Enhancing the Complex-valued Acoustic Spectrograms in Modulation Domain for Creating Noise-Robust Features in Speech Recognition

Hsin-Ju Hsieh[1,2], Berlin Chen[2] and Jeih-weih Hung[1]

[1]National Chi Nan University, Taiwan
[2]National Taiwan Normal University, Taiwan
s101323902@ncnu.edu.tw, berlin@ntnu.edu.tw, jwhung@ncnu.edu.tw

***Abstract*—In this paper, we propose a speech enhancement technique which compensates for the real and imaginary acoustic spectrograms separately. This technique leverages principal component analysis (PCA) to highlight the clean speech components of the modulation spectra for noise-corrupted acoustic spectrograms. By doing so, we can enhance not only the magnitude but also the phase portions of the complex-valued acoustic spectrogram, thereby creating noise-robust speech features. More particularly, the proposed technique possesses two explicit merits. First, via the operation on modulation domain, the long-term cross-time correlation among the acoustic spectrogram can be captured and subsequently employed to compensate for the spectral distortion caused by noise. Next, due to the individual processing of real and imaginary acoustic spectrograms, the proposed method will not encounter a knotty problem of speech-noise cross-term that usually exists in the conventional acoustic spectral enhancement methods especially when the noise reduction process is inevitable. All of the evaluation experiments are conducted on the Aurora-2 and Aurora-4 databases and tasks. The corresponding results demonstrate that under the clean-condition training setting, our proposed method can achieve performance competitive to or better than many widely used noise robustness methods, including the well-known advanced front-end (AFE), in speech recognition.**

I. INTRODUCTION

The performance of automatic speech recognition (ASR) systems often degrades in practical environments riddled with, among others, ambient noise and interferences caused by the recording devices and transmission channels. Such performance degradation is largely due to a mismatch between the acoustic environments for the training and testing speech data in ASR. 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 several decades. Broadly speaking, these noise/interference processing techniques may fall into three main categories [1]: speech enhancement, robust speech features extraction and acoustic model adaptation.

For speech recognition tasks, the Mel-frequency cepstral coefficients (MFCC) approach has been proven to be one of the most effective speech feature representations. The performance of MFCC is quite good under the nearly noise-free laboratory environments, but degrades apparently under the noise-corrupted environments. Therefore, MFCC often requires compensation prior to being used in real-world scenarios. One school of compensation techniques aims to explore the temporal characteristics of MFCC and then regularize the associated statistical moments for both clean and noise-corrupted situations. These techniques include cepstral mean normalization (CMN) [2], cepstral mean and variance normalization (CMVN) [3] and histogram equalization (HEQ) [4], to name but a few. Another stream of work attempts to employ filtering on the temporal sequence of MFCC to emphasize the relatively low time-varying components (except for the DC part), which encapsulates ample linguistic information cues that are part and parcel for speech recognition. Some exemplar methods of this stream include CMVN plus ARMA filtering (MVA) [5] and temporal structure normalization (TSN) [6].

More recently, the technique of deep neural networks (DNN) has been delicately adopted in developing noise robustness methods for ASR, and these methods demonstrate excellent performance under some hypothetical and specific acoustic situations. For example, in [7] a deep recurrent denoising auto encoder (DRDAE) is trained via a series of stereo (clean and noise-corrupted) data, and it helps to reconstruct the clean speech features from the noisy input. In particular, DRDAE outperforms the well-known advanced front-end feature extraction (AFE) [8] scheme under an inside-test module mostly because it employs discriminative training and explicitly learns the difference between the clean and noise-corrupted counterparts. However, DRDAE behaves worse in the outside test mainly because the characteristics of the unseen testing data are not captured very well in the training phase.

