IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 August 2023, accepted 10 September 2023, date of publication 14 September 2023,
date of current version 20 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3315596

RESEARCH ARTICLE

Multi-Channel Bin-Wise Speech Separation Combining Time-Frequency Masking and Beamforming

MOSTAFA BELLA[1,2], HICHAM SAYLANI[2], SHAHRAM HOSSEINI[1], AND YANNICK DEVILLE[1], (Member, IEEE)
[1]IRAP, UPS, CNRS, CNES, Université de Toulouse, 31400 Toulouse, France
[2]MatSim, Faculté des Sciences, Université Ibnou Zohr, Agadir 80000, Morocco

Corresponding author: Mostafa Bella (mbella@irap.omp.eu)

This work was supported in part by Centre National d'Etudes Spatiales (CNES), France; and in part by Centre National pour la Recherche Scientifique et Technique (CNRST), Morocco.

ABSTRACT This paper presents a novel Blind Source Separation method that can handle convolutive mixtures that may be underdetermined. Our method combines TF masking and beamforming and exploits the source signals sparsity in the Time-Frequency (TF) domain. Remarkable performance can be achieved by TF masking-based methods, even in the underdetermined case, although they tend to generate unwanted artifacts at the level of the separated signals. Besides, beamforming techniques can achieve satisfactory performance only in the overdetermined and determined cases without distorting the estimated signals. By combining these two approaches, we can leverage their respective strengths. Firstly, we exploit the source signals sparsity in the TF domain to estimate probabilistic "bin-wise" masks by modeling the frequency observation vectors with a complex Gaussian Mixture Model and using an EM algorithm. However, due to the sensitivity of the EM algorithm to initialization, we propose properly selecting the initial values of the model parameters using Hermitian angles between the frequency observation vectors and a reference vector. Then, we utilize the estimated TF masks to estimate the Relative Transfer Functions of each source. Finally, we propose a new technique to obtain an estimate of the spatial images of the separated sources, which can be regarded as an underdetermined extension of the Linearly Constrained Minimum Power beamformer. Good performance was observed in test results for our method, both in the determined and underdetermined cases, compared to various existing methods with similar working hypotheses.

INDEX TERMS Blind Source Separation, Convolutive mixtures, speech separation, sparsity, TF masking, beamforming.

I. INTRODUCTION

Blind Source Separation (BSS) is a very active research field that aims at recovering a set of N unknown signals,[1] called sources and denoted $s_j(t)$ or their contributions on sensors, called source images and denoted $s_{ij}^{\text{img}}(t)$, knowing only a set of M mixtures of these sources, called observations and denoted $x_i(t)$. This field has gained significant attention due to its wide range of applications in various domains. Among these domains we can mention those of audio [1], [2], [3], [4], [5], [6], [7], telecommunications [8], [9], biomedical applications [10], [11] and astrophysics [12], [13]. BSS methods have been used in the literature to handle several types of linear mixtures. In this work, we are interested in the case of convolutive mixtures, which represent the most general and realistic case of linear mixtures and for which each mixture $x_i(t)$ is expressed as follows [1]:

$$x_i(t) = \sum_{j=1}^{N} \sum_{q=0}^{Q} h_{ij}(q) s_j(t-q);$$

The associate editor coordinating the review of this manuscript and approving it for publication was Ganesh Naik.

[1]In this paper, we suppose that the number of sources N is known.

$$= \sum_{j=1}^{N} h_{ij}(t) * s_j(t);$$
$$= \sum_{j=1}^{N} s_{ij}^{\text{img}}(t), \qquad i \in [1, M], \tag{1}$$

where:

- t is a discrete time index,
- Q represents the order of the finite impulse response (FIR) of the mixing filters,
- $h_{ij}(q)$ is an impulse response coefficient of the mixing filter that connects the source with index j to the sensor with index i,
- "$*$" denotes the discrete-time linear convolution operator,
- $s_{ij}^{\text{img}}(t)$ are the spatial images of each source of index j on each sensor of index i.

