ISCA Archive
http://www.isca-speech.org/archive

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

ROBUST FEATURES FOR SPEECH RECOGNITION USING MINIMUM VARIANCE DISTORTIONLESS RESPONSE (MVDR) SPECTRUM ESTIMATION AND FEATURE NORMALIZATION TECHNIQUES

Yi Chen, Lin-Shan Lee

Graduate Institute of Communication Engineering,
National Taiwan University, Taipei
chenyi@speech.ee.ntu.edu.tw

ABSTRACT

In this paper, feature extraction methods based on frequency-warped Minimum Variance Distortionless Response (MVDR) spectrum estimation are analyzed and tested. The effectiveness of the conventional FFT-based Mel-Frequency Cepstrum Coefficients (MFCCs) and the MVDR-based features are carefully compared. Two normalization techniques are further applied to improve the robustness of the features: the widely used cepstral normalization (CN), and newly proposed progressive histogram equalization (PHEQ). Extensive experiments with respect to the AURORA2 database were performed. The results indicated that both the MVDR-based features and the normalization processes are very helpful.

1. INTRODUCTION

The blueprint for the various applications of automatic speech recognition (ASR) technologies in the future has been extensively laid out and its realization has been highly anticipated by many people [1]. But the recognition accuracy always plays the most dominating role when the real-world applications are considered. It is well known that the recognition accuracy of ASR systems is very often seriously degraded by the mismatch between the acoustic conditions for the training and testing environments and, hence, the robustness for ASR technologies with respect to the changing acoustic environment has always been a key issue in real applications. In this paper we discuss the integration of the Minimum Variance Distortionless Response (MVDR)-based speech feature extraction with the feature normalization techniques which can improve the recognition accuracy and the robustness of the features.

Mel-Frequency Cepstral Coefficients (MFCC) derived from the FFT spectrum have shown consistently satisfactory performance over a wide variety of application tasks, though the optimality for a specific application task is not guaranteed. The MVDR spectral estimator proposed by Capon [2] was shown by Lacoss [3] to provide a minimum variance unbiased estimate of the signal spectral components. MVDR spectrum estimation was previously proposed by Murthi and Rao [4][5] as a spectral envelope estimation technique, and has been applied to speech recognition by Dharanipragada and Rao [6][7]. To improve the perceptual resolution of the MVDR spectral estimate further, it was then proposed to estimate the all-pole model of a speech signal segment on a warped short-term frequency scale, such as the Bark scale or the Mel scale [8][9]. The estimated spectrum gives higher resolution to the low frequency regions and thus mimics the frequency resolution of the human auditory system [8][9]. In this paper, we apply the MVDR spectrum estimation in the feature extraction to alleviate the effect of additive noise, and then process the MVDR-based features with two feature normalization techniques: the widely used cepstral normalization (CN) and the newly proposed progressive histogram equalization (PHEQ)[15][16], to reduce the environmental mismatch and to obtain improved recognition accuracy. Experimental results with respect to the AURORA2 database verified that better performance in the adverse circumstances can be actually achieved.

The remainder of this paper consists of 4 sections. MVDR-based feature extraction, frequency-warped signal processing, and the feature normalization methods (i.e., CN and PHEQ) are very briefly summarized in section 2, the experimental conditions are described in section 3, and extensive experimental results are presented in sections 4. Section 5 gives the concluding remarks.

2. MVDR-BASED FEATURE EXTRACTION AND NORMALIZATION

Here we briefly summarize the algorithms of the feature extraction and normalization schemes discussed in this paper.

2.1. MVDR Spectrum Estimation

Consider an FIR filter with impulse response coefficients $\{h(k), 0 \le k \le M\}$, where M is referred to as the order of this filter, and of the spectral estimator based on this filter. If we pass a speech frame $\{x(n), 0 \le n \le N-1\}$ through this filter, the output y(n) is given by

$$y(n) = \sum_{k=0}^{M} h^*(k)x(n-k) \equiv \mathbf{h}^H\mathbf{x}(n),$$

where

$$\mathbf{x}(n) = [\, x(n)\ x(n\text{-}1)\ \ldots\ x(n\text{-}M)\,] \text{ and } \mathbf{h}(n) = [\, h(0)\ h(1)\ \ldots\ h(M)\,] \quad \textbf{(1)}$$

are respectively the vector representing the input signal and the filter coefficients, and "H" indicates the Hermitian of a matrix. The filter can be designed in such a way that at a specified normalized frequency f_i $(-0.5 \le f_i < 0.5)$ the frequency response is unity, that is,

$$\mathbf{h}^H\mathbf{e}^*(f_i) = 1, \quad \textbf{(2)}$$

where $\mathbf{e}(f_i) \equiv [\,1\ e^{j2\pi f_i}\ e^{j4\pi f_i}\ \ldots\ e^{jM(2\pi f_i)}\,]^T$ Assume that the input signal x(n) is zero-mean, E[x(n)] = 0, the variance of the output signal y(n) is then