In our previous work [9], we proposed to use histogram equalization (HEQ) to compensate the modulation spectra of the real and imaginary portions of the acoustic spectrogram separately, and this process was shown to alleviate noise distortion substantially and promote recognition performance. The new scheme presented in this paper is in fact a variant and extension of the work in [9], and it adopts *principal component analysis* (PCA) [10] to highlight the major speech components in modulation domain of the complex-valued acoustic spectrogram for a speech signal. PCA is expected to reduce the relatively fast-varying anomaly in the modulation

spectrum caused by noise and thus result in noise-robust features for speech recognition. The PCA-based scheme is linear, data-driven and engages unsupervised learning since the underlying principal components together with the spanned subspace are learned by the modulation spectra of all the utterances in the clean training set, regardless of the label (acoustic content) of each utterance. We will show that, this new framework produces highly noise-robust cepstral features, and it behaves better than the HEQ-based method [9] and many state-of-the-art robustness methods.

The remainder of the paper is organized as follows: Section II briefly introduces the concept and operation of PCA. Next, the detail of the presented novel framework is described in Section III. The experimental setup is provided in Section IV, followed by a series of experiments and discussions in Section V. Finally, Section VI concludes this paper and provides some avenues for future work.

II. INTRODUCTION OF PCA

PCA [10] is one of the most celebrated methods in the field of multivariate data analysis, which performs orthogonal transformation for data. The aim of PCA is to obtain the dimension-reduced data with the minimum squared error relative to the original data. Given a real-valued data matrix $\mathbf{V} \in \mathbb{R}^{P\times M}$ with column-wise zero sample mean, where each of the M columns represents an instance (observation) of a random vector of size $P \times 1$ and $M > P$ in general, PCA finds an matrix $\mathbf{A} \in \mathbb{R}^{P\times S}$ consisting of S orthonormal column vectors ($S \leq P$) in order to minimize the difference between the original data $\mathbf{V}$ and the projected data $\mathbf{AA}^T\boldsymbol{V}$ (viz. the projection of $\mathbf{V}$ onto the subspace spanned by the columns of $\mathbf{A}$). It can be shown that the desired S orthonormal column vectors in $\mathbf{A}$ are just the eigenvectors of the covariance matrix for the data $\mathbf{V}$ with respect to the largest S eigenvalues, and these orthonormal vectors are termed the principal components of the data $\mathbf{V}$.

To recap, given a fixed number $S \leq P$, the covariance matrix of the data matrix $\mathbf{V}$, denoted by $\mathbf{C}$, is first calculated, then the matrix $\mathbf{C}$ is passed through the eigen-decomposition to obtain the S unit-length eigenvectors, $\boldsymbol{e_1}, \boldsymbol{e_2}$, ... and $\boldsymbol{e_S}$, associated with the largest S eigenvalues, arranged as the columns of a matrix $\mathbf{A}$ and thus $\mathbf{A} = [\boldsymbol{e_1}\ \boldsymbol{e_2}\ ...\ \boldsymbol{e_S}]$. Finally, the PCA-processed counterpart for each original data instance $\mathbf{v}_j$ of $\mathbf{V}$ is equal to $\mathbf{AA}^T\mathbf{v}_j$.

III. PROPOSED METHODS

This section describes a novel framework in order to create noise-robust speech features. First, in the preprocessed stage, any time-domain utterance in the training and testing sets, denoted by $\{x[\ell]\}$, is passed through a pre-emphasis filter and segmented into a series of frame signals in turn. Then, each 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

$$X[n,k] = X_r[n,k] + jX_i[n,k], \quad (1)$$

$$0 \leq n \leq N-1, 0 \leq k \leq K-1,$$

where $X_r[n,k]$ and $X_i[n,k]$ respectively denote the acoustic real and imaginary spectra, 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. As a side note, $\{X[n,k]\}$ in eq. (1) is usually referred to as the spectrogram of the utterance $\{x[\ell]\}$.

