Modulation Spectrum Factorization for Robust Speech Recognition

Wen-Yi Chu*, Jeih-weih Hung†, and Berlin Chen*
* National Taiwan Normal University, Taipei
E-mail: berlin@csie.ntnu.edu.tw
† National Chi Nan University, Nantou
E-mail: jwhung@ncnu.edu.tw

Abstract— **This paper presents a novel approach to improving the noise robustness of speech features built on top of nonnegative matrix factorization (NMF). To do this, we employ NMF to extract a common set of basis spectral vectors that cover the intrinsic temporal structure inherent in the modulation spectra of clean training speech features. The new modulation spectra of the speech features, constructed by mapping the original modulation spectra into the space spanned by these basis vectors, are demonstrated with good noise-robust capabilities. All experiments were conducted using the Aurora-2 database and task. The results show that the proposed NMF-based approach, together with mean and variance normalization (MVN), can provide average error reduction rates of over 65% and 12% relative as compared with the baseline MFCC system and that using the MVN method alone, respectively.**

I. Introduction

The environmental mismatch caused by additive noise and/or channel distortion often degrades the performance of a speech recognition system seriously. Various robustness methods have been proposed to reduce this mismatch, and one prevalent school of thought aims to refine the modulation spectra of the speech feature sequence. It has been shown in [1] that different modulation frequency components have unequal importance for speech recognition, and most of the useful linguistic information is encapsulated in the modulation frequency components between 1 Hz and 16 Hz, with the dominant component centering around 4 Hz. Accordingly, a number of celebrated temporal processing methods have been proposed to highlight these important frequency components, either explicitly or implicitly, for robust speech recognition. They include, but are not limited to, RelAtive SpecTra (RASTA) [2], mean and variance normalization (MVN) [3] and a series of data-driven temporal filtering methods [4][5].

In this paper, we investigate a novel use of the nonnegative matrix factorization (NMF) [6-8] to learn a parts-based representation of the magnitude modulation spectrum of speech features. NMF is a recently developed method for finding a linear and non-subtractive combination scheme to extract important ingredients that can correspond better with the intuitive notion of the parts of the original data. Compared with the other linear representation methods like principal component analysis (PCA) and independent component analysis (ICA), to name a few, NMF provides nonnegative basis vectors and ensures that the projection of any (nonnegative) data on each basis vector is also nonnegative. Apart from that, the basis vectors obtained by NMF are often sparse and localized. Consequently, NMF appears ideally suited for the purpose of analyzing magnitude modulation spectrum of speech features, which is always non-negative and often possesses a relatively narrow bandwidth. However, as far as we are aware, there is still not much research on leveraging NMF-like methods for analyzing the magnitude modulation spectrum of speech features.

By treating the magnitude modulation spectra of various clean feature sequences as the analyzed data for NMF, we obtain the basis spectral vectors to span a subspace for the magnitude modulation spectrum. Then any clean or noise-corrupted feature sequence is updated in modulation spectrum, in which the magnitude part is replaced by its mapping on the NMF subspace mentioned above. Experiments conducted on the Aurora-2 database show that the updated features via NMF maintain high recognition accuracy for the matched clean condition, and they provide significant accuracy improvements relative to the original features under mismatched noisy conditions. As such, the proposed NMF-based approach for updating the modulation spectrum promotes the noise robustness of speech features.

The remainder of the paper is organized as follows: Section II briefly introduces the principle underlying NMF. Next, we describe the proposed NMF-based modulation spectrum update procedure in Section III. The experimental setup is described in Section IV, followed by a series of experiments and discussions in Section V. Finally, Section VI concludes this paper and discusses avenues for future work.

