INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Exploring Low-Dimensional Structures of Modulation Spectra for Robust Speech Recognition

Bi-Cheng Yan[1], *Chin-Hong Shih*[1], *Shih-Hung Liu*[2], *Berlin Chen*[1]

[1]National Taiwan Normal University, Taiwan
[2]Delta Research Center, Taiwan
{60447055S, 60447003S, berlin}@csie.ntnu.edu.tw, journey.liu@deltaww.com

Abstract

Developments of noise robustness techniques are vital to the success of automatic speech recognition (ASR) systems in face of varying sources of environmental interference. Recent studies have shown that exploring low-dimensional structures of speech features can yield good robustness. Along this vein, research on low-rank representation (LRR), which considers the intrinsic structures of speech features lying on some low dimensional subspaces, has gained considerable interest from the ASR community. When speech features are contaminated with various types of environmental noise, its corresponding modulation spectra can be regarded as superpositions of unstructured sparse noise over the inherent linguistic information. As such, we in this paper endeavor to explore the low dimensional structures of modulation spectra, in the hope to obtain more noise-robust speech features. The main contribution is that we propose a novel use of the LRR-based method to discover the subspace structures of modulation spectra, thereby alleviating the negative effects of noise interference. Furthermore, we also extensively compare our approach with several well-practiced feature-based normalization methods. All experiments were conducted and verified on the Aurora-4 database and task. The empirical results show that the proposed LRR-based method can provide significant word error reductions for a typical DNN-HMM hybrid ASR system.

Index Terms: low-rank representation, sparse representation, deep neural network, robustness, modulation spectrum.

1. Introduction

The performance of automatic speech recognition (ASR) often degrades significantly due to the corruption of acoustic signals by noise and/or channel distortions. Robustness techniques aim to alleviate the negative impact caused by such distortions so as to make ASR systems retain acceptable performance [1]. Representative robustness methods developed so far can be categorized into three broad groups: 1) feature normalization, 2) feature enhancement and 3) model adaptation. Feature normalization methods generally seek for noise resistant and robust speech features [2], while feature enhancement methods aim to remove the environmental effects from the observed speech signal of interest. [3]. In addition, model adaptation methods transform acoustic models from the training (clean) space to the test (noisy) space [4].

Unlike conventional robustness methods that conduct normalization on the speech features directly, recently there has been a school of research thoughts on refining the modulation spectra of a given speech feature sequence (instead of one to several speech frames at a time), which has shown effectiveness to capture linguistic information and improve robustness of ASR in various tasks [2], [5], [6].

On a separate front, a prevalent trend has also been to explore low dimensional structures of speech features or their intermediate representations. In this paper, we investigate the property of low-dimensional structures residing in the modulation spectra of speech features for use in normalization, which is facilitated by leveraging the paradigm of low-rank representation (LRR) [7], [8]. LRR and its variants originally have been applied to the posterior vectors of DNN-HMM based acoustic models to achieve better ASR performance. The fundamental hypothesis is that linguistic cues conveyed in a speech frame predicted by acoustic models, in the form of posterior vectors, are actually embedded in a union of low-dimensional subspaces [9], [10]. To reduce unstructured sparse errors encountered during the estimation of posterior vectors, different approaches can be employed to manipulate the matrix of stacked posterior vectors according to the following assumption: if the uncertainties of posterior vectors are re-occurring patterns, then they encode common linguistic features; otherwise, they may be contaminated with unstructured sparse noise [9]. More specifically, speech features presumably reside on near non-linear manifolds that can be well characterized by union of low-dimensional subspaces, while sparse error can thus be separated from such structured parts through low-rank and sparse decomposition methods [10]. Yet there is another line of research which explores sparse coding to extract spectro-temporal patterns from speech (spectral) features and in turn represent them by a few important atoms of an over-complete dictionary. Through such a kind of dimensionality reduction or its variants, the negative effects of residual noise can be ruled out [11], [12].

Different from the aforementioned methods, we in this paper endeavor to explore the low-dimensional structures of modulation spectra, in the hope to obtain more noise-robust speech features. The main contribution is that we propose a novel use of the LRR-based processing paradigm to discover the subspace structures of modulation spectra, thereby alleviating the negative effects of noise interference. Furthermore, we also extensively compare our method with several well-practiced feature-based normalization methods. To our knowledge, this work is the first attempt to leverage such a modeling paradigm in the modulation frequency domain of speech features for robust ASR.

