Speech Communication 89 (2017) 1–16

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

# Underdetermined blind separation of overlapped speech mixtures in time-frequency domain with estimated number of sources

Haijian Zhang[a,d,*], Guang Hua[b], Lei Yu[a], Yunlong Cai[c], Guoan Bi[d]

[a] School of Electronic Information, Wuhan University, China
[b] School of Electronic Information and Communications, Huazhong University of Science and Technology, China
[c] Department of Information Science and Electronic Engineering, Zhejiang University, China
[d] School of EEE, Nanyang Technological University, Singapore

ARTICLE INFO

*Article history:*
Received 29 February 2016
Revised 16 February 2017
Accepted 22 February 2017
Available online 24 February 2017

*Keywords:*
Underdetermined blind source separation
Noise suppression
Estimation of number of sources
Estimation of mixing matrix
Short-time Fourier transform

ABSTRACT

Noise suppression and the estimation of the number of sources are two practical issues in applications of underdetermined blind source separation (UBSS). This paper proposes a noise-robust instantaneous UBSS algorithm for highly overlapped speech sources in the short-time Fourier transform (STFT) domain. The proposed algorithm firstly estimates the unknown complex-valued mixing matrix and the number of sources, which are then used to compute the STFT coefficients of corresponding sources at each auto-source time-frequency (TF) point. After that, the original sources are recovered by the inverse STFT. To mitigate the noise effect on the detection of auto-source TF points, we propose a method to effectively detect the auto-term location of the sources by using the principal component analysis (PCA) of the STFTs of noisy mixtures. The PCA-based detection method can achieve similar UBSS outcome as some filtering-based methods. More importantly, an efficient method to estimate the mixing matrix is proposed based on subspace projection and clustering approaches. The number of sources is obtained by counting the number of the resultant clusters. Evaluations have been carried out by using the speech corpus NOIZEUS and the experimental results have shown improved robustness and efficiency of the proposed algorithm.

## 1. Introduction

Blind source separation (BSS) is to recover the underlying source signals based on observed mixtures from a sensor array or a single sensor, without knowing the information of the sources and the mixing process.[1] In many practical applications, the challenging case for source separation is when only one sensor is available, which is known as single-channel BSS (Gao et al., 2011; 2013; Tengtrairat et al., 2016). This paper focuses upon the instantaneous underdetermined BSS (UBSS) problem with a sensor array, i.e., the number of sensors is more than one but less than the number of sources. BSS with sensor array is more extensively researched than that with a single sensor. This is simply because more sensors could collect more information from the sources which helps the separation process. The BSS problems have been widely encountered in audio, radar, communication, image processing, and other areas (Naik and Wang, 2014; Xu et al., 2013; Chen et al., 2015; Pertila and Nikunen, 2015; Bofill and Zibulevsky, 2001; Araki et al., 2007; Abed-Meraim et al., 2001). Most of existing BSS algorithms reported in the literature have achieved desirable performance in a high SNR environment. Real-world signals might be contaminated by strong noise, and as a result, many reported algorithms obtain very poor BSS performance or fail to properly handle such severely distorted signals. Better methods to mitigate the noise effect are required to achieve robust solutions of BSS. Another practical problem of BSS algorithms lies in the unknown number of sources which has been often theoretically assumed available. Generally, the information on the number of sources is not available in practical applications (Naik and Wang, 2014; Zhang et al., 2013), and thus blind estimation of the number of sources from the received mixtures becomes crucial in achieving desirable BSS performance. This paper considers the above two practical issues and proposes a noise-robust BSS algorithm with the estimated number of sources by exploiting the spatial time-frequency distribution (STFD) of the sensor array output data.

The application of the STFD for BSS leads to an ever-growing research area. In Belouchrani et al. (1997), a BSS algorithm was introduced based on the joint diagonalization of multiple covariance matrices. Instead of using covariance matrices, Belouchrani et al. in

* Corresponding author.
*E-mail address:* haijian.zhang@whu.edu.cn (H. Zhang).

[1] This work was supported by the National Natural Science Foundation of China under Grant No. 61501335 and the Natural Science Foundation of Hubei Province under Grant No. 2015CFB202.

http://dx.doi.org/10.1016/j.specom.2017.02.003
0167-6393/© 2017 Elsevier B.V. All rights reserved.

Belouchrani and Amin (1998) proposed the spatial time-frequency distribution based BSS (STFD-BSS) algorithm by using the diagonalization of a combined set of spatial STFD matrices, which has been demonstrated to be more robust to noise because the noise power spreads over the entire time-frequency (TF) domain (Zhang et al., 2016; 2015b; 2015a). The main requirement of STFD-BSS is the selection of auto-term or cross-term TF points. Some efficient selection methods have been reported in Fevotte and Doncarli (2004); Belouchrani et al. (2004); Fadaili et al. (2007); Cirillo et al. (2008). In Mu et al. (2003); Linh-Trung et al. (2005), the authors indicated another tendency for efficient BSS by avoiding the problem of TF point selection, i.e., applying signal synthesis techniques in STFDs. However, this signal synthesis method requires the sources to be approximately disjoint in the TF domain. In Aïssa-El-Bey et al. (2007b), Aïssa-El-Bey et al. proposed two efficient STFD-based underdetermined BSS (UBSS) algorithms for TF-nondisjoint source separation by subspace projection and signal synthesis. Compared to the STFD-BSS in Belouchrani and Amin (1998), the STFD-UBSS in Aïssa-El-Bey et al. (2007b) does not require the TF point selection, and is more robust to noise because only the TF features of the localized source are used for signal synthesis. Furthermore, the STFD-UBSS can deal with the underdetermined case. More information on recent research about the STFD-based BSS and UBSS can be found from Peng and Xiang (2009); Xie et al. (2012); Belouchrani et al. (2013).