Since the performance of existing methods still needs improvement, especially in the underdetermined situation (i.e. for $M < N$) [5], [7], [14], the case of convolutive mixtures remains of interest in BSS research. These techniques can be split into two main categories: frequency-domain methods [1], [3], [7], [14], [15], which mainly handle mixtures in the time-frequency (TF) domain, and time-domain methods [6], [16], [17], which handle mixtures in the time domain.

The performance of the latter category is generally quite modest, especially when the reverberation time[2] is large and these methods generally require overdetermined mixtures (i.e. $M > N$) [6], [16], [17]. A detailed survey of these methods has been provided in [6]. As for the frequency-domain methods, they have shown their good performances even in the determined case (i.e. $M = N$) or underdetermined case, and this despite the increasing reverberation time [3], [4], [5], [7], [14], [15], [18], [19], [20], [21], [22], [23], [24], [25], [26]. These frequency-domain methods begin by transposing Eq. (1) to the TF domain using the Short Time Fourier Transform (STFT) as follows:

$$X_i(n,f) = \sum_{j=1}^{N} H_{ij}(f) S_j(n,f), \ n \in [0, T-1], f \in [0, K-1]. \tag{2}$$

where:

- $X_i(n,f)$ and $S_j(n,f)$ are respectively the STFT of $x_i(t)$ and $s_j(t)$; n represents the time dimension and f the frequency dimension,
- K and T are respectively the length of the analysis window[3] and the number of time windows used by the STFT,
- $H_{ij}(f)$ is the discrete Fourier transform of $h_{ij}(t)$ computed on f points for source with index j.

[2]The reverberation time represents the time required for reflections of a direct sound to decay by 60 dB below the level of the direct sound.

[3]Assuming that the length K of the used analysis window is significantly larger than the order of the filters Q.

Using a vector formulation, Eq. (2) yields:

$$X(n,f) = \sum_{j=1}^{N} H_j(f) S_j(n,f), \ n \in [0, T-1], f \in [0, K-1], \tag{3}$$

where:

- $X(n,f) = [X_1(n,f), \ldots, X_M(n,f)]^T$ is the mixture vector at each TF bin (n,f),
- $H_j(f) = [H_{1j}(f), \ldots, H_{Mj}(f)]^T$ is the mixing vector or Acoustic Transfer Function (ATF) at each frequency band f.

In the overdetermined and determined cases, frequency-domain methods based on Independent Component Analysis exploiting source signal independence can be used to separate sources [27], [28]. However, these methods are not applicable in the case of underdetermined mixtures. For the underdetermined case, frequency domain BSS methods generally rely on a sparsity [3], [4], [5], [7], [14], [18], [19], [20], [23], [24] or a non-negativity [21], [22] assumption in the TF plane. Non-negative Matrix Factorization (NMF) methods exploiting the non-negativity of the sources have shown their efficiency in several works [21], [22]. However, these methods suffer from some limitations. Indeed, due to the non-uniqueness of the unconstrained NMF solution and the possible convergence of the algorithm towards spurious local minima, the performances of these methods depend on the parameter initialization. Moreover, these methods usually converge slowly.

The TF clustering-based methods [4], [5], [7], [14], [18], [19], [20] are among the most efficient and robust frequency-domain methods. These techniques rely on the assumption that the sources are W-disjoint orthogonal (WDO) in the TF domain. This implies that, at most, one source is dominant in each TF bin. The basic idea of these methods involves estimating a separation mask denoted as $M_j(n,f)$, which indicates the TF bins where a particular source $S_j(n,f)$ dominates. Each source has its specific mask. By applying the estimated mask $M_j(n,f)$ to a frequency observation $X_i(n,f)$ (TF masking), we retain only the TF bins that belong to the source $S_j(n,f)$. In various methods, channel features such as amplitude ratios and phase differences between frequency observations are computed at each TF bin. These features are then utilized to estimate TF masks using either clustering methods [1], [3], [7], [14], [29] or deep neural networks (DNN) [30], [31], [32], [33]. It is worth noting that DNN-based methods typically require a large amount of training data, and their performance may significantly degrade if the training and test settings differ. In contrast, clustering-based methods directly utilize the extracted features to estimate TF masks and do not require any training procedure.

