A Sparsity-Relaxed Algorithm for the Under-determined Convolutive Blind Source Separation

Junjie Yang*[a], Yi Guo[b], Zuyuan Yang [a], Chao Yang [a]
[a]School of Automation, Guangdong University of Technology, Guangzhou, 510006, China
(yangjunjie1985@gmail.com, yangzuyuan@aliyun.com, chyang513@gdut.edu.cn);
[b]School of Computing, Engineering and Mathematics, Western Sydney University, Parramatta, NSW, 2150, Australia
(Y.Guo@westernsydney.edu.au).

ABSTRACT

Convolutive blind source separation (CBSS) is a kind of signal processing method by separating multiple sources from a convolutive mixing model. The concept of CBSS is to recover the latent sources in a reverberant environment. Usually, a two-stage scheme including the mixing matrix estimation and the source recovery are proposed to fulfill this target. In this paper, we mainly discuss the source recovery problem based on the knowledge of estimated mixing matrix. Specifically, this problem can be categorized as a sparse source construction optimization model, especially for the under-determined case where the number of sources is greater than the number of microphones. Inspirited by the fact that only few source components are active at each time-frequency slot, a new augmented Lagrange method is proposed to find the optimal sparse solution of sources with the ℓp norm ($0 < p < 1$) based measurement function. The proposed method relaxes the strict sparse assumption on sources, hence improve the source separation performance. The experiment results demonstrate that the proposed algorithm is superior than the state-of-the-art methods.

Keywords: Blind source separation, sparsity.

1. INTRODUCTION

Blind source separation refers a kind of signal processing technique of recovering multiple sources from mixture signals recorded by multiple microphones. The application of BSS can be found in various of area, e.g., the speech processing, biomedical signal processing, image processing and so on [1][2][3][4]. Herein, we mainly discuss the problem of convolutive blind source separation (CBSS), where multiple sources are convolved from a multiple delay mixing system model. One class of CBSS method is de-mixing latent sources from the mixture samples based on the time-frequency (TF) domain. Usually, a two-stage scheme is adopted to solve the problem of CBSS in the TF domain, which includes the mixing matrix estimation and the recovery of sources [2]. At the first stage of CBSS, the mixing matrix estimation has been widely studied in the existing works, such as [5] [6]. Considering the non-stationary of sources, the mixing matrix estimation problem is transformed into a sequence of joint-approximate diagonalization optimization model in spatial covariance domain. At the second stage of CBSS, the problem of source recovery is still not fully investigated. First, the solution of source components construction are not unique to the under-determined case, i.e., the number of sources is greater than the number of microphones [7]. Second, the source components ambiguity of scaling and permutation should be carefully corrected before transforming them back to the time-domain [3] [8].

To solve the under-determined CBSS problem, Yilmaz provides a condition called approximately window-disjoint orthogonal (WDO) assumption to solve the source separation problem, which refers that source components are disjoint in each TF point [9]. [10] utilizes a supervised learning method to estimate the complex mask, which is applied to separate the reverberant mixture speech signals. The authors in [7] tries to combine the K-means clustering method and TF masking scheme to separate the source components in the TF domain. This method classifies each TF mixture point into a specific clustering group, which may result in the so-called inter-interference problem. Note that the mentioned works on CBSS are based on a latent assumption, i.e., the sources are strictly sparse in the time domain or TF domain. However, such assumption cannot hold in most circumstance, e.g., the spectral of audio sources are usually overlapped, especially when the number of sources are over two.

In this paper, we try to relax the strict sparse assumption on sources and provide a sparsity-relaxed optimization approach to recover the sources for the under-determined case. The proposed source components are recovered by the proposed sparsity-relaxed optimization method; third, the permutation ambiguity of sources and scaling ambiguity are revised by the existing methods, respectively. The proposed method relaxes the WDO assumption on sources, i.e., source components are no longer require to be disjoint at each time-frequency slot. Moreover, a ℓp norm ($0 < p < 1$) based measurement function is developed based on the sparsity-relaxed assumption, which is demonstrated superior to depict various of sparse sources in the experiments. The rest of paper is organized as follows. First, the system model and assumptions are presented in Section II. Next, the sparsity-relaxed source reconstruction algorithm is provided in Section III. Later, the experimental results are discussed in Section IV. Finally, the conclusion is completed in Section V.

