OPTIMIZATION OF TEMPORAL FILTERS IN THE MODULATION FREQUENCY DOMAIN VIA CONSTRAINED LINEAR DISCRIMINANT ANALYSIS (C-LDA) FOR CONSTRUCTING ROBUST FEATURES IN SPEECH RECOGNITION

Jeih-weih Hung
Dept of Electrical Engineering, National Chi Nan University
Nantou Hsien, Taiwan, Republic of China
jwhung@ncnu.edu.tw

Abstract

Data-driven temporal filtering approaches based on a specific optimization criterion have been shown to be capable of enhancing the discrimination and robustness of speech features in speech recognition. The filters in these approaches are often obtained with the statistics of the features in the temporal domain. 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 approach is based on the constrained version of Linear Discriminant Analysis (C-LDA). It is shown that the proposed C-LDA 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 47% and 30%, when compared with the baseline MFCC processing and CMVN alone, respectively.

Index Terms – temporal filter, modulation frequency, noise robustness, linear discriminant analysis

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 to minimize this mismatch is focused on trying to find a robust feature representation for speech signals so that it is less sensitive to various corrupted acoustic conditions. Cepstral Mean Subtraction (CMS) [1], Cepstral Mean and Variance Normalization (CMVN) [2], and Relative Spectral (RASTA) [3] techniques are typical examples of this category of approaches, in which the time trajectories of speech features are filtered so as to alleviate the harmful effects of various distortions.

In contrast to the above conventional temporal filtering techniques, where the filter form is fixed and is somewhat independent of the applied speech features, 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 the 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 the statistics of the time trajectory of speech features are first extracted and then employed to get the filters. In other words, these temporal filters are 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 this paper we propose that the optimal temporal filters are developed with the characteristics of the modulation spectra of speech features. The modulation spectral power for the output of the filter is viewed as a random variable. Then based on the criterion of constrained Linear Discriminant Analysis (C-LDA), the magnitude-squared response of the temporal filter is obtained in order to optimize the corresponding objective function of this random variable. This criterion is called "constrained" because in its objective function, the parameters, which correspond to magnitude-squared response of the filter, are constrained to be nonnegative. As a result, the optimal solution may not be that of the unconstrained one. After the desired magnitude-squared response of the filter is found through C-LDA, the corresponding finite-impulse-response (FIR) filter coefficients can be obtained approximately via some FIR filter design algorithms like the Parks-McClellan algorithm [6] and the least-squares method [6]. The obtained filter coefficients are symmetric to ensure a linear-phase response.

Besides the proposed approach, a series of experiments on the Aurora-2 database are conducted to obtain the proposed temporal filters and their corresponding performance on the recognition task. It is shown that most of the C-LDA filters are band-pass filters, and they significantly enhance the recognition accuracy of the original MFCC features. Experimental results also show that when the proposed filtering approaches are integrated with some feature normalization process, such as 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 criterion of C-LDA to optimize temporal filters is described. The experimental environment is given in section 4. Section 5 contains the magnitude-squared 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 we wish to design the filter $h_m(n)$ via its frequency response here, all of these sequences, $\{y_m(n)\}$, $\{x_m(n)\}$ and

$\{h_m(n)\}$ 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,

$$\tilde{\mathbf{x}}_m(n) = \left[x_m(n-L+1) \quad \cdots \quad x_m(n-1) \quad x_m(n)\right]. \quad (2)$$

Then, by padding $h_m(n)$ and each of $\tilde{\mathbf{x}}_m(n)$ with $K-L$ zeros, where $K \geq 2L-1$, the linear convolution of $h_m(n)$ and $\tilde{\mathbf{x}}_m(n)$ in equation (3) is equivalent to the circular convolution of the zero-padded versions of them [6]. Therefore, after performing K-point DFT on both sides of equation (1), we obtain

$$Y_m(n,k) = H_m(k) X_m(n,k), \; k = 0,1,\cdots,K-1, \quad (3)$$

where $Y_m(n,k)$, $H_m(k)$ and $X_m(n,k)$ are the K-point DFTs of $y_m(n)$ and 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 modulation spectral power of the filter output is defined as

$$P_Y(n) \triangleq \sum_{k=0}^{\lfloor K/2 \rfloor} \left|Y(n,k)\right|^2 = \sum_{k=0}^{\lfloor K/2 \rfloor} \left|H(k)\right|^2 \left|X(n,k)\right|^2 = \mathbf{H}^T \mathbf{X}(n), \quad (4)$$

where $\mathbf{H} = [\left|H(0)\right|^2 \quad \left|H(1)\right|^2 \quad \cdots \quad \left|H\left(\lfloor K/2 \rfloor\right)\right|^2]^T$, and