Based on the employed clustering procedure for mask estimation, BSS methods can be categorized into "full-band" and "bin-wise" methods. In the case of "full-band" methods, such as those proposed in [4], [5], [15], and [19], the masks are estimated by processing all the

frequency bands simultaneously. In contrast, ''bin-wise'' methods, like those described in [7], [14], [18], and [20], estimate the masks by treating only one frequency band at a time. Among the well-known ''full-band'' methods, noteworthy examples are the ones proposed in [4] and [5]. To estimate the separation masks, these methods rely on clustering the phase differences and amplitude ratios between frequency observations. However, their performance can be compromised in scenarios with high reverberation since the linear phase assumption, which is fundamental to these methods, does not hold under such conditions [4], [5]. Furthermore, when the maximum distance between the sensors reaches half the wavelength of the highest frequency among the source signals involved, an overlapping problem known as ''spatial aliasing'' occurs [4], [5].

As for the best performing and most robust *bin-wise* methods, we can mention those based on the clustering of frequency observation vectors in each frequency band [7], [18], [20]. Since these methods process each frequency band separately, they are robust against the problems that *full-band* methods suffer from. Nevertheless, the order of the clusters is sometimes different when moving from one frequency band to another, which requires an additional step to reorganize them. This classical *permutation problem*, is common to all *bin-wise* methods. The permutation problem has been solved using various approaches [7], [18], [34], [35], [36], [37]. Among these approaches, the two most commonly employed methods are those based on the Time Difference Of Arrival (TDOA) of the sources and those based on the correlation between the different frequency bands [7], [18], [34], [35]. For the first ones, the TDOA of each source on the different sensors is exploited to solve the permutation problem. Indeed, these TDOAs are estimated using the separated signals in each frequency band, and then they are classified to estimate the permutation matrix at each frequency band. These approaches are very simple but inefficient if the reverberation time is high. As for the approaches proposed in [7], [18], [34], and [35], they are based on the fact that there is a substantial correlation between adjacent frequency bands of the same source. To measure this inter-frequency correlation, several activity functions[4] have been proposed in the literature [7], [18], [34], [35], such as the amplitude envelopes [34], the power ratios of the separated signals [35], or the posterior probability sequences of the masks [7], [18].

However, in practice, the WDO assumption made on the sources is not entirely verified, resulting in artifact problems in all BSS methods based on TF clustering (*full-band* and *bin-wise*). These artifacts arise from the TF masking procedure and have more impact on the separated signals when there is a substantial overlap of their spectra in the TF domain. In [3], we have proposed a novel BSS method that addresses these artifact issues. In this method, we estimate the Relative Transfer Function (RTF) for each source using the TF masks of the sources. The RTF represents the ratios between the various mixing filters in the TF domain. By recombining the various mixtures in the TF domain using these RTFs, we can estimate undistorted spatial images instead of their distorted versions as in [5], [7], and [14]. However, despite the encouraging results of this method, it is important to note that it is not applicable in the underdetermined case. Another alternative to solve this ''artifact problem'' is to use adaptive beamforming techniques, which are effective and robust even in complex cases [38], [39], [40]. These techniques aim to form a spatial filter that extracts a target signal from a specific position while attenuating other signals (interference and noise) coming from unwanted positions. It should be noted, however, that the major problem with adaptive beamforming techniques is that their basic version cannot operate blindly. Indeed, these techniques require information about the used microphone array and the target source, such as the ATFs (or RTFs) of the involved sources. In practice, this information is generally unknown and must be estimated from the observations. Besides, these techniques can achieve satisfactory performances only in the overdetermined and determined cases [38], [39], [40].

Due to the problems mentioned above that TF clustering-based BSS methods and adaptive beamforming techniques suffer from, we propose a new method for possibly under-determined convolutive mixtures, that combines these two techniques. Indeed, combining these two techniques makes it possible to benefit from their advantages while reducing their limits. Several existing methods have also combined these two techniques [41], [42], [43], [44]. However, the methods proposed in [41] and [44] suffer from limitations such as scaling indeterminacy in each frequency band or an ideal anechoic space requirement. Similarly, the methods proposed in [42] and [43], which are considered the most efficient, rely on prior knowledge of the RTF of each source, which may not always be available in practical scenarios and yields a significant drawback for these methods. In contrast, our method is completely blind, meaning we do not rely on prior knowledge of the RTFs. Instead, we estimate the RTFs blindly, eliminating the need for prior information.

