Temporal Modulation Spectral Restoration for Robust Speech Recognition

Syu-Siang, Wang
Graduate Institute of Communication Engineering
National Taiwan University
Taipei, Taiwan

Yu Tsao
Research Center for Information Technology Innovation
Academia Sinica
Taipei, Taiwan

***Abstract*—In this paper, we propose a temporal modulation spectral resto-ration (TMSR) approach for robust feature extraction in automatic speech recognition. There were three main function blocks in TMSR. First, mean and variance normalization (CMVN) was applied to the original feature sequence. Second, the noise characteristic was estimated with an analysis of the normalized features. Third, a gain function was designed to attenuate noise and enhance speech components from the normalized features. In this study, a simple high-pass filter noise estimation scheme and a gain function derived by the generalized maximum a posteriori (GMAP) algorithm were employed in TMSR. The proposed method was evaluated on two benchmark databases, Aurora-3 and Aurora-4. Results showed that TMSR outperformed the baseline and several well-known robust feature extraction methods.**

Keywords- temporal modulation spectral restoration, TMSR, noise estimation, generalized maximum a posteriori.

I. INTRODUCTION

Even though the performance of automatic speech recognition (ASR) system has been significantly improved recently, identifying a way to effectively improve ASR in noisy conditions is still an important research direction [1]. Many approaches have been proposed to reach this goal [2] [3]. Among them, a category of approaches develops robust feature extraction algorithms to reduce noise effect and increase system performance. Some robust feature extraction algorithms are proposed to normalize training and testing acoustic features to the referenced distribution [4] [5]. Cepstral mean subtraction (or normalization, CMS, CMN) [6] [7], cepstral mean and variance normalization (CMVN) [8], cepstral statistics compensation (CSC) [9], histogram equalization (HEQ) [10] [11] and higher order cepstral moment normalization (HOCMN) [12] are successful examples. Other than normalization methods, some approaches intend to suppress noise effects in acoustic features. Relative spectral (RASTA) band-pass filter is a well-known approach, which retains the informative speech components around 4 Hz while shrinking components at other frequencies in the modulation frequency domain [13]. A moving-average (MVA) technique adopting a simple low-pass filter to suppress noise components has been confirmed effectively [14] [15].

Recently, some approaches focus on normalizing the speech features in the modulation frequency domain. Spectral histogram equalization (SHE) [16] normalizes the modulation magnitude spectrum of noisy features to a referenced magnitude spectral distribution. Furthermore, considering the various informative speech components in modulation spectrum, sub-band modulation spectrum compensation technique [17] has been proposed to improve the full-band algorithms (CSC and SHE) to sub-band procedures. However, sub-band modulation spectrum compensation technique cannot be applied to mean subtraction on modulation spectra because the magnitudes might become negative. Therefore, generalized logarithmic modulation spectral mean normalization (GLMSMN) [18] incorporates the generalized logarithm operation (q-logarithm) [19] to normalize the magnitude modulation spectra and achieves satisfactory performance. In addition, temporal structure normalization (TSN) [20] algorithm, which designs a dynamic filter from normalized power spectrum densities, has shown good capability to handle the noise issue.

In this paper, we proposed a temporal modulation spectral restoration (TMSR) approach for robust feature extraction. TMSR approach is conducted based on the zero-mean feature sequences. Since a cepstral feature sequence is preprocessed by a mean normalization scheme, some spectral restoration algorithms, which are originally developed for speech enhancement, can be applied to reduce noise effects from the normalized features. Generally, spectral restoration algorithm includes two parts, namely, noise estimation and gain function estimation. In this paper, we simply use a high-pass filter to estimate noise information from the normalized speech features with an assumption that the major noise components are located around the high modulation frequency parts. Then, the gain function is calculated based on the estimated noise information with the generalized maximum a posteriori (GMAP) algorithm [21], [22]. Two standard databases, Aurora-3 [23] [24] and Aurora-4 [25], were used to evaluate the proposed TMSR approach. From our experimental results, we first observe that comparing to the CMVN normalized speech spectra, the TMSR restored speech spectra present smaller magnitude in high modulation frequency parts, and larger magnitude in low modulation frequency bands. Moreover, recognition results on Aurora-3 and Aurora-4 demonstrate that TMSR outperforms the baseline, CMVN, and MVA and several well-known robust feature extraction algorithms.

