Single Channel Source Separation Using Sparse NMF and Graph Regularization

Tuan Pham[a1], Yuan-Shan Lee[a2], Yan-Bo Lin[a3], Tzu-Chiang Tai[b4], and Jia-Ching Wang[a5]
[a]Dept. of Computer Science and Information Engineering, National Central University,Taiwan
[b]Department of Computer Science and Information Engineering, Providence University, Taichung, Taiwan
{[1]103582605, [2]102582003, [3]102502529}@cc.ncu.edu.tw, [4]tctai@pu.edu.tw, [5]jcw@csie.ncu.edu.tw

ABSTRACT
The aim of single channel source separation is to accurately recover signals from mixtures. In supervised case, non-negative matrix factorization (NMF) is a popular method to separate mixed signals from learned dictionaries. These dictionaries can be produced efficiently by sparse NMF to approximate the input signal as closely as possible. However, previous methods neither consider the structure of the data in terms of the similarity between vertices of the input signal nor use state-of-art variants of NMF that are more efficient than conventional ones. This paper presents a method that incorporate graph regularization constraint into a group sparsity NMF to improve the performance of source separation. Experimental results demonstrate that our method is outstandingly effective for speech separation in two representative scenarios.

CCS Concepts
• Computing methodologies→Machine learning→Unsupervised learning→Source separation.

Keywords
Graph regularization; non-negative matrix factorization; sparse coding; source separation.

1. INTRODUCTION
As technology keeps improving in the modern world, human-computer interaction becomes an essential part in our daily life. The speech-based human-computer interface has been gaining increasing interests owing to its accessible, natural, and easy-to-use properties. The interface should provide automatic machine perception of auditory scenes in un-controlled environments. One of the important topics is to develop a speech recognition system capable of performing source separation of the target speech in the presence of multiple competing sound sources in natural environments. Therefore, this paper proposes a sparse non-negative matrix factorization (NMF) based source separation method.

Source separation has been a popular topic of research in the last few decades. It has various potential applications in speech signal processing such as hearing aids, automatic speech recognition and speech coding. A single-channel source separation (SCSS) system is a classical issue in auditory scene analysis. In particular, SCSS is used in the case that only one microphone is available, aims to extracting specific speakers' signals from a single mixed signal.

In recent years, numerous SCSS approaches have been extensively proposed and most fall into two groups: model-based [20, 23, 24] and data-driven [2, 4, 18]. Model-based SCSS attempt to find an efficient model of NMF and dictionary learning to separate mixture signals or enhance a target speaker. However, data-driven methods typically seek discriminative features of mixture signal or use priori information to separate the source signals. Generally, source separation problems can be solved based on numerous approaches in which NMF is used to construct a dictionary learning and to estimate a specific speaker from mixed speech signals.

In single-channel BSS, signals from several unknown sources are usually mixed. Mathematically, mixed speech can be treated as a mixture of N unknown source signals,

$$x(t) = s_1(t) + s_2(t) + \ldots + s_N(t) \quad (1)$$

where t represents time.

Source separation estimates the sources $s_n(t)$, $\forall n \in N$ of length T when from only the mixed signal $x(t)$. Without loss of generality, experiments are carried out herein to test the proposed approach on a mixture of two male and female speech signals. Many methods based on NMF have been developed to solve this problem in unsupervised, semi-supervised or supervised fashion. In the unsupervised case, the separation is conducted by finding a decomposition in which the sources are assumed to be statistically independent. In the semi-supervised case, either speech or the noise dictionary is missing**.** The missing dictionary is derived in the separation phase. In the supervised case, both speech and noise are known and these signals can be utilized to approximate the speech data. Additionally, these dictionaries are learned from a training process and generally fixed in the separation phase.

NMF [1] has many applications, such as audio separation [13, 19, 17, 20, 22, 24], speech enhancement [2, 9], image processing [15], or document clustering [25, 26]. NMF factorizes an original matrix as the product of a basis vector matrix and a coefficient matrix whose elements are all non-negative. In NMF-based blind source separation (BSS), the dictionary matrix is generated using clean speech or noise. The activation matrix is usually a sparse matrix*,* meaning that most of its elements are zero [5, 6, 9]. Many variants of NMF have been developed based on the work of Lee and Seung [1] with additional constraints, such as the sparse constraint that penalizes non-sparse vectors [2, 4, 5, 9, 17], and temporal continuity. Recently, sparse NMF with β-divergence [3, 4, 13] has been utilized to improve efficiently performance of speech separation. However, this approach neglects the structure of the data (the relationship between vertices of a graph).

