INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Adaptive Group Sparsity for Non-negative Matrix Factorization with Application to Unsupervised Source Separation

Xu Li[1], *Ziteng Wang*[1], *Xiaofei Wang*[1], *Qiang Fu*[1] *and Yonghong Yan*[1,2]

[1]Key Laboratory of Speech Acoustics and Content Understanding,
Institute of Acoustics, Chinese Academy of Sciences, Beijing 1001090, China
[2]Xinjiang Laboratory of Minority Speech and Language Information Processing
lixu@hccl.ioa.ac.cn

Abstract

Non-negative matrix factorization (NMF) is an appealing technique for many audio applications, such as automatic music transcription, source separation and speech enhancement. Sparsity constraints are commonly used on the NMF model to discover a small number of dominant patterns. Recently, group sparsity has been proposed for NMF based methods, in which basis vectors belonging to a same group are permitted to activate together, while activations across groups are suppressed. However, most group sparsity models penalize all groups using a same parameter without considering the relative importance of different groups for modeling the input data. In this paper, we propose adaptive group sparsity to model the relative importance of different groups with adaptive penalty parameters and investigate its potential benefit to separate speech from other sound sources. Experimental results show that the proposed adaptive group sparsity improves the performance over regular group sparsity in unsupervised settings where neither the speaker identity nor the type of noise is known in advance.

Index Terms: non-negative matrix factorization, source separation, group sparsity

1. Introduction

Non-negative matrix factorization (NMF) [1, 2] has been widely used in many audio applications, such as automatic music transcription, source separation and speech enhancement. The basic idea of NMF is to decompose the magnitude spectrum of the source into a basis dictionary and a weight matrix which are both constrained non-negative. The basis dictionary conveys meaningful dynamic patterns while the weight matrix represents the activation of different patterns along time.

Sparsity constraints are widely used in the factorization to find a small number of best matching basis vectors [3, 4, 5, 6]. Although common sparsity constraints like an L_1 norm penalty promote discovery of dominant patterns, they do not address the co-occurrence of basis vectors.

Recently, group sparsity [7, 8] has been used for the NMF-based audio source separation [9, 10, 11, 12]. The idea of group sparsity is to allow only a few pre-defined groups of basis vectors to be active. For example, Sun and Mysore proposed the Universal Speech Model(USM) [10] for speaker independent single-channel speech enhancement problem. In the process of USM, regular NMF is first used to learn a large amount of pre-trained speaker-dependent dictionaries, and then during the enhancement stage only a very small number of speakers' dictionaries (groups) that best fit the observed data are active with a group sparsity penalty. Kim proposed the Mixture of Local Dictionaries (MLD) [12] on single-channel speech enhancement and showed improvement over USM. The MLD model learns several small dictionaries for speech source, each of which covers a chunk of similar spectra across all speakers to preserve the source's manifold. And also MLD imposes group sparsity in a frame-by-frame way to activate a small number of dictionaries, which can dynamically find an optimal fit.

Furthermore, some modified group sparsity constraints have been proposed to improve the performance. For example, Badawy proposed relative group sparsity [13] to prevent the activations corresponding to one universal source model from vanishing altogether. Hurmalainen introduced a quadratic penalty function into group sparsity that permits dynamic relationships between basis vectors or groups, since the basic form of group sparsity assumes the independence of different groups without considering which groups will activate, alone or together [14].

However, the group sparsity constraints described above ignore the relative importance of different groups and penalize their activations with a same sparsity parameter. In this paper, we propose adaptive group sparsity to model the relative importance of different groups with adaptive sparsity parameters and investigate its potential benefit for unsupervised source separation based on the USM and the MLD model. In particular, the proposed group sparsity adapts the sparsity parameter according to the activations of each group, since the activations reflect the importance of the particular group for modeling the observed data. Experimental results show that the proposed adaptive group sparsity improves the performance over regular group sparsity for separating speech from other sound sources in the unsupervised setting.

The rest of this paper is organized as follows. Section 2 reviews the standard NMF-based source separation. Section 3 describes the original USM and the MLD algorithms with group sparsity. In Section 4 the proposed adaptive group sparsity for NMF-based source separation is introduced in detail. The experimental setup and evaluation results are presented in Section 5. Finally the paper is concluded in Section 6.