Next, the time series of acoustic real and imaginary spectra, $X_r[n,k]$ and $X_i[n,k]$, $0 \leq n \leq N-1$, in eq. (1) *with respect to any specified frequency bin* k, are updated via PCA in modulation domain, and the updating process consists of the following three steps:

Step 1: Compute the modulation spectrum

Both $X_r[n,k]$ and $X_i[n,k]$ are separately transferred to modulation domain along the n-axis by discrete Fourier transform (DFT). For simplicity, we just show the process of the real component $X_r[n,k]$ hereafter, and the imaginary component $X_i[n,k]$ is processed in the same way. The modulation spectrum of $X_r[n,k]$ is then calculated as:

$$\chi_r[k,m] = \sum_{n=0}^{N-1} X_r[n,k] e^{-j\frac{2\pi nk}{\widetilde{N}}} \quad (2)$$

$$0 \leq m \leq \widetilde{N}-1, 0 \leq k \leq K-1,$$

where m refers to the index of the discrete modulation frequency. Please note that here the DFT size, $\widetilde{N}$, is set to be no less than the number of frames, N. The modulation spectrum shown in eq. (2) can be expressed in polar form as

$$\chi_r[k,m] = A_r[k,m] e^{j\theta_r[k,m]} \quad (3)$$

where $A_r[k,m]$ is the magnitude part of $\chi_r[k,m]$ and $\theta_r[k,m]$ is the phase part of $\chi_r[k,m]$.

Step 2: Update the magnitude modulation spectrum

This step is to modify the magnitude part of the modulation spectra in eq. (3) via PCA, while keeping the phase part unchanged. The details are described as follows:

First, the magnitude modulation spectra, viz. $A_r[k,m]$ in eq. (3), of all utterances *in the training set* are arranged to be the columns of a data matrix $\mathbf{V}$. Then, following the procedures stated in section II, we obtain the matrix $\mathbf{A}$ consisting of the first S eigenvectors associated with the covariance matrix of $\mathbf{V}$. Finally, the magnitude modulation spectrum of each utterance *in both the training and testing sets* are first subtracted by the empirical mean (viz. the mean of the magnitude spectra of the training set), then mapped to the column space of $\mathbf{A}$, and added back by the empirical mean in turn, to obtain the respective PCA-processed new magnitude spectrum.

Step 3: Synthesize the acoustic spectrogram

Combining the updated magnitude part from Step 2, denoted by $\tilde{A}_r[k,m]$, with the original phase part $\theta_r[k,m]$ in eq. (3) can result in the new (complex-valued) modulation spectrum:

$$\tilde{\chi}_r[k,m] = \tilde{A}_r[k,m] e^{j\theta_r[k,m]}. \quad (4)$$

Next, performing an inverse DFT (IDFT) on $\tilde{\chi}_r[k,m]$, we obtain the updated version of the real acoustic spectrum,

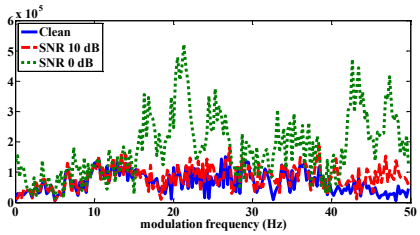

Fig. 1 The magnitude modulation spectral curves of the imaginary acoustic spectrograms at acoustic frequency 375 Hz under three SNR cases (noise type: airport) for the utterance "MFG_52Z7783A.08" in the Aurora-2 database [11].

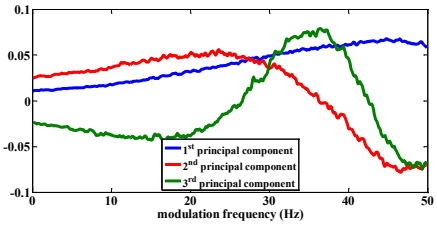

Fig. 2 The first three principal components associated with the magnitude modulation spectrum of the imaginary acoustic spectrograms at acoustic frequency 375 Hz with respect to the clean training set in the Aurora-2 database.