II. Nonnegative Matrix Factorization

The nonnegative matrix factorization (NMF) is a subspace method that approximates data with an additive and linear combination of nonnegative components. Given a nonnegative data matrix $\mathrm{V} \in \mathbf{R}^{L \times M}$, NMF computes another

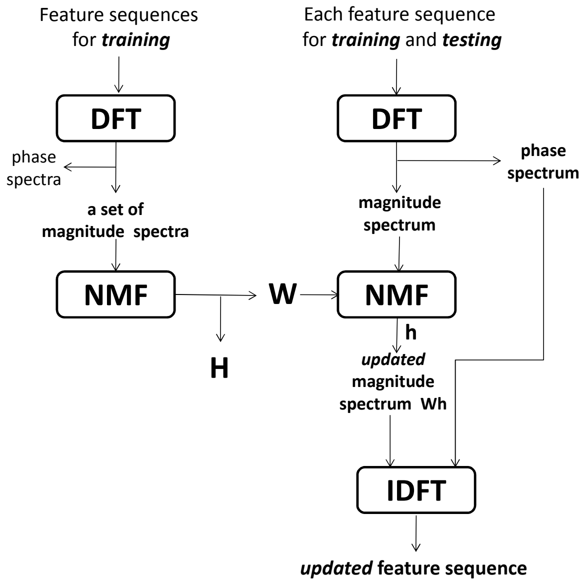

Feature sequences for *training*
DFT
phase spectra
a set of magnitude spectra
NMF
H
W
Each feature sequence for *training* and *testing*
DFT
phase spectrum
magnitude spectrum
NMF
h
updated magnitude spectrum Wh
IDFT
updated feature sequence

Figure 1. The flowchart of the proposed NMF-based approach for updating the modulation spectrum of features.

two nonnegative matrices $\mathrm{W} \in \mathbf{R}^{L\times r}$ and $\mathrm{H} \in \mathbf{R}^{r\times M}$ such that

$$\mathrm{V} \approx \mathrm{WH}. \tag{1}$$

The r columns of W are called basis vectors, and each column of H is often called an encoding, which consists of the coefficients by which the data vector (the column in V) is approximated with a linear combination of basis vectors. The number of basis vectors, r, is often chosen to be fewer than L (the size of each data vector) and M (the total number of data vectors), and thus the product WH is regarded as a compressed form of V. The smaller the rank of V, the better the approximation in (1).

To find an approximate factorization as in (1), we need to define a cost function that quantifies the quality of the approximation. In this paper, the cost function is defined as

$$L = \sum_{i,\mu}\left(\mathrm{V}_{i,\mu} - (\mathrm{WH})_{i,\mu}\right)^2. \tag{2}$$

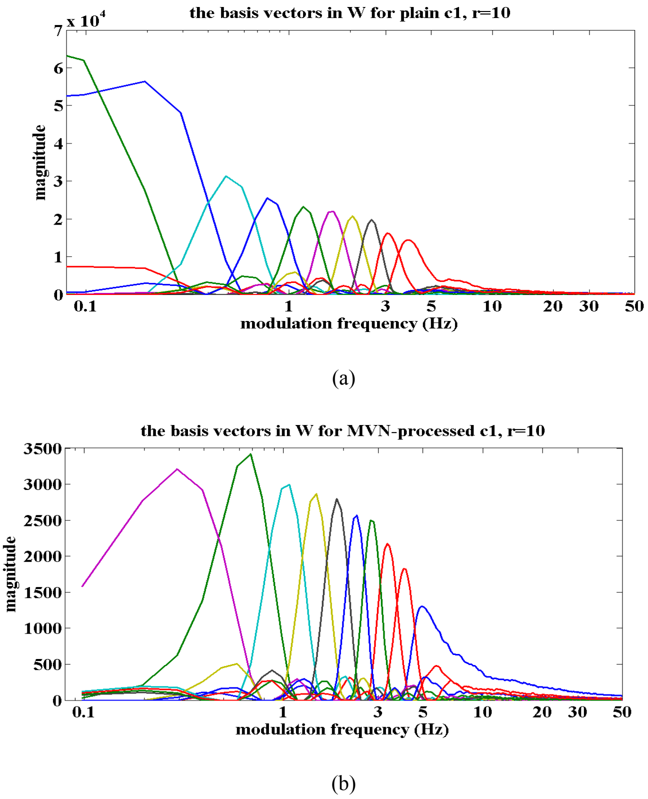
With an initial guess of W and H, the following multiplicative updating rule (see [7] for details) is employed to achieve a local minimum of (2).

