Underdetermined blind source separation of audio sources in time-frequency domain

Abdeldjalil Aissa El Bey, Karim Abed-Meraim, Yves Grenier

UNDERDETERMINED BLIND SEPARATION OF AUDIO SOURCES IN TIME-FREQUENCY DOMAIN

A. Aïssa-El-Bey, K. Abed-Meraim and Y. Grenier

ENST-Paris, 46 rue Barrault 75634, Paris Cedex 13, France

ABSTRACT

This paper considers the blind separation of audio sources in the underdetermined case, where we have more sources than sensors. A recent algorithm applies time-frequency distributions (TFDs) to this problem and gives good separation performance in the case where sources are disjoint in the time-frequency (TF) plane. However, in the non-disjoint case, the reconstruction of the signals requires some interpolation at the intersection points in the TF plane. In this paper, we propose a new algorithm that combines the abovementioned method with subspace projection in order to explicitly treat non-disjoint sources. Another contribution of this paper is the estimation of the mixing matrix in the underdetermined case.

1. INTRODUCTION

Blind source separation (BSS) considers the recovery of unobserved original sources from several mixtures observed at the output of a set of sensors. Each mixture contains a combination of the sources that results from the mixing medium between the sources and the sensors. The term "blind" indicates that no *a priori* knowledge of the sources and the medium structure is available. To compensate for this lack of information, the sources are usually assumed to be statistically independent. Blind source separation has application in different areas, such as communications, speech processing, image processing and biomedical engineering [1].

A challenging problem of BSS occurs when there are more sources than sensors, and this is referred to as *underdetermined* blind source separation (UBSS). A time-frequency based UBSS algorithm has been recently proposed in [2, 3] to successfully separate speech sources using time-frequency distributions (TFDs). This algorithm provides good separation performance when the sources are disjoint in the TF plane. It also provides the separation of TF quasi-disjoint sources, that is the sources are allowed to have a small degree of overlapping in the TF plane. However, the intersection points in the TF plane are not directly treated. More precisely, a point at the intersection of two sources is clustered "randomly" to belong to one of the sources. As a result, the source that picks up this point now contains some information from the other source while the later source loses some information of its own. However, for the other source, there is an interference at this point, hence the separation performance may degrade if no treatment is provided for this. An increasing in the number of intersection points degrades the separation quality. In this paper, we propose another algorithm, combining the TF-UBSS with subspace projection, that allows an explicit treatment of the intersection points. The main assumption used in this algorithm is that the number of sources simultaneously present at an intersection point in the TF plane cannot exceed the number of sensors.

2. PROBLEM FORMULATION

2.1. Data model

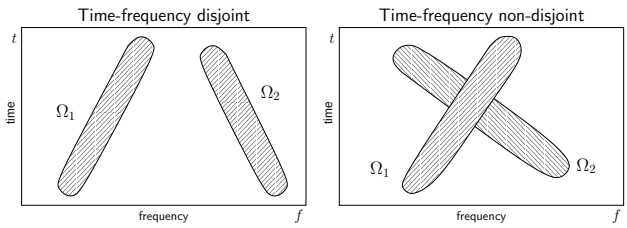
Let $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ represent the N nonstationary source signals. The source signals are transmitted through a medium so that an array of M linear sensors picks up a set of mixed signals represented by an M-dimensional vector $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$. We consider the instantaneous mixing medium that is modeled as follows

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

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ is the mixing matrix and $\mathbf{w}(t) = [w_1(t), \ldots, w_M(t)]^T$ is the observation noise vector. We assume that any M columns of $\mathbf{A}$ are linearly independent. The goal of BSS is to recover $\mathbf{s}(t)$ from $\mathbf{x}(t)$. When $M < N$, the problem becomes UBSS. Let Ω_1 and Ω_2 be the TF supports (i.e. the points of TF plane where the local energy of the considered sources is non-zero) of two sources $s_1(t)$ and $s_2(t)$. If $\Omega_1 \cap \Omega_2 \neq \emptyset$, the sources are said to be non-disjoint in the TF plane. The second assumption is that the sources are not necessarily disjoint, and in particular, there exist, at most, simultaneously $(M-1)$ sources at the same TF point. However, we still assume that there exists for each source signal a region $\mathcal{R}_i$ in the TF plane where it exists alone, i.e. the energy of the other sources are negligible at the TF points within the considered region.