The remainder of this paper is organized as follows. Section 2 introduces the overall TMSR approach. The noise estimation and gain function estimation of TMSR are then presented in

978-1-5090-2179-6/16 $31.00 © 2016 IEEE
DOI 10.1109/BigMM.2016.91

Figure 1. Block diagram of our TMSR processing technique. Abbreviations stand for: FE - feature extraction; CMVN - conventional CMVN process; SR - spectral restoration.

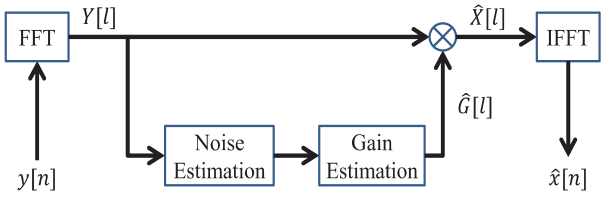


Figure 2 Block diagram of spectral restoration.

Section 3 and 4, respectively. Section 5 presents our experimental setup and results. Finally, a concluding remark is provided in Section 6.

II. Temporal Modulation Spectral Restoration

Figure 1 shows the normalization and spectral restoration stages of the TMSR approach. Assuming that there are N frames in the original noisy feature sequence $\boldsymbol{s}=\{s[n], 0 \le n \le N-1\}$, the main goal of TMSR is to estimate cleaned feature sequence $\boldsymbol{x}=\{x[n], 0 \le n \le N-1\}$. TMSR first per forms normalization on $\boldsymbol{s}$ to get zero-mean feature sequence, $\boldsymbol{y}=\{y[n], 0 \le n \le N-1\}$, and then applies spectral restoration on $\boldsymbol{y}$ to get $\boldsymbol{x}$.

Next, TMSR considers that the noisy feature $y[n]$, is a summation of a clean speech $x[n]$, and noise signal $v[n]$, in the cepstral domain.

$$y[n] = x[n] + v[n]. \tag{1}$$

In the modulation frequency domain, the noisy speech signal with frequency index l, can be expressed as:

$$Y[l] = X[l] + V[l], 0 \le l \le L-1\,, \tag{2}$$

where L is the total number of frequency bin; $X[l]$ and $V[l]$ are the spectra of speech signal and noise signal, respectively.

Figure 2 shows the spectral restoration process, which can be divided into noise estimation and gain function estimation stages. The noise tracking stage computes noise power from the noisy speech $Y[l]$, to obtain necessary statistics, which are then used to calculate a gain function $\hat{G}[l]$. Finally, we obtain enhanced speech $\hat{X}[l]$, by

$$\hat{X}[l] = \hat{G}[l]\, Y[l]. \tag{3}$$

Notably, the calculations of noise power and gain function are derived based on two assumptions: (1) speech and noise signals are additive; (2) speech and noise signals are both random processes and independent of each other.

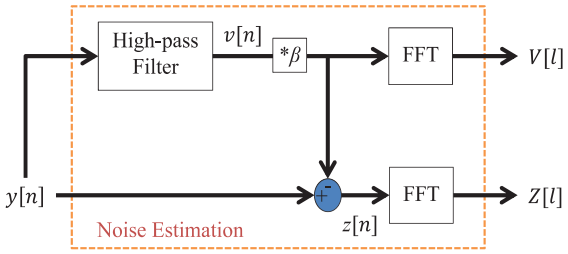


Figure 3 Block diagram of noise estimation.

Finally, an inverse Fourier transform is performed on $\hat{X}[l]$ to get clean spectral feature $\hat{x}[n]$. In the following two sections, we will detail the noise estimation and gain function estimation that used in this study.