This work is partially supported by the SiS Education Foundation.

$$\sigma_y^2 = E[|y(n)|^2] = E[\mathbf{h}^H \mathbf{x}(n)\mathbf{x}^H(n)\mathbf{h}] = \mathbf{h}^H \mathbf{R}_x(n)\mathbf{h}, \quad (3)$$

where $R_x(n)$ is the (M+1)-by-(M+1) sample autocorrelation matrix of the signal segment x(n).

By minimizing the output signal variance in equation (3) subject to the distortionless response constraint in equation (2), we obtain an FIR filter with which the signal component at f_i is undistorted and the other frequencies are attenuated as much as possible, or

$$\min_{\mathbf{h}}[\mathbf{h}^H \mathbf{R}_x(\mathbf{n})\mathbf{h}] \quad \textit{subject to} \quad \mathbf{h}^H \mathbf{e}^*(f_i) = 1, \quad (4)$$

and the solution to this optimization problem was shown to be [3]

$$\mathbf{h}_i(n) = \frac{\mathbf{R}_x^{-1}(n)\,\mathbf{e}^*(f_i)}{\mathbf{e}^T(f_i)\mathbf{R}_x^{-1}(n)\,\mathbf{e}^*(f_i)}, \quad (5)$$

where $h_i(n)$ is the filter that gives unity response for the frequency component f_i, and dependent on both the input data and the specified frequency f_i. The MVDR spectrum estimated at a specified normalized frequency f_i can actually be evaluated directly [12][13],

$$P_{MVDR}(f_i) = \frac{1}{\mathbf{e}^T(f_i)\mathbf{R}_x^{-1}(n)\,\mathbf{e}^*(f_i)}. \quad (6)$$

Thus it is not necessary to find $h_i(n)$ explicitly in equation (5) to estimate the signal spectrum at any frequency of interest. We can also move the normalized frequency f_i in the range of –0.5 to 0.5 freely to sample the signal spectrum. In this way, we have "distortionless response" at the frequency of interest f_i, while minimum leakage power components from all other frequencies.

2.2. Frequency-Warped Signal Processing

In frequency-warped signal processing techniques, the warping process can be achieved by replacing the unit-delay element by a first-order all-pass filter with a transfer function [10][11][8][9]

$$D(z) = (z^{-1} - \lambda)/(1 - \lambda z^{-1}), \quad (7)$$

whose phase response is

$$\tilde{\omega} \equiv \arg[D(e^{-j\omega})] = \omega + 2\cdot\arctan[\lambda\cdot\sin(\omega)/(1 - \lambda\cdot\cos(\omega))], \quad (8)$$

which determines the frequency mapping relations, where λ is a design parameter, ω is the angular frequency ($-\pi \le \omega \equiv 2\pi f_i < \pi$), and $\tilde{\omega}$ is the warped angular frequency. The frequency mapping functions for different values of the parameter λ are plotted in Fig 1. The conversion from the conventional linear frequency f ($0 \le f < 4k, Hz$) to the well-known Mel-frequency f_{mel} is given by

$$f_{mel} = 2595\cdot\log_{10}(1 + f/700). \quad (9)$$

For 8-kHz sampling frequency, the optimal value of λ for the frequency mapping to approximate the Mel-scale warping, in terms of total squared-error sampled at frequencies spaced 1-Hz apart from 0 to 4k Hz, is about 0.362436 (λ^*). However, as will be shown later on in the experiments below, other values of λ may give better recognition performance.