2.2. Time-frequency representation

TF signal processing provides effective tools for analyzing nonstationary signals and linear time-varying systems, whose frequency content varies in time. This concept is a natural



(a) (b)

Fig. 1. *(a)–TF disjoint, (b) TF non-disjoint*

extension of both the time domain and the frequency domain processing, that involves representing signals in a two-dimensional space the joint TF domain, hence providing a distribution of signal energy versus time and frequency simultaneously. For this reason, a TF representation is commonly referred to as a time-frequency distribution (TFD). TFDs have been applied to a wide variety of engineering problems. Specifically, they have been successfully used for signal recovery at low signal-to-noise ratio (SNR), accurate estimation of the instantaneous frequency (IF), signal detection in communication, radar processing and for the design of time-varying filter. For more details on TFDs and related methods, see for example the recent comprehensive reference [5].

The method presented in this paper, uses the Short-Time Fourier Transform (STFT) that is defined as:

$$\mathcal{S}_x(t,f) = \sum_{m=-\infty}^{m=+\infty} h(t-m)x(m)e^{-j2\pi fm} \qquad (2)$$

where $h(t)$ is the Hamming window.

3. CLUSTER-BASED TF-UBSS APPROACH FOR DISJOINT SOURCES

In this section, we briefly review the STFT method in [2], and propose a *cluster-based linear TF-UBSS algorithm* using STFT to avoid some of the drawbacks in [2].

First, under the transformation into the TF domain using the STFT, the model in (1) leads to:

$$\mathcal{S}_{\mathbf{x}}(t,f) = \mathbf{A}\mathcal{S}_{\mathbf{s}}(t,f), \qquad (3)$$

where

$$\mathcal{S}_{x_i}(t,f) = \sum_{m=-\infty}^{m=+\infty} h(t-m)x_i(m)e^{-j2\pi fm} \qquad (4a)$$

$$\mathcal{S}_{\mathbf{x}}(t,f) = [\mathcal{S}_{x_1}(t,f), \ldots, \mathcal{S}_{x_M}(t,f)]^T. \qquad (4b)$$

and $\mathcal{S}_{\mathbf{s}}(t,f)$ is the $N \times 1$ source STFT vector. To avoid processing all TF points (and hence to reduce computational cost), we apply first a noise thresholding as that for each time-slice (t,f):

$$\text{If} \quad \frac{\|\mathcal{S}_{\mathbf{x}}(t,f_0)\|}{\max_f \{\|\mathcal{S}_{\mathbf{x}}(t,f)\|\}} > \epsilon, \quad \text{then keep } (t,f_0), \qquad (5)$$

where ϵ is a small threshold (typically, $\epsilon = 0.05$). Then, the set of all selected points, Ω, is expressed by $\Omega = \bigcup_{i=1}^{N} \Omega_i$, where Ω_i is the TF support of the source $s_i(t)$.

Under the assumption that all sources are disjoint in the TF domain, (3) is reduced to

$$\mathcal{S}_{\mathbf{x}}(t,f) = \mathbf{a}_i \mathcal{S}_{s_i}(t,f), \forall (t,f) \in \Omega_i, \forall i = 1, \cdots, N. \qquad (6)$$

where the source STFT vector has been reduced to only the STFT of the source $s_i(t)$.

Now, in [2], the structure of the mixing matrix is particular as such it has only 2 rows (i.e. the method uses only 2 sensors) and the first row of the mixing matrix contains all 1. Then, (3) is expanded to

$$\begin{bmatrix} \mathcal{S}_{x_1}(t,f) \\ \mathcal{S}_{x_2}(t,f) \end{bmatrix} = \begin{bmatrix} 1 & \ldots & 1 \\ a_{2,1} & \ldots & a_{2,N} \end{bmatrix} \begin{bmatrix} \mathcal{S}_{s_1}(t,f) \\ \vdots \\ \mathcal{S}_{s_N}(t,f) \end{bmatrix}, \qquad (7)$$

and (6) to

$$\begin{bmatrix} \mathcal{S}_{x_1}(t,f) \\ \mathcal{S}_{x_2}(t,f) \end{bmatrix} = \begin{bmatrix} 1 \\ a_{2,i} \end{bmatrix} \mathcal{S}_{s_i}(t,f),$$

which results in

$$a_{2,i} = \frac{\mathcal{S}_{x_2}(t,f)}{\mathcal{S}_{x_1}(t,f)}. \qquad (8)$$

Therefore, all the points for which the ratios on the right-hand side of (8) have the same value form the TF support Ω_i of a single source, say $s_i(t)$. Then, the STFT estimate of $s_i(t)$ is computed by:

$$\hat{\mathcal{S}}_{s_i}(t,f) = \begin{cases} \mathcal{S}_{x_1}(t,f), & \forall (t,f) \in \Omega_i, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, the source estimate $\hat{s}_i(t)$ is obtained by converting $\hat{\mathcal{S}}_{s_i}(t,f)$ to the time domain using inverse STFT [8]. For more details, refer to this paper. It is observed that the structure of the mixing matrix, as expressed in (7) has some limiting factors. First, the extension of the UBSS method in [2] to more than two sensors is not obvious. Second, the division on the right-hand side of (8) is prone to error if the denominator is close to zero.