III. Noise Estimation

The purpose of noise estimation is to obtain the noise statistics to facilitate the following gain function estimation. In this study, we derive a simple high-pass filter noise estimation scheme, as shown in Figure 3.

From Figure 3, an estimated noise features, $v[n]$, is filtered from a full-band input speech feature $y[n]$, by a simple two-point high-pass FIR filter, $h[n]$:

$$h[n] = 0.5\delta[n] - 0.5\delta[n-1]; \tag{4}$$

$$v[n] = y[n] \otimes h[n], \tag{5}$$

where $\otimes$ is a convolutional operator, and $\delta[n]$ is a Kronecker delta function.

Thus, we can obtain the sequence, $z[n]$, by subtracting the weighted high-frequency sequence, $v[n]$, from $y[n]$:

$$z[n] = y[n] - \beta * v[n]. \tag{6}$$

where β is a confident factor determining the scale of estimated noise information. Finally, we apply fast Fourier transformation on noisy, noise, and clean speech sequences and get

$$Y[l] = FFT(y[n])\,, 0 \le l \le L-1, \tag{7}$$

$$V[l] = FFT(v[n])\,, 0 \le l \le L-1, \tag{8}$$

$$Z[l] = FFT(z[n])\,, 0 \le l \le L-1, \tag{9}$$

where l is a modulation frequency index.

In this section, we present the gain function estimation for TMSR. The generalized maximum a posteriori (GMAP) algorithm [21] is used to calculate the gain function in this study. By decomposing noisy and clean speech spectra, $Y[l]$ and $\hat{X}[l]$ in (2), into amplitude and phase parts, we have

$$Y[l] = Y_m exp(j\theta_{Y_l}), \quad (10)$$

$$\hat{X}[l] = \hat{X}_m exp(j\theta_{\hat{X}_l}), \quad (11)$$

where $Y_m = |Y[l]|$, $\hat{X}_m = |\hat{X}[l]|$, $\theta_{Y_l} = \angle Y[l]$ and $\theta_{\hat{X}_l} = \angle \hat{X}[l]$. To restore $\hat{X}[l]$ from $Y[l]$, we directly use the the phase, θ_{Y_l}, for $\theta_{\hat{X}_l}$ (or $\theta_{\hat{X}_l} = \theta_{Y_l}$), and, thus we have

$$\hat{X}[l] = \hat{X}_m exp(j\theta_{Y_l}). \quad (12)$$

Next, we intend to restore the clean speech amplitude, $\hat{X}_m$, and formulate a cost function based on the GMAP criterion:

$$\hat{X}_m = \arg\max_{X_m} J_{GMAP}(X_m), \quad (13)$$

where

$$J_{GMAP}(X_m) = ln\{p[Y[l]|X_m]\,(p[X_m])^{\alpha}\}. \quad (14)$$

Then, by taking differential on Eq. (14) with respect to X_m and equating the result to zero, we obtain

$$\hat{X}_m = \frac{\xi + \sqrt{\xi^2 + (2\alpha - 1)(\alpha + \xi)\xi/\gamma}}{2(\alpha + \xi)} Y_m, \quad (15)$$

with the gain function

$$G_{GMAP} = \frac{\xi + \sqrt{\xi^2 + (2\alpha - 1)(\alpha + \xi)\xi/\gamma}}{2(\alpha + \xi)}, \quad (16)$$

where $\hat{X}_m$ is estimated spectrum of clean speech; ξ and γ, a priori and a posteriori SNR, respectively, which are defined as $\xi = |Z[l]|^2/|V[l]|^2$, $\gamma = |Y[l]|^2/|V[l]|^2$, where $V[l]$ and $Z[l]$ can be estimated from Eqs. (8) and (9).

With the estimated amplitude $\hat{X}_m$ from Eq. (15) and phase $\theta_{\hat{X}_l}$ from Eq. (12), we obtain the restored speech spectra, $\hat{X}[l]$, in Figure 2. Finally, by applying the inverse fast Fourier transform (IFFT) on $\hat{X}[l]$, we obtain the TMSR processed cepstral features, $\hat{x}[n]$.