2.3. Warped MVDR Feature Extraction

The warped-MVDR spectrum estimation performed in this research is based on the warped correlation terms (in the warped autocorrelation matrix) obtained using the system shown in Fig 2 [10]. The warped autocorrelation matrix $\mathbf{R}_X$ thus obtained is Toeplitz and practically invertible, and then the warped linear prediction (LP) coefficients, $a_{M,k}$'s, can be

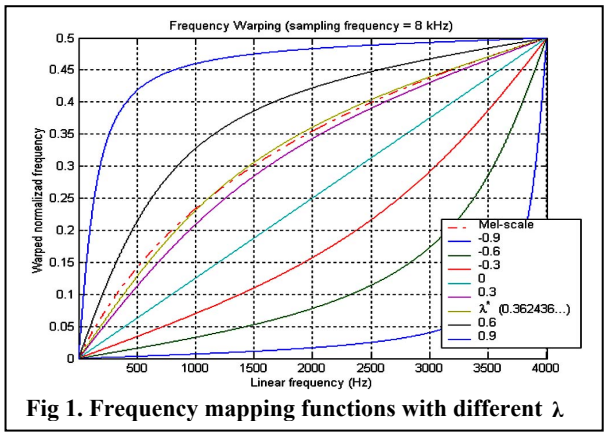

Fig 1. Frequency mapping functions with different λ

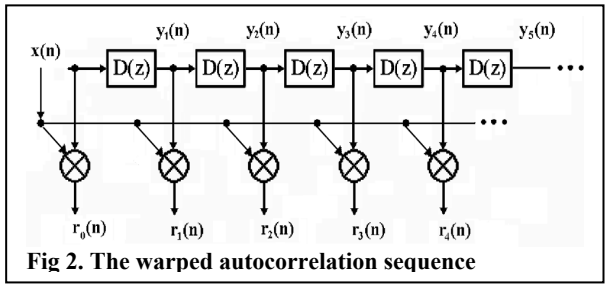

Fig 2. The warped autocorrelation sequence

solved using the Levinson-Durbin algorithm.

The warped MVDR spectrum can be computed directly as

$$P_{MVDRwp}(\omega) \equiv \frac{1}{\sum_{k=-M}^{M} \mu(k) e^{-j2\pi fk}} \quad (10)$$

where the intermediate parameter sequence $\mu(k)$ is defined as

$$\mu(k) = \begin{cases} \frac{1}{P_M}\sum_{i=0}^{M-k}(M+1-k-2i)a_{M,i}a_{M,i+k}, & k = 0, 1, ..., M, \\ \mu^*(-k), & k = -M, -M+1, ..., -1, \end{cases} \quad (11)$$

and $a_{M,0}(=1), a_{M,1}, ..., a_{M,M}$ are the coefficients of the M-th order prediction error filter; P_M is the expected prediction-error power of the filter [14].

We can then apply the Mel-filtering and DCT steps to calculate the feature parameters just as for MFCCs. The frequencies we chose to estimate the MVDR power spectrum are uniformly distributed on the Mel-frequency axis, but converted back to the linear frequency, over the usable frequency range (64 ~ 4k Hz). The total number of spectrum samples is determined by two parameters: (1) the number of channels in the filterbank, and (2) the number of samples in each channel. The usable signal frequency range and the above two parameters determine the actual locations of these sample points. Because the MVDR spectrum samples are taken uniformly in the Mel-frequency domain, the filters used in the filterbank to obtain the feature parameters are designed to be half-overlapped triangular filters with equal width, as shown in Fig 3. It is in fact a "Mel-scaled filterbank". We take 10 samples per channel in the 23-channel filterbank, so there are 120 sample points in total which are comparable in number to the 256-point FFT spectrum samples (126 points in the usable frequency range).

2.4. Feature Normalization Techniques

The cepstral normalization (CN) can be expressed as

$$\hat{\mathbf{c}}_t = \boldsymbol{\Sigma}_c^{-1/2}(\mathbf{c}_t - \boldsymbol{\mu}_c)$$

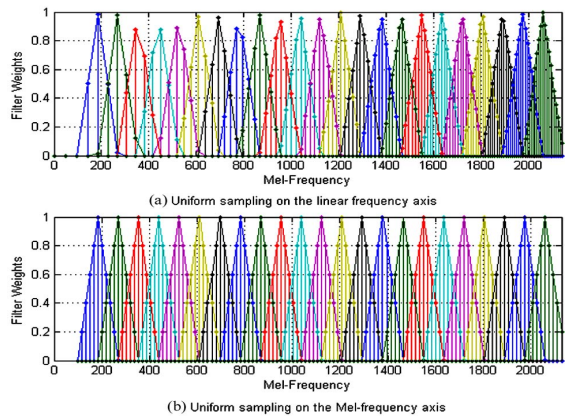

Fig 3. Filterbanks (a) for conventional FFT spectrum, (b) for basic and warped MVDR spectra

where Σ_c is the covariance matrix of the 12-dimensional vector of the cepstral coefficients, either FFT- or MVDR-based, and μ_c is the mean vector. We didn't apply CN on the log energy component in the feature vector to get better recognition results.