2. SYSTEM MODEL AND PROPOSED METHOD

2.1 System Model and Assumptions

Let M , N denote the number of microphones and sources, respectively. Let $\mathbf{x}(t) \in \mathbb{R}^N$ and $\mathbf{s}(t) \in \mathbb{R}^M$ denote as mixture signal and source signal, respectively. With above notations, we can describe the convolutive blind source separation (CBSS) system model as summation of a sequence of MIMO system with orders of L , such as

$$\mathbf{x}(t) = \mathbf{H} \star \mathbf{s}(t) = \sum_{\tau=0}^{L-1} \mathbf{H}(\tau)\mathbf{s}(t-\tau), \tag{1}$$

where $\star$ refers the convolutive operator, and $\mathbf{H}(\tau) \in \mathbb{R}^{M\times N}$ refers to the τ th mixing matrix in time domain. In the time-frequency (TF) domain, (1) can be transformed as a linear system by Short-Time Fourier Transformation (STFT) with length of F, i.e.,

$$\mathbf{x}^{f,d} \approx \mathbf{H}^f \mathbf{s}^{f,d}, f = 0,1,\ldots,F-1, \tag{2}$$

where $\mathbf{x}^{f,d} = [x_1^{f,d},..,x_M^{f,d}]^T \in \mathbb{C}^M$, $\mathbf{s}^{f,d} = [s_1^{f,d},..,s_N^{f,d}]^T \in \mathbb{C}^N$ are the mixture vector and the source vector at TF slot (f,d), respectively; $\mathbf{H}^f$ refers to the f th mixing matrix. To simplify the following discussion, we provide some necessary assumptions on the system model of (2):

A1) The number of sources is greater than the number of microphones, i.e., $M < N$.

A2) The mixing matrices $\{\mathbf{H}^f\}_{f=0}^{F-1}$ are known in advance.

A3) Any $M \times M$ sub-matrix of the mixing matrix are linearly independent.

Assumption A1 is also called as underdetermined case, which is a tough problem in convolutive BSS. If $M \geq N$, the source separation can be easily obtained by applying $\hat{\mathbf{s}}^{f,d} = (\mathbf{H}^f)^{\dagger}\mathbf{x}^{f,d}$. However, in the under-determined case, such method is not applicable as the pseudo-inverse of $\mathbf{H}^f$ is not unique again. Assumption A2 refers that the mixing matrices are given in advance. Since we mainly focus on the performance of source reconstruction, thus, we assume that the knowledge of mixing matrices has already known in advance. Assumption A3 ensures the source reconstruction can be applicable based on the given mixing matrix.

2.2 Proposed Source Reconstruction Method

Here, we present a substitute method to separate sources by exploiting the non-strict sparsity of sources. It is based on the observation that a dynamic number of source components are active at each TF point, i.e., $\mathbf{s}^{f,d}$ is a sparse term satisfying $\left\|\mathbf{s}^{f,d}\right\|_0 > 1$. Thus, the objective of source reconstruction can be transformed as a sparse construction of based on the knowledge of $\mathbf{H}^f$. Next, the source reconstruction problem can be summarized as a sparse optimization model, e.g.,

$$\min_{\mathbf{s}^{f,d}} \left\|\mathbf{s}^{f,d}\right\|_0 \\ \text{s.}t.\ \mathbf{x}^{f,d} \approx \mathbf{H}^f \mathbf{s}^{f,d}, \tag{3}$$

Note that the cardinality minimization model of (3) is a non-deterministic polynomial-time-hard (NP) problem, we can transform the model of (3) as a relaxed ℓp norm ($0 < p < 1$) optimization problem. In this case, the optimization model of (3) can be rewritten as

$$\min_{\mathbf{s}^{f,d}} \left\|\mathbf{s}^{f,d}\right\|_p \\ \text{s.}t.\ \mathbf{x}^{f,d} \approx \mathbf{H}^f \mathbf{s}^{f,d}, \tag{4}$$