$$\mathrm{H}_{a\mu} \leftarrow \mathrm{H}_{a\mu}\frac{\left(\mathrm{W}^T\mathrm{V}\right)_{a\mu}}{\left(\mathrm{W}^T\mathrm{WH}\right)_{a\mu}} \tag{3}$$

and

$$\mathrm{W}_{ia} \leftarrow \mathrm{W}_{ia}\frac{\left(\mathrm{VH}^T\right)_{ia}}{\left(\mathrm{WHH}^T\right)_{ia}}. \tag{4}$$

In general, W and H are further normalized so that the rows of H have unit length.



(a)

(b)

Figure 2. Ten basis spectra learned by NMF from (a) the plain unprocessed $c1$ features (b) the MVN-processed $c1$ features.

III. Updating the Modulation Spectrum

Assume that $\{\mathrm{x}[n]\}$ represents the ordered sequence of feature vectors of an utterance, and $\{x_m[n]\}$ denotes the m^{th} feature channel of $\{\mathrm{x}[n]\}$. Then the discrete Fourier transform (DFT) of the time sequence $\{x_m[n]\}$, denoted by $\{X_m[k]\}$, is often referred to as the modulation spectrum of the utterance (with respect to the feature channel). In this paper, we propose to update the *magnitude part* of $\{X_m[k]\}$, while to keep the phase part unchanged. For the sake of compact notation, we hereafter omit the subscript m, unless otherwise stated.

The procedures to perform the magnitude update via NMF are depicted in Figure 1, and can be generally described as follows. First, the time sequence $\{x[n]\}$ for each utterance in the training set is converted to its spectrum $\{X[k]\}$ via a $2L$-point DFT. Since the property of conjugate symmetry, only the first L+1 points of $\{X[k]\}$ is reserved, and their magnitude parts (which are always nonnegative) form each column of the data matrix V. Accordingly, if the training set consists of M utterances, then V has M columns. Given the data matrix V and a chosen number r, we obtain the two nonnegative matrices W and H, as shown earlier in (2), using NMF.

Next, the (L+1)-point (magnitude) modulation spectra of each utterance in the training and testing sets, denoted by a vector V, is factorized via NMF, i.e., $\mathrm{V} \approx \mathrm{Wh}$, given that W is fixed. The fixed W comes directly from the previous

TABLE I
A SUMMARY OF THE AURORA-2 TASK.

Training Set		(Clean-Condition Training) Training Utterances: 8,440 Channel Effect: G.712
Test Set	Set A	Test Utterances: 28,028 Additive Noses: Subway, Babble, Car, Exhibition SNRs: Clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB Channel Effect: G.712
	Set B	Test Utterances: 28,028 Additive Noses: Restaurant, Street, Airport, Train Station SNRs: Clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB Channel Effect: G.712
	Set C	Test Utterances: 14,014 Additive Noses: Subway, Street SNRs: Clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB Channel Effect: MIRS

TABLE II
RECOGNITION ACCURACY (%) ACHIEVED BY THE NMF-BASED APPROACH CONDUCTED ON THE ORIGINAL MFCC FEATURES, AVERAGED ACROSS THE TEN NOISE TYPES. r IS THE NUMBER OF BASIS SPECTRA USED IN NMF. RR (%) IS THE RELATIVE ERROR RATE REDUCTIONS OVER THE BASELINE.