Since the short-time Fourier transform (STFT) is easy to implement and does not have the cross-terms in the TF domain, this paper is devoted to the development of the spatial STFT based UBSS (STFT-UBSS) algorithm which was originally reported in Aïssa-El-Bey et al. (2007b). The STFT-UBSS in Aïssa-El-Bey et al. (2007b) separates the mixed sources in the STFT domain by assigning the estimated STFT values located at each auto-source TF point to their corresponding sources. Then each source is recovered by the TF synthesis using the estimated STFT values that have been allocated to this source. To minimize the implementation complexity, the STFT-UBSS in Aïssa-El-Bey et al. (2007b) only deals with the auto-source TF points[2] in STFT domain that are the TF points having localized concentration of energy compared with a threshold value. However, due to the inappropriate choice of the threshold value, either certain TF points that are entirely from strong noise are detected as spurious auto-source TF points, or some true auto-source TF points are not detected. In Aziz-Sbaï et al. (2011, 2012), Aziz-Sbaï et al. proposed a method to choose an optimal threshold value and to deal with the noise contribution in the recovery process by estimating the noise standard deviation. Alternatively, we can filter each noisy mixture before the STFT-UBSS is applied. In Andrianakis and White (2009), Andrianakis et al. proposed a speech enhancement algorithm that models the time and frequency dependencies of the speech STFT values with a Markov random field (MRF) prior. This MRF-based method can be used for the enhancement of the STFT of each noisy mixture, and then the auto-source TF points are detected based on the filtered STFT images with a relatively small threshold value at different SNRs. Other noise reduction techniques for speech signals have been reported in Souden et al. (2013) and the references therein. It is expected that a de-noising operation may mitigate the influence of noise and improve the UBSS performance. One objective of our algorithm is to efficiently detect auto-source TF points and achieve performance improvement in a low SNR environment, which is frequently encountered in many practical scenarios.

Another important issue is how to obtain an accurate estimation of the mixing matrix, which is the premise of the STFT-UBSS algorithm. Advanced estimation methods of mixing matrix based on the STFTs of mixtures have been reported in Jourjine et al. (2000); Yilmaz and Rickard (2004); Abrard and Deville (2005); Li et al. (2006); Kim and Yoo (2009); Reju et al. (2009); Thiagarajan et al. (2013). However, they have some limitations. Specifically, the method in Jourjine et al. (2000); Yilmaz and Rickard (2004) is only suitable for two speech mixtures, and requires the sources to be W-disjoint orthogonal in the TF domain. The methods in Abrard and Deville (2005); Li et al. (2006); Kim and Yoo (2009); Reju et al. (2009); Thiagarajan et al. (2013) are designed in the case of real-valued mixing matrix. In Aïssa-El-Bey et al. (2007b), the complex-valued mixing matrix of sources with overlapped (weak-sparseness) spectral contents was estimated by clustering the single-source TF points (i.e., the TF points associated with a single source), which are detected by selecting the TF points having sufficient energy. However, when the number of sources or the observation time increases, more multi-source TF points possessing strong energy will appear, which significantly influences the estimation accuracy of the mixing matrix. In addition, the number of sources is generally assumed to be known and the estimation of the actual number of sources has not been adequately addressed. Two advanced clustering methods have been used for automatically estimating the number of sources in Luo et al. (2006) for the cases that the sources are assumed to be sparse in the TF domain. More sophisticated methods for accurately estimating the complex-valued mixing matrix and the number of highly-overlapped sources are needed.

In this paper, we propose a robust STFT-UBSS scheme by firstly detecting auto-source TF points in low SNR environments. Specifically, the principal component analysis (PCA) technique is applied to compress the STFT images from received mixtures into one noise-removed STFT image, based on which auto-source TF points are detected by assigning a relatively small threshold value. This PCA-based method not only achieves comparable UBSS performance with the filtering-based method in Andrianakis and White (2009), but also requires much less computation time. More importantly, we propose an estimation method of the complex-valued mixing matrix based on subspace analysis and clustering methods (Comaniciu and Meer, 2002). The mixing matrix can be accurately estimated and the number of sources can be obtained by counting the number of columns in the estimated mixing matrix (Zhang et al., 2013). Compared to the algorithm in Aïssa-El-Bey et al. (2007b), the developed STFT-UBSS algorithm is of practical importance, and is especially suitable for the sources which are significantly overlapped in STFT domain. Note that we only consider instantaneous UBSS problems in this paper, while reverberant situations are not in the scope of this paper.

This paper is organized as follows. We describe the proposed STFT-UBSS algorithm in Section 2, where the PCA-based detection of auto-source TF points, the estimation of the mixing matrix as well as the number of sources are elaborated. In Section 3, the STFT-UBSS algorithm is evaluated by simulation with various speech data. The advantages and limitations of the proposed algorithm are discussed in Section 4. Finally, Section 5 concludes this paper.

*Notation:* We use $\{\cdot\}^T$ as the transpose operator, $\{\cdot\}^H$ as the transpose conjugate operator, and $\{\cdot\}^\dagger$ as the Moore-Penrose pseudoinverse operator. $\widehat{\{\cdot\}}$ denotes the estimate of $\{\cdot\}$, $||\cdot||$ denotes the norm operator, $|\cdot|$ denotes the absolute operator, and $\max/\min\{\cdot\}$ means the maximum/minimum function.

## 2. The proposed STFT-UBSS algorithm

Let $s_n(t), n = 1, \ldots, N$, denote the unknown sources, where $N$ is the number of sources impinging on an $M$-dimensional uniform linear array (ULA) from $N$ distinct directions. The output vector

[2] The auto-term location of sources in STFT domain is termed as auto-source TF points in this paper.