The optimization problem of (4) can be transformed as an unstrained optimization problem as follows:

$$\min_{\lambda,\beta,\mathbf{s}^{f,d}} F(\lambda,\beta,\mathbf{s}^{f,d}) = J(\mathbf{s}^{f,d}) + \lambda^H(\mathbf{x}^{f,d} - \mathbf{H}^f\mathbf{s}^{f,d}) + \frac{\beta}{2}\| \mathbf{x}^{f,d} - \mathbf{H}^f\mathbf{s}^{f,d} \|^2, \tag{5}$$

where $J(\mathbf{s}^{f,d}) \triangleq \| \mathbf{s}^{f,d} \|_p$, λ is penalty vector and β is a penalty parameter. By applying augmented Lagrange method, as shown in Appendix A, the solution of (5) can be iteratively calculated from

$$(\mathbf{s}^{f,d})^{(k+1)} = [p\Pi^{(k)}_{\mathbf{s}_{f,d}} + \beta(\mathbf{H}^f)^H \mathbf{H}^f]^{-1}(\mathbf{H}^f)^H[\lambda^{(k)} + \beta\mathbf{x}^{f,d}], \tag{6}$$

where $\Pi_{\mathbf{s}_{f,d}} = diag(\left|s_1^{f,d}\right|^{p-2},\ldots,\left|s_N^{f,d}\right|^{p-2})$, k+1 refers the iteration step. The update rules can be set as follows:

$$\begin{cases} \beta^{(k+1)} = \beta^{(0)} + k\beta, \\ \lambda^{(k+1)} = \lambda^{(k)} - \beta^{(k+1)}[\mathbf{H}^f(\mathbf{s}^{f,d})^{(k)} - \mathbf{x}^{f,d}]. \end{cases} \tag{7}$$

In simulation experiment, we can empirically set a maximum iterative step, e.g., $k_{\max} = 10$, to stop the iteration procedure.

3. EXPERIMENT RESULTS

3.1 Experiment Settings

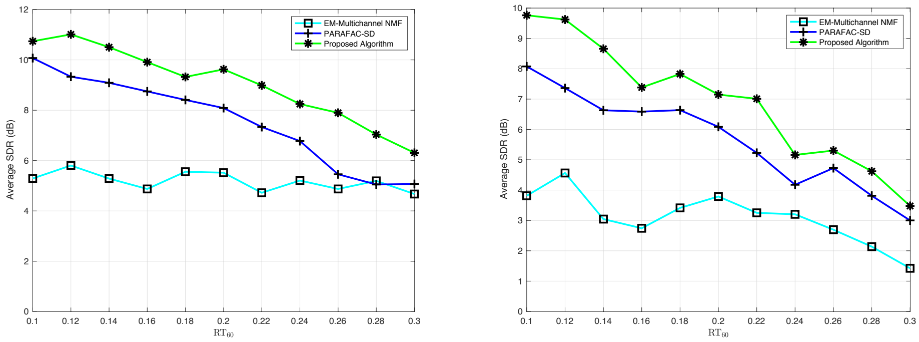
In this part, the provided sparsity-relaxed CBSS algorithm are tested with various experiments. All the experiments were carried out by a Lenovo computer equipped with Intel Core i5-4210U, CPU 1.7 GHz under the system of Windows 10, and the programs were coded by MATLAB R2015b. The proposed algorithm will be compared with two state-of-the-art algorithms [11,12] in the following experiments, which is labeled as 'EM-Multichannel NMF', 'PARAFAC-SD', respectively. We obtain a sequence of various speech sources from the data base [13]. The speech signals are recorded by several female or male speakers, who is reading poem, sentences and numbers with a sampling rate F_s = 16 kHz. We truncate all of recorded speech signals in a selected length of 10 seconds for the convenience of comparison. In the experiments, the speech sources are convolved by synthetic room impulse responses (RIRs) function using the system model of (1). The estimated sources are evaluated by the criterion of [14], which is to calculate signal distortion ratio (SDR) between the target source signal and a series of decomposed terms of source signal including distortion, noises or errors. Specifically, the source signal can be decomposed into a sum of several components, such as

$$\hat{s}_i(t) = s_i^{target}(t) + e_i^{interf}(t) + e_i^{noise}(t) + e_i^{artif}(t) \tag{8}$$

