



Histogram Equalization of Real and Imaginary Modulation Spectra for Noise-Robust Speech Recognition

Hsin-Ju Hsieh^{1,2}, Berlin Chen², Jieh-weih Hung¹

¹National Chi Nan University, Taiwan

²National Taiwan Normal University, Taiwan

s101323902@ncnu.edu.tw, berlin@ntnu.edu.tw, jwhung@ncnu.edu.tw

Abstract

Histogram equalization (HEQ) of acoustic features has received considerable attention in the area of robust speech recognition because of its relative simplicity and good empirical performance. This paper presents a novel HEQ-based feature extraction approach that performs equalization in both acoustic frequency and modulation frequency domains for obtaining better noise-robust features. In particular, the real and imaginary acoustic spectra are first individually transformed to the modulation domain via discrete Fourier transform (DFT). The HEQ process is then carried on the corresponding magnitude modulation spectra so as to compensate for the noise distortions. Finally, the equalized modulation spectra are converted back to form the real and imaginary acoustic spectra, respectively. By doing so, we can enhance not only the magnitude but also the phase components of the acoustic spectra, and thereby create more noise-robust cepstral features. The experiments conducted on the Aurora-2 clean-condition database and task reveal that the presented approach delivers superior recognition accuracy in comparison with some other HEQ-related methods and the well-known advanced front-end (AFE) extraction scheme, which supports the potential utility of this novel approach.

Index Terms: noise robustness, feature extraction, modulation spectrum, histogram equalization, automatic speech recognition

1. Introduction

Varying environmental effects, such as ambient noise and interferences caused by the recording devices and transmission channels, often lead to severe mismatch between the acoustic environments for the training and testing speech data in automatic speech recognition (ASR), and this environmental mismatch inevitably degrades the performance of ASR dramatically [1]. Substantial efforts have been made and also a number of techniques have been developed to address this issue for improving the ASR performance in the past decades. Broadly speaking, these noise/interference processing techniques fall into three main categories: enhancement, normalization and adaptation [2], while these techniques can be conducted either in the speech feature domain or in the acoustic model domain.

Regarding the popular speech feature representation, Mel-frequency cepstral coefficient (MFCC), which reflects the spectral characteristics within a short period of time, has exhibited high discriminating capability for acoustic units and thus gives excellent recognition accuracy in nearly noise-free laboratory environments. However, MFCC is vulnerable to noise/interference and often requires compensation prior to being used in real-world scenarios. The compensation can be carried out in the various intermediate states during the

extraction of the MFCC feature stream for a speech signal. Roughly speaking, according to the MFCC extraction procedure, a time-signal is segmented to a series of overlapping frames, and then each frame signal is transformed into the acoustic spectrum, next into the (linear) critical-band spectrum and the logarithmic critical-band spectrum, and eventually into the cepstrum. First of all, spectral subtraction [3,4], Wiener filtering [5] and MMSE-based log-spectral amplitude estimation (MMSE log-STSA) [6] are exemplary methods that process the frame-based acoustic spectra. Second, a suite of feature moment normalization methods are developed to regulate the statistical moments of the cepstra, such as cepstral mean normalization (CMN) [7], cepstral mean and variance normalization (CMVN) [8] and cepstral histogram normalization (CHN) [2, 9, 10], to name but a few. Since the statistical moments are directly evaluated by the temporal series of cepstra, these moment normalization methods implicitly enhance the cepstra in temporal characteristics. On the other hand, the approaches that employ filtering on the temporal sequence of logarithmic critical-band spectrum or cepstrum include, but are not limited to, RASTA [11], temporal structure normalization (TSN) [12] and CMVN plus ARMA filtering (MVA) [13]. These temporal-filtering approaches in general emphasize the relatively low varying components (except the DC part) of the feature temporal sequence, which encapsulate rich linguistic information cues that are conducive for speech recognition. Additionally, the methods of spectral histogram equalization (SHE) [14] and modulation spectrum replacement/filtering (MSR/MSF) [15] directly modify the modulation spectrum, which is specifically referred to as the Fourier transform of the temporal sequence of cepstra.

