A New Sparse Blind Source Separation Method for Determined Linear Convolutive Mixtures in Time-Frequency Domain

Mostafa Bella(✉) and Hicham Saylani(✉)

Laboratoire d'Électronique, Traitement du Signal et Modélisation Physique, Faculté des Sciences, Université Ibn Zohr, BP 8106, Cité Dakhla, Agadir, Morocco
mostafa.bella@edu.uiz.ac.ma, h.saylani@uiz.ac.ma

Abstract. This paper presents a new Blind Source Separation method for linear convolutive mixtures, which exploits the sparsity of source signals in the time-frequency domain. This method especially brings a solution to the artifacts problem that affects the quality of signals separated by existing time-frequency methods. These artifacts are in fact introduced by a time-frequency masking operation, used by all these methods. Indeed, by focusing on the case of determined mixtures, we show that this problem can be solved with much less restrictive sparsity assumptions than those of existing methods. Test results show the superiority of our new proposed method over existing ones based on time-frequency masking.

Keywords: Blind source separation · Linear convolutive mixtures · Sparsity · Time-frequency masking · Bin-wise clustering · Determined mixtures

1 Introduction

Blind Source Separation (BSS) aims to find a set of N unknown signals, called sources and denoted by $s_j(n)$, knowing only a set of M mixtures of these sources, called observations and denoted by $x_i(n)$. This discipline is receiving increasing attention thanks to the diversity of its fields of application. Among these fields, we can cite those of audio, biomedical, seismic and telecommunications. In this paper, we are interested in so-called linear convolutive (LC) mixtures for which each mixture $x_i(n)$ is expressed in terms of the sources $s_j(n)$ and their delayed versions as follows:

$$x_i(n) = \sum_{j=1}^{N} \sum_{q=0}^{Q} h_{ij}(q) \cdot s_j(n-q) = \sum_{j=1}^{N} h_{ij}(n) * s_j(n), \qquad i \in [1, M], \qquad (1)$$

A. El Moataz et al. (Eds.): ICISP 2020, LNCS 12119, pp. 357–366, 2020.
https://doi.org/10.1007/978-3-030-51935-3_38

where:

- $h_{ij}(q)$ represents the impulse response coefficients of the mixing filter linking the source of index j to the sensor of index i,
- Q is the order of the longest filter,
- the symbol "$*$" denotes the linear convolution operator.

Indeed, in the field of BSS, the case of LC mixtures is still of interest since the performance of existing methods is still modest compared to the particular case of linear instantaneous mixtures for which $Q = 0$. BSS methods for LC mixtures can be classified into two main families. The so-called temporal methods that deal with mixtures in the time domain and the so-called frequency methods that deal with mixtures in the time-frequency (TF) domain. The performance of the former is generally very modest and remains very restrictive in terms of assumptions compared to the latter. Indeed, based mostly on the independence of source signals, most efficient methods are compared to frequency ones only for very short filters (i.e. Q low), and generally require over-determined mixtures (i.e. for $M > N$) [12,16]. Based mostly on the sparsity of source signals in the TF domain, the frequency methods have shown good performance in the determined case (i.e. for $M = N$) or even under-determined case (i.e. for $M < N$), and this despite increasing the filters length [4,8,9,13–15]. These frequency methods start by transposing the Eq. (1) into the TF domain using the short time Fourier transform (STFT) as follows:

$$X_i(m,k) = \sum_{j=1}^{N} H_{ij}(k) \cdot S_j(m,k), \quad m \in [0, T-1], \quad k \in [0, K-1], \qquad (2)$$

where:

- $X_i(m,k)$ and $S_j(m,k)$ are the STFT representations of $x_i(n)$ and $s_j(n)$ respectively,
- K and T are the length of the analysis window[1] and the number of time windows used by the STFT respectively[2],
- $H_{ij}(k)$ is the Discrete Fourier Transform of $h_{ij}(n)$ calculated on K points.

