Effective Modulation Spectrum Factorization for Robust Speech Recognition

Yu-Chen Kao, Yi-Ting Wang and Berlin Chen

Department of Computer Science and Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
berlin@csie.ntnu.edu.tw

10.21437/Interspeech.2014-155

Abstract

Modulation spectrum processing of acoustic features has received considerable attention in the area of robust speech recognition because of its relative simplicity and good empirical performance. An emerging school of thought is to conduct nonnegative matrix factorization (NMF) on the modulation spectrum domain so as to distill intrinsic and noise-invariant temporal structure characteristics of acoustic features for better robustness. This paper presents a continuation of this general line of research and its main contribution is two-fold. One is to explore the notion of sparsity for NMF so as to ensure the derived basis vectors have sparser and more localized representations of the modulation spectra. The other is to investigate a novel cluster-based NMF processing, in which speech utterances belonging to different clusters will have their own set of cluster-specific basis vectors. As such, the speech utterances can retain more discriminative information in the NMF processed modulation spectra. All experiments were conducted on the Aurora-2 corpus and task. Empirical evidence reveals that our methods can offer substantial improvements over the baseline NMF method and achieve performance competitive to or better than several widely-used robustness methods.

Index Terms: automatic speech recognition, robustness, modulation spectrum, nonnegative matrix factorization, normalization

1. Introduction

With the prevalence of handheld electronic devices and wireless communications, there is growing consensus that automatic speech recognition (ASR) will figure prominently in the interaction between people and various smart devices in the near future. Nevertheless, varying environmental effects, such as ambient noise, noises caused by the recording equipment and transmission channels, etc., often lead to a severe mismatch between the acoustic conditions for training and testing. Such a mismatch will no doubt cause substantial degradation in the performance of an ASR system.

In view of this, a wide array of robustness methods has been proposed in the literature to improve the performance of ASR systems. Among the most common streams of research on robust acoustic feature extraction, refining the modulation spectra [1-6] of acoustic feature sequences has received considerable attention because of its relative simplicity and good empirical performance. The modulation spectra of different feature dimensions collectively can be viewed as a holistic representation of the global acoustic structure of a speech utterance. Representative robust acoustic feature extraction methods targeting at normalization of modulation spectra include, but are not limited to, temporal structure normalization (TSN) [4], spectral histogram equalization (SHE) [2, 3], and several other spectral filtering and reweighting techniques [5, 6].

An emerging school of thought is to conduct nonnegative matrix factorization (NMF) [7] to distill intrinsic and noise-invariant temporal structure characteristics of acoustic features. To this end, NMF is employed to extract a common set of spectral basis vectors (or basis spectra) in the modulation spectra of each acoustic feature dimension of the clean training speech utterances. The normalized modulation spectra of acoustic features, constructed by mapping their original modulation spectra into the space spanned by the corresponding basis vectors, are demonstrated with good noise-robust capabilities. However, although NMF manages to yield spectral basis vectors having parts-based representations, this behavior cannot be guaranteed or controlled [8, 9]. Furthermore, since there would exist a high degree of variability among the modulation spectra of different speech utterances, it would be too restrictive to simply assume them all to follow one common factorization.

With the alleviation of the aforementioned deficiencies as motivation, in this paper we propose two novel extensions to the original NMF processing on the modulation spectra. One is to make effective use of the notion of sparsity for NMF so as to ensure the derived basis vectors have sparser and more localized representations of the modulation spectra. The other is to explore an alternative cluster-based NMF processing, in which speech utterances belonging to different clusters have their own set of cluster-specific basis vectors, in order to retain more discriminative information inherent in the NMF-processed modulation spectra. Additionally, we also investigate the feasibility of combining the advantages of these two extensions to further boost the robustness performance of the NMF processing.

The rest of this paper is organized as follows. Section 2 provides the essential fundamentals of the modulation spectrum representation and explains how NMF can be applied to modulation spectrum normalization. Section 3 sheds light on two extensions to the NMF processing on the modulation spectra. After that, the experimental settings and a series of experiments are presented in Sections 4 and 5, respectively. Finally, Section 6 concludes this paper and suggests avenues for future work.