Our work in this paper presents a novel application of histogram equalization (HEQ) [9,10] to reduce the distortion of acoustic spectral features in modulation domain for speech recognition. Unlike the conventional HEQ approaches that often operate on the temporal stream of the cepstra (which is denoted by CHN earlier) or the Mel-filter smoothed logarithmic spectra [9, 10], the presented method performs HEQ on the DFT for the temporal series of acoustic spectra (i.e., the modulation spectra of the acoustic spectra) with respect to each acoustic frequency bin. Furthermore, the real and imaginary parts of the acoustic spectra are treated individually in the presented framework, and it is different from most well-known acoustic spectral-domain robustness techniques, such as spectral subtraction and Wiener filtering, that process the magnitude acoustic spectra directly. By and large, our presented approach has the following three advantages. First, via the HEQ operation the long-term correlation among the acoustic spectra (at the same frequency) can be captured for the compensation of spectral distortion. Second, at a higher cost of computation, the noise effect can be dealt with in a finer (acoustic) frequency resolution. Third, the distortion dwelt in the acoustic spectra can be more

10.21437/Interspeech.2013-277

extensively mitigated due to the independent process for the real and imaginary parts. All of the aforementioned advantages will be confirmed via empirical evaluation.

The rest of this paper is organized as follows. Section 2 provides the essential fundamentals for HEQ and briefly describes how it can be crystallized for robust ASR. Section 3 elucidates our proposed normalization framework. Then, the experimental settings and a series of ASR experiments conducted are presented in Sections 4 and 5, respectively. Finally, Section 6 concludes this paper and suggests avenues for future work.

2. Brief Introduction of HEQ

In the HEQ algorithm, an arbitrary data series, denoted by $\{x_1, x_2, \dots, x_N\}$, is viewed as the sample set of a random variable X with a cumulative distribution function (CDF) $F_X(x)$. Then, via the mapping procedure:

$$y_i = F_T^{-1}(F_X(x_i)), \quad 1 \leq i \leq N \quad (1)$$

the CDF of another random variable Y with the obtained new data series $\{y_1, y_2, \dots, y_N\}$ as samples can approximate a predefined target CDF $F_T(y)$ as long as the number of data, N , is sufficiently large. The target CDF $F_T(y)$ is usually set to be simply a standard Gaussian distribution with zero mean and unity variance, or approximated by the histogram of the training data.

More recently, HEQ has been adopted to compensate for speech features for noise-robust ASR. The CHN [2, 9, 10] and SHE [14] methods mentioned in the previous section are two good instantiations developed along this line of thought, which operate HEQ on the temporal domain and modulation domain of MFCC features, respectively.

3. Proposed Approach

This section describes a novel HEQ-based feature extraction framework in an attempt to improve the noise robustness of speech features. First, in the preprocessed stage, any utterance $x[\ell]$ in the training and testing sets is shaped by a high-pass pre-emphasis filter, and framing as well as windowing operations are performed in turn. Then, each windowed frame signal is transformed to the acoustic frequency domain via short-time Fourier transform (STFT), and the resulting complex-valued acoustic spectrum is denoted by

$$X[n, k] = X_r[n, k] + jX_i[n, k], \quad (2) \\ 0 \leq n \leq N-1, \quad 0 \leq k \leq K-1$$

where $X_r[n, k]$ and $X_i[n, k]$ denote the acoustic real and imaginary spectra, respectively, n and k respectively refer to the indices of frame and discrete frequency, and N and K are respectively the numbers of frames and acoustic frequency bins. By the way, $\{X[n, k]\}$ in Eq. (2) is sometimes called the spectrogram of the utterance. Next, the acoustic real and imaginary spectra, $X_r[n, k]$ and $X_i[n, k]$ in Eq. (2), with respect to a fixed frequency bin k are updated via the subsequent steps.

Step I: Compute the modulation spectrum separately for $X_r[n, k]$ and $X_i[n, k]$ along the n -axis by discrete Fourier transform (DFT) as follows

$$\mathcal{X}_r[k, m] = \sum_{n=0}^{N-1} X_r[n, k] e^{-j\frac{2\pi nk}{N}} \\ \text{and} \quad 0 \leq m \leq N-1, \quad 0 \leq k \leq K-1. \quad (3)$$

$$\mathcal{X}_i[k, m] = \sum_{n=0}^{N-1} X_i[n, k] e^{-j\frac{2\pi nk}{N}}.$$

where m refers to the index of the discrete modulation frequency. The resulting spectra can be expressed in polar form as