2. Standard NMF-based source separation

NMF factorizes a non-negative matrix $\mathbf{X} \in \mathbb{R}_+^{M \times N}$ (the magnitude spectrogram of audio signal) into the product of a dictionary $\mathbf{W} \in \mathbb{R}_+^{M \times K}$ and a weight matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$:

$$\mathbf{X} \approx \mathbf{WH} \tag{1}$$

where K denotes the size of the dictionary $\mathbf{W}$. The columns of $\mathbf{W}$ can typically be interpreted as the spectral basis vectors

http://dx.doi.org/10.21437/Interspeech.2016-321

of the sources in the spectrogram. The matrix $\mathbf{H}$ can then be interpreted as the activity of each vector in a given time frame. Factorization (1) can be achieved by minimizing the cost function:

$$J = d_{\mathrm{KL}}(\mathbf{X} \mid \mathbf{WH}) \tag{2}$$

where $d_{\mathrm{KL}}(\cdot \mid \cdot)$ denotes the generalized Kullback-Leibler (KL) divergence [5] between matrices $\mathbf{A}$ and $\mathbf{B}$:

$$d_{\mathrm{KL}}(\mathbf{A} \mid \mathbf{B}) = \sum_{m,n} \left(A_{m,n} \log \frac{A_{m,n}}{B_{m,n}} - A_{m,n} + B_{m,n} \right) \tag{3}$$

The typical pipeline to perform the standard NMF-based source separation in the presence of two sources, say speech and noise, follows the process detailed in [15]:

1) Compute the spectrograms $\mathbf{X}_S$ and $\mathbf{X}_N$ from the speech and noise training data, as well as the spectrogram $\mathbf{X}$ of the test mixed signal.

2) Factorize the spectrograms $\mathbf{X}_i \approx \mathbf{W}_i \bar{\mathbf{H}}_i$ and form the matrix $\mathbf{W} = [\mathbf{W}_S \quad \mathbf{W}_N]$.

3) Learn the activations $\mathbf{H}$ from the mixture spectrogram $\mathbf{X}$ while keeping $\mathbf{W}$ fixed: $\mathbf{X} \approx \mathbf{WH}$.

4) Partition the activations as $\mathbf{H} = \begin{bmatrix} \mathbf{H}_S \\ \mathbf{H}_N \end{bmatrix}$, and construct the estimated speech spectrogram $\hat{\mathbf{X}}_S = \mathbf{W}_S \mathbf{H}_S$. The estimated speech waveform can be obtained by combining it with the mixture phase and taking the inverse Short Time Fourier Transform (STFT).

This algorithm is called $supervised$ separation. In the $semi\text{-}supervised$ separation either the speech or noise training set is unavailable. If the noise training data is unavailable, $\mathbf{W}_N$ is learned from $\mathbf{X}$ in step 3, while only $\mathbf{W}_S$ is learned in step 2.

3. NMF-based source separation with group sparsity

In the separation of speech and noise with USM, a large amount of pre-trained speaker-dependent dictionaries $\mathbf{W}^{(g)}$ are first learned using regular NMF. The universal speech model is then obtained by concatenating the learned dictionaries into a single large matrix:

$$\mathbf{W}_S = [\mathbf{W}_S^{(1)}, \cdots, \mathbf{W}_S^{(G)}] \tag{4}$$

where G is the number of small dictionaries. In the separation stage, the mixed spectrogram $\mathbf{X}$ is decomposed by minimizing the cost function:

$$J = d_{\mathrm{KL}}(\mathbf{X} \mid \mathbf{WH}) + \lambda \Omega(\mathbf{H}_S) \tag{5}$$

where $\mathbf{W} = [\mathbf{W}_S \quad \mathbf{W}_N]$, $\mathbf{H} = \begin{bmatrix} \mathbf{H}_S \\ \mathbf{H}_N \end{bmatrix}$ and $\mathbf{H}_S = [{\mathbf{H}_S^{(1)}}^T, \cdots, {\mathbf{H}_S^{(G)}}^T]^T$. The first term stands for the KL-divergence from the original NMF algorithm. The function Ω is the group sparsity-inducing penalty which is used to find a small number of speakers' dictionaries that best fit the observed data. In [10], the log/l_1 penalty defined as $\Omega(\mathbf{H}_S) = \sum_g \log(\epsilon + ||\mathbf{H}_S^{(g)}||_1)$ is applied for its monotonicity and induced multiplicative updates. λ is the sparsity parameter that controls the tradeoff between separation and artifacts. An iterative algorithm is derived by majorization-minimization, which is described in [10] in detail.