The proposed method consists of three steps and is partially based on our conference paper [1]. Indeed, in our conference paper [1], we presented the main idea of the proposed method without delving into details. The present paper builds upon the conference paper by making significant improvements to the various steps of the proposed method. It provides a more comprehensive description of the overall method, including a detailed derivation of the parameter update rules, as well as additional experimental results. In the first step, probabilistic bin-wise masks are estimated by exploiting the sparsity of source signals in the TF domain. This is achieved by modeling the frequency-domain observation vectors with a complex Gaussian Mixture Model (cGMM) and using an EM algorithm. To ensure accurate initialization of the

[4]This function represents the evolution (activity) of each estimated signal along time windows. The correlation between these activity functions in adjacent frequency bands will be high if the separated signals come from the same source.

EM algorithm, we suggest selecting initial values based on Hermitian angles between the frequency observation vectors and a reference vector. In the second step, these masks are utilized to estimate the RTFs of each source. Finally, in the third step, by exploiting the estimated masks and RTFs, an underdetermined extension of the Linearly Constrained Minimum Power (LCMP) beamformer is used to yield an estimate of source spatial images.

The rest of this article is organized as follows. We provide a detailed description of our proposed method in Section II. Then, in Section III, we present the results of various tests carried out to measure the performance of our method. Finally, in Section IV, we close with a conclusion and outlook for our work.

II. PROPOSED METHOD

The method proposed in this paper proceeds in three steps (as illustrated in Fig. 1). The details of these three steps are elaborated in Sections II-A, II-B, and II-C, respectively.

A. TIME-FREQUENCY MASK ESTIMATION

In this step, we focus on estimating probabilistic TF masks. Our approach for estimating these masks is primarily based on the method proposed in [20] and [45]. However, we have made some modifications to enhance the mask estimation process, which will be detailed further in this section. To achieve this, we first estimate the posterior probabilities of each source in each frequency band by modeling the mixture vectors with a complex Gaussian Mixture Model (cGMM) and using an expectation-maximization clustering algorithm. These posterior probability sequences in each frequency band are then utilized to solve the permutation problem between different frequency bands, as illustrated in Fig. 2.

1) BIN-WISE CLUSTERING

If we suppose that for each source with an index j, there exists a set of TF bins where this source is dominant (i.e. the magnitude of this source is significantly greater than that of the other sources at these TF bins), then Eq. (3) yields [1]:

$$X(n,f) \approx H_j(f) S_j(n,f), \quad \forall\, n \in E_j(f), \tag{4}$$

where $E_j(f)$ is the set of temporal indices in which the source $S_j(n,f)$ is dominant. In this section, we omit the index f to simplify the notation because each frequency band is processed independently in the first step. Therefore, $X(n,f)$, $S_j(n,f)$ and $E_j(f)$ will be denoted as $X(n)$, $S_j(n)$ and E_j respectively in this step.

We can see in Eq. (4) that when the sources are sufficiently sparse in the TF domain, the clustering can be performed based on the spatial diversity of the sources contained in the mixture vector $X(n)$. Therefore, as in [45] we choose in our method to use the mixture vector $X(n)$ in each frequency band as a feature vector. As in [20] and [45], we assume that each mixture vector $X(n)$ conditioned by $n \in E_j$ follows a zero-mean conditional complex-valued Gaussian distribution. This distribution is described as follows [1]:

$$p(X(n)|j, \phi_j(n)B_j) = \frac{1}{\pi^M \det(\phi_j(n)B_j)} \exp\left(-X(n)^H (\phi_j(n)B_j)^{-1} X(n)\right), \tag{5}$$

where H denotes the Hermitian transpose, $\phi_j(n)$ represents the time-varying spectro-temporal power of source $S_j(n,f)$ and B_j represents the spatial covariance matrix of size $M \times M$, characterizing the time-invariant spatial properties related to $H_j(f)$. It should be noted that the multiplication of B_j by $\phi_j(n)$ in Eq. (5) aims to simultaneously account for both the source's time-invariant spatial properties and its time-varying characteristics.