$$\mathcal{X}_r[k, m] = \mathcal{A}_r[k, m] e^{j\theta_r[k, m]} \quad \text{and} \quad \mathcal{X}_i[k, m] = \mathcal{A}_i[k, m] e^{j\theta_i[k, m]}, \quad (4)$$

where $\mathcal{A}_r[k, m]$ and $\mathcal{A}_i[k, m]$ are, respectively, the magnitude component of $\mathcal{X}_r[k, m]$ and $\mathcal{X}_i[k, m]$, and $\theta_r[k, m]$ and $\theta_i[k, m]$ are, respectively, the phase component of $\mathcal{X}_r[k, m]$ and $\mathcal{X}_i[k, m]$.

Step II: Update the magnitude components of the modulation spectra via HEQ, while keeping the phase components unchanged. The resulting new magnitude modulation spectra are expressed by

$$\tilde{\mathcal{A}}_r[k, m] = F_{\mathcal{A}_r}^{-1}(F_{\mathcal{A}_r}(\mathcal{A}_r[k, m])) \quad \text{and} \quad \tilde{\mathcal{A}}_i[k, m] = F_{\mathcal{A}_i}^{-1}(F_{\mathcal{A}_i}(\mathcal{A}_i[k, m])), \quad (5)$$

where the cumulative distribution functions (CDFs) $F_{\mathcal{A}_r}$ and $F_{\mathcal{A}_i}$ are estimated from those $\mathcal{A}_r[k, m]$ and $\mathcal{A}_i[k, m]$ of the utterance being processed, and the inverse CDFs $F_{\mathcal{A}_r}^{-1}$ and $F_{\mathcal{A}_i}^{-1}$ are from those $\mathcal{A}_r[k, m]$ and $\mathcal{A}_i[k, m]$ of the utterances in the clean training set. As such, combining the updated magnitude components with the original phase components results in the new modulation spectra

$$\tilde{\mathcal{X}}_r[k, m] = \tilde{\mathcal{A}}_r[k, m] e^{j\theta_r[k, m]} \quad \text{and} \quad \tilde{\mathcal{X}}_i[k, m] = \tilde{\mathcal{A}}_i[k, m] e^{j\theta_i[k, m]}. \quad (6)$$

Step III: Construct the new acoustic real and imaginary spectra, denoted by $\tilde{X}_r[n, k]$ and $\tilde{X}_i[n, k]$, respectively, by taking the inverse DFT of $\tilde{\mathcal{X}}_r[k, m]$ and $\tilde{\mathcal{X}}_i[k, m]$, in Eq. (6). Accordingly, we obtain the modified complex-valued acoustic spectrum as

$$\tilde{X}[n, k] = \tilde{X}_r[n, k] + j\tilde{X}_i[n, k]. \quad (7)$$

At the final stage, the processing is the same as in the case of MFCC extraction: the magnitude of the modified acoustic spectrum $\{\tilde{X}[n, k]\}$ in Eq. (7) associated with each frame is weighted by a Mel-frequency filter bank, and the nonlinear compression is achieved by using the logarithmic operation. Lastly, the less correlated MFCC features are derived after the application of the discrete cosine transform (DCT).

Because the main idea of the aforementioned framework is to perform HEQ on the modulation domain of the acoustic spectrum, we will use the short-hand notation ‘‘MAS-HEQ’’ to denote it hereafter.

The MAS-HEQ framework has two remarkable characteristics:

1. In MAS-HEQ, the real and imaginary acoustic spectra are processed individually, which helps to enhance the magnitude and phase parts simultaneously. Note that the modulation-domain HEQ process cannot operate on the magnitude acoustic spectra directly (that is, to perform HEQ on the magnitude modulation spectrum of $|X[n, k]|$ in Eq. (2)) because the resulting new magnitude acoustic spectra are real-valued, but not necessarily nonnegative.
2. As for the comparison between MAS-HEQ and the well-practiced CHN (HEQ performing on the cepstral time series), MAS-HEQ focuses on equalizing the distribution of the data at different modulation frequencies, while CHN equalizes the distribution of the data at different time indices. Furthermore, MAS-HEQ bears some resemblance to the SHE technique [14] since both of them are operated on the modulation domain. However, the modulation spectrum processed by SHE is the DFT of the cepstra rather than the DFT of the acoustic spectra.

