Improved Modulation Spectrum Normalization Techniques for Robust Speech Recognition

Chi-an Pan, Chieh-cheng Wang and Jeih-weih Hung
Dept of Electrical Engineering, National Chi Nan University
Nantou Hsien, Taiwan
s95323544@ncnu.edu.tw, s95323553@ncnu.edu.tw, jwhung@ncnu.edu.tw

Abstract

The modulation spectra of speech features are often distorted due to environmental interferences. In order to reduce this distortion, in this paper we propose several approaches to normalize the power spectral density (PSD) of the feature stream to a reference function. These approaches include least-squares temporal filtering (LSTF), least-squares spectrum fitting (LSSF) and magnitude spectrum interpolation (MSI). It is shown that all the proposed approaches can effectively improve the speech recognition accuracy in various noise corrupted environments. In experiments conducted on the Aurora-2 noisy digits database with a complex back-end, these new approaches provide an average relative error reduction rate of over 40% when compared with the baseline MFCC processing.

Index Terms : robust speech recognition, feature normalization, modulation spectrum

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. Relative Spectral (RASTA) [1], Cepstral Mean Subtraction (CMS) [2], Cepstral Mean and Variance Normalization (CMVN) [3], integration of CMVN and Auto-regressive moving average filtering (MVA) [4] 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 temporal filtering techniques, where the filter form is fixed and is somewhat independent of the applied speech features, the temporal filters can be tuned in order to be suitable for the speech feature characteristics or the environment based on some certain criteria, such as Linear Discriminant Analysis (LDA) [5], Principal Component Analysis (PCA) [6] and Minimum Classification Error (MCE) [6]. These data-driven temporal filters have shown excellent performance in enhancing the robustness of speech features and improving the speech recognition accuracy.

Recently, Haizhou Li et al proposed a new temporal filter design scheme [7], called Temporal Structure Normalization (TSN), which aims to normalize the power spectral density (PSD) of speech feature streams in an utterance to a reference PSD. The reference PSD is obtained by averaging the PSDs of all the clean speech utterances. Different from the temporal filters previously mentioned, here the obtained filter varies utterance by utterance so that the filter can be dynamically adapted to the acoustic environment for each utterance. The TSN filters are reported to be very effective in improving the recognition accuracy when the speech features are first processed by CMVN or MVA.

In the TSN filter design, once the magnitude response of the filter is obtained, the Inverse Discrete Fourier Transform (IDFT) is performed to obtain the initial filter coefficients. Next, only the L central coefficients are extracted, where L is a predefined filter length. These L coefficients are smoothed by a Hanning window, and they are then scaled so that the sum of these coefficients is normalized to one to ensure that the DC gain of the resulting filter is unity.

The algorithm to obtain these TSN filters as mentioned above is very efficient in implementation. However, according to our observations, it has some points which may be further improved. First, the truncated and smoothed IDFT-processed coefficients are just a rough approximation of the optimal ones that well match the desired frequency response. Second, the process to normalize the sum of the filter coefficients to one keeps the DC component of the filtered feature stream unchanged with respect to that of the original one. This implies the PSD of the filtered feature stream approaches a scaled version of the reference PSD, and the scaling factor varies utterance by utterance. Therefore, this process seems to be inconsistent with the original objective that the resulting PSD should approach the reference PSD. Furthermore, since the additive and/or channel noise may cause a scaling effect on the PSD of the original features, to confine the DC gain of the filter to unity seems to fail to deal with this scaling effect. Finally, in [7] the TSN filters are designed only for MVN- or MVA-processed MFCC features. It is wondered if they are also effective for original MFCC features.

Motivated by the above observations for the TSN filter design, in this paper we propose several approaches which attempt to normalize the PSD of the original MFCC feature streams to a reference pattern. In the first approach, least-squares filtering (LSF), given the magnitude response of the filter as in the TSN processes, we design the temporal filter so that it is optimal in a least-squared sense. The feature stream is then filtered by the obtained temporal filter. Next, in the second approach, least-squares spectrum fitting (LSSF), the new feature stream is obtained so that its modulation spectrum has the best approximation to a target spectrum in the least-squares sense, in which the target spectrum is created by the reference PSD and the modulation spectrum of the original feature stream. In the third approach, magnitude spectrum interpolation (MSI), the magnitude part of the target spectrum is obtained by linearly interpolating a reference magnitude spectrum, while the phase part directly comes from the modulation spectrum of the original feature stream. Then the new feature stream is obtained by Inverse Discrete Fourier Transform (IDCT) of the target spectrum. Experimental results conducted on the Aurora-2 database show that the proposed three approaches are capable of improving the recognition accuracy of the original MFCC features under a wide range of noise-corrupted environments. Furthermore, it is shown that they outperform TSN significantly.

