# ESTIMATION OF UNDERDETERMINED MIXING MATRIX WITH UNKNOWN NUMBER OF OVERLAPPED SOURCES IN SHORT-TIME FOURIER TRANSFORM DOMAIN

*A. Haijian Zhang, B. Guoan Bi*

School of EEE
Nanyang Technological University
Singapore

*C. Sirajudeen Gulam Razul, D. Chong-Meng See*

Temasek Laboratories@NTU
and DSO National Laboratories
Singapore

## ABSTRACT

The estimation of the mixing matrix as well as the number of sources in blind source separation are two challenging problems. This paper proposes an effective estimation method to solve these two problems for underdetermined blind separation of overlapped sources in short-time Fourier transform (STFT) domain. Our study considers the blind estimation of the mixing matrix based on subspace projection as well as clustering methods, and the number of sources can be therefore estimated by counting the columns of the estimated mixing matrix. The proposed estimation method is noise-robust and suitable for the sources whose spectral contents are highly overlapped in STFT domain. Numerical results on speech sources are presented to illustrate the effectiveness and robustness of the proposed method.

***Index Terms***— Estimation of mixing matrix, estimation of number of sources, underdetermined blind source separation, short-time Fourier transform

## 1. INTRODUCTION

Recently, the combination of blind source separation (BSS) and time-frequency (TF) distributions has received substantial attention. In [1], the authors proposed a spatial time-frequency distribution BSS (STFD-BSS) algorithm based on the diagonalization of a combined set of spatial TF matrices. The main requirement of STFD-BSS is the selection of auto-term or cross-term TF points. In [2], two STFD based underdetermined BSS (STFD-UBSS) algorithms were proposed for TF overlapped source separation by signal synthesis. Compared to the STFD-BSS [1], the STFD-UBSS does not require TF point selection, and it is more robust to noise since only the localized source TF features are used for signal synthesis.

Our research emphasis is put on the short-time Fourier transform (STFT) based UBSS in [2] since the STFT is easy to implement and no cross-term issue is involved. For the STFT-UBSS algorithm, the essential problem lies in the estimation of the mixing matrix, which is crucial for final BSS performance. The estimation of the mixing matrix based on the STFT can be found in [3, 4, 5]. However, the methods therein have limitations, i.e. the method in [3] is only suitable for two speech mixtures, and the methods in [4, 5] are designed only for real mixing matrix. In [2], the complex mixing matrix estimation of sources with overlapped spectral contents was estimated by clustering the single-source T-F points. The single-source points are detected by selecting the TF points in STFT domain with sufficient strong energy. However, when the number of sources increases, more multi-source TF points satisfying the strong energy requirement will appear, which will significantly influence the estimation accuracy of the mixing matrix.

Besides, the mixing matrix in above references is estimated assuming that the number of sources is a known parameter. However, in many practical situations, the information of number of sources is undetermined, therefore an estimation of the number of sources is indispensable. Relatively little work has focused on the estimation of the number of sources [6]. Two advanced clustering methods were used for automatically estimate the number of sources in [7], nevertheless, the sources are assumed to be nonoverlapped in TF domain.

In this paper, we aim to further develop the STFT-UBSS algorithm proposed in [2] on the estimation of the mixing matrix assuming that the sources are overlapped in TF domain and the number of sources is unknown. We design a method based on subspace analysis and clustering methods to accurately estimate complex mixing matrix without knowing the number of sources, which can be simultaneous estimated by counting the column number of the estimated mixing matrix.

This paper is organized as follows. We briefly introduce the system model in Section 2. In Section 3, the proposed method of estimating the mixing matrix as well as the number of sources is elaborated. In Section 4, the proposed estimation method is evaluated on some speech mixtures and the comparison with the one in the STFT-UBSS algorithm is presented. Finally, Section 5 gives the concluding remarks.