Among most efficient and relatively more recent frequency methods, we can mention those based on TF masking [2,4,6–9,13–15]. The sparsity is often exploited by these methods by assuming that the source signals are W-disjoint orthogonal, i.e. not overlapping[3] in the TF domain. The principle of these methods is to estimate a separation mask, denoted by $M_j(m,k)$ and specific to each source $S_j(m,k)$, which groups the TF points where only this source is present.

[1] Assuming that the length K of the analysis window used is sufficiently larger than the filters order Q (i.e. $K > Q$).

[2] It should be noted however that the equality in Eq. (2) is only an approximation. This equality would only be true if the discrete convolution used was circular, which is not the case here. We also note that this STFT is generally used with an analysis window different than the rectangular window [2,4,6–9,13–15].

[3] Which means, in each TF point at most one source is present.

The application of the estimated mask $M_j(m,k)$ to one of the frequency observations $X_i(m,k)$ allows us to keep from the latter only the TF points belonging to the source $S_j(m,k)$, and then separate it from the rest of the mixture. Depending on the procedure used to estimate the masks, we distinguish between two types of BSS methods based on TF masking. The so-called *full-band* methods [2,4,6,9] for which the masks are estimated integrally using a clustering algorithm that processes all frequency bins simultaneously, and the so-called *bin-wise* methods [7,8,13–15] for which the masks are estimated using a clustering algorithm that processes only one frequency bin at a time.

Among the most popular *full-band* methods we can cite those proposed in [4,9] which are based on the clustering of the level ratios and phase differences between the frequency observations $X_i(m,k)$ to estimate the separation masks. However, this clustering is not always reliable, especially when the order Q of the mixing filters increases [4]. Moreover, when the maximum distance between the sensors is greater than half the wavelength of the maximum frequency of source signals involved, a problem called *spatial aliasing* is inevitable [4]. The *bin-wise* methods [7,8,13–15] are robust to these two problems. However, these methods require the introduction of an additional step to solve a permutation problem in the estimated masks, when we pass from one frequency bin to another, which is a classical problem that is common to all *bin-wise* BSS methods.

However, all of these BSS methods based on TF masking (*full-band* and *bin-wise*) suffers from artifacts problem which affect the quality of the separated signals and due to the fact that the W-disjoint orthogonality assumption is not perfectly verified in practice. Indeed, being introduced by the TF masking operation, these artifacts are more and more troublesome when the spectral overlap of source signals in the TF domain becomes important. In [11] the authors proposed a first solution to this problem which consists of a cepstral smoothing of spectral masks before applying them to the frequency observations $X_i(m,k)$. An interesting extension of this technique, which was proposed in [3], consists in applying cepstral smoothing not to spectral masks but rather to the separated signals, i.e. after applying the separation masks. Knowing that these two techniques [3,11] were have only been validated on a few *full-band* methods, in [5] we have recently proposed to evaluate their effectiveness using a few popular *bin-wise* methods. However, these two solutions could only improve *one particular type of artifact* called *musical noise* [3,5,11]. In the same sense, in order to avoid the artifacts caused by the TF masking operation, we propose in this paper a new BSS method which also exploits the sparsity of source signals in the TF domain for determined LC mixtures. Indeed, by focusing on the case of determined mixtures, we show that we can avoid TF masking and also relax the W-disjoint orthogonality assumption. Note that the case of determined mixtures was also addressed in [1], but with an assumption which is again very restrictive and which consists in having at least a whole time frame of silence[4] for each of the source signals. Thus, our new method makes it possible to carry out the sep-

[4] Of length greater or equal to the length K of the analysis window used in the calculation of the STFT.

aration while avoiding the artifacts introduced by the operation of TF masking, with sparsity assumptions much less restrictive than those of existing methods.

We begin in Sect. 2 by describing our method. Then we present in Sect. 3 various experimental results that measure the performance of our method compared to existing methods, then we conclude with a conclusion and perspectives of our work in Sect. 4.

2 Proposed Method

The sole sparsity assumption of our method is the following.