denoted by $\tilde{X}_r[n,k]$. Furthermore, we follow the same procedure mentioned above to achieve the updated imaginary acoustic spectrum, denoted by $\tilde{X}_i[n,k]$. Then the new complex-valued acoustic spectrum can be obtained as:

$$\tilde{X}[n,k] = \tilde{X}_r[n,k] + j\tilde{X}_i[n,k]. \quad (5)$$

At the final stage, we convert the revised acoustic spectrogram $\{\tilde{X}[n,k]\}$ in eq. (5) to a time series of MFCC features. More specifically, the magnitude of $\tilde{X}[n,k]$ associated with each frame is weighted by a Mel-frequency filter bank, and then compressed nonlinearly via the logarithmic operation. The resulting log-spectrum is further converted via DCT to obtain MFCC features.

Because the main idea of the above framework is to perform PCA on the ***m***odulation domain of the ***a***coustic ***s***pectrum, we will use the short-hand notation "MAS-PCA" to denote the new method hereafter.

Some characteristics of MAS-PCA are as follows:

1. MAS-PCA can revise both the magnitude and phase components of the acoustic spectrograms, while the conventional speech enhancement methods, such as spectral subtraction (SS) and Wiener filtering (WF), deal with the magnitude component only.

2. In general, one defect of PCA is that it is quite sensitive to the outliers of the training set which usually come from the noise inferences. However, this defect does not occur apparently in the proposed MAS-PCA, since the training set that builds the eigenvectors consists of noise-free clean utterances only. The experimental results shown in Section V will also show that MAS-PCA achieves very promising noise robustness.

3. MAS-PCA aims to reduce the relatively fast and large oscillating behavior in the magnitude modulation spectrum caused by noise. To show this, Fig. 1 depicts the magnitude modulation spectral curves at a specific acoustic frequency for an utterance distorted at three SNR levels, and Fig. 2 contains the curves for the first three principal components derived from MAS-PCA associated with the modulation spectrum in Fig. 1. From Fig. 1, we find the 0dB-SNR curve contains larger and sharper fluctuations than the clean noise-free one, and this mismatch can be reduced by the PCA mapping process since the principal components shown in Fig. 2 are rather smooth and slow-varying along the modulation frequency axis.

IV. Experimental Setup

The efficacy of the proposed MAS-PCA method was evaluated on the noisy Aurora-2 [11] and Aurora-4 [12] databases. Aurora-2 is a subset of the TI-DIGITS, and the associated task is to recognize *connected digit utterances* interfered with various noise sources at different signal-to-noise ratios (SNRs).

Compared with Aurora-2, Aurora-4 is a task of *medium to large vocabulary* continuous speech recognition based on the Wall Street Journal (WSJ) database, consisting of clean speech utterances interfered with various noise sources at different SNR levels. In Aurora-4, speech utterances were sampled in both 8 kHz and 16 kHz, while only the 8-kHz sampled utterances were used for our experiments. In particular, there are six noisy environments and one clean environment considered for the evaluation in Aurora-4.

Furthermore, the acoustic model for each digit in the Aurora-2 task was set to a left-to-right continuous density HMM with 16 states, each of which is a 20-mixture GMM. As to the Aurora-4 database, the acoustic model set consisted of state-tied intra-word triphone models, each had 5 states and 16 Gaussian mixtures per state.

In regard to speech feature extraction, each utterance of the training and testing sets was represented by a series of 13 static features (including the zeroth cepstral coefficient) augmented with their delta and delta-delta coefficients, making a 39-dimensional MFCC feature vector. The training and recognition tests used the HTK recognition toolkit [13], 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 with the clean (noise-free) training utterances

V. Experimental Results

At the commencement of this section, the presented MAS-PCA is appraised on the Aurora-2 task in terms of recognition accuracy rates, which are shown in Table I. The number of eigenvectors used in MAS-PCA is varied, and it is labeled in the bracket right after the term "MAS-PCA". For example, MAS-PCA(10) indicates the MAS-PCA method using 10 principal components. Besides, for each MAS-PCA instantiation with different assignments of the number for the eigenvectors, we create the corresponding speech features at the training and testing sets. The new speech features in the training set are then used to rebuild the acoustic models

(HMMs) specific to that instantiation of MAS-PCA for the sub-sequent recognition on the testing set.