As the mixture vector $X(n)$ is modeled by (3), the density function $p(X(n))$ of $X(n)$ can also be described by the following cGMM [20], [45]:

$$p(X(n)|\theta) = \sum_{j=1}^{N} \alpha_j p(X(n)|j, \phi_j(n)B_j), \tag{6}$$

where α_j are the mixture ratios and $\theta = \{\alpha_j, B_j, \phi_j(n)\}_{j=1}^{N}$ (for $n = 1, \ldots, T-1$) is the set of mixing model parameters.

The mixture ratios α_j should satisfy the following conditions:

$$\sum_{j=1}^{N} \alpha_j = 1, \quad 0 \le \alpha_j. \tag{7}$$

In contrast to the approach proposed in [45], which assumed a uniform distribution for all the mixing model parameters θ, we draw inspiration from [7] and model the mixture ratios α_j using a symmetric Dirichlet distribution with the following form:

$$p(\{\alpha_j\}_{j=1}^{N}) = \frac{\Gamma(N\beta)}{\Gamma(\beta)^N} \prod_{j=1}^{N} \alpha_j^{(\beta-1)}, \tag{8}$$

where the constant β is a positive hyper parameter that controls the sparsity of the Dirichlet distribution[5] and Γ is the gamma function. The other parameters of the mixture model are assumed to have a uniform distribution.

Then, as in [45] an iterative *Expectation-Maximization* (*EM*) algorithm is used to estimate the parameters θ, as well as the posterior probabilities $\gamma_j(n)$ at each TF bin which are the desired probabilistic masks. These posterior probabilities are calculated in the *expectation* step, using the *Bayes* theorem [1]:

$$\gamma_j(n) = \frac{\alpha'_j p(X(n)|j, \phi'_j(n)B'_j)}{p(X(n)|\theta')} = \frac{\alpha'_j p(X(n)|j, \phi'_j(n)B'_j)}{\sum_{l=1}^{N} \alpha'_l p(X(n)|l, \phi'_l(n)B'_l)}, \tag{9}$$

[5] After conducting multiple tests to determine the optimal value for β, we found that setting β to 100 resulted in the best performance in terms of the estimated TF masks. As a result, this value will be used in the upcoming tests section.

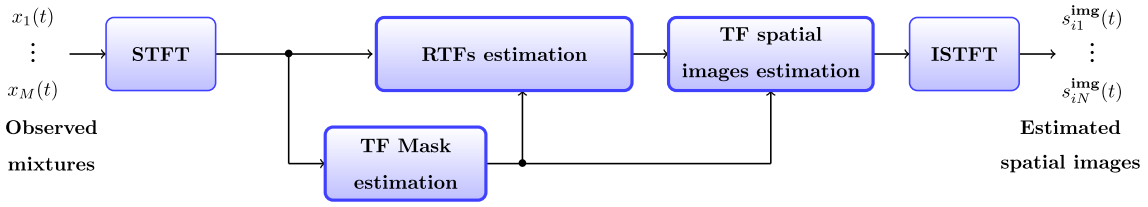


FIGURE 1. Block diagram of the proposed method [1].

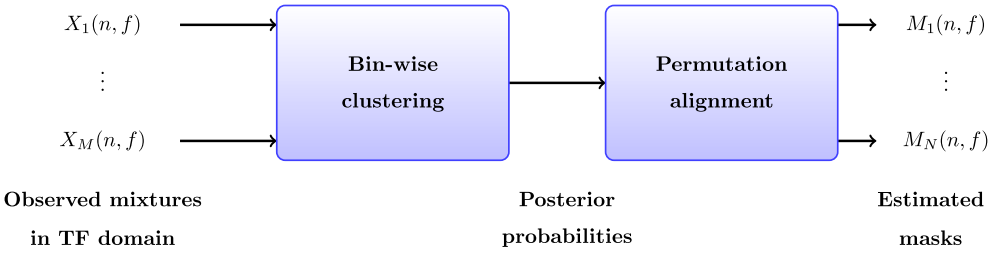


FIGURE 2. Mask estimation process of the proposed method.

where $\theta' = \{\alpha_j', B_j', \phi_j'(n)\}_{j=1}^{N}$ is the set of parameter values of the current iteration.