Assumption: For each source $s_j(n)$ and for each frequency bin k, *there is at least one TF point* (m, k) where it is present alone, i.e:

$$\forall\, j,\ \forall\, k,\ \exists\ m\ /\ S_j(m,k) \neq 0\ \&\ S_i(m,k) = 0,\ \forall\, i \neq j \tag{3}$$

Thus, if we denote by E_j the set of TF points (m, k) that verify the assumption (3), called **single-source points**, then the relation (2) gives us:

$$X_i(m,k) = H_{ij}(k) \cdot S_j(m,k), \quad \forall (m,k) \in E_j. \tag{4}$$

Our method proceeds in two steps. The first step, which exploits the probabilistic masks used by Sawada et al. in [14,15], consists in identifying for each source of index "j" and each frequency bin "k" the index "m_{jk}" such that the TF point (m_{jk}, k) best verifies the Eq. (4), then in estimating the separating filters, denoted $F_{ij}(k)$ and defined by:

$$F_{ij}(k) = \frac{X_i(m_{jk},k)}{X_1(m_{jk},k)} = \frac{H_{ij}(k) \cdot S_j(m_{jk},k)}{H_{1j}(k) \cdot S_j(m_{jk},k)} = \frac{H_{ij}(k)}{H_{1j}(k)}, \quad i \in [2, M] \tag{5}$$

The second step consists in recombining the mixtures $X_i(m, k)$ using the separating filters $F_{ij}(k)$ in order to finally obtain an estimate of the separated sources. The two steps of our method are the subject of Sects. 2.1 and 2.2 respectively.

2.1 Estimation of the Separating Filters

Since the proposed treatment in this first step of our method is performed independently of the frequency, we propose in this section to simplify the notations by omitting the frequency bin index "k". So using a matrix formulation, the Eq. (2) gives us:

$$\mathbf{X}(m) = \sum_{j=1}^{N} \mathbf{H}_j \cdot S_j(m), \tag{6}$$

where $\mathbf{X}(m) = [X_1(m,k), ..., X_M(m,k)]^T$, $\mathbf{H}_j = [H_{1j}(k), ..., H_{Mj}(k)]^T$ and $S_j(m) = S_j(m,k)$. During this first step, we proceed as follows:

1. Each vector $\mathbf{X}(m)$ is normalized and then whitened as follows:

$$\widetilde{\mathbf{X}}(m) = \frac{\mathbf{X}(m)}{||\mathbf{X}(m)||} \quad \text{and} \quad \mathbf{Z}(m) = \frac{\mathbf{W}\widetilde{\mathbf{X}}(m)}{||\mathbf{W}\widetilde{\mathbf{X}}(m)||}, \tag{7}$$

where $\mathbf{W}$ is given by $\mathbf{W} = \mathbf{D}^{-\frac{1}{2}}\mathbf{E}^H$, with $\mathbb{E}\{\widetilde{\mathbf{X}}(m)\widetilde{\mathbf{X}}^H(m)\} = \mathbf{EDE}^H$.

2. Each vector $\mathbf{Z}(m)$ is modeled by a complex Gaussian density function of the form [14]:

$$p(\mathbf{Z}(m)|\mathbf{a}_j, \sigma_j) = \frac{1}{(\pi\sigma_j^2)^M} \cdot \exp\left(-\frac{||\mathbf{Z}(m) - (\mathbf{a}_j^H\mathbf{Z}(m)).\mathbf{a}_j||^2}{\sigma_j^2}\right) \tag{8}$$

where $\mathbf{a}_j$ and σ_j^2 are respectively the centroid (with unit norm $||\mathbf{a}_j|| = 1$) and the variance of each cluster C_j. This density function $p(\mathbf{Z})$ can be described by the following mixing model:

$$p(\mathbf{Z}(m)|\theta) = \sum_{j=1}^{N} \alpha_j \cdot p(\mathbf{Z}(m)|\mathbf{a}_j, \sigma_j), \tag{9}$$

where α_j are the mixture ratios and $\theta = \{\mathbf{a}_1, \sigma_1, \alpha_1, ..., \mathbf{a}_N, \sigma_N, \alpha_N\}$ is the parameter set of the mixing model.
Then, an iterative algorithm of the type *expectation-maximization* (*EM*) is used to estimate the parameter set θ, as well as the posterior probabilities $P(C_j|\mathbf{Z}(m), \theta)$ at each TF point, which are none other than the probabilistic masks used in [14].
In the *expectation step*, these posterior probabilities are given by:

$$P(C_j|\mathbf{Z}(m), \theta) = \frac{\alpha_j p(\mathbf{Z}(m)|\mathbf{a}_j, \sigma_j)}{p(\mathbf{Z}(m)|\theta)}. \tag{10}$$