The remainder of this paper is organized into 4 sections. In section 2, the proposed three approaches for normalizing the PSD of the feature stream are described. Section 3 describes the experimental environment. In section 4, we present the

experimental results and compare the proposed approaches with some other techniques. Finally, a brief concluding remark is given in section 5.

2. Modulation Spectrum Normalization Approaches for MFCC Features

Consider using the Mel-scaled filter-bank cepstral coefficients (MFCC) for speech recognition. Let $x_m[n]$ be the m^{th} cepstral coefficient of the n^{th} frame of an utterance. As a result, we have M feature streams,

$$\{x_m[n] | 0 \le n \le N-1\},\ 0 \le m \le M-1, \quad (1)$$

where M is the number of cepstral coefficients within a frame and N is the number of frames of the utterance. Assume that these features are noise-corrupted, and it is expected to obtain a set of new feature streams $\{y_m[n]\}$ in which the noise effect is alleviated. In our case here, each feature stream $\{x_m[n]\}$ is processed so that the resulting $\{y_m[n]\}$ has a two-sided power spectral density (PSD) close to a reference pattern,

$$\left\{ \left| Z_m(\omega_k) \right| \middle| \omega_k = k\frac{2\pi}{2P}, 0 \le k \le 2P-1 \right\}, \quad (2)$$

where ω_k is the normalized frequency $(0 \le \omega_k < 2\pi)$, and $2P$ is the number of frequency bins. Note that the absolute sign $|\bullet|$ is used on $Z_m(\omega_k)$ to emphasize that each item in the reference pattern is real and nonnegative. As a result, the reference magnitude spectrum of $\{y_m[n]\}$ becomes

$$\left| Y_m(\omega_k) \right| = \left| X_m(\omega_k) \right| \sqrt{\left| Z_m(\omega_k) \right| / P_{XX}(\omega_k)}, \quad (3)$$

where $X_m(\omega_k)$ and $P_{XX}(\omega_k)$ are the 2*P*-point Discrete-Fourier Transform (DFT) and the two-sided PSD of $\{x_m[n]\}$, respectively.

Following [7], here the reference PSD $\{|Z_m(\omega_k)|\}$ is obtained by averaging the PSDs of the m^{th} feature streams for all clean utterances in the training database. For the sake of compact notation, we omit the subscript m in the later discussions.

In the following, we propose three approaches to determine the normalized feature stream $\{y[n]\}$. In the first approach, least-squares temporal filtering (LSTF), a temporal filter is designed and performed on the original feature stream $\{x[n]\}$. That is, we process the features in the *temporal* domain in order to make them well matched to a reference pattern in the *modulation frequency* domain. However, in the next two approaches, least-squares spectrum fitting (LSSF) and magnitude spectrum interpolation (MSI), we perform this pattern matching directly in the *modulation frequency* domain.

2.1 Least-Squares Temporal Filtering (LSTF)

In this approach, a FIR filter with *L*-point impulse response, $\{h[n] | 0 \le n \le L-1\}$, is designed for the feature stream $\{x[n]\}$. First, the two-sided PSD of $\{x[n]\}$ is calculated and denoted as

$$\left\{ P_{XX}(\omega_k) \middle| \omega_k = k\frac{2\pi}{2P}, 0 \le k \le 2P-1 \right\}. \quad (4)$$

Note that $P_{XX}(\omega_k)$ has the same length of the reference pattern $|Z(\omega_k)|$ in eq. (2). To equalize the PSD $P_{XX}(\omega_k)$ to $|Z(\omega_k)|$, the required magnitude response of the filter is

$$\left| H(\omega_k) \right| = \sqrt{\left| Z(\omega_k) \right| / P_{XX}(\omega_k)},\ \omega_k = k\frac{2\pi}{2P}, 0 \le k \le 2P-1, \quad (5)$$

Then, given the magnitude response $\{|H(\omega_k)|\}$, the filter coefficients $\{h[n] | 0 \le n \le L-1\}$ are determined by the least-squares method [8]. The obtained filter coefficients are symmetric to ensure a linear-phase response, and it has the best approximation to the desired frequency response $\{|H(\omega_k)|\}$ in the least-squares sense. Finally, the new feature stream $\{y[n]\}$ is obtained by convoluting $\{x[n]\}$ and $\{h[n]\}$.