ASE BD&SI 2015, October 07 - 09, 2015, Kaohsiung, Taiwan

ACM 978-1-4503-3735-9/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2818869.2818913

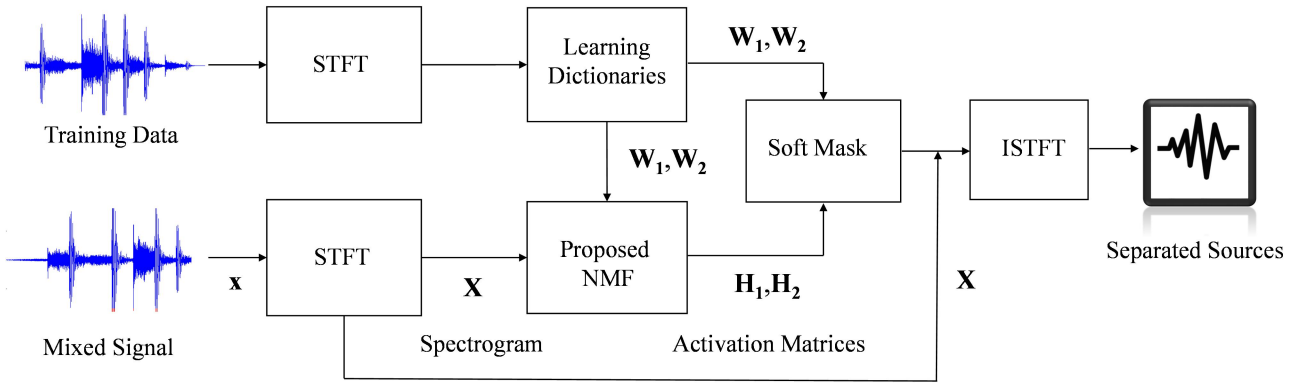

Figure 1. The block diagram of proposed method.

Source separation has been a popular topic of research in the several decades. To solve this problem, a sparseness constraint or a log/l1 penalty has been utilized to make the coefficient vector [2, 5, 6]. Some literature has considered the structure of speech data [2, 7].

These methods add constraints to the original NMF, such as a sparsity constraint to penalize vectors that are not sparse or to measure similarity between a cluster centroid and a basic vector. Another relevant work, mixture local dictionary [2] was introduced to estimate distance between centroid of each local dictionary and dictionary matrix. Each local dictionary, which can be learned beforehand by using clustering techniques, is considered as a priori knowledge of dictionary matrix.

Another method involves model-based source separation, in which used NMF to incorporate a deep neural network (DNN) [20] into NMF. Source separation base on collaboration of NMF and deep neural network comprises three phases: NMF training, DNN training and speech separation phases. In the NMF training phases, conventional NMF is utilized to produce basis matrix for each source. To reconstruction each source form mixture, optimal encoding vector (activation vector), which was traditionally produced by using NMF, must be estimated. However, [20] used a deep neural network to estimate the encoding vectors faithfully reconstruct the desired source data vectors. This framework can be enhanced by including sparse coding and graph regularization to produce better dictionary in the first step. These past approaches focus majorly on adding constraint or incorporating other modern techniques; however, the structure of the data was not considered.

Lastly, another relevant work [27] used graph regularized sparse coding and NMF in image presentation. The important contribution of our work is to use β-divergence NMF, group sparsity and graph regularization to improve upon sparse NMF (SNMF) by imposing graph regularization [10, 15], which represents the closeness of speech samples, and using group sparsity instead of l1-norm. This paper is extension work of [19] with advanced algorithm, more experiments on two datasets and more scenarios. Figure 1 shows the block diagram of proposed method, including source-specific dictionary learning [18] and activation matrix that corresponds to learned dictionaries. Experiments in two representative scenarios reveal that the proposed algorithm is outstandingly effective in speech separation.

The rest of this paper is organized as follows. Section 2 presents the proposed model and its related algorithms. Section 3 presents experimental results concerning speech separation. Finally, Section 4 draws conclusions.

2. SOURCE SEPARATION USING SPARSE NMF WITH BETA-DIVERGENCE AND GRAPH REGULARIZATION

This section briefly reviews the sparse NMF with β-divergence and then develops the proposed group sparsity NMF with β-divergence and graph regularization in source separation.