Method		clean	Noisy (SNR: 20 dB~0 dB)	RR
MFCC baseline		99.79	72.07	—
NMF	r = 5	99.61	84.65	45.04
	r = 10	99.64	84.11	43.11
	r = 15	99.69	83.87	42.25
	r = 20	99.68	83.04	39.24

TABLE III
RECOGNITION ACCURACY (%) ACHIEVED BY MVN AND THE NMF-BASED METHOD PROCESSED ON THE MVN-PROCESSED MFCC FEATURES, AVERAGED ACROSS THE TEN NOISE TYPES. r IS THE NUMBER OF BASIS SPECTRA USED IN NMF. RR$_1$ (%) AND RR$_2$ (%) ARE THE RELATIVE ERROR RATE REDUCTIONS OVER THE MFCC BASELINE AND MVN, RESPECTIVELY.

Method		clean	Noisy (SNR: 20 dB~0 dB)	RR$_1$	RR$_2$
MFCC baseline		99.79	72.07	—	—
MVN		99.83	88.82	59.97	—
NMF+ MVN	r = 5	99.69	90.42	65.70	14.31
	r = 10	99.73	90.57	66.24	15.65
	r = 15	99.78	90.60	66.34	15.92
	r = 20	99.76	90.52	66.06	15.21

step, and the encoding vector h can be obtained via the updating rule (cf. (3)). The vector $\tilde{\mathrm{V}} \approx \mathrm{Wh}$ is a linear combination of the basis vectors involved in W, which is created via the clean utterances. Therefore we expect that the vector $\tilde{\mathrm{V}}$, representing the new magnitude spectrum, can highlight the important information for speech recognition and alleviate the effect of noise from the original V.

Finally, a $2L$-point inverse DFT is performed on the new modulation spectrum (with the conjugate symmetric last-half part being appended), which consists of the *updated* magnitude parts and the original phase parts, to obtain the new time sequence.

Some details about the NMF-based approach mentioned above are listed below:

(1) The length of the time sequence (i.e., the number of frames) varies from utterance to utterance. However, here the DFT-size $2L$ is set to be constant, which results in the same length of modulation spectra for different utterances. In addition, the value of $2L$ is assigned to be greater than the length of each utterance to be processed.

(2) The $2L$-point inverse DFT for the updated modulation spectrum produces a new time sequence with $2L$ in length. However, only the first N points of this sequence are reserved, where N is the length of the original time sequence.

Figures 2(a) and 2(b), respectively, depict the NMF basis spectra for the modulation spectrum of the original $c1$ (the first MFCC feature) and the MVN-processed $c1$, corresponding to the clean training set of the Aurora-2 database [9]. Consulting Fig. 2(a) and Fig. 2(b) we notice three particularities. First, the basis spectra shown in the two sub-figures reveal localized and sparse characteristics, which coincide with the fact that NMF often learns a parts-based representation of data. Next, these basis spectra are primarily located in the frequency region below 10 Hz, and thus they, to some extent, can capture or emphasize the lower modulation frequency components of the speech features, which have been shown to correspond to important linguistic information essential for speech recognition. Finally, one significant difference between the two sub-figures is that there is no "low-pass" basis spectrum in Fig. 1(b), reflecting the effect of removing the near-DC components made by MVN.

IV. EXPERIMENTAL SETUP

The proposed NMF-based method has been tested on the Aurora-2 database [9], which contains a database designed to evaluate the performance of speech recognition algorithms in noisy conditions. For the recognition environment, three different subsets are defined: Test Sets A and B are each affected by four types of noise, and Test Set C is affected by two types. Each noise instance is added to the clean speech at six SNR levels (ranging from 20 dB to -5 dB). Each utterance in the clean training set and three noise-corrupted testing sets is first converted into a sequence of 39-dimensional feature vectors (MFCC, $c0$-$c12$, plus their first and second order derivatives). Next, following the procedures described in Section III, we update the modulation spectrum of each feature sequence of each utterance in both the training and testing sets. The DFT size $2L$ is set to 1,024, and the number of basis vectors, r, is varied from 5 to 20, with an interval of 5. With these new feature vectors in the clean training set, the

Table IV

Detailed word recognition accuracy (%) and relative WER reduction (%) as compared to the MFCC baseline for the NMF-based approach ($r = 5$) conducted on the original MFCC features and the NMF-based approach ($r = 15$) conducted on the MVN-processed MFCC features at different SNR values but averaged over all the noise types in test set A of the Aurora-2 database.