IV. EXPERIMENT

The proposed TMSR approach was tested on two standardized databases, Aurora-3 and Aurora-4. This section first presents the experimental setup and then shows the trajectory and modulation spectral analyses of the normalized features and that further processed by MVA or TMSR. Finally, the recognition results on Aurora-3 and Aurora-4 are reported.

A. Experimental Setup

This subsection introduces the Aurora-3 and Aurora-4 tasks and setups of speech recognition evaluations.

1) Database1: Aurora-3

Speech utterances in Aurora-3 were recorded by close-talking (channel-0) and hands-free (channel-1) microphones in in-car environments with various driving conditions. Aurora-3 comprises four different languages, Danish, Spanish, Finnish, and German, with each language including three matching-conditions: (1) Well-matched condition (WM): For each language, 70% of the speech utterances recorded in channel-0 and channel-1 were used as training set, and the remaining 30% speech data were used as testing set. (2) Medium-mismatched condition (MM): Only speech utterances recorded by channel-1 were used. Noise-contaminated speech data under less noise driving condition was applied as training data. The remaining speech utterances in channel-1 were applied as testing data. (3) Highly-mismatched condition (HM): The entire set of speech utterances in channel-0 was used as the training set. The testing set was constructed by the speech data in channel-1 except the quietest noise condition.

In this study, the acoustic model includes 11 digit models in addition to silence and short pause models. Each digit is modeled with a hidden Markov model (HMM) containing 16 states and three Gaussian mixtures per state. Silence and short pause models include three and one states, respectively, both with six Gaussian mixtures per state. Hidden Markov toolkit (HTK) [26] was used to train model and test recognitions. For all the experiments, an end pointing process (i.e., presegmentation) was performed to improve performance [15]. Each utterance was characterized by standard 39 dimensional MFCC components, consisting of 13 static coefficients, and their first and second derivatives.

2) Database1: Aurora-4

Aurora-4 is a benchmark database for large vocabulary continuous speech recognition (LVCSR) [25] under noisy conditions. The clean speech utterances in Aurora-4 were obtained from the Wall Street Journal (WSJ0) corpus [27] and then contaminated by different types of noise artificially to generate noisy speech data. The Aurora-4 database provides two sampling rates, 8k Hz and 16k Hz, and 8k Hz data was selected here for training and testing. Aurora-4 contains two sets of training data, clean-condition and multi-condition training sets. Aurora-4 provides four test sets: set A (clean speech in the same channel condition as in training; set 1), set B (noisy speech in the same channel condition as in training; sets 2-7), set C (clean speech in a different channel condition as in training; set 8), and set D (noisy speech in a different channel condition as in training; sets 9-14).

In this study, the multi-condition training set was used to train acoustic models. We constructed context-dependent tri-phone acoustic models, where each model was characterized by an HMM. Each HMM consists of 3 states, with 8 Gaussian mixtures per state. A tri-gram language model was computed based on the reference transcription of the training data. Each utterance was characterized by standard 39 dimensional MFCC components, consisting of 13 static coefficients, and their first and second derivatives.

B. Temporal Envelop Analysis

A speech sample was chosen from the Aurora-4 database and used for temporal analysis. Figure 4 shows a part of C0 coefficients from the speech sample, processed by CMVN, MVA and TMSR techniques. Figure 4 (a) shows the CMVN and MVA processed C0 trajectories, and Figure 4 (b) shows the CMVN and TMSR processed C0 trajectories.

From Figure 4 (a) and (b), MVA processed C0 presents smoother trajectory than that of CMVN processed C0. In the meanwhile, comparing to CMVN processed C0, TMSR processed C0 presents larger amplitude variances while the smoothing capability of TMSR is still observable. We observe similar results for other MFCC coefficients. Notably, different from the filtering approaches, which focus on designing a filter to smooth the acoustic features, TMSR intends to enhance the signal-to-noise ratio (SNR) of acoustic features. Accordingly, the characteristic of TMSR and MVA are rather different, as shown in Figure 4.