In the maximization step, the update of centroid $\mathbf{a}_j$ is given by the eigenvector associated with the largest eigenvalue of the matrix $\mathbf{R}_j$ defined by:

$$\mathbf{R}_j = \sum_{m=0}^{T-1} P(C_j|\mathbf{Z}(m), \theta) \cdot \left\{\mathbf{Z}(m)\mathbf{Z}^H(m)\right\}. \tag{11}$$

The parameters σ_j^2 and α_j are updated respectively via the following relations:

$$\sigma_j^2 = \frac{\sum_{m=0}^{T-1} P(C_j|\mathbf{Z}(m), \theta) \cdot ||\mathbf{Z}(m) - (\mathbf{a}_j^H\mathbf{Z}(m)).\mathbf{a}_j||^2}{M \cdot \sum_{m=0}^{T-1} P(C_j|\mathbf{Z}(m), \theta)} \tag{12}$$

$$\alpha_j = \frac{1}{T}\sum_{m=0}^{T-1} P(C_j|\mathbf{Z}(m), \theta). \tag{13}$$

However, since the *EM* algorithm used in [14,15] is sensitive to the initialization[5], we propose in our method to initialize the masks with those obtained

[5] Which is done randomly in [14,15] and can lead to terrible performance.

by a modified version of the MENUET method [4]. Indeed, we replaced, in the clustering step for the estimation of the masks, the *k-means* algorithm used in [4] by the fuzzy c-means (*FCM*) algorithm used in [13], in order to have probabilistic masks.

3. After the convergence of the *EM* algorithm, the classical permutation problem between the different frequency bins is solved by the algorithm proposed in [15], which is based on the inter-frequency correlation between the time sequences of posterior probabilities $P(C_j|\mathbf{Z}(m),\theta)$ in each frequency bin k. In the following we denote these posterior probabilities by $P(C_j|\mathbf{Z}(m,k))$.
4. Unlike the approach adopted in [14,15] which consists in using all the TF points of the estimated probabilistic masks $P(C_j|\mathbf{Z}(m,k))$, we are interested in this step only in identifying one **single-source** TF point for each source of index "j" and for each frequency bin "k", therefore a single time frame index that we denote by "m_{jk}", which best verifies our working assumption (4). We then define this index m_{jk} as being the index "m" for which the presence probability of the corresponding source is maximum[6]:

$$m_{jk} = \underset{m}{\operatorname{argmax}}\, P(C_j|\mathbf{Z}(m,k)), \quad m \in [0, T-1] \tag{14}$$

5. After having identified these "best" **single-source** TF points (m_{jk}, k), we finish this first step of our method by estimating the separating filters $F_{ij}(k)$ defined in (5) by:

$$F_{ij}(k) = \frac{X_i(m_{jk},k)}{X_1(m_{jk},k)} = \frac{H_{ij}(k)}{H_{1j}(k)},\ i \in [2, M] \tag{15}$$

2.2 Estimation of the Separated Sources

In this section, for more clarity, we provide the mathematical bases for the second step of our method for two LC mixtures of two sources, i.e. for $M = N = 2$. The generalization to the case $M = N > 2$ can be derived directly from this in an obvious way. In this case, the mixing Eq. (1) gives us:

$$\begin{cases} x_1(n) = h_{11}(n) * s_1(n) + h_{12}(n) * s_2(n) \\ x_2(n) = h_{21}(n) * s_1(n) + h_{22}(n) * s_2(n) \end{cases} \tag{16}$$