Progressive Histogram Equalization (PHEQ) was recently discussed extensively [15, 16]. Histogram equalization (HEQ) has been proved to provide significant improvements in speech recognition under noisy environment, in which the cumulative distribution function of the feature parameters, usually evaluated from an utterance, is normalized to a reference distribution. Instead, PHEQ performs the histogram equalization (HEQ) over a reference interval progressively moving with the frame being considered. Based on the assumption that the additive noise is time-varying, equalization with respect to some short interval near the present frame makes better sense than to a whole utterance. For a feature parameter of a noisy speech frame located at time t, the corresponding reference interval for feature normalization can be defined to be between time indices s and e, where $s \le t \le e$. So a total of $n = e-s+1$ temporally neighboring feature parameters are collected from the reference interval, on which histogram equalization (HEQ) can then be performed. The length of the reference interval, n, can be carefully chosen [15][16]. In this research, the reference distribution is assumed to be the standard normal.

There are also advantages of PHEQ over the utterance-wise feature normalization approaches in terms of real-time processing requirements. For the utterance-wise approach, the processing can be performed only when the complete utterance is received. With the progressive approach, the processing can be in parallel with the waveform read-in and feature extraction processes. The time delay can thus be substantially reduced.

3. EXPERIMENTAL CONDITIONS

The experiments reported in this paper were conducted on the database AURORA2. Ten different types of noise, as representatives of real-world noise, were included in this database. We use both the clean training and multi-condition training data sets to construct two sets of HMM models. All the three sets of testing conditions in AURORA2 tasks were tested, i.e., set A (subway, babble, car, and exhibition noise), set B (restaurant, street, airport, and train station noise), and set C (subway and street noise, with channel effect). Each of the two training data sets in the AURORA2 database consists of 8440 utterances of English connected digit strings. The MFCC feature extraction follows the WI007 front-end, which gives 14-coefficient (C0~C12 and log energy) feature vector [17]. Our MVDR-based 13-dimensional feature vectors consist of the 12 MVDR-based cepstral coefficients (C1 to C12) and the log energy, which were then used to obtain the delta and delta-delta components to form the 39-dimensional features for the following recognition tasks. For the MVDR spectrum estimation, M = 40 and N = 200 are defined previously in sec. 2.1 and in equation (1).

4. EXPERIMENTAL RESULTS

The experimental results are presented in the following sections.

4.1. Baseline Experiments

In Table 1, the word accuracies with FFT-based MFCCs are taken as the baseline and the results are listed in the first row. The following rows are for the MVDR features. $\lambda = 0$ is for original MVDR, while other values of λ are for frequency-warped cases. The results in Table 1 show that in each column a good choice of λ for MVDR-based features can actually offer better results than the baseline FFT-based MFCCs, when $\lambda \le 0.5$.

Training Sets		Clean Training (0~20dB)				Multi-Condition Training (0~20dB)			
Test Sets		A	B	C	Avg	A	B	C	Avg
FFT-based MFCCs		61.3	55.7	66.1	61.1	87.8	86.3	83.8	86.0
MVDR-based Features, M=40	$\lambda=0$	60.2	53.6	66.6	60.1	88.0	**86.92**	83.5	86.1
	$\lambda=0.1$	63.5	58.4	**68.5**	**63.4**	88.2	86.86	**84.0**	**86.4**
	$\lambda=0.2$	59.6	55.9	63.4	59.6	88.1	86.1	83.6	85.9
	$\lambda=0.3$	60.9	58.1	64.4	61.1	88.2	85.4	83.8	85.8
	$\lambda=\lambda^*$	59.5	57.0	62.9	59.8	88.4	86.1	83.5	86.0
	$\lambda=0.4$	62.5	58.3	64.2	61.6	**88.5**	86.1	83.2	85.9
	$\lambda=0.5$	**64.5**	**60.0**	64.6	63.0	87.3	84.6	82.4	84.8
	$\lambda=0.6$	58.8	55.5	57.7	57.3	85.8	82.5	81.6	83.3
	$\lambda=0.7$	52.5	48.5	58.1	53.0	82.8	78.9	79.4	80.4
	$\lambda=0.8$	46.0	43.5	52.2	47.2	77.3	72.8	73.2	74.5
	$\lambda=0.9$	26.8	29.9	33.1	30.0	63.1	59.0	56.8	59.6