In the NMF-based source separation using MLD , several small dictionaries are learned for the speech source and each dictionary covers similar spectra across all speakers to preserve the manifold of speech source. And during the separation MLD activates only a small number of dictionaries for a given noisy input spectrum in the frame-by-frame way, which can model the dynamics of spectra. Results have shown that MLD performs better than USM for unsupervised source separation. The cost function of the MLD approach is defined as:

$$J = d_{\mathrm{KL}}(\mathbf{X} \mid \mathbf{WH}) + \lambda \sum_t \Omega(\mathbf{h}_{S,t}) \tag{6}$$

where $\mathbf{H}_S = [\mathbf{h}_{S,1}, \cdots, \mathbf{h}_{S,N}]$ and $\mathbf{h}_{S,t} = [{\mathbf{h}_{S,t}^{(1)}}^T, \cdots, {\mathbf{h}_{S,t}^{(G)}}^T]^T$. $\Omega(\mathbf{h}_{S,t}) = \sum_g \log(\epsilon + ||\mathbf{h}_{S,t}^{(g)}||_1)$ is the group sparsity-inducing penalty that penalizes the activations in the frame-by-frame way. λ is the penalty parameter the same as used in Equation (5). The differences between Equation (5) and Equation (6) are the way to obtain $\mathbf{W}_S$ and the way the group sparsity function penalizes the activations, which make the separation performance different.

USM and MLD are mainly used for unsupervised and semi-supervised cases. In the unsupervised scenario, the speech dictionary are learned using third-party speech signals. Then, the noise dictionary is learned form the mixture. In the semi-supervised case where only the type of noise is known, it can be solved in a supervised case with the noise dictionary and the suboptimal speech dictionary.

4. Adaptive group sparsity for NMF-based source separation

According to the procedure described above, the group sparsity functions of both the USM and the MLD algorithms penalize different groups with a same sparsity parameter λ, so they rely heavily on the iterative algorithm to find the best fit groups for modeling the observed data. While these models promote general group sparsity, for many purposes it would be beneficial to have more control on how to select the groups in the iterative algorithm. To this end, we propose adaptive group sparsity, where the parameter λ is different and adaptive for each group in the iterative algorithm according to the values of activations.

4.1. NMF-based source separation using USM with adaptive group sparsity

In the NMF-based source separation using USM with adaptive group sparsity, the cost function in Equation (5) is modified as:

$$J = d_{\mathrm{KL}}(\mathbf{X} \mid \mathbf{WH}) + \sum_g \lambda^{(g)} \log(\epsilon + ||\mathbf{H}_S^{(g)}||_1) \tag{7}$$

where $\lambda^{(g)}$ is the sparsity parameter for the g-th group. In the proposed adaptive group sparsity, $\lambda^{(g)}$ is different for each group and adaptive in the iterative algorithm. The intuition is that if a group is very important to model the observed data, then the corresponding sparsity parameter should be small to preserve the group. If a group is negligible for the observed data, then the corresponding sparsity parameter should be large to penalize the group. To this end, we adapt the sparsity parameter according to the activations of each group as:

$$\lambda^{(g)} = \lambda_0 \frac{\max_g \{H_{S,sum}^{(g)}\}}{H_{S,sum}^{(g)}} \tag{8}$$

where $H_{S,sum}^{(g)} = ||\mathbf{H}_S^{(g)}||_1$ and λ_0 is the regularization parameter. In Equation (8) we can see that the sparsity parameter $\lambda^{(g)}$

is negative correlated with the activations $H^{(g)}_{S,sum}$. Therefore, $\lambda^{(g)}$ is able to be preserve groups with large activations and neglect groups with small activations, which is beneficial to the separation performance. Starting from the USM approach in Section 3 and Equation (7)(8), we derive adaptive group sparsity for the NMF-based source separation as presented in Algorithm 1. In order to find the groups more accurately and avoid trapped in a bad local minimum, regular USM method is used to initialize the activations $\mathbf{H}_S$ and $\mathbf{H}_N$.

