $\tilde{\mathbf{x}}(n)$ are treated as the samples of a single random vector $\mathbf{X}$. Thus, the covariance of $\mathbf{X}$ can be calculated and denoted by $\boldsymbol{\Sigma}$. By denoting σ^2 as the global variance of $\mathcal{E}_Y$, the objective function with C-PCA is

$$J_{PCA}(\mathbf{H}) = \sigma^2 = \mathbf{H}^T\boldsymbol{\Sigma}\mathbf{H}, \text{ subject to } \mathbf{H} \ge 0. \quad (16)$$

Note that if the components of $\mathbf{H}$ are not constrained to be nonnegative, then the optimal solution of $\mathbf{H}$ will be the eigenvector of the covariance matrix $\boldsymbol{\Sigma}$ that corresponds to the largest eigenvalue. Now the nearly optimal $\mathbf{H}$ is determined by the iteration process in equations (13), (14) and (15). As a result, the $\mathbf{H}$ obtained is optimal in the sense that it maximizes the variance of the filter output spectral energy among all possible magnitude-squared responses of the temporal filter.

3.2 Constrained Maximum Class Distance

At first, the squared magnitude spectrum, $\mathbf{X}(n)$, of each windowed segment, $\tilde{\mathbf{x}}(n)$ for a specific time trajectory in the training set is first labeled as one of the J classes or speech models, where J is the total number of classes or speech models. Then the mean, $\boldsymbol{\mu}^{(j)}$, and covariance matrix, $\boldsymbol{\Sigma}^{(j)}$, for those $\mathbf{X}(n)$ labeled as belonging to each class j, $\mathbf{X}^{(j)}(n)$, are calculated and denoted by $\boldsymbol{\mu}^{(j)}$ and $\boldsymbol{\Sigma}^{(j)}$, respectively. For simplicity, we assume that $\mathbf{X}^{(j)}$, the random vector representing the squared magnitude spectrum of class j, is multivariate Gaussian distributed with mean $\boldsymbol{\mu}^{(j)}$ and covariance matrix $\boldsymbol{\Sigma}^{(j)}$. Subsequently, with equation (10), the spectral energy of the filter output, $\mathcal{E}_Y$, for the j-th class, denoted as $\mathcal{E}_Y^{(j)}$, is a univariate Gaussian random variable with mean $\mathbf{H}^T\boldsymbol{\mu}^{(j)}$ and variance $\mathbf{H}^T\boldsymbol{\Sigma}^{(j)}\mathbf{H}$. Next, the distance between two different classes i and j of the filter output spectral energy $\mathcal{E}_Y$ is defined as

$$d_{ij} = \log\frac{\mathbf{H}^T\boldsymbol{\Sigma}^{(j)}\mathbf{H}^T}{\mathbf{H}^T\boldsymbol{\Sigma}^{(i)}\mathbf{H}^T} + \frac{\mathbf{H}^T\left(\boldsymbol{\mu}^{(i)}-\boldsymbol{\mu}^{(j)}\right)\left(\boldsymbol{\mu}^{(i)}-\boldsymbol{\mu}^{(j)}\right)^T\mathbf{H}}{\mathbf{H}^T\boldsymbol{\Sigma}^{(j)}\mathbf{H}^T} + \frac{\mathbf{H}^T\boldsymbol{\Sigma}^{(i)}\mathbf{H}^T}{\mathbf{H}^T\boldsymbol{\Sigma}^{(j)}\mathbf{H}^T} - 1, \quad (17)$$

which is the Kullback-Leibler divergence between two Gaussian probability distributions. Consequently, with C-MCD, the objective function to be maximized is the sum of all class distances. That is,

$$J_{MCD}(\mathbf{H}) = \sum_i \sum_{j\ne i} d_{ij} \text{ ,subject to } \mathbf{H} \ge 0. \quad (18)$$

Again, the nearly optimal $\mathbf{H}$ is determined by the iteration process in equations (13), (14) and (15). Therefore, the $\mathbf{H}$ obtained is optimal in the sense that it maximizes the overall class distances of the filter output spectral energy among all possible magnitude-squared responses of the temporal filter.