$\mathbf{X}(n) = [\left|X(n,0)\right|^2 \quad \left|X(n,1)\right|^2 \quad \cdots \quad \left|X\left(n, \lfloor K/2 \rfloor\right)\right|^2]^T$,

where $\lfloor K/2 \rfloor$ is the largest integer smaller than $K/2$. Thus $\mathbf{H}$ and $\mathbf{X}(n)$ are the vector form of the magnitude-squared response for the filter, and the squared magnitude spectrum for the filter input, respectively. Here, $P_Y(n)$ and $\mathbf{X}(n)$ are viewed as the samples of a random variable P_Y, and a random vector $\mathbf{X}$, respectively, and it can be written as

$$P_Y = \mathbf{H}^T \mathbf{X}. \quad (5)$$

Then the optimal vector $\mathbf{H}$ is found to maximize a specific objective function of P_Y, and is related to the statistics of $\mathbf{X}$. The objective function is determined by the chosen optimization criterion, constrained Linear Discriminant Analysis (C-LDA), which will be 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 $\mathbf{X}(n)$ are regarded as the samples of $\mathbf{X}$, with which the statistics of $\mathbf{X}$ can be obtained.

Furthermore, it is worthwhile to note that choosing P_Y as the parameter for optimizing the temporal filter possesses many advantages. First, it is much easier to evaluate the performance of a scalar variable P_Y than that of a sequence of variables, $\{\left|Y(k)\right|^2\}$ or $\{Y(k)\}$. Secondly, the square sum P_Y is "physically meaningful" since it represents the power of the filter output. Finally, P_Y is directly related to $\{\left|H(k)\right|^2\}$ rather than $\{H(k)\}$. As we know, $\{\left|H(k)\right|^2\}$ are always real and nonnegative while $\{H(k)\}$ are often complex.

Once the optimal $\mathbf{H}$ is approximately obtained and thus the magnitude-squared response of the temporal filter is found, the corresponding impulse response $\{h[n], 0 \leq n \leq L-1\}$ can be approximately obtained by a number of filter design algorithms, such as the Parks-McClellan algorithm and the least-squares method [6]. With the help of these filter design techniques, an FIR filter with symmetric impulse response can be designed, i.e.,

$$h[n] = h[L-n]. \quad (6)$$

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 in the frequency band of interest. 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 Filters Design in the Modulation Frequency Domain by Constrained Linear Discriminative Analysis

Linear Discriminative Analysis (LDA) has been widely applied in pattern recognition. Its goal is to find the most "discriminative" representation of the data. In this approach, a function representing the discriminative nature among different classes within the data is maximized by finding an optimal linear transform to be applied to the data. Here we use it to derive the optimal squared magnitude responses of the temporal filters. That is, the temporal filter is obtained in the modulation frequency domain to maximize the Fisher discriminating function of the power variable P_Y in equation (5), which is the ratio of the between-class variance and the within-class variance of P_Y. However, different from the conventional LDA, the linear transform parameters, $\mathbf{H}$, which correspond to the squared magnitude response of the temporal filters are constrained to be real and nonnegative. Thus the algorithm to derive $\mathbf{H}$ stated here is called constrained LDA (C-LDA), and is described below.