As we pass to the TF domain, we get:

$$\begin{cases} X_1(m,k) = H_{11}(k) \cdot S_1(m,k) + H_{12}(k) \cdot S_2(m,k) \\ X_2(m,k) = H_{21}(k) \cdot S_1(m,k) + H_{22}(k) \cdot S_2(m,k) \end{cases} \tag{17}$$

We use the separating filters $F_{ij}(k)$, with $i = 2$ and $j = 1, 2$, estimated in the first step to recombine these two mixtures as follows:

$$\begin{cases} X_2(m,k) - F_{22}(k) \cdot X_1(m,k) = \widetilde{S}_1(m,k) \\ X_2(m,k) - F_{21}(k) \cdot X_1(m,k) = \widetilde{S}_2(m,k) \end{cases} \tag{18}$$

[6] Note however that in practice, only the indices "m" with an energy $||\mathbf{X}(m)||^2$ which is not negligible are concerned by the Eq. (14).

Since we have $F_{21}(k) = \frac{H_{21}(k)}{H_{11}(k)}$ and $F_{22}(k) = \frac{H_{22}(k)}{H_{12}(k)}$, based on the Eq. (15), we get after all simplifications have been made:

$$\begin{cases} \widetilde{S}_1(m,k) = \frac{H_{21}(k).H_{12}(k) - H_{22}(k).H_{11}(k)}{H_{12}(k)} \cdot S_1(m,k) \\ \widetilde{S}_2(m,k) = \frac{H_{22}(k).H_{11}(k) - H_{21}(k).H_{12}(k)}{H_{11}(k)} \cdot S_2(m,k) \end{cases} \tag{19}$$

In order to ultimately obtain the contributions of sources in one of the sensors, we propose to add a post-processing step (as in [1]) which consists in multiplying the signals $\widetilde{S}_j(m,k)$ by filters, denoted by $G_j(k)$, as follows:

$$G_j(k) \cdot \widetilde{S}_j(m,k) = Y_j(m,k), \qquad j \in \{1,2\}, \tag{20}$$

$$\text{where} \qquad G_1(k) = \frac{1}{F_{21}(k) - F_{22}(k)} \qquad \text{and} \qquad G_2(k) = \frac{1}{F_{22}(k) - F_{21}(k)}. \tag{21}$$

After all the simplifications are done, we get:

$$\begin{cases} Y_1(m,k) = H_{11}(k) \cdot S_1(m,k) \\ Y_2(m,k) = H_{12}(k) \cdot S_2(m,k) \end{cases} \tag{22}$$

By denoting $y_j(n)$ the inverse STFT of $Y_j(m,k)$ we get:

$$\begin{cases} y_1(n) = h_{11}(n) * s_1(n) \\ y_2(n) = h_{12}(n) * s_2(n) \end{cases} \tag{23}$$

These signals are none other than the contributions of source signals $s_1(n)$ and $s_2(n)$ on the first sensor (see the expression of the mixture $x_1(n)$ in (16)).

3 Results

In order to evaluate the performance of our method and compare it to the most popular *bin-wise* methods known for their good performance, that is the method proposed by *Sawada* et al. [15] and the *UCBSS* method [13], we performed several tests on different sets of mixtures. Each set consists of two mixtures of two real audio sources, which are sampled at 16 KHz and with a duration of 10 s each, using different filter sets. Generated by the toolbox [10], which simulates a real acoustic room characterized by a reverberation time denoted by RT_{60}[7], the coefficients $h_{ij}(n)$ of these mixing filters depend on the distance between the two sensors (microphones), denoted as D and on the absolute value of the difference between directions of arrival of the two source signals, denoted as $\delta\varphi$. For the calculation of the STFT, we used a 2048 sample Hanning window (as analysis window) with a 75% overlap. To measure the performance we used two of the most commonly used criteria by the BSS community, called *Signal to*

[7] RT_{60} represent the time required for reflections of a direct sound to decay by 60 dB below the level of the direct sound.