2.1 Sparse NMF with β-divergence

Given a non-negative dimensional data matrix $\mathbf{V}_{signal} \in \mathbb{R}_{+}^{m \times n}$. NMF proposed by [1] aims to decompose original matrix into basis and coefficient matrix. To estimate $\mathbf{W}^{m \times k} \in \mathbb{R}_{+}^{m \times k}$ and $\mathbf{H}^{k \times n} \in \mathbb{R}_{+}^{m \times k}$, an objective function $D(\mathbf{V} \mid \mathbf{WH})$ was used to estimate reconstruction error and it is iteratively minimized through a multiplicative update rule,

$$(\mathbf{W},\mathbf{H}) = \underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}} F = \underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}} D(\mathbf{V} \mid \mathbf{WH}) \tag{2}$$

To generalize reconstruction metric, β-divergence introduced by [3] has been proved that it can produce better separation performance [6, 16] and the β-divergence [3] is defined as follows,

$$D_{\beta}(x \mid y) = \begin{cases} \dfrac{1}{\beta(\beta-1)}(x^{\beta} + (\beta-1)y^{\beta} - \beta x y^{\beta-1}) & \beta \in \mathbb{R} \setminus \{0,1\} \\ x \log \dfrac{x}{y} + (y - x) & \beta = 1 \\ \dfrac{x}{y} - \log \dfrac{x}{y} - 1 & \beta = 0 \end{cases} \tag{3}$$

In β-divergence, $\beta = 0$ yields the Itakura-Saito (IS) distance [16]; $\beta = 1$ yields the generalized Kullback-Leibler (KL) divergence, and $\beta = 2$ yields the Euclidean distance. In particular, the cost function to be minimized is,

$$(\mathbf{W},\mathbf{H}) = \underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}} F = \underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}} D_{\beta}(\mathbf{V} \mid \mathbf{WH}) + \mu \|\mathbf{H}\|_1 \tag{4}$$

where the first term denotes NMF with β-divergence [3] and the second term is an sparse constraint which aims to enhance quality of produced dictionary. The multiplicative update rules are widely used to minimize (4) because of their simplicity and efficiency. The multiplicative update rules that preserve the non-negativity of $\mathbf{H}$ and $\mathbf{W}$ are given by,

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\mathbf{\Lambda}^{\beta-2} \otimes \mathbf{V} + \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{T}\mathbf{\Lambda}^{\beta-1})\mathbf{H}^{T}}{(\mathbf{\Lambda}^{\beta-1} + \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{T}(\mathbf{\Lambda}^{\beta-2} \otimes \mathbf{V}))\mathbf{H}^{T}} \tag{5}$$

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\tilde{\mathbf{W}}^T (\mathbf{V} \otimes \boldsymbol{\Lambda}^{\beta-2})}{\tilde{\mathbf{W}} \boldsymbol{\Lambda}^{\beta-1} + \mu} \tag{6}$$

where the matrix $\boldsymbol{\Lambda} = \tilde{\mathbf{W}}\mathbf{H}$, and $\tilde{\mathbf{W}}$ is the column-wise normalized version of $\mathbf{W}$ Operation $\otimes$ is element-wise multiplication, and the division operation is also carried out in an element-wise fashion.

2.2 Sparse NMF with β-divergence and Graph Regularization

According to manifold learning theory [10], the geometric structure of input data can be efficiently modeled by graph regularization. A weight matrix **U,** used to measure the closeness of two points, can be constructed for *n* nearest neighbors. Given a graph with *N* vertices, each of which corresponds to a data point x_i and $n \in N$ The Laplacian Eigenmap algorithm typically has two steps. The first step is the construction of a graph from input samples, and the second step is the definition of the weight matrix **U**. Nodes *i* and *j* are connected by an edge if *i* is among *n* nearest neighbors of *j* or *j* is among *n* nearest neighbors of *i*. There are two variations to obtain weighted edges in **U** [10]. The first variation is heat kernel was estimated by (7) if nodes *i* and *j* are connected,

$$\mathbf{U}_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{t}} \tag{7}$$

where parameter $t \in \mathbb{R}$.

Another simplification of weighting **U** is simple-minded without parameter *t* . $\mathbf{U}_{ij}$ = 1 if and only if vertices *i* and *j* are connected by an edge,

$$\mathbf{U}_{ij} = \begin{cases} 1 \\ 0 \end{cases} \tag{8}$$

A graph Laplacian matrix is then calculated as,

$$\mathbf{L}=\mathbf{D}-\mathbf{U} \tag{9}$$

where **D** is a diagonal matrix whose entries are column sums of **U**. Finally, the graph regularization term is given by,

$$\begin{aligned} R &= \frac{1}{2}\sum_{i,j=1}^{N} \left\| \mathbf{x}_j - \mathbf{x}_i \right\|^2 \mathbf{U}_{ji} \\ &= \sum_{j=1}^{N} \mathbf{x}_j^T \mathbf{x}_j \mathbf{D}_{jj} - \sum_{j,i=1}^{N} \mathbf{x}_j^T \mathbf{x}_i \mathbf{U}_{ji} \\ &= Tr\left(\mathbf{HDH^T}\right) - Tr\left(\mathbf{HUH^T}\right) \\ &= Tr\left(\mathbf{HLH^T}\right) \end{aligned} \tag{10}$$

The objective function in the proposed method is defined as (11),

$$\underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}} D_\beta(\mathbf{V}|\mathbf{WH}) + \mu\Omega(\mathbf{H_s}) + \alpha Tr(\mathbf{HLH^T}) \tag{11}$$

where the first terms represent NMF with β-divergence, the second term is block-sparsity-inducing which used in our approach instead of l_1-norm, and the third term is the graph regularization term. The terms α and μ control the degree of regularization. Additionally, we have several choices of Ω which relate to monotonicity, convergence or complex of multiplicative update rule [17].