C. Spectral Analysis

Figure 5 illustrates the temporal modulation spectral analyses of TMSR and MVA. The modulation spectra of C_0 coefficient are plotted using the same speech sample that were used in Figure 4. From Figure 5, both MVA and TMSR suppress the magnitude in the higher modulation bands (25~50 Hz). However, for the lower modulation frequency bands (0~16 Hz), TMSR further amplifies the magnitude while MVA doesn't. This set of analyses confirms that TMSR attenuates noise components (attenuation in higher modulation frequency bands) and enhances speech information (magnification in lower modulation frequency bands).

D. Recognition Results

The recognition performances in the experiments were presented in word error rate (WER). For Aurora-3, we present recognition results of the four language sets and three matching types along with an average (set Avg) over the three types. For Aurora-4, we show the results of set A, set B, set C, and set D also with an average (set Avg) of the four test sets.

1) Recognition rates for Aurora-3

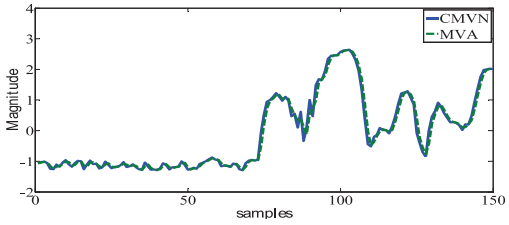
Table I lists the WERs of the proposed TMSR approach along with the baseline (MFCC), CMVN, and MVA for four languages in Aurora-3. For the TMSR experiments, β in Eq. (6) and α in Eq. (14) are empirically optimized for each language set. Notably, the high-pass filter noise estimation scheme presented in Section 3.3 calculates $v[n]$ by

$$v[n] = 0.5y[n] - 0.5y[n-1]. \tag{18}$$

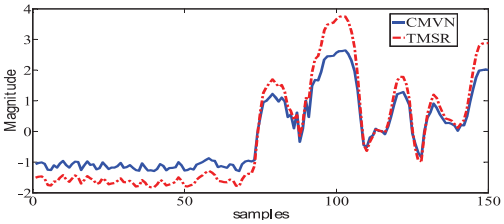
MVA calculates features by

$$\begin{aligned} x_{MVA}[n] &= 0.5y[n] + 0.5y[n-1] \\ &= y[n] - \gamma * v[n], \quad \gamma = 1.0. \end{aligned} \tag{19}$$

Thus, $x_{MVA}[n]$ equals $z[n]$ in Eq. (6), with $\beta = 1.0$.



(a) Trajectories of C_0 processed by CMVN and MVA.



(b) Trajectories of C_0 processed by CMVN and TMSR.

Figure 4 Trajectories of CMVN, MVA and TMSR processed C_0 coefficient of a speech sample in Aurora-4.

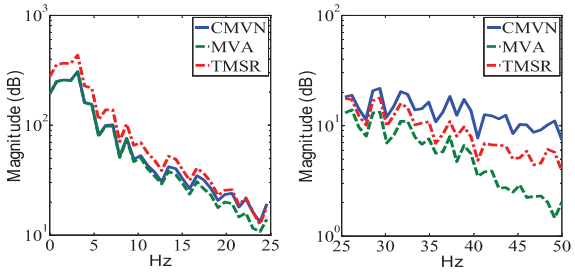


Figure 5 Modulation spectral analyses of CMVN, MVA, and TMSR processed C_0 of a speech sample in Aurora-4.

From Table I, TMSR outperforms the baseline, CMVN, and MVA in most conditions for all of the four language sets. Notably, the improvements from MFCC to CMVN comes from the cepstral normalization, where the speech features are normalized to zero-mean and one-variance signal sequence. On the other hand, the improvement from CMVN to MVA is from a direct subtraction of noise components from the CMVN features. Finally, the improvements from CMVN to TMSR is from applying spectral restoration to attenuate noise components in the modulation frequency domain. Since MVA and TMSR uses a same noise estimation scheme, better results achieved by TMSR suggest that utilizing noise information to design a gain function (dynamic filter) for spectral restoration can be a more effective way then a direct subtraction.