4. Experimental Setup

We perform recognition experiments on the AURORA-2 database. For the recognition environment, three sets of utterances artificially contaminated by different types of noise (subway, babble, etc.) and different SNR levels (from 20dB to -5dB) were prepared. Since the proposed approach only involves the front-end feature extraction, all the procedures for training and testing are identical to the reference experiments stated in the Aurora-2 documentations.

For the clean training database, each of the 8440 strings is first converted into a sequence of 13-dimensional Mel-frequency cepstral coefficients (12 MFCCs + logarithmic energy coefficient). For the parameters in the procedures of the C-PCA and C-MCD approaches, the filter length, L, in equation (1), the DFT size, K, in equation (3), and the exponent, P, in equation (12) are set to be 101, 256, and 4, respectively. For C-MCD optimization process, the 8440 training feature strings are classified into 13 digit classes, i.e., oh, zero to nine, short pause and silence. For C-PCA, however, no such classification is needed. In addition, the Parks-McClellan algorithm is used to obtain the final filter coefficients. These filters are respectively applied on the time trajectories of the MFCC feature vectors for the clean training database. The resulting 13-dimensional new features, plus their first and second order derivatives, are then the components of the finally used 39-dim feature vectors. With these feature vectors, the HMMs for each digit are trained.

For the testing phase, the digit strings in three test sets were also first converted to MFCCs, processed by the C-PCA or C-MCD temporal filters obtained with the training data, and then augmented with their first and second order derivatives to form various sets of feature vectors.

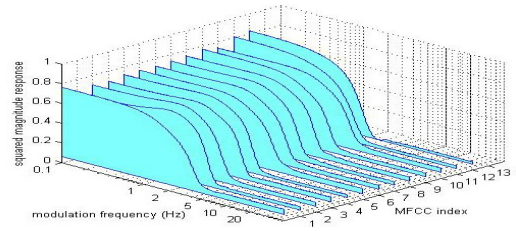
5. Experimental Results

The magnitude-squared responses over the modulation frequencies of the C-PCA and C-MCD temporal filters for 13 MFCC coefficients are shown in Figures 1(a) and 1(b), respectively. From the two figures, we have some observations and discussions as given below.

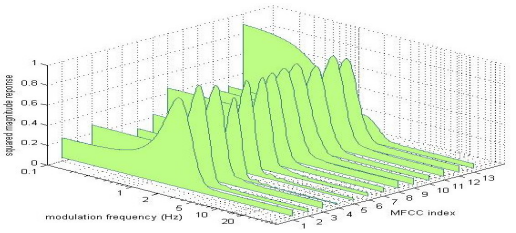
1. Most of the C-MCD temporal filters are band-pass, and the modulation frequency components between 1 Hz and 5 Hz are relatively emphasized. In other words, the pass-band is roughly between 1 Hz and 5 Hz.
2. The C-PCA temporal filters are mostly low-pass, and the pass-band width is about 5 Hz. In the very low modulation

frequency band, C-PCA filters are quite different from C-MCD ones. That is, C-PCA temporal filters do not attenuate the near-DC component.

3. Although the C-MCD temporal filters have a band-pass characteristic, they do not completely eliminate the very low modulation frequency components below 1 Hz. As we know, the very useful CMS and CMVN are high-pass filters, while the well-known RASTA is a band-pass filter. This implies that eliminating the very low modulation frequency components of the signals should be very helpful, and it leads to the fact that such processing, in addition to the proposed filtering approaches, may be of further use, as will be discussed later.