## 2. SYSTEM MODEL

Let $s_n(t), n = 1, \ldots, N$, denote the sources ($N$ is the number of sources), and $x_m(t), m = 1, \ldots, M$, be the received

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

mixtures ($M$ is the sensor number):

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{1}$$

where the $M \times N$ matrix $\mathbf{A}$ denotes the mixing matrix. We assume $M < N$ since we are in underdetermined cases. $\mathbf{n}$(t) is additive noise.

The sources are allowed to be non-disjoint in STFT domain, and the same assumption as in [2] is made: the number of sources at any TF point is less than $M$.

## 3. ESTIMATION OF "A" AND "N"

### 3.1. Estimation method in the STFT-UBSS algorithm

In [2], the single-source TF points are firstly selected by detecting the TF points with strong energy, and then the mixing matrix is estimated by using the *k-means* clustering method assuming that the number of sources is known, and assuming for each source, we can always find some TF points, where only this source occurs. We mark the estimation method of $\mathbf{A}$ in [2] *Method 1*, which is described as follows:

*i) The TF point set $\Omega$ which corresponds the single-source points is firstly found by using the criterion below for each time-slice:*

$$\frac{||\boldsymbol{S}_x(t,f)||}{max_{\xi}\,||\boldsymbol{S}_x(t,\xi)||} > \epsilon_0 \tag{2}$$

*where $||\cdot||$ denotes the norm operator, $\boldsymbol{S}_x(t,f)$ is the STFT value vector of the mixtures $\boldsymbol{x}(t)$, and $\epsilon_0$ is an empirical threshold value which selects the TF points with very strong energy. All the TF points which satisfy this criterion will be included in the set $\Omega$.*

*ii) Secondly, we compute the spatial direction vector for each TF point in set $\Omega$:*

$$\boldsymbol{v}(t,f) = \frac{\boldsymbol{S}_x(t,f)}{||\boldsymbol{S}_x(t,f)||}, \quad (t,f) \in \Omega. \tag{3}$$

*And then, the k-means clustering method is conducted on the spatial vectors of all TF points in $\Omega$ to estimate the mixing matrix assuming $N$ is known.*

The *Method 1* has two imperfections. Firstly, the selection of single-source TF points according to (2) has a significant influence on the mixing matrix estimation. In many cases, however, many TF points with strong energy in $\Omega$ may contain multiple sources, which will significantly impact the estimation accuracy of the mixing matrix. Secondly, the desirable TF points of estimating the mixing matrix are not actually the TF points where only one source exists. In the following proposed method, we will prove that the desirable TF points are actually the points where the energy of one source is dominant over those of other sources and noise power.

### 3.2. The proposed estimation method

In our study, we firstly demonstrate that the dominant TF points are the appropriate points for mixing matrix estimation by analyzing the mean square error (MSE) of the spatial vectors of the estimated mixing matrix.

The *Method 1* regards all the points in $\Omega$ as single-source TF points. Instead, we assume the TF points in $\Omega$ are either single-source points or double-source points [1]. Defining two source STFT values $S_{s_i}$ and $S_{s_j}$ for each double-source TF point in $\Omega$, and we define a dominance parameter $\lambda$:

$$\lambda = \frac{||S_{s_i}||}{||S_{s_j}||}, \quad i,j \in \{1,2,\ldots,N\}. \tag{4}$$

When $\lambda$ is a big value at this double-source TF point, we say this point is dominant by source $i$. Let $[\mathbf{a}_i, \mathbf{a}_j]$ denotes the steering vectors of source $i$ and $j$ at each double-source point in $\Omega$, which gives:

$$\mathbf{S}_x = \begin{bmatrix} S_{x_1} \\ \vdots \\ S_{x_M} \end{bmatrix} = \begin{bmatrix} a_{i1} & a_{j1} \\ \vdots & \vdots \\ a_{iM} & a_{jM} \end{bmatrix} \begin{bmatrix} S_{s_i} \\ S_{s_j} \end{bmatrix}. \tag{5}$$