Test	Method	Clean	20dB	15dB	10dB	5dB	0dB	Avg. (0-20dB)	Relative WER Reduction
Set A	MFCC	99.80	96.29	89.33	75.41	57.17	44.11	72.46	
	NMF	99.62	98.00	95.63	89.65	77.26	58.44	83.80	41.18
	NMF+CMVN	99.78	98.94	97.74	94.75	87.58	72.53	90.31	64.81

Table V

Detailed word recognition accuracy (%) and relative WER reduction (%) as compared to the MFCC baseline for the NMF-based approach ($r = 5$) conducted on the original MFCC features and the NMF-based approach ($r = 15$) conducted on the MVN-processed MFCC features at different SNR values but averaged over all the noise types in test set B of the Aurora-2 database.

Test	Method	Clean	20dB	15dB	10dB	5dB	0dB	Avg. (0-20dB)	Relative WER Reduction
Set B	MFCC	99.80	93.81	84.12	69.38	52.86	41.38	68.31	
	NMF	99.62	98.24	96.64	92.02	80.30	61.11	85.66	54.74
	NMF+CMVN	99.78	99.14	98.20	95.77	88.95	74.11	91.23	72.33

Table VI

Detailed word recognition accuracy (%) and relative WER reduction (%) as compared to the MFCC baseline for the NMF-based approach ($r = 5$) conducted on the original MFCC features and the NMF-based approach ($r = 15$) conducted on the MVN-processed MFCC features at different SNR values but averaged over all the noise types in test set C of the Aurora-2 database.

Test	Method	Clean	20dB	15dB	10dB	5dB	0dB	Avg. (0-20dB)	Relative WER Reduction
Set C	MFCC	99.77	97.34	93.63	84.19	67.64	51.32	78.82	
	NMF	99.60	97.78	95.19	89.47	78.67	60.72	84.36	26.16
	NMF+CMVN	99.76	98.87	97.52	94.48	87.45	71.34	89.99	52.74

HMMs for each word (digit) and silence are trained, following the Microsoft complex back-end training scripts [10]. Each HMM has 16 states and 20 Gaussian mixtures per state. Table I gives a summary of the Aurora-2 task.

V. Experimental Results and Discussions

In the first set of experiments, we compare the recognition performance of the MFCC baseline and the proposed NMF-based approach conducted on MFCC. The corresponding results are shown in Table II, from which several observations can be made:

(1) Under the matched clean condition, NMF slightly worsens the recognition accuracy compared to the MFCC baseline. However, the accuracy degradation is relatively insignificant (0.17% and 0.11% for the cases $r = 5$ and $r = 20$, respectively), which clearly confirm our intuition that NMF provides a very efficient coding for the magnitude (modulation) spectrum since using only a small number of basis vectors can preserve most discriminative information for recognition.

(2) For mismatched noisy conditions, NMF behaves better than MFCC obviously, and the optimal accuracy improvement provided by NMF with $r = 5$ is 12.58%, which corresponds to 45.04% in relative error reduction. Therefore, the results reveal that MFCC is enhanced in noise robustness via NMF. Looking at the basis vectors shown in Fig. 1, the higher modulation spectral components that probably correspond to non-speech distortions are attenuated, and thus NMF gives rise to less noise-contaminated MFCCs.

(3) In contrast to the clean case, increasing the value of r in NMF does not always provide better accuracy rates for noisy conditions. However, the differences among the accuracy rates obtained with different r are in fact relatively slight, and the maximum accuracy degradation is just1.61% (from 84.65% to 83.04%).