Table 1. List of common group sparsity

Penalty	$\Omega(\mathbf{H}_s)$
l_1/l_∞	$\sum_{g=1}^{M} \left\|\mathbf{H}_g\right\|_\infty$
l_1/l_2	$\sum_{g=1}^{M} \left\|\mathbf{H}_g\right\|_2$
$\log/l_1$	$\sum_{g=1}^{M} \log(\varepsilon + \left\|\mathbf{H}_g\right\|_1)$

In particular, $\log/l_1$ penalty is suggested in [2, 6, 17] because of its monotonicity and induced multiplicative updates. The equation of second term is described below,

$$\Omega(\mathbf{H}_s) = \sum_{g=1}^{M} \log(\varepsilon + \left\|\mathbf{H}_g\right\|_1) \tag{12}$$

where *g* is group index

The objective function (11) can also be minimized by applying multiplicative update rules. The obtained multiplicative update rules in the proposed algorithm are given by (13), (14) and (15),

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\boldsymbol{\Lambda}^{\beta-2} \otimes \mathbf{V} + \tilde{\mathbf{W}}\tilde{\mathbf{W}}^T \boldsymbol{\Lambda}^{\beta-1})\mathbf{H}^T}{(\boldsymbol{\Lambda}^{\beta-1} + \tilde{\mathbf{W}}\tilde{\mathbf{W}}^T (\boldsymbol{\Lambda}^{\beta-2} \otimes \mathbf{V}))\mathbf{H}^T} \tag{13}$$

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\tilde{\mathbf{W}}^T (\mathbf{V} \otimes \boldsymbol{\Lambda}^{\beta-2}) + \alpha\mathbf{HU}}{\tilde{\mathbf{W}} \boldsymbol{\Lambda}^{\beta-1} + \alpha\mathbf{HD}} \tag{14}$$

$$\mathbf{h}_g \leftarrow \frac{1}{1 + \mu / (\varepsilon + \left\|\mathbf{h}_g\right\|_1)} \mathbf{h}_g \tag{15}$$

2.3 Source Separation Using Sparse NMF with β-divergence and Graph Regularization

The short-time Fourier transform (STFT) is used to generate the spectrogram of a signal. Taking the magnitude of each time-frequency point in a spectrogram yields a non-negative matrix. Let $\mathbf{V}_{mix} \in \mathbb{R}_+^{m \times n}$, $\mathbf{V}_{source1} \in \mathbb{R}_+^{m \times n}$, and $\mathbf{V}_{source2} \in \mathbb{R}_+^{m \times n}$ denote the non-negative matrixes corresponding to the mixed signal and two source signals, respectively. In the training mode, the proposed method finds the dictionary matrix $\mathbf{W}_{source1}$ and $\mathbf{W}_{source2}$ that corresponds to the source signals. Here **W** and **H** are initialized to random numbers and both are updated to approximate the original signal through the iterations. The separation stage can be segmented into unsupervised mode, supervised mode and semi-supervised mode, respectively. Algorithm 1 describes the proposed speech separation method.

In the unsupervised mode, separation of the unknown source signals is done by learning the activations of the corresponding dictionaries in training mode. The activation matrix is estimated by applying (14). On the other hand, learned dictionaries must be fixed if all sources are known. This case is called the supervised mode in which the dictionary $\mathbf{W}_{dic}$ is:

$$\mathbf{W}_{dic} = [\mathbf{W}_{source1}, \mathbf{W}_{source2}] \tag{16}$$

Algorithm 1 Speech separation using sparse NMF and graph regularization for training and testing

Input: $\mathbf{V}_{mix} \in \mathbb{R}_{+}^{m \times n}$, $\mathbf{V}_{source1} \in \mathbb{R}_{+}^{m \times n}$, and $\mathbf{V}_{source2} \in \mathbb{R}_{+}^{m \times n}$ (In the training mode, the inputs are $\mathbf{V}_{source1}$ and $\mathbf{V}_{source2}$. Otherwise, the input is $\mathbf{V}_{mix}$)

Output: **W** or **H** (In the training mode, the output is dictionary **W**. Otherwise, the output is **H**)

1: Initialize missing matrix **W** or **H** to a random number

2: Construct matrix **U** using (7) or (8) and derive matrix **D** and **L** using **U** and (9)

3: **repeat**

4: Update **H** using (14), and $\mathbf{h}_g$ using (15)

5: **if** training mode

6: Update **W** using (13)

7: **end if**

8: Obj_cost = error sparse NMF + error sparsity + laplacian

9: **until** Convergence

In the semi-supervised case, one source is known. If the source1 is available and the source2 is unknown, then the learned dictionary of source1 is fixed and the dictionary of source2 is updated using (13).