2. Modulation Spectrum Factorization

2.1. Principle and Formulation

Given an ordered acoustic feature sequence $\{x_\ell\}$ of a specific acoustic feature dimension of an utterance, the modulation spectrum of this sequence can be defined by

$$V[k] = \sum_{\ell=0}^{L-1} x_\ell e^{\frac{-k\ell 2\pi i}{L}}, \quad (1)$$

where k is the index of modulation frequency components and $i = \sqrt{-1}$. Eq. (1) can be interpreted as the discrete Fourier

transform (DFT) of $\{x_\ell\}$, which is equivalent to treating the acoustic feature sequence as a signal and rendering its dynamic patterns along the temporal axis.

Modulation spectra constitute an efficient vehicle for analyzing the temporal-domain behaviors of acoustic feature sequences. For example, it has been reported in [10] that different modulation frequency components are of unequal importance for speech recognition, while most of the useful linguistic information is encapsulated in the modulation frequency components lying between 1 Hz and 16 Hz, with the dominant components centering around 4 Hz.

2.2. Nonnegative Matrix Factorization (NMF)

NMF is a subspace method that approximates data with an additive linear combination of nonnegative basis vectors. Given a nonnegative matrix $\boldsymbol{V} \in \mathbb{R}^{L\times M}$ with L being the dimensionality of each data vector and M the number of sample instances, NMF manages to find an unique rank-r factorization of $\boldsymbol{V}$ [11]:

$$\boldsymbol{V} \cong \boldsymbol{W}\boldsymbol{H}\,, \tag{2}$$

where the columns of $\boldsymbol{W} \in \mathbb{R}^{L\times r}$ represent the derived set of nonnegative basis vectors and the columns of $\boldsymbol{H} \in \mathbb{R}^{r\times M}$ store the corresponding nonnegative encodings. As such, each instance (column) in $\boldsymbol{V}$ can be represented as a linear combination of all the basis vectors in $\boldsymbol{W}$, weighted by its corresponding encoding (column) vector in $\boldsymbol{H}$. The rank r here is often chosen to be far fewer than M and L. Consequently, the product $\boldsymbol{WH}$ can be interpreted as a compressed form of $\boldsymbol{V}$, seeking to retain intrinsic and important building blocks embedded in $\boldsymbol{V}$. The compactness of the set of basis vectors is fundamentally different from other sparse-representation-based methods (for example in [12] and [13]) which normally capitalize on over-complete or exemplar-based dictionaries.

In order to find the factorization expressed in Eq. (2), a least square criterion that minimizes the reconstruction error (or the Euclidean distance) between the original data matrix and the approximate one can be used:

$$L = \|\boldsymbol{V} - \boldsymbol{W}\boldsymbol{H}\|_2^2\,. \tag{3}$$

The factorization should satisfy the constraint that $\boldsymbol{W}$ and $\boldsymbol{H}$ both be nonnegative. While Eq. (3) can be minimized with direct gradient descent methods, an expectation-maximization (EM) approach can also be adopted to get around the need for tuning the learning rate. With some initial guess of $\boldsymbol{W}$ and $\boldsymbol{H}$, the corresponding iterative update rule (also called the multiplicative update rule) can be expressed as following, which was derived and proved in [14, 15]:

$$\boldsymbol{H}_{ij} \leftarrow \boldsymbol{H}_{ij}\frac{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{V})_{ij}}{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{H})_{ij}}, \qquad \boldsymbol{W}_{ij} \leftarrow \boldsymbol{W}_{ij}\frac{(\boldsymbol{V}\boldsymbol{H}^{\mathrm{T}})_{ij}}{(\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})_{ij}}\,. \tag{4}$$

The notion of NMF has been recently introduced and crystalized for analyzing the modulation spectrum domain of acoustic features [7], with the goal to extract intrinsic and noise-invariant temporal structure characteristics of speech utterances for better robustness. In the training phase, the columns $\boldsymbol{v}$ of $\boldsymbol{V}$ are set to represent the modulation frequency components of each acoustic feature dimension of all training utterances. As such, each acoustic feature dimension has its own set of $\boldsymbol{W}$ and $\boldsymbol{H}$, while we only preserve $\boldsymbol{W}$ for the subsequent processing of the testing speech utterances. More specifically, in the testing phase, the following three steps are performed to reconstruct (normalize) the acoustic feature vector sequences of the testing speech utterances:

Table 1. The algorithm of S-NMF employed in this paper.