4.2. NMF-based source separation using MLD with adaptive group sparsity

In the NMF-based source separation using MLD with adaptive group sparsity, the cost function in Equation (6) is modified as:

$$J = d_{\mathrm{KL}}(\mathbf{X} \mid \mathbf{WH}) + \sum_t \sum_g \lambda_t^{(g)} \log(\epsilon + ||\mathbf{h}_{S,t}^{(g)}||_1) \quad (9)$$

where $\lambda_t^{(g)}$ is the adaptive sparsity parameter for the g-th group in the t-th frame. Here a approach similar with Equation (8) is used to adapt $\lambda_t^{(g)}$ in the frame-by-frame manner, which is defined as:

$$\lambda_t^{(g)} = \lambda_0 \frac{\max_g\{h_{S,sum,t}^{(g)}\}}{h_{S,sum,t}^{(g)}} \quad (10)$$

where $h_{S,sum,t}^{(g)} = ||\mathbf{h}_{S,t}^{(g)}||_1$ and λ_0 is the regularization parameter. Details of adaptive group sparsity with the MLD approach is presented in Algorithm 2. And also, regular MLD method is used to initialize the activations $\mathbf{H}_S$ and $\mathbf{H}_N$.

Algorithm 1 The source separation algorithm using USM with adaptive group sparsity

1) Input: $\mathbf{X} \in \mathbb{R}_+^{M\times N}$, $\{\mathbf{W}_S^{(g)} \in \mathbb{R}_+^{M\times R_S} | 1 \le g \le G\}, \mathbf{W}_N$(optional, semi-supervised) or R_N(optional, unsupervised)
2) Output: $\mathbf{H}_S, \mathbf{H}_N$
3) Initialize $\mathbf{H}_S$ and $\mathbf{H}_N$ with output of regular USM algorithm
4) **Repeat**
  $\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T(\frac{\mathbf{X}}{\mathbf{WH}})}{\mathbf{W}^T\mathbf{1}}$
5) for $g = 1, \cdots, G$
  $H_{S,sum}^{(g)} = ||\mathbf{H}_S^{(g)}||_1$
  end for
  for $g = 1, \cdots, G$
  $\lambda^{(g)} = \lambda_0 \frac{\max_g\{H_{S,sum}^{(g)}\}}{H_{S,sum}^{(g)}}$
  $\mathbf{H}_S^{(g)} \leftarrow \mathbf{H}_S^{(g)} / \{1 + \frac{\lambda^{(g)}}{\epsilon + ||\mathbf{H}_S^{(g)}||_1}\}$
  end for
  if Unsupervised then
  $\mathbf{W}_N \leftarrow \mathbf{W}_N \otimes \frac{(\frac{\mathbf{X}}{\mathbf{WH}})\mathbf{H}_N^T}{\mathbf{1}\mathbf{H}_N^T}$
  end if
6) **until convergence**

Algorithm 2 The source separation algorithm using MLD with adaptive group sparsity

1) Input: $\mathbf{X} \in \mathbb{R}_+^{M\times N}$, $\{\mathbf{W}_S^{(g)} \in \mathbb{R}_+^{M\times R_S} | 1 \le g \le G\}, \mathbf{W}_N$(optional, semi-supervised) or R_N(optional, unsupervised)
2) Output: $\mathbf{H}_S, \mathbf{H}_N$
3) Initialize $\mathbf{H}_S$ and $\mathbf{H}_N$ with output of regular MLD algorithm
4) **Repeat**
  $\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T(\frac{\mathbf{X}}{\mathbf{WH}})}{\mathbf{W}^T\mathbf{1}}$
  for $g = 1, \cdots, G, t = 1, \cdots, N$
  $h_{S,sum,t}^{(g)} = ||\mathbf{h}_{S,t}^{(g)}||_1$
  end for
  for $g = 1, \cdots, G, t = 1, \cdots, N$
  $\lambda_t^{(g)} = \lambda_0 \frac{\max_g\{h_{S,sum,t}^{(g)}\}}{h_{S,sum,t}^{(g)}}$
  $\mathbf{h}_{S,t}^{(g)} \leftarrow \mathbf{h}_{S,t}^{(g)} / \{1 + \frac{\lambda_t^{(g)}}{\epsilon + ||\mathbf{h}_{S,t}^{(g)}||_1}\}$
  end for
  if Unsupervised then
  $\mathbf{W}_N \leftarrow \mathbf{W}_N \otimes \frac{(\frac{\mathbf{X}}{\mathbf{WH}})\mathbf{H}_N^T}{\mathbf{1}\mathbf{H}_N^T}$
  end if
5) **until convergence**