During the Maximization step, inspired by [45], the parameters of the set θ are estimated with a Maximum Likelihood (ML) approach by maximizing the auxiliary function $Q(\theta, \theta')$ defined by:

$$Q(\theta, \theta') = \sum_{n=0}^{T-1} \sum_{j=1}^{N} \gamma_j(n) \log\left(\alpha_j p(X(n)|j, \phi_j(n) B_j)\right) + \log(p(\theta)), \tag{10}$$

where $\log(p(\theta))$ is expressed according to Eq. (8) by[6]:

$$\log(p(\theta)) = (\beta - 1) \sum_{j=1}^{N} \log(\alpha_j) + \text{const.} \tag{11}$$

After maximizing the function $Q(\theta, \theta')$ with respect to each parameter in θ as described in Appendix A, we find that the update of the mixing ratios is given by the following equation:

$$\alpha_j = \frac{\sum_{n=0}^{T-1} \gamma_j(n) + \beta - 1}{T + N(\beta - 1)}. \tag{12}$$

The two parameters $\phi_j(n)$ and B_j are updated respectively via the following two equations:

$$\phi_j(n) = \frac{1}{M} X^H(n) B_j^{-1} X(n), \tag{13}$$

$$B_j = \frac{\sum_{n=0}^{T-1} \gamma_j(n) X(n) X^H(n) / \phi_j(n)}{\sum_{n=0}^{T-1} \gamma_j(n)}. \tag{14}$$

These two *Expectation* and *Maximization* steps are iterated until convergence, i.e. until the difference between two successive values of the parameters θ is lower than a given threshold.

The EM algorithm is sensitive to initialization, meaning that the initial chosen values can impact the final results. The reference method [20], [45] randomly initializes the EM algorithm, which may not always yield the best results. In our method, we suggest a different approach to initialize the EM algorithm. We first initialize the probabilistic masks $\gamma_j(n)$, then estimate the initial values of the parameters θ. More specifically, we initialize the masks based on the Hermitian angle between the observation vector $X(n)$ and the reference vector $h = [1, \ldots, 1]^T$ of size $M \times 1$. The use of the Hermitian angle is inspired by [14], and this angle can be calculated using the following equation:

$$\Psi_H(n) = \cos^{-1}(|\cos(\Phi(n))|), \tag{15}$$

where:

$$\cos(\Phi(n)) = \frac{X(n)^H h}{||X(n)|| \, . \, ||h||}. \tag{16}$$

Next, we use a Fuzzy c-means clustering algorithm [46] to classify the vectors of Hermitian angles, denoted as $\Psi_H = [\Psi_H(0), \ldots, \Psi_H(T-1)]$. The membership functions obtained with this clustering algorithm are then used to initialize the masks $\gamma_j(n)$. It should be noted that using Hermitian angles as an initialization method in the EM algorithm is a novel approach, and it shows promise in improving the convergence and accuracy of the algorithm.

We had previously omitted the frequency index f to simplify notations. In the following, we propose reintroducing this index to represent the corresponding frequency band. Consequently, the sequences of posterior probabilities will be denoted as $\gamma_j(n, f)$ in the following.

[6]The shape of the distribution $p(\theta)$ is determined by the distribution of $p(\{\alpha_j\}_{j=1}^{N})$, as the other parameters of the mixture model are assumed to follow a uniform distribution.

2) PERMUTATION ALIGNMENT

Once the *EM* algorithm has converged, the classical permutation problem between different frequency bands is handled using the method presented in [7] and [18]. This method is based on the inter-frequency correlation between the sequences of posterior probabilities $\gamma_j(n,f)$ in each frequency band. The first step of the method involves a global optimization step, which maximizes a cost function based on the correlation between the sequences $\gamma_j(n,f)$ and their centroids [7]. This cost function, denoted as $\mathcal{J}$, is formulated as follows [7], [18]:

$$\mathcal{J}(\{c_k\},\{\Pi_f\}) = \sum_{f}\sum_{k=1}^{N} \text{corr}(v_j^f, c_k)|_{j=\Pi_f(k)}, \tag{17}$$

where v_j^f corresponds to the vector of posterior probabilities $\gamma_j(n,f)$, and centroids c_j are calculated for each source $S_j(n,f)$ based on the current permutation vector Π_f at frequency band f [7].