In this paper, we also leverage a polynomial-fitting scheme (denoted by PHEQ) [10] to efficiently approximate the inverse CDFs, F_T^{-1} and $F_{T_c}^{-1}$ in Eq. (5), to work in concert with the presented MAS-HEQ. PHEQ provides the advantages of lower storage and time consumption when compared with the existing HEQ methods. It makes effective use of data fitting (or so-called least squares error regression) to estimate the inverse CDFs of the training data.

The notion of processing real and imaginary components of acoustic spectra in the modulation domain for speech enhancement has been investigated recently [16]. However, to our knowledge, there is still a dearth of work investigating the effectiveness of normalizing the real and imaginary components of acoustic spectra in the modulation domain for speech recognition. As will be shown in Section 5, such a joint normalization paradigm shows promise and performs quite well.

4. Experimental Setup

The speech recognition experiments were conducted under various noise conditions using the Aurora-2 database and task [17]. The Aurora-2 database is a subset of the TI-DIGITS, which contains a set of connected digit utterances spoken in English; while the task consists of the recognition of the connected digit utterances interfered with various noise sources at different signal-to-noise ratios (SNRs), in which the Test Sets A and B are artificially contaminated with eight different types of real world noises (e.g., the subway noise, street noise, etc.) in a wide range of SNRs (-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB and Clean) and the Test Set C additionally includes the channel distortion.

As for the baseline experiment, each utterance of the training and testing sets were converted to a series of 39-dim MFCC feature vectors (c_0, c_1-c_{12} plus their delta and delta-delta). The frame length and shift were set to 25 ms and 10 ms, respectively. In particular, each of the robustness algorithms to be evaluated is to produce the 13 static cepstra (c_0, c_1-c_{12}) only, and then the 26 dynamic cepstra are computed accordingly.

More specifically, the acoustic model for each digit was a left-to-right continuous density HMM with 16 states, and each

state has a 20-mixture diagonal GMM. The training and recognition tests used the HTK recognition toolkit [18], which followed the setup originally defined for the ETSI evaluations. All the experimental results reported below are based on clean-condition training, i.e., the acoustic models were trained only with the clean (uncontaminated) training utterances.

5. Experimental Results

At the outset, we evaluate the utility of MAS-HEQ in terms of recognition accuracy. For the purpose of comparison, the results of some well-known feature robustness methods are also reported here. These methods are roughly divided into two categories depending on the feature type to be adjusted directly:

1. Acoustic spectrum processing methods: ETSI advanced front-end (AFE) [19], MMSE-based log-spectral amplitude estimation (MMSE log-STSA) [6], Wiener filtering (WF) based on a priori signal-to-noise-ratio estimation [5] and two versions of spectral subtraction (SS) [3,4], denoted by SS_{Boll} and SS_{Berouti} for short, respectively, in which the author names are represented by the subscripts.
2. Cepstrum processing methods: cepstral mean normalization (CMN) [7], cepstral mean and variance normalization (CMVN) [8], cepstral histogram normalization (CHN) [2], cepstral gain normalization (CGN) [20], CMVN plus ARMA filtering (MVA) [13], spectral histogram equalization (SHE) [14] and temporal structure normalization (TSN) [12].

In particular, we additionally perform CMN on the cepstral features produced by any of the acoustic spectrum processing methods, including the presented MAS-HEQ. Note that the CMN procedure has been inherently embedded in all of the cepstrum processing methods.

Table 1 shows the recognition accuracy rates for the various methods, from which we notice several particularities:

1. It comes as no surprise that every method can give rise to significant improvement in recognition rates for all the three test sets as compared to the MFCC baseline. The simple CMN process can achieve a relative error rate reduction of 32.12%, and all the other methods that integrate the CMN process produce even better results relative to CMN alone.
2. As for the cepstrum processing methods, SHE behaves the best, followed by TSN, MVA, CHN, CGN and then CMVN. There are several noteworthy points. First, CHN outperforms CMVN due to its further normalization on the statistical moments higher than the second order. Second, the constraint of unity dynamic range for CGN eliminates the outliers in the resulting data and makes it behaves as well as CHN. Next, MVA explicitly enhances the low time-varying components of CMVN features with a fixed ARMA filter and performs very well, while TSN, which employs a data-driven temporal filter, produces better results than MVA. Finally, the better outcome of SHE compared with CHN implies histogram equalization (HEQ) conducted in the modulation domain of cepstra provides superior robustness than in the temporal domain.
3. Regarding the acoustic spectrum processing methods, WF and two variants of SS behave less effective than the other spectral-domain methods possibly because they are initially designed for speech enhancement. MMSE log-STSA performs specifically well for Set C and the corresponding overall results are close to the best possible ones achieved