5. Experiments and Results Analysis

5.1. Preparation of the Dataset

The proposed algorithms were evaluated in separating speech from other sound sources. All signals were sampled at 16kHz. To calculate spectral vectors, STFT was performed with 64ms analysis hanning window and 16ms window shift. Twenty speakers (10 sentences each) randomly chosen from the training set of TIMIT corpus were used as general data to learn speaker independent dictionaries. Each of 5 held-out speakers from the test set of TIMIT corpus and each of 10 noise examples were mixed for a total of 50 test examples. The noise examples were from [16], which included nonstationary noises, such as computer keyboards and birds. The test sets were mixed with varying signal-to-noise ratio (SNR) (-10, -5, 0, 5 and 10dB), along with the corresponding clean speech utterances. The performance was evaluated using signal-to-distortion ratio (SDR), signal-to-artifact ratio (SAR) and signal-to-interference ratio (SIR), which were calculated by BSS-EVAL [17], to measure the suppression of the noise and the artifacts of speech signal that introduced by the separation process.

5.2. Algorithms

We compared the proposed algorithms with the USM and the MLD methods with regular group sparsity. A speaker-dependent NMF method was also used as the baseline [15]. For the speaker-dependent NMF algorithm, the left 9 sentences of each speaker in the test set were used to train the speaker-dependent dictionary with 20 basis vectors. For the USM, the spectrogram of each speaker was factorized to learn a speech dictionary and each dictionary held $R_S = 10$ basis vectors. For the MLD, 20 local dictionaries, each of which held $R_S = 10$ bases were learned using the algorithm described in [12]. The number of MM iterations and the parameter λ in both USM and MLD methods were chosen that lead to the best average SDR score. In the proposed algorithms, regular USM and MLD algorithms were implemented by running 10 MM iterations first. Then, the estimated activations were used as initial values in Algorithm 1 and Algorithm 2, respectively. And also, the number of MM iterations and the parameter λ_0 were chosen to obtain the best average SDR score. In all of the algorithms, the number

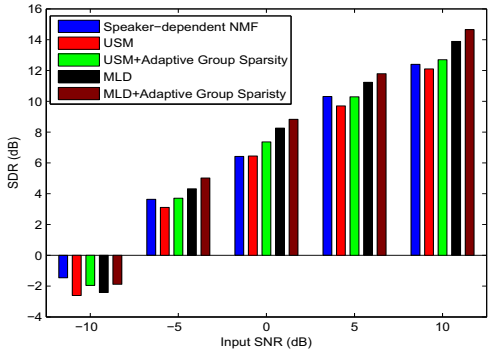

Figure 1: *The averaged SDR of the separated speech as a function of input SNRs.*

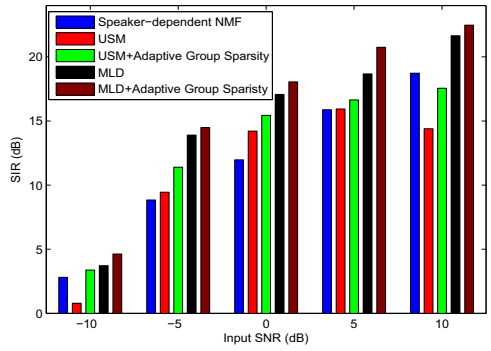

Figure 2: *The averaged SIR of the separated speech as a function of input SNRs.*

of noise bases were fixed with the optimal ones investigated in [16].

5.3. Results Analysis

The averaged SDR, SIR and SAR of the separated speech signal as a function of the input SNRs in the unsupervised case are illustrated in Figs 1-3, respectively. For all the input SNRs in the experiment, the proposed algorithms have an improvement in the SDR compared with the baseline approaches. Specifically, there is a significant improvement in terms of SIR and comparable SAR. This means that the proposed methods are able to suppress more noise without introducing more artifacts than the baselines. The reason is that the proposed algorithms penalize different groups with different and adaptive parameters. Therefore, the proposed algorithms are able to preserve the most important groups and suppress nonessential groups better, which is the reason to suppress more noise and do not produce more artifacts.