where $s_i^{target}(t)$, $e_i^{interf}(t)$, $e_i^{noise}(t)$, $e_i^{artif}(t)$ are the target source with allowed distortion, interferences, noises and artifacts error terms, respectively. The provided experiments were conducted in the virtual room with artificial RIRs in order to simulate a more real convolutive environment. Such RIRs are generated by the method in [15], which can simulate various situations with arbitrary settings, e.g., the physical sizes of the room, the locations of the sources and microphones, and the reverberation time RT_{60} which is defined as the time decay by 60 dB. RT_{60} varies from 0.1 to 0.3 with an interval of 0.02 in the following experiments. The sizes of the virtual room were selected as 12m × 9m × 3m. Three under-determined cases by setting various number of sources and microphones were introduces in the following experiments where M=3,4 and N=4,5. According to the various number of sources and microphones, these two cases were labeled as *CaseM*3*N*4 and *CaseM*4*N*5, respectively. The x, y, z coordinates of each source are set as (2m, 1m, 1.6m), (2m, 1.4m, 1.6m), (2m, 1.8m, 1.6m) and (2m, 2.2m, 1.6m), while those of each sensor are set as (3m,1m,1.6m), (3m, 1.5m,1.6m), (3, 2,1.6), respectively.

3.2 Experiment Results

The settings of proposed method is offered as follows: the STFT length is $F = 2048$, the overlapping factor $\alpha = 0.25$, the ℓp norm parameter is $p = 0.8$, penalty parameter $\beta^{(0)} = 10^3$ and Lagrange vector $\lambda^{(0)} = [1,...,1]^T$. As shown in Fig.1 (a), three algorithms are tested under *CaseM*3*N*4 and *CaseM*4*N*5, respectively. It can observe that the SIR performance degrades slowly with the increase of reverberation RT_{60}. As shown in Fig.1 (a), the proposed method is superior than EM-Multichannel NMF and PARAFAC-SD by approximately 5 dB and 1dB for each case, respectively. Similar results of $\mathrm{Case}M3N4$ and *CaseM*4*N*5 also can be observed from Fig.1 (b). It is concluded that our method can find a non-strict sparse solution of source reconstruction from the mixtures, which is a more applicable scheme to enhance the source separation performance.



(a) *CaseM*3*N*4

(b) *CaseM*4*N*5

Figure 1. SDR performance with the provided algorithms under various cases.

4. CONCLUSION

The under-determined source recovery problem for CBSS has been discussed in this paper. This problem was converted as a sparsity-relaxed optimization model based on a $\ell p\ (0 < p < 1)$ norm measurement function. A new augmented Lagrange method has been proposed to find the optimal sparse solution of sources from the established sparsity-relaxed optimization model. The proposed method tried to relax the strict sparse assumption on sources, which is suitable to real source separation problem. The experiment results have demonstrated the validity of proposed method in various under-determined CBSS cases.

5. ACKNOWLEGEMENT

This work was supported in part by China Postdoctoral Science Foundation under grants 2018M643022. This work was also supported in part by the National Natural Science Foundation of China under grants 61722304, 61333013, 61603099. This work was also supported in part by Guangdong Natural Science Funds under grants 2018A0303130080, and Pearl River S&T Nova Program of Guangzhou under grants 201806010176.

APPENDIX

The gradient of objective function in (5) can be derived as:

$$\frac{\partial F(\lambda,\beta,\mathbf{s}^{f,d})}{\partial \mathbf{s}^{f,d}} = \frac{\partial J(\mathbf{s}^{f,d})}{\partial \mathbf{s}^{f,d}} - (\mathbf{H}^{f})^{H}\lambda + \beta(\mathbf{H}^{f})^{H}(\mathbf{H}^{f}\mathbf{s}^{f,d} - \mathbf{x}^{f,d}), \tag{9}$$

where $\frac{\partial J(\mathbf{s}^{f,d})}{\partial \mathbf{s}^{f,d}} = p\,\Pi_{\mathbf{s}^{f,d}}\mathbf{s}^{f,d}, \, and \; \Pi_{\mathbf{s}^{f,d}} = diag(\left|s_1^{f,d}\right|^{p-2},...,\left|s_N^{f,d}\right|^{p-2})$. Let $\frac{\partial F(\lambda,\beta,\mathbf{s}^{f,d})}{\partial \mathbf{s}^{f,d}} = \mathbf{0}$, we have

$$p\,\Pi_{\mathbf{s}^{f,d}}\mathbf{s}^{f,d} - (\mathbf{H}^f)^H\lambda + \beta(\mathbf{H}^f)^H(\mathbf{H}^f\mathbf{s}^{f,d} - \mathbf{x}^{f,d}) = \mathbf{0}. \tag{10}$$