Finally, source signals are recovered using the soft mask as follows [17]:

$$\begin{aligned} \mathbf{V}_{source1} &= \mathbf{V}_{mix} \otimes \left(\mathbf{W}_{source1}\mathbf{H}_{source1}\right) / \left(\mathbf{W}_{dic}\mathbf{H}_{mix}\right) \\ \mathbf{V}_{source2} &= \mathbf{V}_{mix} \otimes \left(\mathbf{W}_{source2}\mathbf{H}_{source2}\right) / \left(\mathbf{W}_{dic}\mathbf{H}_{mix}\right) \end{aligned} \quad (17)$$

where $\mathbf{H}_{mix} = \left[\mathbf{H}_{source1}\ \mathbf{H}_{source2}\right]^T$, operation $\otimes$ is element-wise multiplication, and the division is also carried out in the element-wise operation.

3. EXPERIMENTAL RESULT

The experimental results that were obtained using the proposed algorithm, sparse NMF [4], and sparse-coding-based NMF (SC-NMF) [9, 13] were compared. All speech signals that were used in the experiments were obtained from two datasets, GRID [21] and TIMIT [11], with a sampling rate of 16 kHz. A subset of speakers (eight female and eight male) were chosen at random from the TIMIT dataset. From GRID dataset, two female and two male speakers were selected randomly. Eighty randomly chosen utterances by each speaker were used for training and 20 from each speaker were used for testing. The BSS Eval toolbox [12] was utilized to evaluate the quality of the separated signals in terms of signal-to-distortion (SDR), signal-to-interference ratio (SIR), and signal-to-artifacts ratio (SAR). SDR measures the overall quality of the separated speech while SIR and SAR are proportional to the degree of noise reduction and the inverse of speech distortion, respectively.

The same setting was utilized in the proposed algorithm and the baseline systems. For the proposed algorithm and the baseline systems, the following parameters were set; number of iterations = 400, α = 0.1 and μ = 5, β = 0, M=20 and base number K = 1024. The sparse coding and NMF were implemented according to [13]. The base number K = 50 is set for these two systems and the result became worse as K was increased. In addition, as α and $\mu \to \infty$, worse results are obtained.

The experiments involved in two scenarios. In the first scenario, signals from each speaker were used in both training and testing. This scenario is used for the particular environment in which a priori information is available. For each speaker, a local dictionary was constructed. Therefore, a total of sixteen local dictionaries $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_{16}]$ were constructed.

Table 2. Source separation performance using eight male and eight female speakers for training (TIMIT dataset)

Measure	SC-NMF	Sparse NMF	Proposed
SDR	4.98	8.69	8.95
SIR	6.77	13.19	13.88
SAR	10.77	11.41	11.19

Table 3. Source separation performance using two male and two female speakers for training (GRID dataset)

Measure	SC-NMF	Sparse NMF	Proposed
SDR	5.05	7.22	7.48
SIR	5.33	12.36	13.19
SAR	11.88	9.49	9.37

Tables 2 and 3 present the experimental results on TIMIT and GRID dataset respectively. According to Table 2 and 3, both the proposed algorithm and sparse NMF performed significantly better than the SC-NMF. The proposed algorithm yielded higher values than the sparse NMF for both SIR and SDR, but a slightly worse SAR. The most widely used evaluation measure is SIR and SDR.

In the second scenario, the speaker identities used in training differ from those used in testing. This scenario reflects the real environment in which the priori information is usually unavailable and only a dictionary trained by other speakers can be used. In this scenario, the speech from one female and one male speaker was used to generate a dictionary $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2]$. Mixed sentences spoken by other speakers were generated for testing. Table 4 and 5 presents the experimentally obtained average results. The proposed algorithm outperforms both SC-NMF and sparse NMF from TIMIT dataset. From GRID dataset, the performance of our algorithm is slightly better than baselines because structure of GRID dataset is more suitable for speaker-specific than multi-speaker separation.