Input: $\boldsymbol{V}$, the data matrix which contains all modulation spectrum of training utterances, and a learning rate α
Output: $\boldsymbol{W}$, the set of basis vectors, and $\boldsymbol{H}$, the encoding
Algorithm:

1. Randomly (and nonnegatively) initialize $\boldsymbol{W}$ and $\boldsymbol{H}$.
2. Update $\boldsymbol{W} \coloneqq \boldsymbol{W} - \alpha(\boldsymbol{W}\boldsymbol{H} - \boldsymbol{V})\boldsymbol{H}^{\mathrm{T}}$.
3. For each column $\boldsymbol{w}$ in $\boldsymbol{W}$, perform the nonnegative projection algorithm described in [9].
4. Update $\boldsymbol{H}$ using Eq. (4).
5. Go to step 2 unless convergence.

1) **Analysis:** DFT is operated on the original acoustic feature vector sequence to obtain the corresponding modulation spectrum of each feature vector dimension.
2) **Processing:** Given a set of noise-free basis vectors (or spectra) $\boldsymbol{W}$ of a specific feature vector dimension and a column vector $\boldsymbol{v}$ representing the magnitude modulation spectrum of this dimension for a testing speech utterance, we can reconstruct $\boldsymbol{v}$ by using $\boldsymbol{W}$, estimated based on the clean training speech utterances, and the following equation:

$$\tilde{\boldsymbol{v}} = \boldsymbol{W}\boldsymbol{h}, \tag{5}$$

where $\boldsymbol{h}$ is an encoding vector whose estimate can be obtained in an iterative way via Eq. (4) and based on some initial guess for its values.
3) **Synthesis:** Perform inverse DFT on the modified magnitude modulation spectrum $\tilde{\boldsymbol{v}}$ and the phase of the original modulation spectrum to obtain the reconstructed (normalized) acoustic feature vector sequence.

Since $\boldsymbol{W}$ is trained using the clean training speech utterances in this paper, we expect that the reconstructed magnitude modulation spectrum $\tilde{\boldsymbol{v}}$ could properly highlight the most critical information for speech recognition, while eliminating the adverse effects of noise from the original magnitude modulation spectrum $\boldsymbol{v}$.

3. Proposed Extensions

3.1. Sparse NMF (S-NMF)

Similar to the original NMF, sparse NMF (denoted by S-NMF hereafter) tries to decompose data into the product of two nonnegative matrices [9]. However, sparse NMF further controls the degree of sparsity when analyzing the original data. More specifically, in S-NMF, Eq. (3) is minimized subject to the constraint that each column of $\boldsymbol{W}$ or each column of $\boldsymbol{H}$ could meet a certain sparsity measure. The sparsity measure $s(\cdot)$ of a vector $\boldsymbol{x}$ with length L can be expressed as following [9]:

$$s(\boldsymbol{x}) = \frac{(\sqrt{L} - \|\boldsymbol{x}\|_1)/\|\boldsymbol{x}\|_2}{\sqrt{L} - 1}\,, \tag{6}$$

where $\|\cdot\|_p$ represents the ℓ^p-norm. This constraint can be applied to the optimization of both of $\boldsymbol{W}$ and $\boldsymbol{H}$ (or, either $\boldsymbol{W}$ or $\boldsymbol{H}$). In this paper, since we are intended to obtain from the training procedure the set of basis vectors $\boldsymbol{W}$ to represent distinct and important ingredients of modulation spectra, we therefore apply the sparseness constraint on the estimation of $\boldsymbol{W}$. The algorithm used in this paper is outlined in Table 1.

NMF with the sparsity constraint has recently attracted much attention and been applied with success to many multimedia processing tasks. However, as far as we are aware, this notion has never been extensively explored to extract more localized representations of utterance-level modulation spectra for robust speech recognition, to the best of our knowledge.

3.2. Cluster-based NMF (C-NMF)

Although NMF can be used to extract distinct and important ingredients from the modulation spectra of a given feature vector dimension, there would exist high degree of variability among the modulation spectra of different speech utterances. As such, it would be too restrictive to simply assume that all modulation spectra of a given feature vector dimension follow merely one common factorization. In this paper, we propose a cluster-based approach (denoted by C-NMF hereafter) for relaxing such an assumption.