The remainder of the paper is organized as follows: Section 2 introduces the notion of processing in the modulation domain. The principles of sparse representation and low-rank representation will be elucidated in Section 3. After that, the experimental setup is described in Section 4, followed by a series of experiments and associated discussions in Section 5. Finally, Section 6 concludes this paper.

http://dx.doi.org/10.21437/Interspeech.2017-611

2. The Modulation Spectrum

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

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

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 in a holistic fashion [6]. Furthermore, it has been reported 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 [13].

3. Low-Dimensional Structures

The notion of exploring low-dimensional structures inherent in data instances of interest has its root that the data instances presumably bear certain intrinsic properties of low dimensionality. That is, they may lie in a low-dimensional subspace which is spanned by some basis, or lie on a specific low-dimensional manifold. The associated methods developed so far are often characterized into three processing paradigms: LRR, sparse representation and manifold learning. In this paper, we explore the former two paradigms (i.e., LRR and sparse representation) to normalize the magnitude modulation spectra of speech features in order to produce more noise-resistant speech features.

3.1. Sparse Representation

In the context of magnitude modulation spectrum normalization of speech features, we assume that magnitude modulation spectra of a given noisy speech feature sequence can be viewed as dense vectors, while reconstruction of a magnitude modulation spectrum by some important atoms spanning a low-dimensional subspace could remove residual noise. Sparse representations of data instances (i.e., the magnitude modulation spectra of a given noisy speech feature sequence) are basically achieved through two processing components working in tandem, i.e., dictionary learning and sparse coding [14]. Two representative methods of sparse representation are the K-SVD method and the NN-K-SVD method.

3.1.1. Dictionary learning

Dictionary learning aims to learn a set of basis vectors (or dictionary atoms) from training data instances by some specifically devised algorithms which iteratively minimize the reconstruction error and maximize the sparsity of weights [15], [16], [17]. The various methods developed for dictionary learning generally fall into three common families according to their associated learning criteria: the probabilistic methods [18], the clustering-based methods [15] and the specific structure-based methods [19].

K-SVD is arguably one of the most celebrated clustering-based methods and is considered to be a generalization of the K-means clustering algorithm. The K-SVD algorithm performs alternately between two stages: one is the sparse coding stage and the other is the dictionary update stage. At the sparse coding stage, K-SVD attempts to exclude residual noise from a data instance based on the current dictionary. At the dictionary update stage, it groups training data instances to their respective subspaces (or clusters). In the context of magnitude modulation spectrum normalization of speech features, given $\mathrm{Y} \in R^{m \times n}$ (a matrix composed of stacked magnitude modulation spectra of training utterances), m (the number of indices in a modulation spectrum), n (the number of training utterances), the K-SVD objective function can be expressed as:

$$\min_{\mathrm{D,X}} \|\mathrm{Y} - \mathrm{DX}\|_F^2 \ \text{s.t.} \ \|x_i\|_0 \le T_0, \ i = 1,2,\dots,n \qquad (2)$$

where T_0 is the upper bound value of the sparsity constraint. The dictionary D groups the magnitude modulation spectra into T_0 subspaces (clusters) and X represents a matrix containing the weight vector for each training utterance. On the other hand, considering the nonnegativity property of the amplitude modulation spectrum, we can impose a nonnegativity constraint on both the dictionary and the weight vectors of training utterance [20], leading to the so-called nonnegativity K-SVD (NN-K-SVD) method.

3.1.2. Sparse coding

The sparse coding methods address the problem of how to sparsely represent an input instance as a linear combination of atoms from a given pre-specified dictionary. By doing so, the residual noise can be naturally excluded from this kind of sparse representation. There exist several effective algorithms to achieve the purpose of sparse coding, including matching pursuit (MP) [21], orthogonal matching pursuit (OMP) [22], non-negativity sparse coding (NNSC) [20], etc.