Note that in this approach, most of the procedures are identical to those in TSN, except that here the filter coefficients are obtained by the least-squares method and the sum of the filter coefficients is *not* normalized to one.

2.2 Least-Squares Spectrum Fitting (LSSF)

In this approach, the new feature stream $\{y[n]\}$ is obtained so that the squared error between the 2*P*-point DFT of $\{y[n]\}$ and a reference spectrum is minimized. The 2*P*-point reference spectrum is constructed by

$$\widehat{Y}(\omega_k) = \left| Y(\omega_k) \right| \exp\left(j\theta_X(\omega_k) \right),\quad 0 \le k \le 2P-1, \quad (6)$$

where $|Y(\omega_k)|$ is defined in eq. (3), and $\{\theta_X(\omega_k)\}$ are the phase parts of the 2*P*-point DFT of $\{x[n]\}$. That is, we attempt to keep the phase spectrum of $\{x[n]\}$ while update the PSD of $\{x[n]\}$ so that it is close to the reference PSD $\{|Z(\omega_k)|\}$ in eq. (2). Since in general $N \ne 2P$, we cannot obtain the *N*-point $\{y[n]\}$ directly from the IDFT of $\{\widehat{Y}(\omega_k)\}$. Here, by imposing a constraint that $2P \ge N$, we obtain the new feature stream $\{y[n]\}$ so that the corresponding spectrum fits the reference spectrum in eq. (6) in the least-squares sense:

$$\{y[n]\} = \min_{\{\widehat{y}_m[n] | 0 \le n \le N-1\}} \sum_{k'=0}^{2P-1} \left| \sum_{n=0}^{N-1} \widehat{y}[n] e^{-j\frac{2\pi nk'}{2P}} - \widehat{Y}(\omega_{k'}) \right|^2 \quad (7)$$

The above equation can be re-written in vector-matrix form as

$$\mathbf{y} = \min_{\widehat{\mathbf{y}}} \left\| W\widehat{\mathbf{y}} - \widehat{\mathbf{Y}} \right\|^2, \quad (8)$$

where

$$\mathbf{y} = \begin{bmatrix} y[0] & y[1] & \cdots & y[N-1] \end{bmatrix}^T,$$

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}[0] & \widehat{y}[1] & \cdots & \widehat{y}[N-1] \end{bmatrix}^T$$

$$\widehat{\mathbf{Y}} = \begin{bmatrix} \widehat{Y}(\omega_0) & \widehat{Y}(\omega_1) & \cdots & \widehat{Y}(\omega_{2P-1}) \end{bmatrix}^T,$$

and W is an $2P \times N$ matrix whose (m,n)-th term is

$$W_{mn} = \exp\left(-j\frac{2\pi mn}{2P} \right) \quad (9)$$

Note that each component of the vector $\widehat{\mathbf{y}}$ is real, and thus eq. (8) can be re-written as

$$\mathbf{y} = \min_{\tilde{\mathbf{y}}} \left\| \left(W_R \tilde{\mathbf{y}} - \hat{\mathbf{Y}}_R \right) + j \left(W_I \tilde{\mathbf{y}} - \hat{\mathbf{Y}}_I \right) \right\|^2$$
$$= \min_{\tilde{\mathbf{y}}} \left(\left\| W_R \tilde{\mathbf{y}} - \hat{\mathbf{Y}}_R \right\|^2 + \left\| W_I \tilde{\mathbf{y}} - \hat{\mathbf{Y}}_I \right\|^2 \right) \quad (10)$$

where W_R and $\hat{\mathbf{Y}}_R$ are the real parts of W and $\hat{\mathbf{Y}}$, respectively, and W_I and $\hat{\mathbf{Y}}_I$ are the imaginary parts of W and $\hat{\mathbf{Y}}$, respectively. As a result, solving $\mathbf{y}$ in eq. (10) becomes a classical least-squares (LS) problem, and the solution is

$$\mathbf{y} = \left(W_R^T W_R + W_I^T W_I \right)^{-1} \left(W_R^T \hat{\mathbf{Y}}_R + W_I^T \hat{\mathbf{Y}}_I \right). \quad (11)$$

As a result, the vector $\mathbf{y}$ in eq. (11) constitutes the final feature stream $\{y[n]\}$.