To avoid the above mentioned problems, we propose here a modified version of the method valid for any number of sensors. This method is now referred to as the cluster-based linear TF-UBSS algorithm. The clustering method proceeds as follows: first compute the spatial direction vectors by:

$$\mathbf{v}(t,f) = \frac{\mathcal{S}_{\mathbf{x}}(t,f)}{\|\mathcal{S}_{\mathbf{x}}(t,f)\|}, \quad (t,f) \in \Omega, \qquad (9)$$

and force them, without loss of generality, to have the first entry real and positive.

Next, we cluster these vectors into N classes $\{C_i \,|\, i = 1, \cdots, N\}$, using the k-mean clustering algorithm [7]. The collection of all points, whose vectors belong to the class C_i, now forms

Table 1. Cluster-based TF-UBSS algorithm

1. Mixture STFT computation by (4).
2. Vector clustering by (9) and [7].
3. Mixing matrix and source STFT estimation by (10) and (11).
4. Source TF synthesis by [8].

the TF support Ω_i of the source $s_i(t)$. Then, the column vector $\mathbf{a}_i$ of $\mathbf{A}$ is estimated as the centroid of this set of vectors:

$$\hat{\mathbf{a}}_i = \frac{1}{|C_i|} \sum_{(t,f)\in\Omega_i} \mathbf{v}(t,f), \tag{10}$$

where $|C_i|$ is the number of vectors in this class.
Therefore, we can estimate the STFT of each source $s_i(t)$ (up to scalar constant) by:

$$\hat{\mathcal{S}}_{s_i}(t,f) = \hat{\mathbf{a}}_i^H \mathcal{S}_{\mathbf{x}}(t,f), \quad \forall\,(t,f) \in \Omega_i, \tag{11}$$

since, from (6), we have

$$\hat{\mathbf{a}}_i^H \mathcal{S}_{\mathbf{x}}(t,f) = \hat{\mathbf{a}}_i^H \mathbf{a}_i \mathcal{S}_{s_i}(t,f) \propto \mathcal{S}_{s_i}(t,f), \quad \forall\,(t,f) \in \Omega_i.$$

This algorithm is summarized in Table 1.

4. SUBSPACE-BASED TF-UBSS APPROACH FOR NON-DISJOINT SOURCES

We propose here to use an appropriate subspace projection to estimate the TFDs of the individual sources, under the previously stated data assumptions. Under the TF non-disjoint condition, consider a source point $(t,f) \in \Omega$ at which there are K contributing sources $s_{\alpha_1}(t), \ldots, s_{\alpha_K}(t)$, with $K < M$. Then, (3) is reduced to the following

$$\mathcal{S}_{\mathbf{x}}(t,f) = \tilde{\mathbf{A}} \mathcal{S}_{\tilde{\mathbf{s}}}(t,f), \quad \forall (t,f) \in \Omega \tag{12}$$

where $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{s}}$ are defined by:

$$\tilde{\mathbf{s}} = [s_{\alpha_1}(t), \ldots, s_{\alpha_K}(t)]^T, \tag{13a}$$

$$\tilde{\mathbf{A}} = [\mathbf{a}_{\alpha_1}, \ldots, \mathbf{a}_{\alpha_K}]. \tag{13b}$$

Let $\mathbf{Q}_{\tilde{\mathbf{A}}}$ be the orthogonal projection matrix onto the noise subspace of $\tilde{\mathbf{A}}$. Then, $\mathbf{Q}_{\tilde{\mathbf{A}}}$ can be computed by:

$$\mathbf{Q}_{\tilde{\mathbf{A}}} = \mathbf{I} - \tilde{\mathbf{A}} \left(\tilde{\mathbf{A}}^H \tilde{\mathbf{A}} \right)^{-1} \tilde{\mathbf{A}}^H. \tag{14}$$

We have the following observation:

$$\begin{cases} \mathbf{Q}_{\tilde{\mathbf{A}}} \mathbf{a}_i = 0, & i \in \{\alpha_1, \ldots, \alpha_K\} \\ \mathbf{Q}_{\tilde{\mathbf{A}}} \mathbf{a}_i \neq 0, & otherwise \end{cases}. \tag{15}$$

Table 2. Subspace-based TF-UBSS algorithm

1. STFT computation.
2. Single-source point selection; mixing matrix estimation by, k-mean algorithm.
3. For all source points, perform subspace-based TFD estimation of sources by (14), (16) and (17).
4. Source TF synthesis by [8].