by cepstrum processing methods. The cepstra derived from the well-known AFE without further CMN processing achieves an accuracy rate of 87.17%, higher than those obtained by any other methods discussed before. Nevertheless, CMN is not well additive to AFE probably due to the effect of over-normalization on the AFE-derived features. Finally, the presented MAS-HEQ turns out to be the best-performing one among all of the tested methods in terms of the overall averaged recognition accuracy. Compared with AFE, MAS-HEQ is better for Test Set B and worse for Test Sets A and C. In brief, MAS-HEQ shows excellent performance in creating noise-robust speech features.

- MAS-HEQ outperforms SHE consistently over different Test Sets, and on average, the respective accuracy improvement is around 3%. These results indicate that when considering the effectiveness of processing speech features in modulation domain via HEQ, the acoustic spectra seem to be a better choice than the cepstra. However, MAS-HEQ is less efficient than SHE in implementation since the number of the (discrete) acoustic spectra in MAS-HEQ is larger than that of the cepstra in SHE.

Apart from recognition performance, we also examine the presented MAS-HEQ with regard to its capability of reducing the mismatch in the power spectral density (PSD) of the cepstral sequence caused by noise. Figs. 1(a) to 1(d) depict the averaged PSD curves of the first MFCC feature c_1 for the 1001 utterances in the Test Set B of the Aurora-2 database for three SNR levels, clean, 10 dB and 0 dB (with airport noise) before and after various processes (CMN, AFE and MAS-HEQ), respectively. First, for the unprocessed case as in Fig. 1(a), it shows that the noise causes a significant PSD mismatch over the entire modulation frequency band [0, 50 Hz]. Second, by comparing Fig. 1(b) with Fig. 1(a) we find that CMN just eliminates the distortion at the DC component and provides nearly no benefit for the PSD mismatch at any other frequency (even so, CMN can bring about significant accuracy improvement, as evident in Table 1). Finally, Figs. 1(c) and 1(d) show that both AFE and the presented MAS-HEQ can considerably reduce the PSD distortion, while MAS-HEQ appears more effective than AFE to mitigate the PSD mismatch at higher frequencies. These results may partly explain why MAS-HEQ outperforms AFE for processing Test Set B, as shown in Table 1, and they also reveal that MAS-HEQ can provide a more noise-robust feature representation.

6. Conclusions

In this study, we have proposed a novel noise-robustness framework, termed MAS-HEQ, for equalization of the acoustic spectra in modulation domain. Applying histogram equalization on the magnitude parts of the DFTs for the real and imaginary acoustic spectra separately enables MAS-HEQ to reduce the noise effect effectively and refine the resulting features elaborately. The experimental results conducted on Aurora-2 demonstrate that MAS-HEQ can provide superior performance over many state-of-the-art robustness methods, including the ETSI advanced front-end (AFE). As to future work, we envisage several directions, including extending the idea of our work to process the Mel-filter smoothed (complex-valued) spectra, analyzing the possible addition of our work with more other robustness methods and further confirming our observations on larger-scale ASR experiments.

Table 1. Recognition accuracy rates (%) averaged over different noise types and different SNRs for the baseline MFCC and various robustness methods. RR (%) is the relative error rate reduction over the MFCC baseline.