The ideal steering vector of the source $i$ can be computed from (5):

$$\mathbf{a}_i = \begin{bmatrix} a_{i1} \\ \vdots \\ a_{iM} \end{bmatrix} = \begin{bmatrix} \frac{S_{x_1}}{S_{s_i}} - a_{j1}\frac{S_{s_j}}{S_{s_i}} \\ \vdots \\ \frac{S_{x_M}}{S_{s_i}} - a_{jM}\frac{S_{s_j}}{S_{s_i}} \end{bmatrix} \tag{6}$$

where $a_{im} = \frac{1}{\sqrt{M}} e^{-j\frac{2\pi}{\lambda}d(m-1)sin(\theta_i)}, m \in \{1,\ldots,M\}$, and $d$ is the interelement spacing, $\lambda$ is the wavelength, and $\theta_i$ denotes the direction of arrival (DOA) of source $i$.

The normalized STFT observation value in (3) at each point in $\Omega$ is approximated as the steering vector estimation:

$$\widehat{\mathbf{a}}_i = \begin{bmatrix} \widehat{a}_{i1} \\ \vdots \\ \widehat{a}_{iM} \end{bmatrix} = \begin{bmatrix} \frac{S_{x_1}}{||\mathbf{S}_x||} \\ \vdots \\ \frac{S_{x_M}}{||\mathbf{S}_x||} \end{bmatrix} \cdot \frac{||S_{x_1}||}{S_{x_1}}. \tag{7}$$

The estimation error of $\mathbf{a}_i$ can be evaluated by computing the MSE according to (6) and (7):

$$MSE_i = ||\mathbf{a}_i - \widehat{\mathbf{a}}_i||^2 \approx \frac{\sqrt{1+\lambda^2}+1-\lambda}{\sqrt{1+\lambda^2}} \tag{8}$$

which is an increasing function signifying that the estimation accuracy highly depends on the dominance parameter $\lambda$, i.e., the larger $\lambda$ is, the lower MSE we obtain. The estimation error will be further mitigated because the final estimation of

[1] It is possible some TF points in $\Omega$ may contain more than two sources, but this seldom happens and is not considered in this paper.

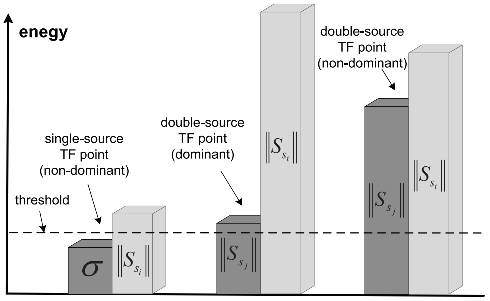
enegy
double-source TF point (non-dominant)
double-source TF point (dominant)
single-source TF point (non-dominant)
threshold
$\sigma$
$\|S_{s_i}\|$
$\|S_{s_j}\|$
$\|S_{s_i}\|$
$\|S_{s_j}\|$
$\|S_{s_i}\|$

**Fig. 1**. Three possible situations of TF points in set $\Omega$.

the steering vector of the source $i$ will be averaged over all the detected points dominated by the source $i$. The above analysis process is also effective for single-source points by replacing $S_{s_j}$ with $\sigma$, i.e., $\lambda = \frac{||S_{s_i}||}{\sigma}$, and $\sigma$ is noise deviation.

The aforementioned analysis proves that the TF points used for the mixing matrix estimation are those with dominant energy but the single-source TF points. Fig. 1 illustrates three possible situations of TF points in set $\Omega$: one single-source TF point and two double-source TF points, where we use an energy threshold to discriminate source and noise. Three observations can be made. Firstly, single-source TF points with a low value of $\lambda$ are undesirable for mixing matrix estimation. Secondly, the TF points with strong energy in $\Omega$ may contain double-source TF points. Lastly, some double-source TF points are desirable for estimation as long as these points have a high value of $\lambda$.