If $\mathbf{A}$ is known or a priori estimated, then this observation gives us the criterion to detect the indices $\alpha_1, \ldots, \alpha_K$; and hence, the contributing sources at the considered TF point (t,f). In practice, to take into account noise, one detects the column vectors of $\tilde{\mathbf{A}}$ minimizing:

$$\{\alpha_1, \ldots, \alpha_K\} = \arg \min_{\beta_1,\ldots,\beta_K} \left\{ \|\mathbf{Q}_{\tilde{\mathbf{A}}} \mathcal{S}_{\mathbf{x}}(t,f)\| \mid \tilde{\mathbf{A}}_\beta \right\} \tag{16}$$

where $\tilde{\mathbf{A}}_\beta = [\mathbf{a}_{\beta_1}, \ldots, \mathbf{a}_{\beta_K}]$.
Next, TFD values of the K sources at TF point (t,f) are estimated by:

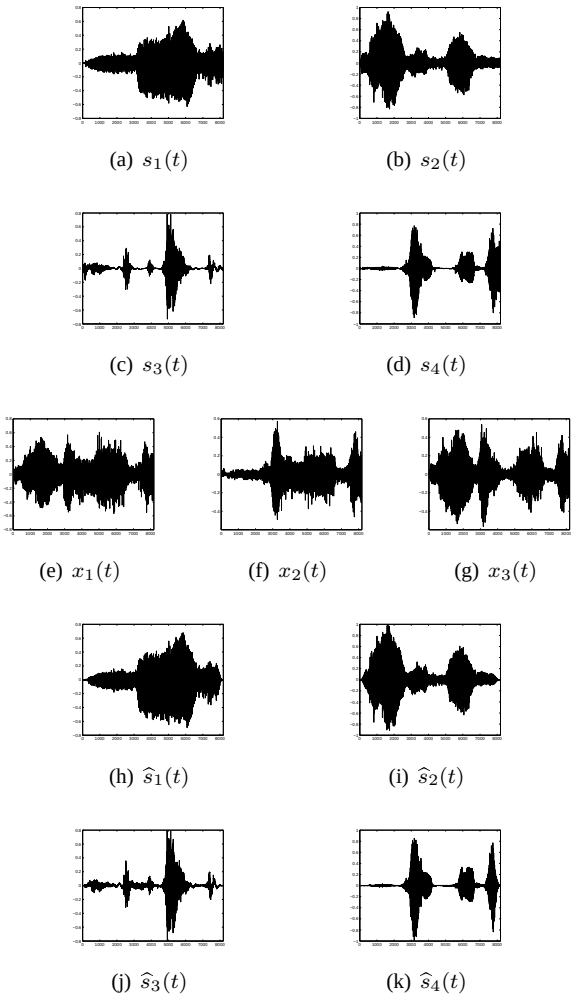
$$\hat{\mathcal{S}}_{\tilde{\mathbf{s}}}(t,f) \approx \tilde{\mathbf{A}}^{\#} \mathcal{S}_{\mathbf{x}}(t,f). \tag{17}$$

where $\tilde{\mathbf{A}}^{\#}$ represents the pseudo-inverse of $\tilde{\mathbf{A}}$.
Now, to apply the above procedure, we need to estimate $\mathbf{A}$ first. This is performed here by clustering all the spatial direction vectors in (9) as for the preview TF-UBSS algorithm. Then within each class C_i we estimate the far-located vectors from the centroid (in the simulation we estimate vectors $\mathbf{v}(t,f)$ such that: $\|\mathbf{v}(t,f) - \hat{\mathbf{a}}_i\| > 0.8 \max_{\mathbf{v}(t,f)\in\Omega_i} \|\mathbf{v}(t,f) - \hat{\mathbf{a}}_i\|$ leading to a reduced size class $\tilde{C}_i$.
This is to essentially keep the vectors corresponding to the TF region $\mathcal{R}_i$ (which are ideally equal to the spatial direction $\mathbf{a}_i$ of the considered source signal). Finally, the i^{th} column vector of $\mathbf{A}$ is estimated as the centroid of $\tilde{C}_i$.
Table 2 provides a summary of the subspace projection based TF-UBSS algorithm.