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. This labeling process can be performed by means of the time alignment with pre-trained models. Then the mean $\boldsymbol{\mu}^{(j)}$ and covariance matrix $\boldsymbol{\Sigma}^{(j)}$ for those $\mathbf{X}(n)$ labeled as belonging to class j are calculated,

$$\boldsymbol{\mu}^{(j)} = \frac{1}{N_j} \sum_{n=1}^{N_j} \mathbf{X}^{(j)}(n), \quad (7)$$

$$\boldsymbol{\Sigma}^{(j)} = \frac{1}{N_j} \sum_{n=1}^{N_j} \left(\mathbf{X}^{(j)}(n) - \boldsymbol{\mu}^{(j)}\right)\left(\mathbf{X}^{(j)}(n) - \boldsymbol{\mu}^{(j)}\right)^T, \quad (8)$$

where $\mathbf{X}^{(j)}(n)$ denotes those $\mathbf{X}(n)$ labeled as belonging to the j-th class, and N_j is the total number of such $\mathbf{X}^{(j)}(n)$. With these parameters, the between-class matrix S_B and within-class matrix S_W of $\mathbf{X}$ can be defined as

$$\mathbf{S}_B = \sum_{j=1}^{J} N_j \left(\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}\right)^T, \quad (9)$$

$$\mathbf{S}_W = \sum_{j=1}^{J} N_j \boldsymbol{\Sigma}^{(j)}, \quad (10)$$

where $\boldsymbol{\mu} = \sum_{j=1}^{J} N_j \boldsymbol{\mu}^{(j)} \Big/ \sum_{j=1}^{J} N_j$.

Therefore, by denoting σ_B^2 and σ_W^2 as the between-class and within-class variances of P_Y, respectively, the objective function to be maximized with the constrained LDA criterion is

$$\mathbf{H}^* = \arg\max_{\mathbf{H}} J_{LDA}(\mathbf{H}), \; J_{LDA}(\mathbf{H}) = \frac{\sigma_B^2}{\sigma_W^2} = \frac{\mathbf{H}^T \mathbf{S}_B \mathbf{H}}{\mathbf{H}^T \mathbf{S}_W \mathbf{H}}, \quad (11)$$

subject to every component of $\mathbf{H}$ is nonnegative. That is,

$$H_k \geq 0, \; k = 0,1,\cdots,K-1, \quad (12)$$

where H_k denotes the k –th component of $\mathbf{H}$. Note that if $\{H_k\}$ are not constrained to be nonnegative, then the optimal solution $\mathbf{H}^*$ will be the eigenvector of the matrix $\mathbf{S}_W^{-1}\mathbf{S}_B$ that corresponds to the largest eigenvalue. In order to solve this constrained optimization problem, we introduce an intermediate variable vector $\bar{\mathbf{H}} = \left[\bar{H}_0 \quad \bar{H}_1 \quad \cdots \quad \bar{H}_{K-1}\right]^T$,

where $\bar{H}_k, k = 0,1,\cdots,K-1,$ can be any real number, and the relationship between $\mathbf{H}$ and $\bar{\mathbf{H}}$ is

$$H_k = \left(\exp\left(\bar{H}_k\right)\Big/\sum_{m=0}^{K-1}\exp\left(\bar{H}_m\right)\right)^{\frac{1}{P}},\ k = 0,1,\cdots,K-1\,, \qquad (13)$$

where P is a positive integer. With equation (13), apparently the nonnegative condition for H_k in equation (12) is satisfied, and

$$\sum_{m=0}^{K-1} H_k^P = 1\,. \qquad (14)$$

Equation (14) is used to keep the parameters $\{H_k\}$ from unbounded increasing. Therefore, the optimization function in equation (11) with the constrained $\mathbf{H}$ becomes unconstrained through the unconstrained $\bar{\mathbf{H}}$. The next step is to find the optimal $\mathbf{H}$ that maximize $J_{LDA}(\mathbf{H})$ through the intermediate vector $\bar{\mathbf{H}}$. We use the gradient-descent algorithm to iteratively update the value of $\bar{\mathbf{H}}$, and then $\mathbf{H}$. By arbitrarily choosing an initial guess $\bar{\mathbf{H}}^{(0)}$, the iterative procedure is as follows,

$$\bar{\mathbf{H}}^{(\theta+1)} = \bar{\mathbf{H}}^{(\theta)} + \varepsilon \left.\frac{\partial J_{LDA}}{\partial \bar{\mathbf{H}}}\right|_{\bar{\mathbf{H}}=\bar{\mathbf{H}}^{(\theta)}}, \qquad (15)$$

where ε is the step size, and

$$\frac{\partial J_{LDA}}{\partial \bar{\mathbf{H}}} = \frac{\partial \mathbf{H}}{\partial \bar{\mathbf{H}}}\frac{\partial J_{LDA}}{\partial \mathbf{H}}\,, \qquad (16)$$