Table 1. Word accuracies for baseline FFT-based MFCCs and MVDR-based features

4.2. Applying Cepstral Normalization

The results of applying CN on the FFT-based MFCCs and MVDR-based features are shown in Table 2. As compared to those shown in Table 1, we can see CN offers significant improvements for all cases, and again in each column, a good choice of λ for MVDR-based features can provide better accuracy than the baseline FFT-based MFCCs.

4.3. Progressive Histogram Equalization (PHEQ)

Here PHEQ was used for feature normalization to replace CN as in sec. 4.2, and the corresponding results are shown in Table 3. The length of the reference interval in PHEQ, i.e. n in sec. 2.4, is empirically chosen to be 100, which roughly corresponds to a speech segment of one-second time span.

Comparing the data in Table 3 to those in Table 2, for clean-training very significant improvements are evident in each case. But the improvements become relatively limited for multi-condition-training. Apparently PHEQ is a very powerful feature normalization technique, not only for FFT-based MFCCs, but equally applicable for MVDR-based features as well. Also it is again apparent that with PHEQ, in most cases a good choice of λ for MVDR features can offer better results than FFT-based MFCCs, although here the improvements become less significant. It seems that in the cases tested here with a powerful normalization process, the choice of a good feature parameter becomes less important. Note that for the multi-condition-training case, regardless of FFT-based MFCCs or MVDR-based features, the improvements that PHEQ can offer (Table 3) as compared to CN (Table 2) are relatively limited, except for test set C. This may be explained as follows. PHEQ is applied to the cepstrum coefficients to alleviate the residual mismatch, and can deal with the channel mismatch in test set C efficiently. The goal of applying (progressive) histogram equalization is to normalize the feature distributions, or to limit the feature dynamics to be within some statistical range. With multi-condition training, the feature distribution may be much more complicated than a single standard normal as we assumed here.

Training Sets		Clean Training (0~20 dB)				Multi-Condition Training (0~20dB)			
Test Sets		A	B	C	Avg	A	B	C	Avg
FFT-based MFCCs		74.7	76.8	73.3	74.9	89.3	87.7	86.8	88.0
MVDR-based Features, M=40	λ =0	75.2	77.7	73.6	75.5	89.3	88.0	86.8	88.0
	λ =0.1	74.9	77.3	73.4	75.2	89.0	88.0	86.7	87.9
	λ =0.2	76.5	78.4	75.1	76.6	**89.6**	**88.4**	86.7	88.22
	λ =0.3	**78.0**	**79.0**	76.4	**77.8**	89.4	88.1	86.9	88.15
	$\lambda = \lambda^*$	77.2	78.4	76.3	77.3	89.4	88.0	87.2	88.20
	λ =0.4	77.5	78.6	**76.5**	77.6	89.5	88.0	**87.3**	**88.24**
	λ =0.5	76.2	77.0	75.3	76.2	88.7	87.2	86.1	87.3
	λ =0.6	75.1	75.6	73.9	74.8	87.7	85.7	84.5	86.0
	λ =0.7	72.7	72.4	70.8	72.0	85.1	82.8	81.1	83.0
	λ =0.8	65.0	64.3	62.2	63.8	79.8	77.4	74.0	77.0
	λ =0.9	45.7	45.4	42.1	44.4	64.9	61.2	57.4	61.2

Table 2. Word accuracies for cepstral-normalized FFT-based MFCCs and MVDR-based features