Table 4. Source separation performance using one male and one female speaker for training (TIMIT dataset)

Measure	SC-NMF	Sparse NMF	Proposed
SDR	3.42	6.52	7.08
SIR	5.40	9.32	10.41
SAR	9.03	10.80	10.61

Table 5. Source separation performance using one male and one female speaker for training (GRID dataset)

Measure	SC-NMF	Sparse NMF	Proposed
SDR	4.42	6.54	6.52
SIR	4.57	10.37	10.68
SAR	10.78	8.37	8.19

Figures 2 and 3 compare sparse NMF with our proposed algorithm in term of β-divergence. The relevant experiments were conducted on two scenarios, using all available training speakers and using only one male and female speakers, from TIMIT dataset. From Fig. 2, the performance of proposed approach outperforms by SDR, SIR but not SAR. General characteristics of proposed algorithm and sparse NMF are loss of stability of source separation that lead to decrease quality of output signals when $\beta \geq 2$. Both algorithms perform best when $\beta = 0$ and performance worsens as β approaches 2. When β is between 0 and 1.8, the two major evaluation metrics exceed the baseline for the proposed algorithm. In particular, when $\beta = 2$, all of three metrics dropped dramatically and convergence rate is rapid.

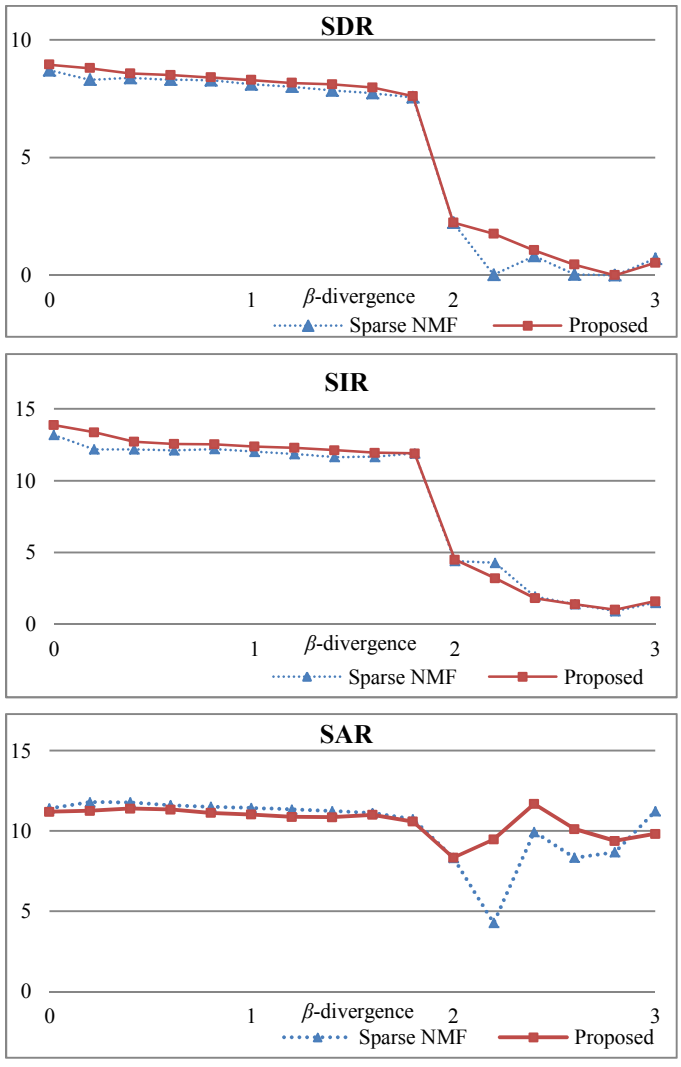

Figure 2. β-divergence study using all of available dictionary

Figure 3 shows β-divergence study of speech separation using one random-dictionary. In this experiment, our proposed algorithm is remarkable when $\beta = 0$, SDR and SIR metric of our method is significantly higher than baseline. Generally, SAR is less vulnerable than SDR and SIR by the value of beta. When β is form 1.8 to 2, all of three metrics also reduce intensely. Lastly, our line graph is above baseline on the most of β value.