(a)



(b)

Figure 1. The squared magnitude responses (vs. modulation frequency) of the (a) 13 C-PCA temporal filters (b) 13 C-MCD temporal filters

Table 1 shows the recognition accuracy rates obtained using various filtering approaches on the plain MFCC features under different noisy conditions. From this table, some observations can be made:

1. The two data-independent filtering approaches, RASTA and CMVN, can improve the MFCC's performance in most cases in Test Sets A and B. However, neither of them brings significant improvements for Test Set C.
2. It is shown that, similar to C-LDA [6], the new proposed C-PCA and C-MCD approximately preserve the high recognition accuracy of MFCC features at the clean matched condition. Under the mismatched noisy conditions, C-PCA and C-MCD provide significant recognition improvements for almost all the cases in Test Sets A and B.
3. C-MCD and C-LDA are apparently better C-PCA, which is possibly because that C-PCA does not attenuate the near-DC component that may correspond to the time-invariant additive noise or channel distortions. Even so, C-PCA helps to filter out the high modulation frequency spikes caused by additive noise, and thus improves the robustness of MFCC.
4. When the proposed temporal filters are integrated with CMVN, the recognition performance can be further improved for all three test sets, and the performance difference among C-LDA, C-PCA and C-MCD becomes insignificant. For example, compared with CMVN alone, C-PCA and C-MCD achieve 31.13% and 29.51% averaged word-error-rate (WER) reduction for Test Set A, respectively. In particular, for Test Set C, C-PCA and C-MCD become effective in promoting the recognition performance by integrating CMVN.

Test	System/SNR	clean	average (0~20dB)	relative WER reduction
Test Set A	MFCC baseline	98.91	61.13	
	RASTA	98.72	67.17	15.54
	CMVN	98.98	70.38	23.80
	C-LDA	98.80	70.87	25.06
	C-PCA	98.66	65.39	10.96
	C-MCD	98.18	69.24	20.86
	CMVN+C-LDA	98.85	78.46	44.58
	CMVN+C-PCA	98.51	79.60	47.52
	CMVN+C-MCD	98.28	79.12	46.28
Test Set B	MFCC baseline	98.94	55.57	
	RASTA	98.72	71.46	35.76
	CMVN	98.98	71.10	34.95
	C-LDA	98.80	69.52	31.40
	C-PCA	98.66	59.45	8.73
	C-MCD	98.18	71.02	34.77
	CMVN+C-LDA	98.85	78.98	52.69
	CMVN+C-PCA	98.70	82.46	60.52
	CMVN+C-MCD	98.12	81.27	57.84
Test Set C	MFCC baseline	99.00	66.68	
	RASTA	98.69	66.38	-0.90
	CMVN	99.12	66.80	1.26
	C-LDA	98.87	71.99	15.94
	C-PCA	98.77	67.22	1.62
	C-MCD	98.36	65.63	-3.15
	CMVN+C-LDA	98.83	75.84	27.49
	CMVN+C-PCA	98.66	78.20	34.57
	CMVN+C-MCD	98.25	78.06	34.15

Table 1. Word recognition accuracies (%) and relative word-error-rate (WER) reduction (%) for various temporal filtering approaches as compared to the MFCC baseline and CMVN

6. References

[1] Atal, B.S. "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", J. Acoust. Soc. Am. 55 (6), 1974

[2] S. Tibrewala and H. Hermansky, "Multiband and adaptation approaches to robust speech recognition," Eurospeech 1997.

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

[4] C. Avendano, S. Vuuren and H. Hermansky, "Data-based filter design for RASTA-like channel normalization in ASR", ICSLP 1996

[5] Jeih-weih Hung and Lin-shan Lee, "Optimization of temporal filters for constructing robust features in speech recognition", IEEE Trans on Audio, Speech and Language Processing, 2006

[6] Jeih-weih Hung, "Optimization of temporal filters in the modulation frequency domain via constrained linear discriminant analysis (C-LDA) for constructing robust features in speech recognition", ICASSP 2007

[7] Sanjit K. Mitra, "Digital Signal Processing, a computer-Based Approach", 2nd version, McGraw-Hill