The second step, which aims at improving the estimation of the permutation matrix obtained in the first step, involves a local optimization procedure that maximizes the correlation between a specific set of frequencies (harmonic and adjacent frequencies). These two steps are iterated until convergence to estimate the permutation matrix Π of size $N \times K$. After convergence, the desired masks $M_j(n,f)$ are obtained as the posterior probability $\gamma_j(n,f)$ with the estimated permutation matrix Π, as the following equation shows:

$$M_j(n,f) = \gamma_{j'}(n,f)|_{j'=\text{find}(\Pi(:,f)=j)},\ j \in [1,N]. \tag{18}$$

B. RELATIVE TRANSFER FUNCTION ESTIMATION

The accurate estimation of the mixing vectors $H_j(f)$ (or ATF) is the most crucial element to successfully apply beamforming techniques. These vectors are traditionally estimated using the DOA of the sources with a plane wave model [47] or using the signals that have been pre-separated via TF masking [41], [44]. However, these estimating approaches are not always reliable due to limitations such as the scaling indeterminacy present in each frequency band or requirement of an ideal anechoic space, which can result in modest performance of beamforming in certain scenarios.

In this paper, we are rather interested in the estimation of the Relative Transfer Functions (RTFs), denoted $F_{ij}(f)$, defined by:

$$F_{ij}(f) = \left[\frac{H_{1j}(f)}{H_{ij}(f)}, \frac{H_{2j}(f)}{H_{ij}(f)}, \ldots, \frac{H_{Mj}(f)}{H_{ij}(f)}\right]^T,$$
$$j \in [1,N],\ \ i \in [1,M]. \tag{19}$$

In fact, we demonstrated in [3] that once estimated, these RTFs allow us to reconstruct spatial images, while avoiding artifacts caused by the TF masking operation.

It is important to note that certain methods, such as those proposed in [42] and [43], also utilize the RTFs as steering vectors for beamforming. However, these methods rely on prior knowledge of the RTFs, which may not always be readily available in practical scenarios, yielding a significant drawback for these methods. In contrast, our method does not rely on such prior knowledge. Instead, we present an effective approach for estimating these vectors, which enables us to subsequently estimate the spatial images $s_{ij}^{\text{img}}(t)$ without relying on prior information. Indeed, for TF bins "(n,f)" that satisfy Assumption (4) for a source with index j, we have [1]:

$$\frac{X_p(n,f)}{X_i(n,f)} = \frac{H_{pj}(f)S_j(n,f)}{H_{ij}(f)S_j(n,f)} = \frac{H_{pj}(f)}{H_{ij}(f)} = F_{ij}^{(p)}(f),\ \ p \in [1,M], \tag{20}$$

where $F_{ij}^{(p)}(f)$ represents the p-th element of the vector $F_{ij}(f)$. In order to determine these filter ratios, we first identify a set of TF bins for each source with index j and in each frequency band with index f that best validates our working hypothesis (4). These selected indices are denoted as n_{jf} and they correspond to the indices "n" that satisfy the following condition [1]:

$$n \in \{n_{jf}\}\ \ if\ \ M_j(n,f) \geq \eta \cdot \max(M_j(:,f)), \tag{21}$$

where η is a threshold to be set and $\max(M_j(:,f))$ is the maximum value of the mask $M_j(n,f)$ at each frequency band. These best single-source TF bins (n_{jf},f) are then utilized to estimate the filter ratios $F_{ij}^{(p)}(f)$, $p \in [1,M]$, using the following relation, where the average $\hat{E}[.]$ is computed over the index n_{jf} [1]:

$$\hat{E}\left[\frac{X_p(n_{jf},f)}{X_i(n_{jf},f)}\right] \approx \frac{H_{pj}(f)}{H_{ij}(f)} = F_{ij}^{(p)}(f). \tag{22}$$

C. ESTIMATION OF SOURCE SPATIAL IMAGES

Unlike the approach adopted in [4], [5], [7], [14], [15], [18], [19], [20], and [29], which consists of estimating the source spatial images by directly applying the estimated masks $M_j(n,f)$ on the observations (TF masking), the proposed method adopts a different approach based on beamforming techniques to estimate the spatial images of the sources. One advantage of the approach proposed in this paper for estimating the source images $\hat{S}_{ij}^{\text{img}}(n,f)$ is that it can significantly reduce the presence of artifacts in the separated signals, which are common issues encountered in TF masking-based methods [4], [5], [7], [14], [15], [18], [19], [20], [29].