Distortion Ratio (SDR) and *Signal to Artifacts Ratio* (SAR) provided by the *BSSeval* toolbox [17] and both expressed in decibels (*dB*). The SDR measures the global performance of any BSS method, while the SAR provides us with a specific information on its performance in terms of artifacts presented in the separated signals.

For each test we evaluated the performance of the three methods, in terms of SDR and SAR, over **4 different realizations** of the mixtures related to the use of different sets of source signals cited above. Thus, the values provided below for SDR and SAR represent the average obtained over these 4 realizations[8].

In the first experiment, we evaluated the performance as a function of the parameters D and $\delta\varphi$ for an acoustic room characterized by $RT_{60} = 50\,\text{ms}$. Table 1 groups the performance for $D \in \{0.3\,\text{m}, 1\,\text{m}\}$ and $\delta\varphi \in \{85°, 55°, 30°\}$, where the last column for each value of the parameter D represents the average value of SDR and SAR over the three values $\delta\varphi_i$ of $\delta\varphi$.

Table 1. SDR (dB) and SAR (dB) as a function of D and $\delta\varphi$ for $RT_{60} = 50$ ms.

Method	Performance	$D = 0.3\,\text{m}$				$D = 1\,\text{m}$			
		$\delta\varphi_1$	$\delta\varphi_2$	$\delta\varphi_3$	Mean	$\delta\varphi_1$	$\delta\varphi_2$	$\delta\varphi_3$	Mean
Sawada	SDR	11.75	11.76	12.25	11.92	12.46	12.20	11.88	12.18
	SAR	12.16	12.17	12.72	12.35	12.74	12.59	12.34	12.56
UCBSS	SDR	5.02	8.68	5.82	6.51	8.65	8.71	10.73	9.36
	SAR	7.71	9.99	8.30	8.67	9.67	9.88	11.73	10.43
Proposed method	SDR	16.55	17.40	17.17	**17.04**	16.17	15.08	16.03	**15.76**
	SAR	17.74	18.56	18.57	**18.29**	17.83	16.13	17.81	**17.26**

According to Table 1, we can see that our method is performing better than the other two methods, and this over the 4 realizations of mixtures tested. Indeed, the proposed method shows superior performance over these two methods by about 5 dB for $D = 0.3\,\text{m}$ and 3.5 dB for $D = 1\,\text{m}$ in terms of SDR. This performance difference is even more visible in terms of SAR, which confirms that the artifacts introduced by our method are less significant than those introduced by the other two methods.

In our second experiment we were interested in the behavior of our method with regard to the increase of the reverberation time while fixing the parameters D and $\delta\varphi$ respectively to $D = 0.3\,\text{m}$ and $\delta\varphi = 55°$. Table 2 groups the performance of the three methods in terms of SDR, for RT_{60} belonging to the interval $\{50\,\text{ms}, 100\,\text{ms}, 150\,\text{ms}, 200\,\text{ms}\}$[9].

According to Table 2, we can see again that the best performance is obtained by using our method whichever the reverberation time. However, we note that

[8] We have indeed opted for these 4 realizations instead of only one in order to approach as close as possible to a statistical validation of our results.

[9] I.e. the mixing filters length ($Q + 1 = f_s \cdot RT_{60}$) varies from 800 coefficients (for $RT_{60} = 50\,\text{ms}$) to 3200 coefficients (for $RT_{60} = 200\,\text{ms}$).

Table 2. SDR (dB) as a function of RT_{60} for $D = 0.3$ m and $\delta\varphi = 55°$.

Method	RT_{60}			
	50 ms	100 ms	150 ms	200 ms
Sawada	11.76	11.42	9.26	7.65
UCBSS	8.68	5.12	3.83	3.04
Proposed method	**17.40**	**13.60**	**11.02**	**8.12**

this performance is degraded when RT_{60} increases. This result, which is common to all BSS methods, is expected and is mainly explained by the fact that the higher the reverberation time, the less the assumption (here of sparseness in the TF domain) assumed by these methods on source signals is verified.