Training Sets		Clean Training (0~20dB)				Multi-Condition Training (0~20dB)			
Test Sets		A	B	C	Avg	A	B	C	Avg
FFT-based MFCCs		81.8	**82.85**	82.7	82.5	89.4	88.8	88.8	89.0
MVDR-based Features, M=40	λ = 0	81.1	82.2	81.8	81.7	89.1	88.5	88.6	88.7
	λ =0.1	81.7	82.82	82.4	82.3	**89.6**	**89.0**	88.92	**89.2**
	λ =0.2	81.2	82.6	82.8	82.2	89.3	88.6	88.90	88.9
	λ =0.3	**82.0**	82.83	**83.1**	**82.6**	89.0	88.0	**88.94**	88.7
	$\lambda = \lambda^*$	81.6	82.2	82.3	82.0	88.8	87.9	88.90	88.5
	λ =0.4	81.4	81.9	82.0	81.8	88.5	87.6	88.7	88.3
	λ =0.5	81.5	81.7	81.6	81.6	87.9	86.9	87.9	87.6
	λ =0.6	80.1	79.9	80.0	80.0	85.8	84.6	85.7	85.3
	λ =0.7	77.5	76.9	77.8	77.4	82.7	81.4	82.5	82.2
	λ =0.8	70.3	68.8	71.8	70.3	76.1	74.3	76.0	75.5
	λ =0.9	55.0	53.5	55.0	54.5	60.6	57.7	58.5	58.9

Tables 3. Results of applying PHEQ for FFT-based MFCCs and MVDR-based features

5. CONCLUSIONS

In this paper, we refine the feature extraction scheme for speech recognition by the warped MVDR-based approaches. Cepstral normalization (CN) and progressive histogram equalization (PHEQ) were further applied to improve the robustness of the warped MVDR-based features. The results of the experiments showed that the MVDR-based features can offer better recognition accuracy than the conventional FFT-based MFCCs if a good parameter λ can be chosen.

6. REFERENCES

[1] L.-S. Lee and Y. Lee, "Voice Access of Global Information for Broad-band Wireless: Technologies of Today and Challenges of Tomorrow", Proceedings of the IEEE, Jan. 2001.

[2] J. Capon, "High Resolution Frequency Wavenumber Spectral Analysis," Proc. IEEE, vol. 57, Aug. 1969.

[3] R. T. Lacoss, "Data Adaptive Spectral Analysis Methods," Geophysics, vol. 36, Aug. 1971.

[4] M. N. Murthi, B. D. Rao, "All-pole Model Parameter Estimation for Voiced Speech," IEEE Workshop Speech Coding Telecommunications Proc., 1997.

[5] M. N. Murthi, B. D. Rao, "All-Pole Modeling of Speech Based on the Minimum Variance Distortionless Response Spectrum," IEEE Transactions on Speech and Audio Processing, vol. 8, no. 3, May 2000.

[6] S. Dharanipragada, B. D. Rao, "MVDR Based Feature Extraction for Robust Speech Recognition," ICASSP 2001.

[7] S. Dharanipragada, "Feature Extraction for Robust Speech Recognition," IEEE International Symposium on Circuits and Systems 2002, May 2002.

[8] M. Wolfel, J. McDonough, A. Waibel, "Minimum Variance Distortionless Response on a Warped Frequency Scale," Eurospeech 2003.

[9] M. Wolfel, J. McDonough, A. Waibel, "Warping and Scaling of the Minimum Variance Distortionless Response," ASRU 2003.

[10] A. Harma, U. K. Laine, "A Comparison of Warped and Conventional Linear Predictive Coding," IEEE Transactions on Speech and Audio Processing, vol. 9, no. 5, Jul. 2001.

[11] H. W. Strube, "Linear Prediction on a Warped Frequency Sacle," J. Acoust. Soc. Amer., vol. 68, Oct. 1980.

[12] S. L. Marple Jr., *Digital Spectral Analysis with Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[13] S. Haykin, *Adaptive Filter Theory*, 4th ed., Prentice-Hall Inc., 2002.

[14] B. R. Musicus, "Fast MLM Power Spectrum Estimation from Uniformly Spaced Correlations," IEEE Trans. ASSP, vol. ASSP-33, no. 4, Oct. 1985.

[15] S.-N. Tsai, L.-S. Lee, "A New Feature Extraction Front-End for Robust Speech Recognition Using Progressive Histogram Equalization and Multi-Eigenvector Temporal Filtering," ICSLP 2004.

[16] S.-N. Tsai, "Improved Robustness of Time-Frequency Principal Components (TFPC) by Synergy of Methods in Different Domains," ICSLP 2004.

[17] D. Pearce, H.-G. Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," ICSLP 2000.