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

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

$$0 \le i,j \le K-1 \qquad (17)$$

and

$$\frac{\partial J_{LDA}}{\partial \mathbf{H}} = \frac{2\left(\mathbf{H}^T\mathbf{S}_W\mathbf{H}\right)\mathbf{S}_B\mathbf{H} - 2\left(\mathbf{H}^T\mathbf{S}_B\mathbf{H}\right)\mathbf{S}_W\mathbf{H}}{\left(\mathbf{H}^T\mathbf{S}_W\mathbf{H}\right)^2}\,. \qquad (18)$$

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 (13) is used to obtain the final magnitude-squared response $\mathbf{H}^*$, which is optimal in the sense that it maximizes the Fisher discriminating function of the filter output power, P_Y , among all possible squared magnitude 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, with an interval of 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 documentation [7].

For the clean training database, each of the 8440 strings was first converted into a sequence of 13-dimensional Mel-frequency cepstral coefficients (12 MFCCs + logarithmic energy feature). Then these training feature strings are segmented into 13 digit classes, i.e., oh, zero to nine, short pause and silence. For each time trajectory, the C-LDA temporal filters were constructed using these 13 classes. For the parameters in the procedures of the C-LDA approach, the filter length, L , in equation (2), the DFT size, K , in equation (3), and the exponent, P , in equation (13) are set to be 101, 256, and 4, respectively. In addition, the Parks-McClellan algorithm was used to obtain the final filter coefficients. These filters were 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-dimensional feature vectors. With these new 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-LDA 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

5.1 Frequency Response of the C-LDA temporal filters

The magnitude-squared responses over the modulation frequencies of the C-LDA temporal filters for 13 MFCC coefficients are shown in Figure 1. From this figure, we have some observations and discussions as given below.

1. Most of the temporal filters have similar magnitude-squared responses and show band-pass characteristics, even though they are for different MFCC components.
2. For most of the temporal filters, 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.
3. Although the C-LDA temporal filters have band-pass characteristics, they do not completely eliminate the very low modulation frequency components below 1 Hz. As we know, the very useful CMS and its derivative, 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 should be very helpful, and it leads to the fact that such processing, in addition to the proposed filtering approach, may be of further use, as will be discussed later.

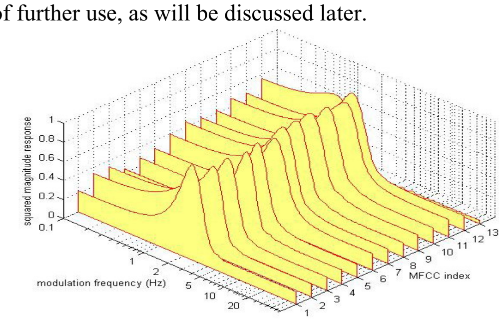

Figure 1. The squared magnitude responses (vs. modulation frequency) of the 13 C-LDA temporal filters

5.2 Recognition Accuracy

Table 1 shows the recognition accuracy rates obtained using various filtering approaches on the plain MFCC features under different noisy conditions. The results are for different signal-to-noise ratio conditions but are averaged over all the noise types in the same set. From this table, some observations can be made:

1. The two data-independent filtering approaches, RASTA and CMVN, are capable of improving the MFCC's performance in most cases of Test Sets A and B. CMVN performs better than RASTA for Test Set A, while the situation is converse for Test