The MP and OMP methods are popular instantiations with the l_0-norm constraint to sparsely reconstruct an input instance based on a greedy strategy. The MP algorithm iteratively searches for the best matching atom from the pre-specified dictionary, and then utilizes the residual vector representation as the next approximation target until the termination condition of iteration is satisfied. The OMP algorithm performs almost as same as the MP method, except that OMP guarantees the residual is orthogonal to the atoms that have been chosen. It has been verified that the OMP algorithm can converge within a limited number of iterations. In addition, The NNSC method is a kind of the pursuit methods with an ancillary nonnegativity constraint, which combines the merits of the sparse coding method and nonnegative matrix factorization (NMF) [20]. Its algorithm utilizes a multiplicative update rule to estimate the nonnegative weight vector for a given dictionary. The NNSC method is more amenable to the property of the magnitude modulation spectra of speech features.

3.2. Low-Rank Representation (LRR)

Low-rank representation considers the intrinsic dimensions of linguistic information encapsulated in the modulation spectrum which are located in some low-dimensional subspace [7], [8]. The history of developing the LRR based method can be dated back to robust principal component analysis (RPCA) [23], which enables principal component analysis (PCA) to maintain the ability of dimensionality reduction and be unaffected by the undesired effects outliers (e.g., data instances corrupted by noise) simultaneously. One drawback of RPCA is that it assumes that data instances are approximately drawn from a single low-rank subspace. Thus, the LRR method is proposed

Table 1: *The statistics of the Aurora-4 database.*

Sampling rate	8000 Hz
Speech content	Wall Street Journal (WSJ) 5000-word dictionary
Speech length	roughly 7 seconds in average
Training data	7138 clean utterances Recorded by Sennheiser Mic.
Test data	Set A:330 utterances, No noise added. Recorded by Sennheiser Mic.
	Set B:1980 utterances, including six types of additive noise: car, babble, restaurant, street, airport and train station (each noise added at SNRs between 5 and 15 dB) Recorded by Sennheiser Mic.
	Set C:330 uttrtances, No noise added. Recorded by Secondary Mic.
	Set D : 1980 utterances, including six types of additive noise: car, babble, restaurant, street, airport and train station (each noise added at SNRs between 5 and 15 dB) Recorded by Secondary Mic.

to address the issue and extend the idea that the data instances are approximately drawn from a mixture of several low-rank subspaces:

$$\mathrm{Y} = \mathrm{L} + \mathrm{E} \tag{3}$$

Eq.(3) can be regarded as magnitude modulation spectra Y that contains two parts: one is strictly drawn from the linguistic information subspaces L and the other is E that is composed of unstructured noise. The PCA based method unveils L via singular value decomposition (SVD) with an appropriate rank of Y and assumes E is small perturbation error; however, this assumption may not hold here. The noise encountered in ASR might be unstable perturbation error such as addition noise and/or channel distortions. The RPCA based method further assumes that E is sparse noise and be characterized by l_1-norm which can be solved by the objective function as follows:

$$\min_{\mathrm{L,E}} \|\mathrm{L}\|_* + \lambda\|\mathrm{E}\|_1 \quad \mathrm{s.t.}\ \mathrm{Y} = \mathrm{L} + \mathrm{E} \tag{4}$$

where $\|.\|_*$ denotes the nuclear norm, which is sum of singular values. The rank is quantified by nuclear norm and can be solved with Eq. (4) by convex optimization algorithms. Though the RPCA based method can alleviate the effect of outliers, it fails to produce robust low-rank representation since the magnitude modulation spectrum amplitudes Y may drawn from a union of multiple subspaces. Specifically, if we denote multiple subspaces as S_1,S_2,…,S_k, the RPCA based method treats the dimensionality reduction problem as $\mathrm{Y} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{S}_\mathrm{i}$. Previous work proposed to use LRR to segment data instances drawn from a union of multiple linear subspaces, $\mathrm{Y} = \bigcup_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{S}_\mathrm{i}$, and redefine LRR to seek the lowest-rank representation among all the candidates that represent all vectors as the linear combination of the bases in a dictionary [8]. The corresponding objective function is thus expressed as:

$$\min_{\mathrm{Z,E}} \|Z\|_* + \lambda\|\mathrm{E}\|_1 \quad \mathrm{s.t.}\ \mathrm{Y} = \mathrm{DZ} + \mathrm{E} \tag{5}$$

where dictionary D is assumed to linearly span the space of Y. However the unstructured noise E quantified by l_1-norm is unclear, which often contains some type of structure. Here we can use $l_{2,1}$-norm to characterize E as sample-specific noise, which means that noise only happens on a small fraction of input sample instances, such as nonstationary noise [7]. The objective function of LRR with sample-specific error can be defined as follows:

$$\min_{\mathrm{Z,E}} \|Z\|_* + \lambda\|\mathrm{E}\|_{2,1} \quad s.t.\ \mathrm{Y} = \mathrm{DZ} + \mathrm{E} \tag{6}$$

where $\|\mathrm{E}\|_{2,1} = \sum_j \sqrt{\sum_i (E_{ij})^2}$. Here we can employ the lowest-rank representation Z to approximate the magnitude modulation spectra Y with respect to a dictionary D that stands for a union of multiple subspaces. The unstructured noise E is sample-specific error which may bring negative impact to the ASR system.