2.3 Magnitude Spectrum Interpolation (MSI)

In this approach, the new features stream $\{y[n]\}$ is obtained as the N-point IDFT of an N-point reference spectrum, denoted as

$$\left\{ \tilde{Y}\left(\tilde{\omega}_{k'}\right) \middle| \tilde{\omega}_{k'} = \frac{2\pi k'}{N}, 0 \le k' < N \right\}. \quad (12)$$

The phase part of the $\left\{\tilde{Y}\left(\tilde{\omega}_{k'}\right)\right\}$ can be directly taken from the N-point DFT of $\{x[n]\}$. However, since in general $N \neq 2P$, the $2P$-point $\left\{\left|Y\left(\omega_k\right)\right|\right\}$ in eq. (3) cannot be directly used as the magnitude part of $\left\{\tilde{Y}\left(\tilde{\omega}_{k'}\right)\right\}$. Here we obtain an approximate estimate of $\left\{\left|\tilde{Y}\left(\tilde{\omega}_{k'}\right)\right|\right\}$ by linearly interpolating $\left\{\left|Y\left(\omega_k\right)\right|\right\}$. Note that since $\left\{\left|\tilde{Y}\left(\tilde{\omega}_{k'}\right)\right|\right\}$ must be symmetric with its central tap (except for $\tilde{Y}(0)$), the $\left(\left\lfloor N/2 \right\rfloor + 1\right)$ interpolating points are first obtained as $\left\{\left|\tilde{Y}\left(\tilde{\omega}_{k'}\right)\right| 0 \le k' \le \left\lfloor 2/N \right\rfloor\right\}$ from the first $P+1$ points of $\left\{\left|Y\left(\omega_k\right)\right|\right\}$, and then the remaining $\left(N - \left\lfloor N/2 \right\rfloor - 1\right)$ points of $\left\{\left|\tilde{Y}\left(\tilde{\omega}_{k'}\right)\right|\right\}$ are obtained to meet the symmetry requirement. That is,

$$\left|\tilde{Y}\left(\tilde{\omega}_{k'}\right)\right| = \left|\tilde{Y}\left(\tilde{\omega}_{N-k'}\right)\right|, \quad \left\lfloor N/2 \right\rfloor + 1 \le k' \le N-1. \quad (13)$$

As a result, the N-point reference spectrum is constructed by

$$\tilde{Y}\left(\tilde{\omega}_{k'}\right) = \left|\tilde{Y}\left(\tilde{\omega}_{k'}\right)\right| \exp\left(j\theta_X\left(\tilde{\omega}_{k'}\right)\right), \quad 0 \le k' \le N-1, \quad (14)$$

where $\left\{\theta_X\left(\tilde{\omega}_{k'}\right)\right\}$ are the phase part of the N-point DFT of $\{x[n]\}$. The new feature stream $\{y[n]\}$ is then obtained as the N-point IDFT of $\left\{\tilde{Y}\left(\tilde{\omega}_{k'}\right)\right\}$:

$$y[n] = \frac{1}{N} \sum_{k'=0}^{N-1} \tilde{Y}\left(\tilde{\omega}_{k'}\right) e^{j\frac{2\pi n k'}{N}}, \quad 0 \le n \le N-1. \quad (15)$$