In the next of experiments, we carry out the NMF-based approach on the MVN-processed MFCC features. The MVN preprocessing normalizes the first- and second-order statistics of the feature sequence and is very helpful in reducing the effect of noise. Here we are interested to investigate if NMF can provide MVN with further improvement in recognition accuracy. The corresponding recognition results are listed in Table III. First, all the methods appear to be on par with each other for the clean case. Next, MVN outperforms the MFCC baseline significantly in noisy conditions as expected. The corresponding accuracy improvement is about 16.75%. Finally, the pairing of NMF with MVN can further promote the recognition accuracy relative to MVN alone, and the improved performance is quite similar for different assignments of the parameter r in NMF. These results agree with our previous observations in Table II that a small number of basis vectors in NMF suffice to give a robust feature representation. For easier comparison, the detailed recognition results of the MFCC baseline, and the NMF-based approach on the MVN-processed MFCC features for Test Sets A, B and C are also reported in Tables IV, V and VI, respectively.

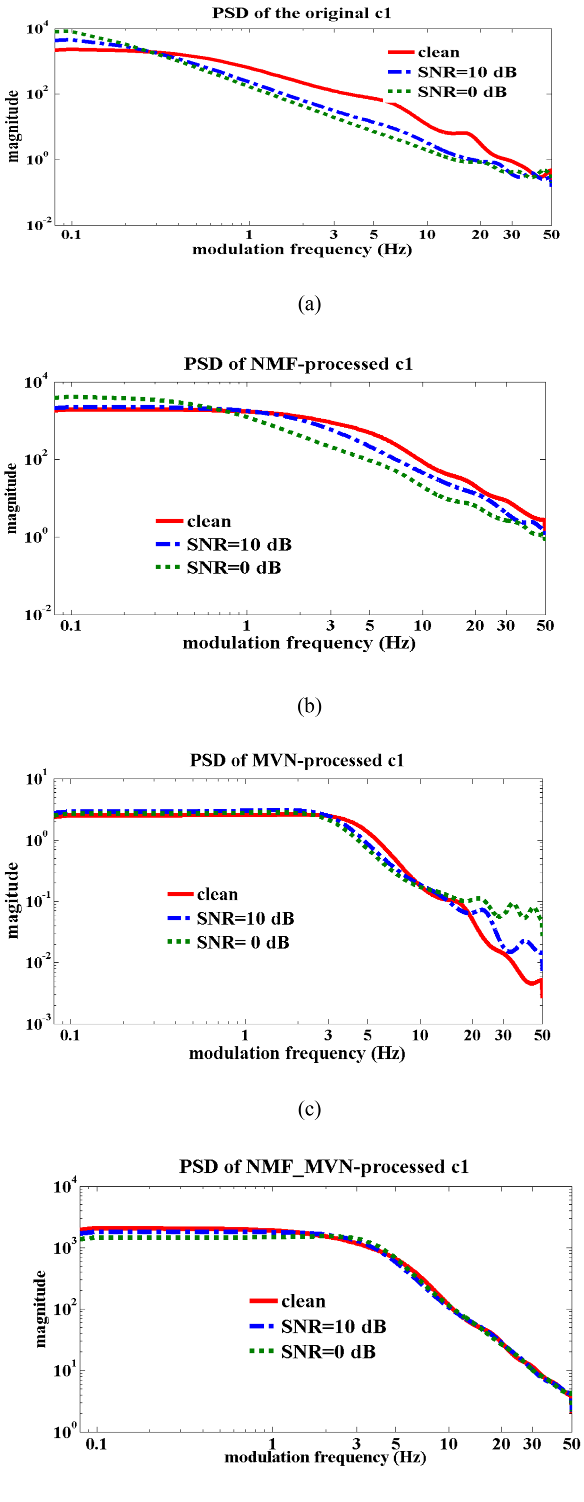
As a final point, in addition to the recognition accuracy, we examine NMF by the capability of reducing the modulation spectrum distortion caused by noise. Fig. 3(a)-(d) show the power spectral density (PSD) curves of the first MFCC feature $c1$ of an utterance (the file "MAR_5376869A.08" in the Aurora-2 database) for three SNR levels, clean, 10 dB and 0 dB (with subway noise), before and after various processes, respectively. First, for the unprocessed case as in Fig. 3(a), it shows that the additive noise results in a significant PSD mismatch over the entire frequency range [0 50 Hz]. Second, from Fig. 3(b), we see that NMF conducted on the original $c1$ can reduce the PSD mismatch. However, the mismatch reduction is less pronounced for the low SNR case. Third, Fig. 3(c) shows that MVN can effectively normalize the PSDs under different SNR conditions, while it seems to provide no significant benefit for reducing the PSD mismatch located in the higher frequency region above 10 Hz. Finally, Fig. 3(d) shows that the pairing of NMF with MVN can considerably reduce the PSD distortion over the entire band. Therefore, the above observations again imply that, the proposed NMF-based approach can provide a more noise-robust feature representation, and it can be conducted additively to MVN to reduce the effect of noise further.