5. SIMULATIONS AND RESULTS

Simulation results are illustrated in the figures below. In this simulation, we have used uniform linear array of $M = 3$ sensors. It receives signals from $N = 4$ independent speech sources, lasting 8192 samples. In figure 2, the upper line represents the original source signals, the second line represents the M mixtures and the bottom one represents the sources estimates by our algorithm. In figure 3, we compare the performance of our method with the TF-UBSS method of Table 1 (i.e. modified method of [2]). The plots represents the average normalized MSE (NMSE) of the estimated sources versus the SNR in dB. For the subspace-based method we have used

$K = 2$ for all TF points. As can be observed, a significati gain is obtained, thanks to our subspace projection.



(a) $s_1(t)$ (b) $s_2(t)$

(c) $s_3(t)$ (d) $s_4(t)$

(e) $x_1(t)$ (f) $x_2(t)$ (g) $x_3(t)$

(h) $\widehat{s}_1(t)$ (i) $\widehat{s}_2(t)$

(j) $\widehat{s}_3(t)$ (k) $\widehat{s}_4(t)$

Fig. 2. *Blind source separation example for 4 speech sources and 3 sensors in instantaneous mixture case: the upper line represents the original source signals, the second line represents the M mixtures and the bottom one represents estimates of sources by our algorithm.*

6. CONCLUSION

This paper introduces a new approach for blind separation of non-disjoint and nonstationary sources using TFDs. The proposed method can separate more sources than sensors and provides, in the case of non-disjoint sources, a better separation quality than the method proposed in [2]. This method is based on a vector clustering procedure that estimates the mixing matrix $\mathbf{A}$, and subspace projection to separate the sources at the intersection points in the TF plane.

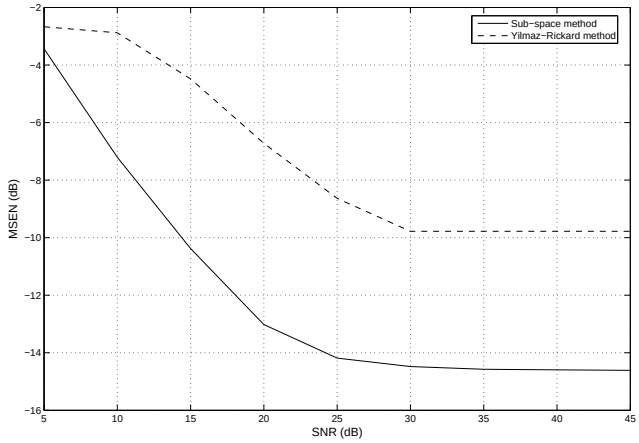


Fig. 3. *NMSE versus SNR for 4 speech sources and 3 sensors: comparison of the performance of our algorithm with the modified TF-UBSS*

7. REFERENCES

[1] A.K. Nandi (editor), "Blind estimation using higher-order statistics." *Kluwer Acadimic Publishers*, Boston 1999.

[2] O. Yilmaz, S. Rickard, "Blind separation of speech mixtures via time-frequancy masking", *IEEE Transaction on Signal Processing*, Vol. 52, no. 7, July 2004.

[3] N. Linh-Trung, A. Belouchrani, K. Abed-Meraim and B. Boashash, "Separating more sources than sensors using time–frequency distributions," *EURASIP J. of Applied Sig. Proc.*, Nov. 2004, in press.

[4] A. Belouchrani and M. G. Amin, "Blind source separation based on time-frequency signal representations," *IEEE Trans. on Sig. Proc.*, vol. 46, no. 11, pp. 2888–2897, Nov. 1998.

[5] B. Boashash, Ed., *Time Frequency Signal Analysis and Processing: Method and Applications*, Elsevier, Oxford, 2003.

[6] L. De Lathauwer, B. Moor, J. Vandewalle, "ICA techniques for more sources than sensors", *Higher-oder statistic Proc. of the IEEE Sig. Proc. Workshop*, pp. 116–120, 1999.

[7] I.E. Frank and R. Todeschini, "The data analysis handbook", *Elsevier,Sci. Pub. Co.*, 1994.

[8] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acous., Speech, Sig. Proc.*, vol. ASSP-32, no. 2, pp.236–243, Apr. 1984.