4 Conclusion and Perspectives

In this paper, we have proposed a new Blind Source Separation method for linear convolutive mixtures with a sparsity assumption in the time-frequency domain that is much less restrictive compared to the existing methods [1,2,4,6–9,13–15]. Indeed, by focusing on the case of determined mixtures, we have shown that our method avoids the problem of artifacts at the separated signals from which suffers most of these methods [2,4,6–9,13–15]. According to the results of the several tests performed, the performance of our new method, in terms of SDR and SAR, is better than that obtained by using the method proposed by *Sawada* et al. [15] and the *UCBSS* method [13], which are known for their good performance within existing methods. Nevertheless, considering that these results were obtained over 4 different realizations of the mixtures and only for some values of the parameters involved, a larger statistical performance study including all these parameters is desirable to confirm this results. Furthermore, it would be interesting to propose a solution to this problem of artifacts also in the case of under-determined linear convolutive mixtures.

References

1. Albouy, B., Deville, Y.: Alternative structures and power spectrum criteria for blind segmentation and separation of convolutive speech mixtures. In: Fourth International Conference on Independent Component Analysis and Blind Source Separation (ICA2003), pp. 361–366, April 2003
2. Alinaghi, A., Jackson, P.J., Liu, Q., Wang, W.: Joint mixing vector and binaural model based stereo source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(9), 1434–1448 (2014)
3. Ansa, Y., Araki, S., Makino, S., Nakatani, T., Yamada, T., Nakamura, A., Kitawaki, N.: Cepstral smoothing of separated signals for underdetermined speech separation. In: 2010 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 2506–2509 (2010)

4. Araki, S., Sawada, H., Mukai, R., Makino, S.: Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. Sig. Process. **87**(8), 1833–1847 (2007)
5. Bella, M., Saylani, H.: Réduction des artéfacts au niveau des sources audio séparées par masquage temps fréquence en utilisant le lissage cepstral. In: Colloque International TELECOM 2019 and 11^{emes} JFMMA, pp. 58–61, June 2019
6. Ito, N., Araki, S., Nakatani, T.: Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3238–3242, May 2013
7. Ito, N., Araki, S., Nakatani, T.: Modeling audio directional statistics using a complex bingham mixture model for blind source extraction from diffuse noise. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 465–468, March 2016
8. Ito, N., Araki, S., Yoshioka, T., Nakatani, T.: Relaxed disjointness based clustering for joint blind source separation and dereverberation. In: 2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 268–272, September 2014
9. Jourjine, A., Rickard, S., Yilmaz, O.: Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In: 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings, vol. 5, pp. 2985–2988, June 2000
10. Lehmann, E.A., Johansson, A.M.: Prediction of energy decay in room impulse responses simulated with an image-source model. J. Acoust. Soc. Am. **124**(1), 269–77 (2008)
11. Madhu, N., Breithaupt, C., Martin, R.: Temporal smoothing of spectral masks in the cepstral domain for speech separation. In: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 45–48, March 2008
12. Pedersen, M.S., Larsen, J., Kjems, U., Parra, L.C.: A survey of convolutive blind source separation methods. In: Springer Handbook of Speech Processing. Springer, November 2007
13. Reju, V.G., Koh, S.N., Soon, I.Y.: Underdetermined convolutive blind source separation via time-frequency masking. IEEE Trans. Audio Speech Lang. Process. **18**(1), 101–116 (2010)
14. Sawada, H., Araki, S., Makino, S.: A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures. In: 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 139–142, October 2007
15. Sawada, H., Araki, S., Makino, S.: Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans. Audio Speech Lang. Process. **19**(3), 516–527 (2011)
16. Saylani, H., Hosseini, S., Deville, Y.: Blind separation of convolutive mixtures of non-stationary and temporally uncorrelated sources based on joint diagonalization. In: Elmoataz, A., Mammass, D., Lezoray, O., Nouboud, F., Aboutajdine, D. (eds.) ICISP 2012. LNCS, vol. 7340, pp. 191–199. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31254-0_22
17. Vincent, E., Gribonval, R., Fevotte, C.: Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)