In the training phase, for each acoustic feature dimension, a cosine-based *K*-means clustering [16] algorithm is used to generate different clusters of modulation spectra from the training data in an unsupervised fashion. Following that, a set of fine-grained NMF basis vectors is trained for each cluster according to the training instances assigned to that cluster. It should also be pointed out that in this method, we also derive a global set of basis vectors $\boldsymbol{W}$ based on all training data so as to extract a more coarse-grained modulation spectrum characteristics of speech utterances.

In the testing phase, for each feature vector dimension, the cluster that is most close to the current speech utterance (in terms of the modulation spectrum cosine distance) is selected; and the set of basis vectors belonging to this cluster is used in conjunction with the global set of basis vectors to reconstruct the modulation spectrum of the utterance. More specifically, suppose that the modulation spectrum of a specific dimension of the current speech utterance is most close to the centroid of cluster *i* in accordance with the cosine distance, we can simply construct the modulation spectrum using Eq. (5) but with $\boldsymbol{W}$ replaced by $\boldsymbol{W}_i$. To further improve the robustness of C-NMF, in this paper the modulation spectrum processed by C-NMF is smoothed with the modulation spectrum processed by NMF using the global set of basis vectors:

$$\tilde{\boldsymbol{v}} = \lambda\tilde{\boldsymbol{v}}_{\text{global}} + (1-\lambda)\tilde{\boldsymbol{v}}_{\text{local}} \cong \lambda\boldsymbol{W}\boldsymbol{h} + (1-\lambda)\boldsymbol{W}_i\boldsymbol{h}_i, \quad (7)$$

where λ is a weighting factor controlling how much should we rely on the cluster-specific set of basis vectors instead of the global one.

4. Experimental Setup

The speech recognition experiments were conducted under various noise conditions with the widely-used Aurora-2 corpus and task [17]. The Aurora-2 corpus is a subset of the TI-DIGITS [18], 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 in a wide range of SNRs and the Test Set C additionally includes the channel distortion.

The acoustic model for each digit in the Aurora-2 task was a left-to-right continuous density HMM with 16 states, and each state has a 20-mixture diagonal GMM. Two additional silence models were defined. One had three states with a 20-mixture GMM per state for modeling the silence at the beginning and at the end of each utterance. The other one had one state with a 20-mixture GMM for modeling the inter-word short pause.

The baseline acoustic features are 39-dimension Mel-frequency cepstral coefficients (MFCC) [19]. The feature consists of 12 cepstral coefficients and an additional 0th cepstral term, along with their corresponding delta and acceleration coefficients. The training and recognition tests use the HTK toolkit [20]. All results reported below are based on clean-condition training, i.e., the acoustic models were trained only with the clean training utterances.

Table 2. Recognition accuracy rates (%) for the MFCC baseline system and NMF under different number of bases.

Method	Number of Bases	Test Set			Average
		Set A	Set B	Set C	
MFCC	-	54.98	49.6	63.07	54.44
NMF	5	67.09	70.98	68.22	68.87
	10	66.58	71.36	65.35	68.24
	15	64.07	69.32	61.85	65.72
	20	64.70	69.89	62.04	66.24

Table 3. Recognition accuracy rates (%) for S-NMF based on MFCC feature under different degrees of sparsity.

$s(\boldsymbol{x})$	0.4	0.5	0.6	0.7	0.8
S-NMF	76.73	77.89	78.71	78.91	75.27

Table 4. Recognition accuracy rates (%) for C-NMF based on MFCC feature under different numbers of clusters.

# of clusters	1	3	10	20	30
C-NMF	68.87	70.62	69.05	72.31	68.98

5. Experimental Results

In this section, we begin by reporting on the performance of the standard MFCC system and the baseline NMF method. The corresponding results are shown in Table 2, where NMF is evaluated with respect to different numbers of basis vectors being used. We can see from Table 2 that NMF provides consistent and significant improvements over the MFCC systems across all test sets, especially when the number of basis vectors is small. The improvements, however, seem to soon level off or decrease when the number of basis vector is set to be equal to or larger than 10. One possible explanation is that, since most of the important linguistic cues are encapsulated in the lower frequency components of the modulation spectra of acoustic feature vectors, a small number of basis vectors thus would be sufficient to represent the most important inherent linguistic clues. In the following

experiments, the number of basis vectors for the various NMF-derived methods is set to 5, unless otherwise stated.