4. Experimental Setup

4.1. Speech Corpus and DNN-HMM Modeling

The proposed robustness methods are evaluated via a series of speech recognition experiments on the Aurora-4 corpus database [25], which is a benchmark database designed to evaluate the performance of robustness methods for ASR. In more detail, Aurora-4 is a medium to large vocabulary recognition task originated from the Wall Street Journal (WSJ) database; it consists of clean speech utterances interfered with various noise sources at different SNR levels ranging from 5 dB to 15 dB. In Aurora-4, speech utterances were sampled in both 8 kHz and 16 kHz, while only 8-kHz sampled speech utterances were used for our experiments here. The clean-condition training set consisting of 7,137 utterances was recorded with a Sennheiser microphone. The test sets are totally composed of 14 subsets, each of which contains 330 utterances contaminated with various types of environmental noise at different SNR levels. A half of the test set was recorded by a primary microphone, while the rest was recorded by a secondary microphone. These 14 test sets were grouped into 4 subsets: clean, noisy, clean with channel distortion and noisy with channel distortion, which will be referred to as Tests A, B, C and D henceforth. Table 1 summarizes the statistics of the Aurora-4 database.

The acoustic models used in this study were configured with the DNN-HMM (Deep Neural Network-Hidden Markov Model) modeling framework, following the setup provided in [4]. Here DNN is used in place of GMM (Gaussian Mixture Model) for modeling the state emission probabilities in HMM. The DNN architecture is set to have 7 hidden layers with 2,048 units per layer. Each utterance is first converted into a sequence of 39-dimensional feature vectors (MFCC, c0-c12) with a context window of 11 consecutive frames, plus their first- and second-order derivatives.

In this paper, we enhanced the static part of the MFCC features (i.e., c0-c12) in their modulation spectrum domain. When feature normalization had been accomplished, we added the first- and second-order derivatives of the normalized static features to form an enhanced MFCC feature vector [2].

4.2. Experimental Procedure

In the training phase, we utilized the magnitude modulation spectra of clean-condition speech utterance in the training set to train the dictionary. Next, the dictionary learning and sparse coding method and the LRR method were employed, respectively, to obtain a newly updated pair of dictionary and weight matrix. After that, in the test phase, we performed in an utterance-by-utterance manner to obtain the corresponding weight vectors of test noisy utterances in accordance with the dictionary atoms obtained in the training phase. The multiplication of the dictionary atoms learned from the training phase and the corresponding weight vector estimated in the test phase constitutes an updated magnitude modulation spectrum

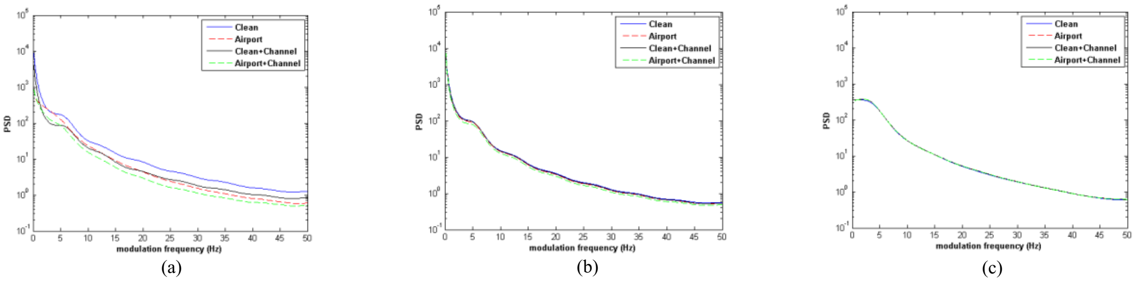

Figure 1: *The average c1 PSD curves for Aurora-4 test utterances with various noise types, i.e., clean, airport noise, clean with channel distortion and airport noise with channel distortion,which were processed by two normalization methods: (a) the MFCC baseline (without normalization), (b) NN-K-SVD+NNSC and (c) LRR.*

of a test utterance with respect to a certain speech feature dimension, which was expected to be more resistant to noise interference.