VI. Conclusion and Future Work

We have presented a novel use of NMF for deriving noise-robust speech features, showing that the basis spectra constructed on top of NMF correspond well with the intuitive notion of the important components (or parts) of modulation frequency. This NMF-based method conducted on the conventional MFCC features can yield an accuracy improvement of up to 12% absolute on average over all test conditions of the Aurora-2 task. Besides, incorporating the NMF-based method with MVN behaves better than MVN alone, and has the added advantage of offering an extra improvement of 2% absolute. As to future work, we envisage the following two directions. First, we will explore whether further normalizing the encoding vector h in the mapping process of NMF can bring better recognition accuracy. Moreover, we will examine if some possible extensions of NMF, such as probabilistic NMF [11], and other compressed sensing methods [12] can further enhance the modulation spectrum and make the derived speech features more noise-robust for complicated speech recognition tasks.

VII. 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 98-2221-E-260-012-



(a)

(b)

(c)

(d)

Figure 3. The $c1$ PSD curves of an utterance ("MAR_5376869A.08" in the Aurora-2 database) after various processing methods with three SNR levels, clean, 10 dB and 0 dB: (a) no processing, (b) NMF, (c) MVN, (d) NMF+MVN.

MY2, NSC 99-2221-E-003-017-MY3, NSC 100-2515-S-003-003, and NSC 100-2631-S-003 -006.

VIII. References

[1] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," *in European Conf. Speech Communication and Technology* (*Eurospeech*), 1997.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 2, no. 4, 1994.

[3] P. Chen, K. Filaliy and J. A. Bilmes, "Frontend post-processing and backend model enhancement on the Aurora 2.0/3.0 databases," *in Proc. Int. Conf. Spoken Lang. Process. ICSLP*), 2002.

[4] X. Xiao, E. S. Chng, and Haizhou Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Trans. Acoust., Speech, Lang. Process.*, 2008.

[5] J-W. Hung and W-Y. Tsai, "Constructing modulation frequency domain based features for robust speech recognition," *IEEE Trans. Acoust., Speech, Lang. Process.*, 2008.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, 401:788–791, 1999.

[7] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing Systems* 13, 2000.

[8] B. Schuller, F. Weninger, M. Wöllmer, Y. Sun, and G. Rigoll, "Non-negative matrix factorization as noise-robust feature extractor for speech recognition," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* (*ICASSP*), 2010.

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

[10] J. Droppo, L. Deng, and A. Acero, "Evaluation of SPLICE on the AURORA 2 and 3 tasks," *in Proc. Int. Conf. Spoken Lang. Process*. (*ICSLP*), 2002.

[11] Eric Gaussier and Cyril Goutte, "Relation between PLSA and NMF and implications," *in Int. ACM SIGIR conf. on Research and development in information retrieval,* 2005.

[12] David L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, 52(4): 1289-1306, 2006.