In the next set of experiments, we evaluate the utility of our proposed two variants of NMF (i.e., S-NMF and C-NMF). The corresponding results are shown in Tables 3 and 4, from which we can draw two noteworthy observations. First, both S-NMF and C-NMF can further boost the performance of NMF, while S-NMF stands out in performance as compared to C-NMF, leading to more than 10% absolute recognition accuracy improvements over NMF. Second, the performance of S-NMF is steadily improved when the degree of sparsity $s(\boldsymbol{x})$ becomes larger; nonetheless, the improvements seem to reach a plateau when $s(\boldsymbol{x})$ is set to 0.7. To get a better sense of the utility of S-NMF and C-NMF, we further compare them with a few widely-practiced acoustic feature normalization methods, including spectral histogram equalization (SHE) [2, 3], principle component analysis (PCA) [21], cepstral mean and variance normalization (CMVN) [22] and histogram equalization (HEQ) [23, 24]: the former two methods operate on the modulation spectrum domain, while the latter two on the acoustic feature domain. The corresponding results of these methods are shown in Table 5. Inspection of Table 5 reveals three noteworthy points. First, S-NMF is the best-performing one among the four modulation spectrum normalization methods (i.e., SHE, PCA, S-NMF and C-NMF) compared here, while C-NMF delivers slightly better results than PCA but is inferior to SHE. Second, HEQ that manages to normalize all moments of the probability distributions for each acoustic feature dimension yields slightly better results than S-NMF. Third, CMVN that normalizes only the first two moments of the probability distributions for each acoustic feature dimension turns out to perform worse than S-NMF.

In addition, because S-NMF and C-NMF perform normalization on the modulation spectrum domain, they thus focus exclusively on normalizing the dynamic patterns of all the acoustic feature vector dimensions. In contrast, HEQ and CMVN directly normalize the values of acoustic feature vector components at each speech frame. Hence, there is reason to combine S-NMF and C-NMF with either CMVN or HEQ to further enhance the noise-invariant capabilities of acoustic feature vectors. It is evident from the lower part of Table 5 that all possible combinations of methods belonging to these two families can offer significant improvements as compared to that using each one of them in isolation, while CMVN+C-NMF achieve the best-performing one among all possible combinations. Note that since the performance of C-NMF depends on the accuracy in the clustering stage (*cf.* Section 3.2), C-NMF can substantially benefit from employing either CMVN or HEQ to pre-process the acoustic feature vectors. Furthermore, as S-NMF and C-NMF actually present two distinct extensions to NMF, it seems that we can also combine them together (denoted by CS-NMF; i.e., to first cluster modulation spectra of speech utterances and then conduct the S-NMF processing based on the corresponding cluster assignments) for better robustness; here we take the CMVN-processed acoustic feature vectors as an example (*cf.* CMVN+CS-NMF in Table 5) due to space constraints.

In the final set of experiments, we compare the NMF-based normalization methods proposed in this paper with the ETSI advanced front-end (referred to hereafter as AFE) [25]. AFE is believed to be one of the most elaborated and effective robustness methods, which leads to an average WAR of 87.17% on the Aurora-2 task, as shown in Table 6. It is obvious that CMVN+C-NMF and CMVN+CS-NMF (*cf.* Table 5) can perform on par with AFE, even though they only operates on the MFCC feature vector components without explicitly using any online noise estimation or reduction process. We also attempt to conduct CS-NMF on the output of AFE, in order to see if it can still offer complementary normalization capability for the AFE-processed features. The corresponding results are also shown in Table 6. As compared with the results obtained by using AFE in isolation, such integration can offer moderate speech recognition error reductions.

Table 5. Recognition accuracy rates (%) averaged over different noise types and different SNRs for a few acoustic feature normalization methods.