3. Experimental Setup

We perform recognition experiments on the AURORA-2 database [9]. 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) are prepared. Each utterance in the clean training set and three noise-corrupted testing sets is first converted into a sequence of 13-dimensional mel-frequency cepstral coefficients (MFCC, c0-c12). The PSD of each feature stream for each of the 8440 utterances in the clean training set is estimated. The reference PSD in equation (1) is obtained by averaging all the 8440 PSDs for each feature stream. Following the specifications in [7], the PSD is estimated using the Yule-Walker method with the autoregressive model order being 15. The number of frequency bins, $2P$ in eq. (1), is set to 256 in most approaches, except that in the LSSF approach, $2P$ is set to 1024 so that it can be always greater than N, the number of frames in an utterance. The filter length L for the filters designed by LSTF and TSN is set to 21. The 13-dimensional feature sequences in the clean training set and the three noise-corrupted testing sets are individually processed by various post-processing approaches mentioned previously. The resulting 13 new features, plus their first and second order derivatives, are then the components of the finally used 39-dimensional feature vector. With these new feature vectors in the clean training set, the HMMs for each digit and silence are trained following the Microsoft complex back-end training scripts [10]. Each digit HMM has 16 states and 20 Gaussian mixtures per state.

3. Experimental Results

Fig. 1 (a)-(e) shows the normalized PSD of the first MFCC feature $c1$ of an utterance after various PSD normalization schemes for three SNR levels, clean, 10dB and 0dB. TSN-1 in Fig. 1(a) is the original version of TSN in [7], in which the sum of the filter coefficients is normalized to one. TSN-2 in Fig. 1(b) follows all the procedures in TSN-1 except that the last step to normalize the filter coefficients is skipped. Fig. 1(a) shows that only the normalized PSD for the clean case approaches the reference PSD,

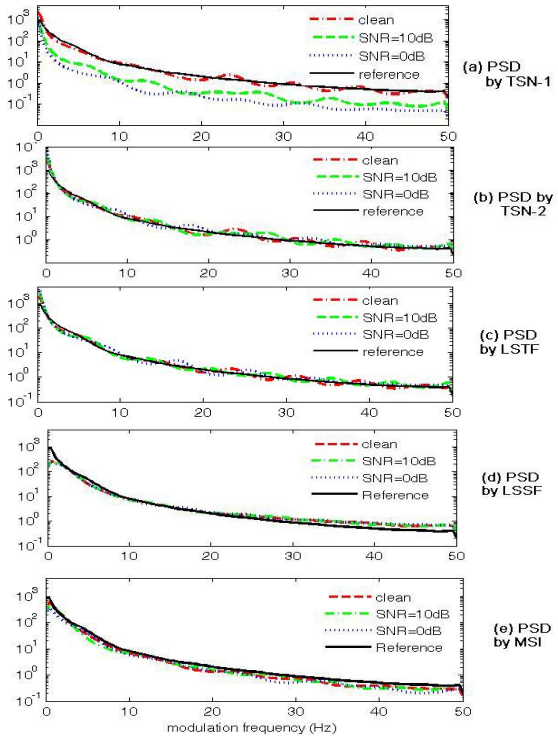

Figure 1. The normalized $c1$ PSD curves of an utterance ("FAK_6654599A.08" in the Aurora-2 database) after various PSD normalization schemes for three SNR levels, clean, 10dB and 0dB.

while the other two for the SNR=10dB and 0dB cases, respectively, significantly deviate from the reference PSD. This phenomenon supports our comments previously that the original TSN fails to deal with the scaling effect on the PSD caused by noise for plain MFCC features. However, in Fig. 1(b), all the PSDs for different SNR cases are very close to the reference PSD. Similar results can be also observed in Fig. 1(c)-(e), which shows that all the proposed PSD normalization approaches, LSTF, LSSF, and MSI, are capable of alleviating the PSD deviation caused by noise.

The recognition results for various approaches performed on the plain MFCC features are summarized in Table 1. For the purpose of comparison, the results of CMVN and RASTA are also shown in this table. From this table, some observations can be made as follows:

1. Although TSN-1 reveals very good improvement for MVN- and MVA- processed MFCC features as reported in [7], it does not improve the plain MFCC very significantly. The possible reason is that the scaling effect on the PSD by noise has been removed or alleviated in the MVN- and MVA- processed MFCC features, while it still exists in the plain MFCC, and TSN-1 does not process it properly as shown in Fig.1 (a).
2. By skipping the filter coefficient normalizing procedure in TSN-1, TSN-2 outperforms TSN-1 in most cases, and provides an absolute improvement rate of 8.09%.
3. The proposed LSTF provides better results than TSN-2, which implies the filter designed by LSTF, which possess a closer approximation to the desired magnitude response in (5), acts better than that designed by TSN-2.
4. The proposed LSSF and MSI give very outstanding improvements for MFCC. MSI performs better than RASTA, TSN-1, TSN-2 and LSTF, and it acts as well as CMVN. LSSF achieves the highest accuracy among all approaches and it provides an absolute improvement rate of 15.35%. This implies performing the PSD normalization directly in the modulation frequency domain brings about better results.