Test	System	clean	20dB	15dB	10dB	5dB	0dB	-5dB	average (0~20dB)	relative WER reduction
Test Set A	MFCC baseline	98.91	94.99	86.93	67.28	39.36	17.07	8.40	61.13	
	RASTA	98.72	96.09	91.24	76.80	48.27	23.46	11.09	67.17	15.54
	CMVN	98.98	95.98	91.66	80.48	57.40	26.40	10.96	70.38	23.80
	T_LDA	98.63	94.49	83.67	60.43	30.60	11.11	6.56	56.06	-13.04
	MF_C-LDA	98.80	96.67	92.94	81.34	57.17	26.24	8.92	70.87	25.06
	CMVN+T_LDA	97.32	93.95	89.51	81.72	65.94	40.41	18.35	74.30	33.88
	CMVN+MF_C-LDA	98.51	96.48	93.42	86.47	72.87	48.76	22.68	79.46	47.16
Test Set B	MFCC baseline	98.94	92.35	80.79	58.06	32.04	14.63	7.92	55.57	
	RASTA	98.72	96.42	92.62	81.68	56.85	29.71	12.51	71.46	35.76
	CMVN	98.98	96.41	92.15	81.78	58.69	26.47	10.98	71.10	34.95
	T_LDA	98.63	93.72	87.06	66.18	36.54	17.85	9.70	60.27	10.58
	MF_C-LDA	98.80	96.20	91.74	80.28	55.81	23.57	5.08	69.52	31.40
	CMVN+T_LDA	97.32	93.86	90.42	85.13	66.40	40.68	16.88	75.30	44.41
	CMVN+MF_C-LDA	98.70	96.92	94.44	88.49	74.08	48.62	20.87	80.51	56.13
Test Set C	MFCC baseline	99.00	94.83	88.66	75.23	50.85	23.83	11.4	66.68	
	RASTA	98.69	95.70	90.41	73.79	47.43	24.58	13.41	66.38	-0.90
	CMVN	99.12	95.51	88.71	74.21	51.25	24.30	10.49	66.80	0.36
	T_LDA	98.71	89.57	78.18	58.28	36.56	17.28	9.52	55.97	-32.14
	MF_C-LDA	98.87	96.00	91.70	80.66	59.34	32.28	15.52	71.99	15.94
	CMVN+T_LDA	97.48	93.65	89.10	79.00	62.52	39.75	18.09	72.80	18.37
	CMVN+MF_C-LDA	98.66	96.10	92.43	84.67	69.79	48.04	23.04	78.21	34.60

Table 1. Word recognition accuracies (%) and relative word-error-rate (WER) reduction (%) for various temporal filtering approaches as compared to the MFCC baseline for different SNR values but averaged over all noise types in the same test set

Set B. However, neither of them brings significant improvements for Test Set C.

2. In the table, the new proposed C-LDA temporal filtering approach in the *m*odulation *f*requency domain is denoted as MF_C-LDA, while the previously proposed LDA temporal filtering approach in the *t*emporal domain is denoted as T_LDA. It is found that in some cases T_LDA degrades the performance of MFCC. One possible reason of this degradation is that most of the temporal filters obtained by T_LDA are low-pass filters as in [5], which amplify the near-DC components that correspond to the slow-varying distortions. However, with the new proposed MF_C-LDA, significant improvements can be achieved in almost every condition. Compared with the baseline MFCC processing, it results in 9.74%, 13.95% and 5.31% absolute accuracy improvement for Test Sets A, B, and C, respectively. In particular, MF_C-LDA performs very close to RASTA and CMVN for Test Sets A and B, and it is significantly outstanding for Test Set C. As a result, we may roughly conclude that it seems better to design the temporal filters in the modulation frequency domain than in the temporal domain.
3. As mentioned previously, CMVN and RASTA remove the very low frequency components, and are shown to be very helpful in dealing with additive noise. Thus, here we integrate CMVN and each of the two data-driven filtering approaches, T_LDA and MF_C-LDA, which results are shown in the last two rows for each test set in Table 1. First, it is shown that T_LDA becomes effective by integrating CMVN. For example, it gives better recognition accuracy than the baseline MFCC processing and CMVN alone by 19.73% and 4.20%, respectively, for Test Set B. Secondly, integrating CMVN and MF_C-LDA also brings further improvements when compared with the results of each individual approach, and this integration performs the best among all approaches listed here for almost all cases in three test sets. It brings about 47%, 56% and 35% relative word error rate reduction for Test Sets A, B, and C, respectively, when compared with the baseline MFCC processing.

6. Concluding Remarks

In this paper, we have proposed a new temporal filtering approach based on the statistics of the modulation spectra of speech features. The criterion of LDA is modified with constraints to obtain the magnitude-squared responses of the temporal filters. Significant improvements in recognition accuracy, as compared with the plain MFCC, have demonstrated the effectiveness of the proposed approach. Furthermore, it can be integrated with the well-known Cepstral Mean and Variance Normalization (CMVN) to obtain extra improvements.

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

[7] H.-G Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," ISCA ITRW ASR 2000, Paris, France, 2000