Next, we design a method to accurately estimate the mixing matrix and the number of sources. The essential idea of the proposed method is to extract all the dominant TF points from the set $\Omega$. Compared to *Method 1*, the proposed method defines the following relaxed assumptions based on $\Omega$:

- We assume there are two sources at each TF point of the set $\Omega$ instead of assuming only single-source TF points;

- For each source, we can always find some TF points, where the energy of this source is dominant over those of other sources and noise power;

The proposed method is called *Method 2* in the rest of the paper, which is described as follows:

*i) Applying k-means clustering method to classify all the spatial direction vectors in the set $\Omega$ given a fixed cluster number $N_0$, which is generally set to be larger than* 10. *Then, a column vector is estimated by averaging all the direction vectors in each cluster, thus we obtain a $M \times N_0$ mixing matrix $\mathbf{A}_0$.*

*ii) The STFT values of the two sources at each point can be estimated by:*

$$\widehat{\boldsymbol{S}}_s(t,f) = \begin{bmatrix} S_{s_{\alpha_1}} \\ S_{s_{\alpha_2}} \end{bmatrix} = \mathbf{A}_2^{\sharp} \boldsymbol{S}_x(t,f) \qquad (9)$$

*where $\sharp$ denotes the Moore-Penrose's pseudoinversion operator. $\mathbf{A}_2 = [\mathbf{a}_{\alpha_1}, \mathbf{a}_{\alpha_2}]$ are the steering vectors of two sources present at each point in $\Omega$. For each point, we try to find out the optimal $\mathbf{a}_{\alpha_1}$ and $\mathbf{a}_{\alpha_2}$ from the estimated $\mathbf{A}_0$ by minimizing the subspace projection:*

$$\{\mathbf{a}_{\alpha_1}, \mathbf{a}_{\alpha_2}\} = arg \min_{\mathbf{a}_{\beta_1}, \mathbf{a}_{\beta_2}} \left\{ \boldsymbol{P}\boldsymbol{S}_x(t,f) \right\} \qquad (10)$$

*where $\boldsymbol{P} = \boldsymbol{I} - \widetilde{\mathbf{A}}_2(\widetilde{\mathbf{A}}_2^H \widetilde{\mathbf{A}}_2)^{-1} \widetilde{\mathbf{A}}_2^H$ is the orthogonal projection matrix onto noise subspace of $\widetilde{\mathbf{A}}_2$, and $\widetilde{\mathbf{A}}_2 = [\mathbf{a}_{\beta_1}, \mathbf{a}_{\beta_2}]$, $\beta_1, \beta_2 \in \{1, \ldots, N_0\}$. The dominant TF points are selected by defining a threshold $\lambda_0$:*

$$\frac{max[||S_{s_{\alpha_1}}||, ||S_{s_{\alpha_2}}||]}{min[||S_{s_{\alpha_1}}||, ||S_{s_{\alpha_2}}||]} > \lambda_0. \qquad (11)$$

*We define the set $\Omega_d$ which contains all the TF points in $\Omega$ which satisfy (11).*

*iii) Finally, the mean-shift clustering method [8] without knowing $N$ is implemented on the set $\Omega_d$. The number of sources is determined by the number of clusters, and the mixing matrix can be obtained by averaging all the direction vectors in each cluster.*