2) Recognition rates for Aurora-4

Table II shows the recognition results of Aurora-4. The parameters, α and β, used in TMSR are set to 8.0 and 0.4, respectively. In addition to CMVN and HEQ, and MVA (γ = 1.0 in Eq. (19)), we test recognition using a modified MVA by setting $\gamma = 0.4$ in Eq. (19) and compare its performance with TMSR ($\beta = 0.4$).

Table I. WER values of Aurora-3.

Danish	*MFCC*	*CMVN*	*MVA*	*TMSR*
WM	11.13	6.86	6.69	**6.42**
MM	28.11	15.25	15.82	**14.83**
HM	48.86	24.53	24.73	**22.30**
Avg	26.51	14.21	14.40	**13.33**
Finnish				
WM	6.09	4.22	3.95	**3.92**
MM	17.17	**9.30**	10.12	9.64
HM	33.85	13.07	11.13	**10.78**
Avg	16.91	8.21	7.90	**7.64**
German				
WM	8.03	4.49	4.29	**4.27**
MM	18.45	10.54	10.03	**9.52**
HM	24.24	9.53	9.07	**8.79**
Avg	15.73	7.87	7.49	**7.24**
Spanish				
WM	5.14	3.43	3.39	**3.36**
MM	12.72	7.68	8.10	**7.59**
HM	46.53	11.64	9.98	**9.95**
Avg	18.14	6.97	6.69	**6.49**

Table II. WER values of Aurora 4

Set	*A*	*B*	*C*	*D*	*Avg*
MFCC	10.83	20.66	16.57	28.34	22.95
CMVN	11.16	16.82	13.37	23.18	18.89
HEQ	10.46	17.23	13.33	22.91	18.91
***MVA*($\gamma = 1.0$)**	11.23	16.99	13.04	23.10	18.91
***MVA*($\gamma = 0.4$)**	11.57	17.09	13.08	22.66	18.80
TMSR	**10.72**	**16.68**	**12.63**	**22.32**	**18.38**

From Table II, we observed that TMSR achieves the best performances for all of the four testing sets and set Avg comparing to other approaches, including the baseline, CMVN, HEQ, and the conventional MVA ($\gamma = 1.0$). Additionally, we again verify that utilizing the estimated noise characteristics to design a gain function for spectral restoration (for TMSR) provides better performance than a direct subtraction process (for modified MVA with $\gamma = 0.4$).

V. CONCLUSIONS

This paper proposes a novel TMSR approach for robust feature extraction to improve speech recognition under noisy conditions. To validate the capability of TMSR, we used a simple high-pass filter (HF) noise estimation scheme and a GMAP-derived gain function to enhance CMVN normalized

MFCC features. Notably, the well-known MVA algorithm directly subtracts the noise components estimated by HF from the CMVN normalized features. On the other hand, TMSR applies the GMAP gain function, based on the estimated noise components, to perform spectral restoration in the modulation frequency domain. Based on the trajectory and modulation spectral analyses, TMSR suppresses the noise components and increases the SNR of the original features. The recognition results on Aurora-3 and Aurora-4 confirm that TMSR provides superior performance to CMVN and MVA.

This paper presents our first study on applying spectral restoration techniques to enhance acoustic features. In the future, we will investigate algorithms to optimize α and β in Eqs. (6) and (14) according to the types of noise and SNR levels. Additionally, we will explore different noise estimation schemes and gain function estimation algorithms.

REFERENCES

[1] M. Seltzer, D. Yu, and E. Wang, "An investigation of deep neural networks for noise robust speech recognition," i*n Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 7398–7402, 2013.

[2] S. Molau, D. Keysers and H. Ney, "Matching training and test data distributions for robust speech recognition," *Speech Communication*, vol. 41, pp. 579–601, 2003.