5. Experimental Results

5.1. Magnitude Modulation Spectrum Normalization based on MFCC Features

In the first set of experiments, we compare several low-dimensional structure methods used to model the magnitude modulation spectra of the standard MFCC features. The corresponding results are shown in Table 2, from which we can draw three noteworthy observations. First, K-SVD and its variants, as well as LRR, can boost the performance of the baseline MFCC system significantly. They, respectively, only used very few dictionary atoms (the number of atoms is set to 5 in this paper) to linearly reconstruct the MFCC-based modulation spectrum. It means that the linguistic information conveyed in the modulation spectra may be lying in unions of these linear subspaces. Second, NN-K-SVD+NNSC can reduce average word error rate (WER) of the two variants of K-SVD (i.e., K-SVD+MP and K-SVD+OMP) by 0.86% and 0.91%, respectively. This implies that the nonnegativity constraint should be imposed when unveiling the low-dimensional structures and sparse representations of the magnitude modulation spectra of the MFCC features. Third, LRR stands out in performance as compared to NN-K-SVD+NNSC, achieving a further WER reduction of 3.84%. It should be mentioned here that LRR adopts a low-rank representation the noisy magnitude modulation spectrum, which reconstructs the clean magnitude modulation spectrum by the union of a few important atoms, simultaneously squeezing residual noise out to the sparse error component. In contrast, NN-K-SVD+NNSC utilizes non-negative sparse coding to diminish the redundant information residing in the noisy modulation spectrum.

5.2. Analysis of PSD Curves

Apart from recognition performance, we also compare the presented NN-K-SVD+NNSC and LRR with regard to their capabilities of reducing the mismatch in the power spectral density (PSD) of the MFCC-based cepstral feature sequence. Figs. 1(a) to 1(c) depict the average PSD curves of the unprocessed, NN-K-SVD+NNSC processed and LRR processed first MFCC feature component (c1) for the Aurora-4 test utterances contaminated with four types of environmental noise, with SNR levels varying from 5 dB to 15 dB. First, for the unprocessed case shown in Fig. 1(a), the various noise sources cause a significant PSD mismatch over the entire modulation frequency band [0, 50 Hz]. Figs. 1(b) and 1(c) show that both NN-K-SVD+NNSC and LRR can considerably reduce the PSD distortion, while LRR appears to be more effective than NN-K-SVD+NNSC with regard to the mitigation of the PSD mismatch at all frequency bands.

Table 2: *Word error rate (%) for the MFCC baseline, various K-SVD based methods and the LRR based method.*

Method	Set A	Set B	Set C	Set D	Avg.
MFCC	3.75	49.93	22.55	60.32	34.14
K-SVD+MP	5.66	35.05	20.72	46.19	26.91
K-SVD+OMP	5.90	35.09	20.38	46.45	26.96
NN-K-SVD +NNSC	6.05	33.48	19.91	44.76	26.05
LRR	5.40	29.24	13.79	39.70	22.03

Table 3: Word error rates (%) for the synergy of the LRR based method and some state-of-the-art methods.

Method	Set A	Set B	Set C	Set D	Avg.
CMVN	3.92	33.96	9.81	46.22	23.48
CMVN+LRR	3.34	29.64	10.29	42.00	21.32
AFE	3.70	21.14	9.45	30.36	16.16
AFE +LRR	3.85	20.41	9.06	29.31	15.66

5.3. Magnitude Modulation Spectrum Normalization based on CMVN- and AFE-Processed Features

In the last set of experiments, we investigate the synergy of the proposed LLR method with two state-of-the-art robustness methods that directly perform normalization on the MFCC components at each time frame instead of the modulation spectra; they are cepstral mean and variance normalization (CMVN) and the ETSI advanced front-end based method (AFE) [24]. As evident from Table 3, the synergy of LLR and the two existing methods that directly enhance the MFCC features can bring considerable additional gains for the two latter methods, thereby confirming the complementary robustness capability of additionally normalizing the magnitude modulation spectra of speech features.