In addition, we further compared the proposed algorithms with the baselines in the semi-supervised setting as the input SNR was 0dB. In the semi-supervised setting, a noise dictionary for each noise type was learned, and then Algorithm 1 and Algorithm 2 were implemented without the unsupervised option. The results are shown in Table 1. Compared with the USM and MLD methods with regular group sparsity, the proposed methods produce comparable results. A possible explanation is that

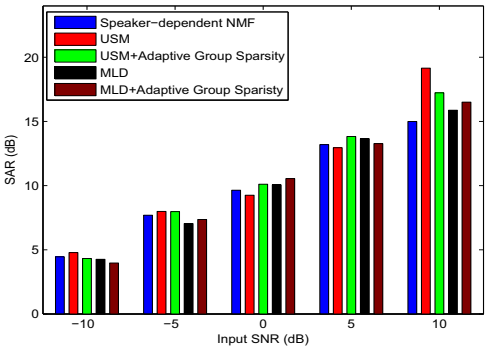

Figure 3: *The averaged SAR of the separated speech as a function of input SNRs.*

Table 1: *The averaged SDR, SIR and SAR of the separated speech in the semi-supervised case.*

	SDR(dB)	SIR(dB)	SAR(dB)
Speaker-dependent NMF	9.54	15.97	11.57
USM	9.98	19.15	11.36
USM+Adaptive Group Sparsity	9.87	20.12	11.02
MLD	10.47	21.22	11.24
MLD+Adaptive Group Sparsity	10.51	21.67	11.16

once the speech dictionary and noise dictionary are both learned in advance, regular group sparsity can find the best groups well and imposing adaptive group sparsity does not help much.

6. Conclusions

In this paper we proposed the notion of adaptive group sparsity for single-channel source separation. In particular, the proposed adaptive group sparsity penalized different groups with different and adaptive sparsity parameters according to the relative importance of this groups for modeling the observed data. We investigated it for separating speech and other sound sources especially in the unsupervised setting where neither the speaker identity nor the type of noise was known in advance. Experiments with mixtures containing various noise types showed that the proposed adaptive group sparsity outperformed conventional group sparsity in the unsupervised case. However, in the semi-supervised case where the noise type was known in advance, the proposed adaptive group sparsity did not improve the performance. For the future work, we expect to apply the proposed method for other separation tasks.

7. Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Nos. 11461141004, 11590770, 11590771, 11590772, 11590773, 11590774), the Strategic Priority Research Program of the Chinese Academy of Sciences (Nos. XDA06030100, XDA06030500), National 863 Program (No. 2015AA016306) and the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No. 201230118-3).

8. References

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] ——, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[3] P. O. Hoyer, "Non-negative sparse coding," in *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*. IEEE, 2002, pp. 557–565.

[4] M. Schmidt and R. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," 2006.

[5] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, 2007.

[6] T. Virtanen, B. Raj, J. F. Gemmeke *et al.*, "Active-set newton algorithm for non-negative sparse coding of audio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3092–3096.

[7] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2011.

[8] S. Bengio, F. C. N. Pereira, Y. Singer, and D. Strelow, "Group sparse coding." in *Advances in Neural Information Processing Systems 22: Conference on Neural Information Processing Systems 2009. Proceedings of A Meeting Held 7-10 December 2009, Vancouver, British Columbia, Canada*, 2009, pp. 82–89.

[9] A. Lefvre, F. Bach, and C. Fvotte, "Itakura-saito nonnegative matrix factorization with group sparsity," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 2011, pp. 21–24.

[10] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 2013, pp. 141–145.

[11] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation," in *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*, 2014, pp. 1–6.

[12] M. Kim and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 22, pp. 288–292, 2015.

[13] D. El Badawy, A. Ozerov, and N. Q. K. Duong, "Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015.

[14] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Similarity induced group sparsity for non-negative matrix factorisation," in *40th IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2015.

[15] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Independent Component Analysis and Signal Separation, 7th International Conference, ICA 2007, London, UK, September 9-12, 2007.*, 2007, pp. 414–421.

[16] Z. Duan, G. J. Mysore, and P. Smaragdis, "Online plca for real-time semi-supervised source separation," in *Proceedings of the 10th international conference on Latent Variable Analysis and Signal Separation*, 2012, pp. 34–41.

[17] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.