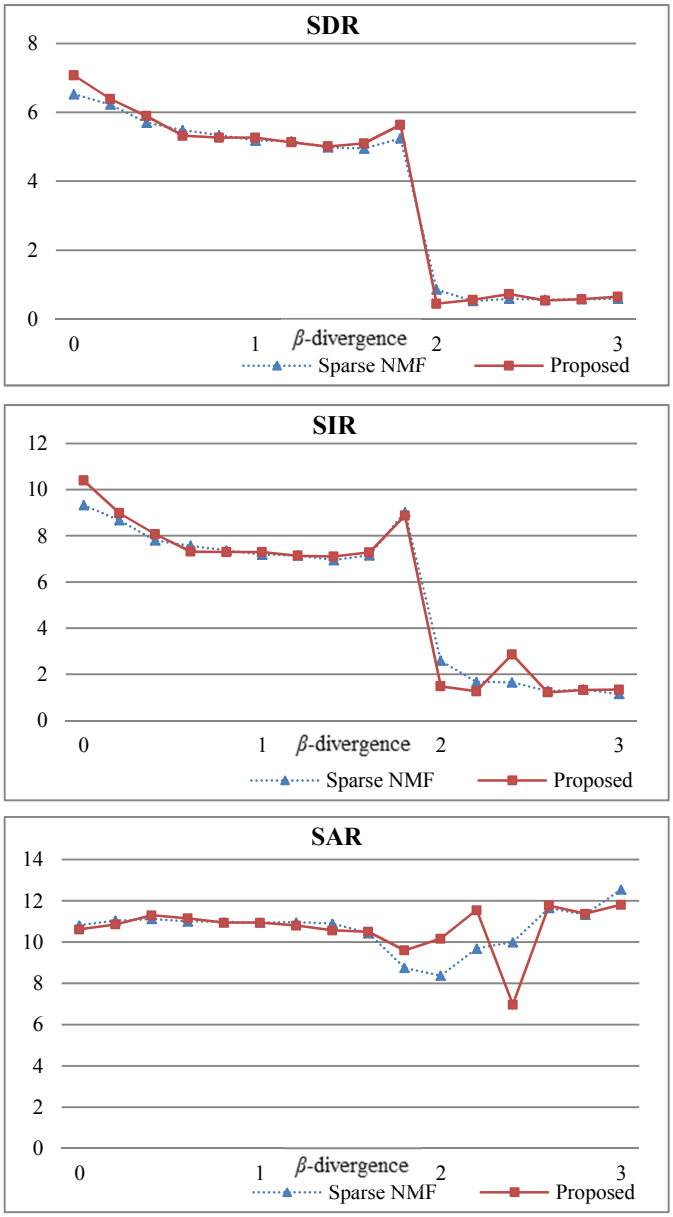

Figure 3. β-divergence study using one random-dictionary

Figure 4 shows the mixed signal and the separated sources using all of dictionaries and sparse NMF β-divergence with graph regularization. Visually, it can be seen that the mixture has been separated efficiently comparing with the original sources.

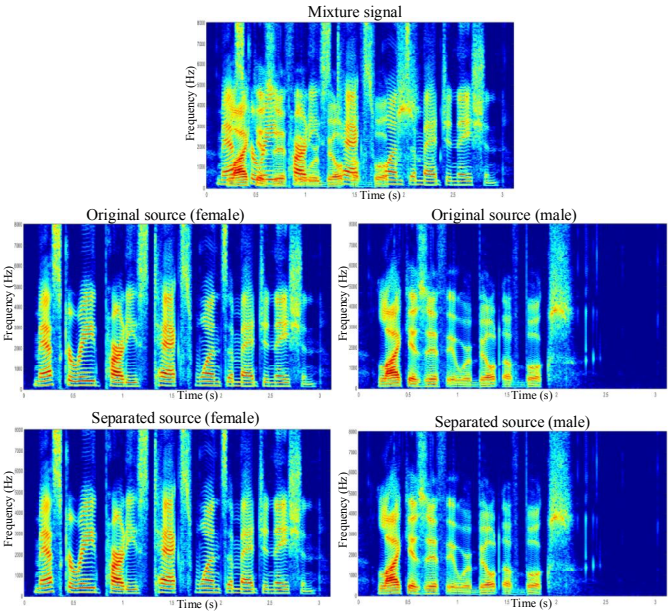

Figure 4. Two original sources, observed mixture and two separated sources.

4. CONCLUSIONS

This work proposed source separation method using group sparsity NMF with β-divergence and graph regularization. The proposed method is an extension of the state-of-art group sparsity NMF with β-divergence, which imposes graph regularization to take into account the structure of signals. The experimental results demonstrate that the proposed algorithm improves the overall quality of the separated speeches above that in previous studies. Our future work will extend the proposed algorithm by taking into account the correlation of training and testing phases i.e. objective function of separation process will include objective function of training (bi-level optimization). Other techniques, such as DNN or probabilistic latent component analysis (PLCA), can be incorporated into our framework. Accordingly, the problem of speech separation can be solved using group sparsity NMF with β-divergence and underlying manifold of input data.

5. REFERENCES

[1] Lee, D. D. and Seung, H. S. 2001. Algorithms for non-negative matrix factorization, *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 13.

[2] Minje, K. and Smaragdis, P. 2015. Mixtures of local dictionaries for unsupervised speech enhancement, *IEEE Signal Processing Letters*. 22, 3 (March. 2015), 293 - 297.