Next, we discuss why $\mathbf{A}_2$ for each TF point in $\Omega$ can be effectively obtained from $\mathbf{A}_0$ via the minimization process in (10). The estimated $\mathbf{A}_0$ by $k$-means clustering method in the first step of *Method 2* can be expressed as follows:

$$\mathbf{A}_0 = \left[ \overbrace{\mathbf{a}_1 \mathbf{a}_2 \cdots \mathbf{a}_N} \quad \overbrace{\mathbf{a}_{12} \mathbf{a}_{13} \cdots \mathbf{a}_{(N-1)N}} \quad \overbrace{others} \right] \qquad (12)$$

which denotes that $\mathbf{A}_0$ is generally comprised of three parts: pure steering vectors from $N$ sources, mixed steering vectors by $N$ sources and other cases, e.g. distorted steering vectors due to noise. The optimal $\mathbf{a}_{\alpha_1}$ and $\mathbf{a}_{\alpha_2}$ will be detected from the first part of $\mathbf{A}_0$ since the minimization process will always choose the purer steering vectors. Specifically, for the TF points with two source $i$ and $j$, $i, j \in \{1, \ldots, N\}$, the resultant $\mathbf{a}_{\alpha_1}$ and $\mathbf{a}_{\alpha_2}$ by implementing (10) will be the purest vectors of source $i$ and source $j$ among the first part of $\mathbf{A}_0$. For the single-source TF points with source $i$ in $\Omega$, one of $\mathbf{a}_{\alpha_1}$ and $\mathbf{a}_{\alpha_2}$ comes from the first part of $\mathbf{A}_0$, whereas the other one could be any column vector of $\mathbf{A}_0$ due to random noise. However, this random spatial vector will not cause detrimental effect on the ratio computation in (11).

## 4. SIMULATION

In this section, numerical results are given to show the effectiveness and robustness of *Method 2* on the estimation of $\mathbf{A}$

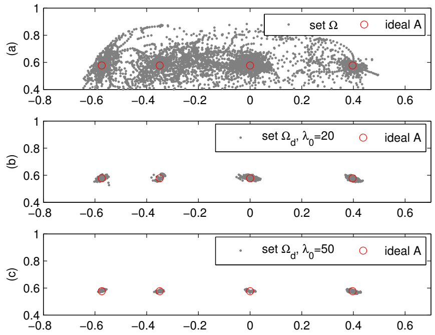

**Fig. 2**. Two-dimensional scattered view of $\Omega$ and $\Omega_d$ by plotting the first two elements of $\mathbf{S}_x(t, f)$ (SNR=20dB).

**Table 1**. Estimation of $N$ using *Method 2* on $\Omega_d$ ($\lambda_0 = 50$)

| SNR \ $\widehat{N}$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 20dB | 0 | 0 | **100%** | 0 | 0 | 0 | 0 |
| 15dB | 0 | 1% | **98%** | 1% | 0 | 0 | 0 |
| 10dB | 0 | 3% | **93%** | 4% | 0 | 0 | 0 |
| 5dB | 0 | 2% | **82%** | 15% | 1% | 0 | 0 |
| 0dB | 0 | 15% | **75%** | 9% | 1% | 0 | 0 |

and $N$. We use a uniform linear array, and 4 speech sources and 3 sensors are used. The $N$ sources are from different DOAs: $\theta_1 = 15°$, $\theta_2 = 30°$, $\theta_3 = 45°$, and $\theta_4 = 75°$. The speech sources with $2.5s$ duration are highly overlapped in STFT domain. The value of $N_0$ in *Method 2* is set to 12.

Fig. 2(a) shows the set $\Omega$ obtained from (2) by setting $\epsilon_0 = 0.3$. The corresponding $N$ ideal steering vectors marked by red circles in Fig. 2 reveal that the set $\Omega$ via the criterion (2) cannot provide an accurate estimation of $A$. The detected dominant set $\Omega_d$ using *Method 2* is shown in Fig. 2(b) and (c) by setting $\lambda_0 = 20$ and $\lambda_0 = 50$, respectively. It is seen that the detected dominant points are very close to the ideal centroids of the steering vectors. Based on the set $\Omega$, the number of sources will be overestimated by using clustering methods. We implement the *mean-shift* clustering method on the set $\Omega_d$, and the estimation performance of number of sources for different SNRs is displayed in Table 1. We can note that *Method 2* can accurately estimate the number of clusters since the clusters in $\Omega_d$ are clearly separated.