[3] X. Huang, A. Acero, and H.-W. Hon, "Spoken language processing: a guide to theory," Algorithm and System Development. New Jersy: Prentice Hall PTR, 2001.

[4] F. Hilger, H. Ney, and L. F. I. Vi, "Quantile based histogram equalization for noise robust speech recognition," *in Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1135–1138, 2001.

[5] S.-H. Lin, Y.-M. Yeh, and B. Chen, "Exploiting polynomial-fit histogram equalization and temporal average for robust speech recognition," *in Proc. International Conference on Spoken Language Processing (ICSLP)*, pp. 1135–1138, 2006.

[6] O. Viikki and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise," *Acoustics, Speech and Signal Processing*, vol. 2, pp. 733-736, 1998.

[7] H. Kim and R. C. Rose, "Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 435-446, 2003.

[8] S. Tibrewala and H. Hermansky, "Multiband and adaptation approaches to robust speech recognition," *in Proc. Eurospeech*, pp. 2619-2622, 1997.

[9] J.-W. Hung, "Cepstral statistics compensation using online pseudo stereo codebooks for robust speech recognition in additive noise environments," *in Proc. ICASSP*, pp. 513-516, 2006.

[10] D. P. Ibm, S. Dharanipragada, and M. Padmanabhan, "A nonlinear unsupervised adaptation technique for speech recognition," *in Proc. International Conference on Spoken Language Processing (ICSLP)*, pp. 556–559, 2000.

[11] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Prez-Crdoba, M. C. Bentez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 355–366, 2005.

[12] C.-W. Hsu and L. S. Lee, "Higher order cepstral moment normalization (HOCMN) for robust speech recognition," i*n Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 197-200, 2004.

[13] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*,vol. 2, pp. 578-589, 1994.

[14] C.-P. Chen, J. A. Bilmes, and K. Kirchhoff, "Lowresource noise-robust feature post-processing on aurora 2.0," *in Proc. International Conference on Spoken Language Processing (ICSLP)*, pp. 2445-2448, 2002.

[15] C.-P. Chen, K. Filali, and J. A. Bilmes, "Frontend post-processing and backend model enhancement on the aurora 2.0/3.0 databases," *in Proc. International Conference on Speech and Language Processing (ICSLP)*, pp. 241–244, 2002.

[16] L.-C. Sun and L.-S. Lee, "Modulation spectrum equalization for improved robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.20, no.3, pp.828-843, March 2012.

[17] W.-H. Tu, S.-Y. Huang and J.-W. Hung, "Sub-band modulation spectrum compensation for robust speech recognition," *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop*, pp.261-265, 2009

[18] H.-T. Fan, C.-H. Hsu and J.-W. Hung, "The study of q-logarithmic modulation spectral normalization for robust speech recognition," *International Conference on System Science and Engineering (ICSSE)*, pp.183-186, 2012.

[19] H. F. Pardede, and K. Shinoda, "Generalized-log spectral mean normalization for speech recognition," *in Proc. Interspeech*, pp 1645-1648, 2011.

[20] X. Xiao, E.-S. Chng, and H.-Z. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.16, no.8, pp.1662-1674, 2008.

[21] Y.-C. Su, Y. Tsao, J.-E. Wu and F.-R. Jean, "Speech enhancement using generalized maximum a posteriori spectral amplitude estimator," *in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[22] Y. Tsao, and Y.-H. Lai. "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement." *Speech Communication*, 2015

[23] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *ISCA ITRW ASR2000*, Paris, September 2000.

[24] Motorola Au/374/01, "Small vocabulary evaluation: baseline Mel-cepstrum performances with speech endpoints," 2001.

[25] N. Parihar and J. Picone, "Aurora working group: Dsrfront end lvcsr evaluation au/384/02," *in Institute for Signal and Information Processing Report*, 2002.

[26] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P.Woodland, *The HTK Book* (*for HTK Version 3.3*). Cambridge University Engineering Department, 2005.

[27] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," *in Proc. International Conference on Spoken Language Processing (ICSLP)*, 1992.