Method	Test Set			Average
	Set A	Set B	Set C	
SHE	74.82	77.44	76.47	76.20
PCA	70.90	73.34	71.39	71.97
HEQ	80.13	82.37	80.17	81.04
CMVN	75.56	76.48	76.48	76.11
S-NMF	77.53	80.33	78.84	78.91
C-NMF	70.98	74.20	71.18	72.31
HEQ+S-NMF	82.39	84.60	83.24	83.44
HEQ+C-NMF	87.12	86.40	86.28	86.66
CMVN+S-NMF	83.24	85.40	83.94	84.24
CMVN+C-NMF	87.13	88.15	87.54	87.62
CMVN+CS-NMF	87.20	88.65	87.49	87.84

Table 6. Recognition accuracy rates (%) averaged over different noise types and different SNRs for AFE and its combinations with S-NMF and C-NMF.

Method	Test Set			Average
	Set A	Set B	Set C	
AFE	87.68	87.10	86.27	87.17
AFE+CS-NMF	88.23	87.82	86.85	87.79

6. Conclusion and Outlook

In this paper, we have presented two novel extensions of the NMF processing on the modulation spectra of acoustic features (i.e., S-NMF and C-NMF), showing that normalizing the modulation spectra of acoustic features with these two methods can yield more noise-invariant acoustic features. S-NMF and C-NMF conducted on the conventional MFCC features can yield significant recognition accuracy rate improvements, while their combinations with CMVN can provide further improvements, arriving at a performance level on par with AFE which is believed to be one of the most elaborated and effective robustness methods. As to future work, we would like to adopt our methods to larger speech recognition tasks and other multimedia-related applications.

7. Acknowledgement

This research is supported in part by the "Aim for the Top University Project" of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, and by the Ministry of Science and Technology, Taiwan, under Grants NSC 101-2221-E-003-024-MY3, NSC 102-2221-E-003-014-, NSC 101-2511-S-003-057-MY3, NSC 101-2511-S-003-047-MY3 and NSC 103-2911-I-003-301.

8. References

[1] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *The Journal of Acoustical Society of America*, vol. 95, no. 3, pp. 1593–1602, 1994.

[2] Y.-C. Kao and B. Chen, "Leveraging distributional characteristics of modulation spectra for robust speech recognition," in *Proceedings of Internetional Conference on Information Scirnce, Signal Processing and their Applications*, pp. 120–125, 2012.

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

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

[5] N. Wada, N. Hayasaka, S. Yoshizawa, and Y. Miyanaga, "Direct control on modulation spectrum for noise-robust speech recognition and spectral subtraction," in *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 2532–2536, 2006.

[6] J.-W. Hung, W.-H. Tu, and C.-C. Lai, "Improved modulation spectrum enhancement methods for robust speech recognition," *Signal Processing*, vol. 92, no. 11, pp. 2791–2814, 2012.

[7] W.-Y. Chu, J.-W. Hung, and B. Chen, "Modulation spectrum factorization for robust speech recognition," in *Proceedings of APSIPA Annual Summit and Conference*, pp. 194–206, 2011.

[8] H.-T. Fan, Y.-C. Tsai, and J.-W. Hung, "Enhancing the sub-band modulation spectra of speech features via nonnegative matrix factorization for robust speech recognition," in *Proceedings of International Conference on System Science and Engineering*, pp. 179–182, 2012.

[9] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Resources*, vol. 5, pp. 1457–1469, 2004.

[10] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," in *Proceedings of European Conference on Speech Communication and Technology*, pp. 1079–1082, 1997.

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

[12] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210 –227, 2009.

[13] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions of Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067 –2080, 2011.

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems* (T. K. Leen, T. G. Dietterich, and V. Tresp, eds.), pp. 556–562, MIT Press, 2001.

[15] C.-J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1589–1596, 2007.

[16] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1-2, pp. 143–175, 2001.

[17] D. Pearce, H. G. Hirsch, and D. Gmbh, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of ISCA Workshop on ASR*, pp. 29–32, 2000.

[18] R. Leonard, "A database for speaker-independent digit recognition," in *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, vol. 9, pp. 328–331, 1984.

[19] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustic Speech Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[20] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[21] J.-Y. Lee and J.-W. Hung, "Exploiting principal component analysis in modulation spectrum enhancement for robust speech recognition," in *IEEE International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 3, pp. 1947–1951, 2011.

[22] 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.

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

[24] S.-H. Lin, B. Chen, 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*, vol. 17. no.1, pp. 84–94, 2009.

[25] D. Macho, L. Mauuary, B. Noe, 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 Annual Conference of the International Speech Communication Association*, pp. 17–20, 2002.