	Set A	Set B	Set C	Avg	RR
MFCC baseline	54.87	48.87	63.95	54.29	-
Cepstrum processing methods					
CMN	66.81	71.79	67.64	68.97	32.12
CMVN	75.93	76.76	76.82	76.44	48.46
CHN	80.03	82.05	80.10	80.85	58.11
CGN	80.08	81.48	80.20	80.66	57.69
MVA	80.89	82.00	81.49	81.45	59.42
TSN	83.26	84.50	82.83	83.67	64.27
SHE	83.37	85.08	83.47	84.08	65.17
Acoustic spectrum processing methods					
SS _{Boll}	73.03	76.84	73.00	74.55	44.32
SS _{Berouti}	78.70	82.81	79.69	80.54	57.43
WF	79.64	81.39	80.29	80.47	57.27
MMSE log-STSA	82.96	83.95	84.60	83.68	64.30
[†] AFE ₍₁₎	87.68	87.10	86.27	87.17	71.93
[†] AFE ₍₂₎	85.53	86.59	85.47	85.94	69.24
MAS-HEQ	86.89	88.66	85.33	87.29	72.19

[†]AFE₍₁₎ denotes the original AFE, and AFE₍₂₎ denotes the pairing of AFE and CMN. Note: The CMN process is integrated with all of the methods except for AFE₍₁₎.

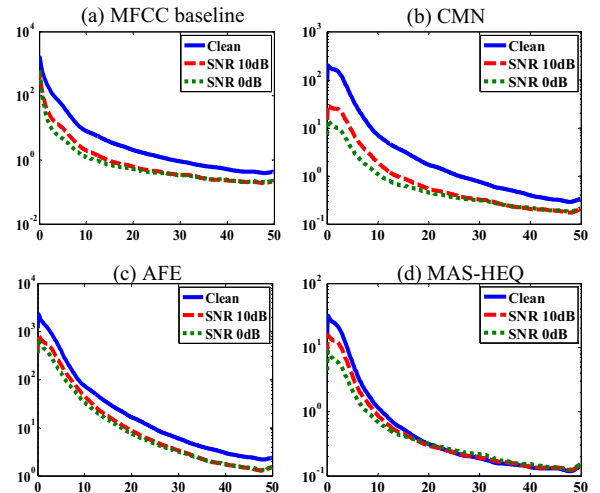


Figure 1: The MFCC c_1 PSD curves processed by various compensation methods: (a) the MFCC baseline (no compensation), (b) CMN, (c) AFE and (d) MAS-HEQ

7. Acknowledgements

This work was sponsored in part by ‘‘Aim for the Top University Plan’’ of National Taiwan Normal University and Ministry of Education, Taiwan, and the National Science Council, Taiwan, under Grants NSC 101-2221-E-003 -024 -MY3 and NSC 99 -2221-E-003 -017 -MY3.

8. References

- [1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communications*, 16, pp. 261-291, 1995.
- [2] J. Droppo and A. Acero, "Environmental robustness," in *Springer Handbook of Speech Processing*, Chapter 33, pp. 653-679, 2008.
- [3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2), pp. 113-120, 1979.
- [4] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 208-211, 1979.
- [5] C. Plapous, C. Marro and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, 14(6), pp. 2098-2108, 2006.
- [6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2), pp. 443-445, 1985.
- [7] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2), pp. 254-272, 1981.
- [8] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communications*, 25(1-3), pp. 133-147, 1998.
- [9] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3), pp. 845-854, 2006.
- [10] S. H. Lin, B. Chen and Y. M. Yeh, "Exploring the use of speech features and their corresponding distribution characteristics for robust speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, 17(1), pp. 84-94, 2009.
- [11] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, 2(4), pp. 578-589, 1994.
- [12] 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*, 16(8), pp. 1662-1674, 2008.
- [13] C. P. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1), pp. 257-270, 2007.
- [14] L. C. Sun and L. S. Lee, "Modulation spectrum equalization for improved robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3), pp. 828-843, 2012.
- [15] J. W. Hung, W. H. Tu and C. C. Lai, "Improved modulation spectrum enhancement methods for robust speech recognition," *Signal Processing*, 92(11), pp. 2791-2814, 2012.
- [16] Y. Zhang and Y. Zhao, "Spectral subtraction on real and imaginary modulation spectra," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4744-4747, 2011.
- [17] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the 2000 Automatic Speech Recognition: Challenges for the new Millennium*, pp. 181-188, 2000.
- [18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, Cambridge, UK, 2006.
- [19] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouvét, H. Kelleher, D. Pearce and F. Saadoun, "Evaluation of a noise-robust DSR front-end on Aurora databases," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 17-20, 2002.
- [20] S. Yoshizawa, N. Hayasaka, N. Wada and Y. Miyanaga, "Cepstral gain normalization for noise robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 209-212, 2004.