[3] Févotte, C. and Idier, J. 2011. Algorithms for nonnegative matrix factorization with the beta-divergence, *Neural Computation*.

[4] Roux, J. L. , Weninger, F., and Hershey, J. R. 2015. Sparse NMF – half-baked or well done?, *Mitsubishi Electric Research Laboratories Technical Report*. (Mar. 2015).

[5] Hoyer, P. 2004. Non-negative matrix factorization with sparseness Constraints, *J. Mach. Learn. Res.* 5, 1457-1469.

[6] Lefèvre, A., Bach, F., and Févotte, C., Itakura-Saito non-negative matrix factorization with group sparsity, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*.

[7] Hurmalainen, A., Saeidi, R., and Virtanen, T. 2015. Similarity induced group sparsity for non-negative matrix factorization, in *Proc. ICASSP 2015*. (Brisbane, Australia, April. 2015).

[8] Eguchi, S. and Kano, Y. 2001. Robustifying maximum likelihood estimation, *ISM Research Memo*. (June. 2001).

[9] Eggert, J. and Körner, E. 2004. Sparse coding and NMF, in *Proc. IEEE International Joint Conference on Neural Networks*. 4, 2529 - 2533.

[10] Belkin, M. and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering, *Advances in neural information processing systems*. Cambridge, MA: MIT Press.

[11] Seneff, S., Glass, J., Zue, V. 1990. Speech database development at MIT: Timit and beyond, *Speech Communication*. 9, 4 (Aug. 1990), 351-356.

[12] Vincent, E. , Gribonval, R., and Févotte, C. 2006. Performance measurement in blind audio source separation, *IEEE Trans. Audio, Speech and Language Processing*. 14, 1462-1469.

[13] Mikkel, N. 2007. Speech separation using non-negative features and sparse non-negative matrix factorization, *Elsevier*.

[14] Reddy, A. M. and Raj, B. 2004. Soft mask estimation for single channel speaker separation, in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-04*. (October. 2004).

[15] Cai, D., He, X., Han, J. and Huang, T. 2010. Graph regularized nonnegative matrix factorization for data representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 8, 1548-1560.

[16] Fevotte, C., Bertin, N. and Durrieu, J. L. 2009. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music, *Neural Comput.* 21, 3, 793-830.

[17] Sun, D. L. and Mysore, G. J. 2013. Universal speech models for speaker independent single channel source separation, in *Proc. IEEE Int.Conf. Acoustics, Speech, and Signal Process.* Vancouver.

[18] Bao, G., Xu, Y. and Ye, Z. 2014. Learning a discriminative dictionary for single-channel speech separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 22, 7 (April. 2014), 1130 - 1138.

[19] Lin, Y. B., Pham, T., Lee, Y. S. and Wang, J. C. 2015. Monaural source separation using nonnegative matrix factorization with graph regularization constraint, *Conference on Computational Linguistics and Speech Processing,* (Oct 2015).

[20] Gyoon, K. T., Kwon, K., Shin, J. W. and Soo, K. N. 2015. NMF-based target source separation using deep neural network, *IEEE Signals Processing Letters*, 22, 2, (Feb. 2015), 229-233.

[21] Cooke, M., Barker, J., Cunningham, S. and Shao, X. 2006. An audio-visual corpus for speech perception and automatic speech recognition, *J. of the Acoustical Society of America*. 120, 2421-2424.

[22] Schmidt, M. and Olsson, R. 2006. Single-channel speech separation using sparse non-negative matrix factorization, in *Proc. Interspeech*. 2614–2617.

[23] Radfar, M. H. and Dansereau, R. M. 2007. Single-channel speech separation using soft mask filtering, *IEEE Trans. Audio Speech Lang. Process.*15, 8 (Nov. 2007), 2299–2310.

[24] Mowlaee, P., Saeidi, R., Christensen, M. G., Tan, Z. H., Kinnunen, T., Franti, P. and Jensen, S. H. 2012. A joint approach for single-channel speaker identi fication and speech separation, *IEEE Trans . Audio Speech Lang. Process* . 20, 9 (Nov. 2012), 2586–2601.

[25] Xu, W., Xin, L. and Yihong, G. 2003. Document clustering based on non-negative matrix factorization.in *Proc. of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003.DOI= http://doi.acm.org/10.1145/860435.860485.

[26] Pauca, V. P., Farial, S., Berry, M. W. and Plemmons, R. J. 2004. Text mining using non-negative matrix factorizations.. *Society for Industrial and Applied Mathematics*. 4.

[27] Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G. and Cai, D. 2011. Graph regularized sparse coding for image representation, *IEEE Trans. Image Process.* 20, 5, 1327-133.