Method	Set A	Set B	Set C	Avg.	AR	RR
Baseline	72.46	68.31	78.32	71.97	-	-
RASTA	76.05	79.58	76.48	77.55	5.58	19.89
CMVN	85.07	85.57	85.63	85.38	13.41	47.85
TSN-1	73.64	70.48	77.18	73.08	1.11	3.97
TSN-2	80.06	82.24	75.68	**80.06**	8.09	28.86
LSTF	83.04	84.47	81.35	**83.02**	11.30	40.32
LSSF	86.72	87.88	87.43	**87.33**	15.35	54.78
MSI	84.65	86.63	85.70	**85.65**	13.68	48.81

Table 1. Accuracy (%) achieved by various approaches for Aurora-2 task averaged across the SNR between 0 and 20dB, where TSN-1 is TSN *with* filter coefficient normalization, and TSN-2 is TSN *without* filter coefficient normalization. AR (%) and RR (%) are the absolute and relative error rate reductions over the baseline.

Since CMVN performs very well as shown in Table 1, we attempt to integrate our proposed LSTF, LSSF, and MSI with CMVN to see if further improvements can be achieved. That is, the MFCC features are first normalized by CMVN, and then processed by either of the three approaches. Note that here the reference PSDs used in the three approaches are created by the CMVN-processed features.

Table 2 shows the recognition results for the proposed approaches performed on CMVN-processed features. The results of MVA, which combines ARMA filtering and CMVN, are also listed in the table for comparison. From this table, it is shown that integrating either of our proposed approaches with CMVN brings very excellent recognition performance. The three approaches enhance CMVN by improving the accuracy of about 4.5%. Also, they perform better than the ARMA filtering in all cases. Finally, MSI becomes the best of the three approaches, although the performance differences among them are relatively insignificant.

Method	Set A	Set B	Set C	Avg.	AR	RR
CMVN	85.07	85.57	85.63	85.38	-	-
CMVN+LSTF	89.65	90.49	89.11	**89.88**	4.50	30.76
CMVN+LSSF	89.12	90.17	89.16	**89.55**	4.17	28.50
CMVN+MSI	89.83	90.80	89.66	**90.18**	4.80	32.85
CMVN+ARMA	88.25	88.82	88.61	88.55	3.17	21.67

Table 2. Accuracy (%) achieved by various approaches for Aurora-2 task averaged across the SNR between 0 and 20dB. AR (%) and RR (%) are the absolute and relative error rate reductions over CMVN.

5. Concluding Remarks

In this paper, we follow the concept of temporal structure normalization (TSN) in [7] and propose several new modulation spectrum normalization techniques for speech features. Significant improvements in recognition accuracy have demonstrated the effectiveness of these proposed approaches. In addition, experimental results show that further improvement can be achieved when these newly proposed approaches are integrated with the technique of cepstral mean and variance normalization (CMVN).

References

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

[2] Atal, B.S. "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," Journal of the Acoustical Society of America vol. 55, 1304-1312, 1974

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

[4] C-P. Chen and J. Bilmes, "MVA processing of speech features", IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 1, pp. 257-270, January 2007

[5] S. van Vuuren and H. Hermansky, "Data-driven design of RASTA-like filters," Eurospeech 1997

[6] J-W. Hung and L-S. Lee, "Optimization of temporal filters for constructing robust features in speech recognition", IEEE Trans. on Audio, Speech and Language Processing, Vol 14, 2006

[7] X. Xiao, E-S. Chng, and Haizhou Li, "Temporal structure normalization of speech feature for robust speech recognition", IEEE Signal Processing Letters, vol. 14, 2007

[8] Sanjit K. Mitra, "Digital Signal Processing, a computer-based approach", 3rd version, McGraw-Hill

[9] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", ISCA ITRW ASR2000

[10] J. Droppo, L. Deng, and A. Acero, "Evaluation of SPLICE on the AURORA 2 and 3 tasks", ICSLP 2002