The solution of (10) can be transformed as follows:

$$[p\,\Pi_{\mathbf{s}^{f,d}} + \beta(\mathbf{H}^f)^H\mathbf{H}^f]\mathbf{s}^{f,d} = (\mathbf{H}^f)^H(\lambda + \beta\mathbf{x}^{f,d}). \tag{11}$$

Furthermore, we have

$$\mathbf{s}^{f,d} = [p\,\Pi_{\mathbf{s}^{f,d}} + \beta(\mathbf{H}^f)^H\mathbf{H}^f]^{-1}(\mathbf{H}^f)^H[\lambda + \beta\mathbf{x}^{f,d}]. \tag{12}$$

To obtain the solution of , an iterative scheme can be applied as follows:

$$(\mathbf{s}^{f,d})^{(k+1)} = [p\Pi^{(k)}_{\mathbf{s}^{f,d}} + \beta(\mathbf{H}^f)^H\mathbf{H}^f]^{-1}(\mathbf{H}^f)^H[\lambda^{(k)} + \beta\mathbf{x}^{f,d}], \tag{13}$$

where *k* refers the iterative step. In addition, the vector λ and β can be updated by the following strategy,

$$\begin{cases} \beta^{(k+1)} = \beta^{(0)} + k\beta, \\ \lambda^{(k+1)} = \lambda^{(k)} - \beta^{(k+1)}[\mathbf{H}^f(\mathbf{s}^{f,d})^{(k)} - \mathbf{z}^{f,d}], \end{cases} \tag{14}$$

and $\lambda^{(0)} = p[\mathbf{H}^f(\Pi^{(k)}_{\mathbf{s}^{f,d}})^{-1}(\mathbf{H}^f)^H]^{-1}\mathbf{x}^{f,d}$.

REFERENCE

[1] A. Holobar and D. Zazula, “Surface EMG decomposition using a novel approach for blind source separation,” Informatica Medica Slovenica, 2–14,2003.

[2] K. Rahbar, J. Reilly, and J. H. Manton, “Blind identification of MIMO-FIR systems driven by quasi-stationary sources using second order statistics: A frequency domain approach,” IEEE Trans. Signal Process., 52(2), 406–417,2004.

[3] Z. He, S. Xie, S. Ding, and A. Cichocki, “Convolutive blind source separation in the frequency domain based on sparse representation,” IEEE Trans. Audio, Speech, Lang. Process., 15(5), 1551-1563, 2007.

[4] L. Mesin, A.Holobar, and R. Merletti, “Blind source separation: application to biomedical signals,” Advanced Methods of Biomedical Signal Process.,2011.

[5] K.RahbarandJ.P.Reilly,“A frequency domain method for blind source separation of convolutive audio mixtures,” IEEE Trans. Speech Audio Process., 13(5), 832–844,2005.

[6] D. Nion and N. D. Sidiropoulos, “Adaptive algorithms to track the PARAFAC decomposition of a third-order tensor,” IEEE Trans. Signal Process., 57(6), 2299–2310,2009.

[7] V. G. Reju, S. N. Koh, and I. Y. Soon, “Underdetermined convolutive blind source separation via time-frequency masking,” IEEE Trans. Audio, Speech, Lang. Process., 18(1), 101–116,2010.

[8] H. Sawada, R. Mukai, S. Araki, and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” IEEE Trans. Speech Audio Process., 12(5), 530–538, 2004.

[9] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” IEEE Trans. Signal Process., 52(7), 1830-1847,2004.

[10] D. S. Williamson and D. Wang, “Time-frequency masking in the complex domain for speech dereverberation and denoising,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(7),1492–1501,2017.

[11] A. Ozerov and C. Fe´votte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” IEEE Trans. Audio, Speech, Lang. Process., 18(3), 550-563,2010.

[12] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, “Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures,” IEEE Trans. Audio, Speech, Lang. Process., 18(6), 1193–1207, 2010.

[13] D. Nion, “Blind source separation (BSS) of convolutive speech mixtures,” http://dimitri.nion.free.fr/bss/BSS.html, 2010, [Online].

[14] E. Vincent, H. Sawada, P. Bofill, S. Makino and J.P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results", in Proc. Int. Conf. on Independent Component Analysis and Signal Separation, 2007.

[15] J. Allen and D. Berkley, “Image method for efficiently simulating small- room acoustics,” J. Acoust. Soc. Amer., 65(4), 1979.