For comparison, Table I further contains the results of several well-known feature robustness methods. Please note that, we additionally perform CMN on the cepstral features derived from MAS-PCA, for the reason that the CMN procedure has been also inherently embedded in all of the other methods listed in Table I, except for MFCC baseline and $AFE_{(1)}$.

From Table I, some observations can be made: First, every method can give rise to significant improvements in recognition accuracy as compared to the MFCC baseline. Next, as for the cepstral processing methods, spectral histogram equalization (SHE) [14] behaves the best, followed by TSN, MVA, HEQ, CMVN and CMN. After that, the well-known AFE without further CMN processing (denoted by $AFE_{(1)}$) achieves an accuracy rate of 87.17%, higher than the results of any other aforementioned methods. Nevertheless, the results of $AFE_{(2)}$ indicate that CMN is not well additive to AFE, probably due to the over-normalization effect brought by CMN to the AFE features. In addition, our recently proposed MAS-THEQ [9] behaves better than SHE and close to AFE without further CMN processing. Lastly, the results of MAS-PCA show that:

1. All instantiations of MAS-PCA give very promising results in recognition accuracy. All of them behave better than the cepstral processing methods. In particular, MAS-PCA with 3, 5 and 6 principal components outperforms $AFE_{(1)}$ and $AFE_{(2)}$ and MAS-THEQ.
2. The performance of MAS-PCA is improved by increasing the number of principal components from 3 to 6. However, further increasing the number of principal components more than 6 degrades MAS-PCA gradually.

To take a step forward, the effectiveness of MAS-PCA is validated on Aurora-4. The experiments are conducted on one clean test set and six noisy test sets (viz. Sets 8 to 14) of the Aurora-4 task, where each of the noisy test was interfered with both additive noise and channel distortion. The corresponding results of MFCC baseline, two forms of AFE-related methods mentioned in Table 1 and MAS-PCA demonstrated in Table 2. From this table, we have the following observations:

1. Similar to the situation shown in Table 1, the four robustness methods behave much better than the MFCC baseline for all seven Test Sets.
2. AFE followed by CMN (denoted by $AFE_{(2)}$) outperforms AFE (denoted by $AFE_{(1)}$) alone, shows that CMN can further enhance AFE in improving the recognition accuracy at Aurora-4, while this effect is not clearly shown at Table 1 for the Aurora-2 case.
3. Compared with the two AFE-related methods, MAS-PCA behaves better for four noise situations (babble, restaurant, street and airport) while they are worse for the clean noise-free condition and the other two noise situations (car and train). In average, these four methods perform very close to one another. These results confirm that MAS-PCA can provide noise-robust features to improve recognition accuracy in a large-scale speech recognition task.

TABLE I
WORD ACCURACY RATES (%) ON THE AURORA-2 TASK, ACHIEVED BY 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	-
CMN	66.81	71.79	67.64	68.97	32.12
CMVN	75.93	76.76	76.82	76.44	48.46
HEQ	80.03	82.05	80.10	80.85	58.11
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
†$AFE_{(1)}$	87.68	87.10	86.27	87.17	71.93
†$AFE_{(2)}$	85.53	86.59	85.47	85.94	69.24
MAS-THEQ	86.49	88.13	84.98	86.84	71.21
MAS-PCA(3)	87.04	88.68	85.12	87.31	72.24
MAS-PCA(5)	87.16	88.69	85.50	87.44	72.52
MAS-PCA(6)	87.31	88.80	85.93	87.63	72.94
MAS-PCA(9)	86.76	88.19	85.13	87.01	71.58
MAS-PCA(10)	86.69	88.11	85.03	86.92	71.38
MAS-PCA(15)	86.11	87.72	84.56	86.44	70.33

†$AFE_{(1)}$ 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 $AFE_{(1)}$.