6. Conclusions

In this paper, we have explored a novel use of the low-rank representation (LRR) to discover the intrinsic subspace structures residing in the modulation spectra of speech features for alleviating the negative effects of environmental noise. In addition, we empirically compare our methods with state-of-the-art methods. The experimental results demonstrate that LRR-based feature normalization conducted in the modulation spectrum domain can significantly improve the baseline MFCC system, as well as the CMVN- and the AFE-based systems.

7. References

[1] J. Droppo, and A. Acero, “Environmental robustness,” springer handbook of speech processing. Springer Berlin Heidelberg, pp. 653–680, 2008.

[2] Y.-C. Kao et al., “Effective modulation spectrum factorization for robust speech recognition,” in Proc. INTERSPEECH, pp. 2724–2728, 2014.

[3] Y. He, S. Guanglu, and H. Jiqing, “Spectrum enhancement with sparse coding for robust speech recognition,” Digital Signal Processing, 43, 59–70, 2015.

[4] M.L. seltzer et al., “An investigation of deep neural networks for noise robust speech recognition,” in Proc. ICASSP, pp. 7398–7402, 2013.

[5] J.W. Hung et al.,“Robust speech recognition via enhancing the complex-valued acoustic spectrum in modulation domain,” IEEE/ACM Transactions on Audio, Speech and Language Processing, 24(2), pp. 236–251, 2016.

[6] N. Kanedera et al., “On the importance of various modulation frequencies for speech recognition,” in Proc. EUROSPEECH, pp. 1079–1082, 1997

[7] G. Liu et al., “Robust recovery of subspace structures by low-rank representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1), pp. 171–184, 2013.

[8] G. Liu et al., “Robust subspace segmentation by low-rank representation” in Proc. ICML, 2010.

[9] G. Luyet, et al., “Low-rank representation of nearest neighbor phone posterior probabilities to enhance DNN acoustic modeling,” No. EPFL-REPORT-218116. Idiap, 2016.

[10] P. Dighe, et al, “Exploiting low-dimensional structures to enhance dnn based acoustic modeling in speech recognition.” in Proc. ICASSP, pp. 5690-5694, 2016.

[11] G.S.V.S. Sivaram, et al. "Sparse coding for speech recognition,” in Proc. ICASSP, 2010.

[12] J. F. Gemmeke et al., “Exemplar-based sparse representations for noise robust automatic speech recognition,” IEEE Transactions on Audio, Speech, and Language Processing, 19(7), pp. 2067–2080, 2011.

[13] S. Greenberg, “On the origins of speech intelligibility in the real world,” in Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, 1997.

[14] Z. Zhang et al., “A survey of sparse representation: algorithms and applications,” IEEE Transactions on Content Mining, 3, pp. 490–530, 2015.

[15] M. Aharon et al., “K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation,” IEEE Transactions on Signal Processing, vol. 54, no. 11, pp. 4311–4322, 2006.

[16] J. Mairal, et al, “Online learning for matrix factorization and sparse coding,” Journal of Machine Learning Research, 11, pp. 19–60, 2010.

[17] C. Lu et al., “Online robust dictionary learning,” in Proc. CVPR, pp. 415–422, 2013.

[18] D.P. Wipf and B.D. Rao, “Sparse Bayesian learning for basis selection,” IEEE Transactions on Signal Processing, 52(8), pp. 2153–2164, 2004.

[19] Y, Mehrdad et al., “Parametric dictionary design for sparse coding,” IEEE Transactions on Signal Processing, 57(12), pp. 4800–4810, 2009.

[20] P.O. Hoyer, “Non-negative sparse coding,” in Proc. of Neural Networks for Signal Processing, 2002.

[21] M. Stéphane and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” IEEE Transactions on signal processing, 41(12), pp. 3397–3415, 1993.

[22] P. Yagyensh et al., “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” in Proc. of Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, 1993.

[23] J.E. Candès, et al. “Robust principal component analysis?,” Journal of the ACM, 58, 3:11, 2011.

[24] D. Machoet et al., “Evaluation of a noise-robust dsr front-end on aurora databases,” in Proc. INTERSPEECH, pp. 17–20, 2002.

[25] D. Pearce, Aurora working group: DSR front end LVCSR evaluation AU/384/02. Diss. Mississippi State University, 2002.