Lastly, the estimation performance of $\mathbf{A}$ using *Method 1* and *Method 2* is shown in Fig. 3. It is observed that the performance of *Method 1* is limited due to the inaccurate detection of single-source TF points. In contrast, the good performance of *Method 2* verifies that the dominant TF points are the desirable TF points for accurate estimation of $\mathbf{A}$.

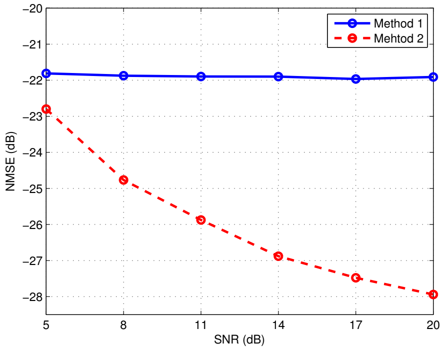

**Fig. 3**. Normalized MSEs of estimation of $\mathbf{A}$ vs. different SNRs using *Method 1* and *Method 2*.

## 5. CONCLUSION

In this paper, a robust estimation method for the mixing matrix as well as the number of sources is proposed for underdetermined blind separation of nondisjoint sources in STFT domain. The proposed estimation method for complex mixing matrix can be also applied for other applications with real mixing matrix, e.g. communication signal separation. Furthermore, the accurate estimation results of the mixing matrix and the number of sources can well improve the source separation performance in a totally blind environment.

## 6. RELATION TO PRIOR WORK

The study in this paper is an extension of the STFT-UBSS algorithm in [2]. The number of sources in [2] is assumed to be known. In addition, the estimation performance of mixing matrix in [2] is limited by the cases where the spectral contents of sources are highly overlapped in TF domain. The novelty of this paper lies in the proposed estimation method, which can more accurately estimate mixing matrix for highly TF overlapped sources by detecting dominant TF points with unknown number of sources. Another limitation of the method in [2] lies in the optimal selection of $\epsilon_0$ in (2) for different applications. The proposed method avoids the optimal selection of $\epsilon_0$ by detecting dominant TF points from an initial set $\Omega$, which can be obtained by setting a relatively small value of $\epsilon_0$. Although the dominant points are detected by evaluating a value of $\lambda_0$, however, final estimation performance is not sensitive to this parameter, and it can allow an evaluation of wide dynamic range as denoted in Fig. 2.

# Acknowledgment

The authors would like to thank Prof. Abdeldjalil Aïssa-El-Bey et al. for providing us the program code of their paper.

## 7. REFERENCES

[1] A. Belouchrani and M.G. Amin, “Blind source separation based on time-frequency signal representations,” *IEEE Transactions on Signal Processing*, vol. 46, no. 11, pp. 2888–2897, Nov. 1998.

[2] A. Aïssa-El-Bey, N. Linh-Trung, K. Abed-Meraim, A. Belouchrani, and Y. Grenier, “Underdetermined blind separation of nondisjoint sources in the time-frequency domain,” *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 897–907, Mar. 2007.

[3] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[4] Y. Li, S. Amari, A. Cichocki, D.W.C. Ho, and S. Xie, “Underdetermined blind source separation based on s-parse representation,” *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 423–437, Feb. 2006.

[5] S. Kim and C.D. Yoo, “Underdetermined blind source separation based on subspace representation,” *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2604–2614, Jul. 2009.

[6] M. Wax and T. Kailath, “Detection of signals by information theoretic criteria,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 387–392, Apr. 1985.

[7] Y. Luo, W. Wang, J.A. Chambers, S. Lambotharan, and I. Proudler, “Exploitation of source nonstationarity in underdetermined blind source separation with advanced clustering techniques,” *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2198–2212, Jun. 2006.

[8] D. Comaniciu and P. Meer, “Mean shift: a robust approach toward feature space analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002.