TABLE II
WORD ACCURACY RATES (%) ON THE AURORA-4 TASK, ACHIEVED BY BASELINE MFCC AND VARIOUS ROBUSTNESS METHODS.

	MFCC	$AFE_{(1)}$	$AFE_{(2)}$	MAS-PCA(5)
Clean	63.45	79.34	80.77	77.16
Car	37.27	72.56	74.77	69.65
Babble	30.31	61.58	61.07	62.43
Rest.	34.30	55.65	55.29	60.04
Street	26.13	58.31	57.38	59.01
Airport	31.85	60.55	60.85	64.75
Train	26.95	60.88	59.74	60.15
Avg.	35.75	64.12	64.27	64.74

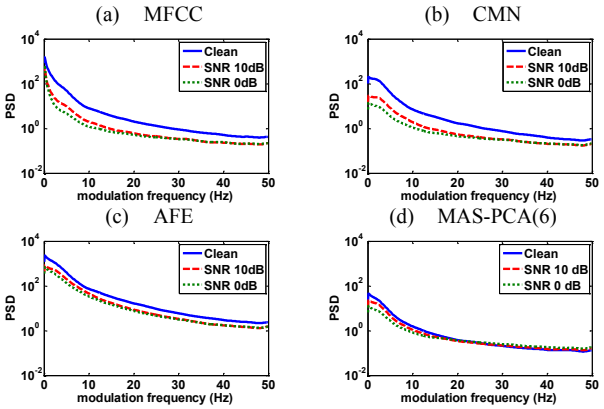

Fig. 3 The MFCC c1 PSD curves processed by various compensation methods: (a) the MFCC baseline (no compensation), (b) CMN, (c) AFE and (d) MAS-PCA(6).

Lastly, we examine the proposed method by the capability of reducing the cepstral modulation spectrum distortion caused by noise. Figs. 3(a) to 3(d) depict the averaged power spectral density (PSD) curves of the first MFCC feature c1 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 CMN, AFE and MAS-PCA(6), respectively. First, for the unprocessed case, as shown in Fig. 3(a), the environmental noise results in a significant PSD mismatch over the entire frequency range [0 50 Hz]. Second, from Figs. 3(b) to 3(d), we see that the PSD mismatch can be considerably suppressed after performing any of the three methods, CMN, AFE and MAS-PCA(6). As a result, MAS-PCA is shown to be effective in producing noise-robust cepstral features.

VI. Conclusions

In this paper, we presented a novel use of PCA for enhancing the complex-valued acoustic spectrograms of speech signals in modulation domain for noise-robust speech recognition. Different from the state-of-the-art deep neural network schemes, the proposed framework does not adopt any prior knowledge of the actual distortions caused by noise, while it still behaves quite well when evaluated in unseen noise environments. As to future work, we will explore the possible addition of our work with other robustness methods to further enhance the speech features.

References

[1] J. Droppo and A. Acero, "Environmental robustness," *in Springer Handbook of Speech Processing*, Chapter 33, pp. 653-679, 2008.

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2), pp. 254-272, 1981.

[3] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133-147, 1998.

[4] A. Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Bentez and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 355-366, 2005.

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

[6] 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.

[7] A. L. Maas, Q. V. Le, T. M. ONeil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," *in Proc. Interspeech*, 2012.

[8] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouvet, 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.

[9] H. J. Hsieh, B. Chen, J. W. Hung, "Histogram equalization of real and imaginary modulation spectra for noise-robust speech recognition," *in Proc. Interspeech*, 2013.

[10] C. Bishop, "Pattern Recognition and Machine Learning" *Springer*, 2007

[11] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *in Proc. ICSA ITRW ASR*, pp. 181-188, 2000.

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

[13] 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.

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