Beamforming techniques can be applied in both overdetermined and underdetermined cases. However, in the underdetermined case, the interference suppression capability of beamforming is limited, as a beamformer with M sensors can only attenuate a maximum of $M-1$ interferences at each TF bin [39], [40].

Several recent techniques have extended the MVDR beamformer [38] to address the underdetermined case [41], [42], [43]. Similarly, we propose a novel technique in this step of our method to estimate the spatial images of the source. This technique can be seen as an extension of

the LCMP beamformer [40] to handle the underdetermined scenario. The objective of the LCMP beamformer is to estimate an optimal filter, denoted as $\mathbf{w}(f)$, that minimizes the output power while satisfying the following constraints $w^H(f)C(f) = g^H$. The desired filter is obtained as the solution to the following minimization problem [1]:

$$w(f) = \underset{w(f)}{\operatorname{argmin}} \left\{ w^H(f) R_X(f) w(f) \right\} \quad \text{s.t. } w^H(f)C(f) = g^H, \tag{23}$$

where $C(f)$ is the constraint matrix, $R_X(f)$ is the covariance matrix of the observed data, and $g = [1, 0, \cdots, 0]^T$ is a vector with N elements. The beamforming coefficients, assuming that the N sources in the TF domain are uncorrelated, are computed as follows [40]:

$$w(f) = R_X^{-1}(f)C(f)(C^H(f)R_X^{-1}(f)C(f))^{-1}g. \tag{24}$$

Depending on whether the situation is overdetermined, determined or underdetermined, the approach to estimating the coefficients of these filters $w(f)$ and the strategy we adopt for using them will vary.

In the next two subsections, we assume that the sources are uncorrelated and that the observations are centered in each frequency band.

1) OVERDETERMINED AND DETERMINED CASES

Our approach utilizes the RTFs $F_{ij}(f)$ and the LCMP beamformer to estimate the spatial images $s_{ij}^{\text{img}}(t)$ in the overdetermined and determined cases. For each source within each frequency band, we generate a beamformer $w_{ij}(f)$ that filters the mixture $X(n,f)$. This beamformer is determined by:

$$w_{ij}(f) = R_X^{-1}(f)C_{ij}(f)(C_{ij}^H(f)R_X^{-1}(f)C_{ij}(f))^{-1}g, \tag{25}$$

where $g = [1, 0, \ldots, 0]^T$ represents the constraint vector of size $N \times 1$, and $C_{ij}(f)$ is the constraint matrix of size $M \times N$. The first column of $C_{ij}(f)$ contains the RTF $F_{ij}(f)$ of the source $s_j(t)$, while the remaining columns contain the RTFs $F_{ir}(f)$ of the other sources (where $r \neq j$). The covariance matrix $R_X(f)$ is defined as:

$$R_X(f) = E\{X(n,f)X^H(n,f)\}. \tag{26}$$

Here, statistical expectations are approximated by averaging over T time frames. The final estimation of the TF spatial images is given by:

$$\hat{S}_{ij}^{\text{img}}(n,f) = w_{ij}^H(f)X(n,f). \tag{27}$$

2) UNDERDETERMINED CASE

In the underdetermined case, we introduce a novel technique for estimating the spatial images of sources. This technique can be seen as an extension of the LCMP beamformer [40] specifically designed to handle the underdetermined case. The fundamental idea behind this technique is to classify the TF bins in each frequency band into $D = \frac{(N-1)!}{(M-1)!(N-M)!} + 1$ groups, where M distinct sources dominate each group. We use the estimated masks $M_j(n,f)$, $j = 1, \ldots, N$, to classify these bins. Next, to estimate a specific source, we construct a unique beamformer for each group where the desired source is one of the dominant sources. Each beamformer is designed to suppress a set of $M - 1$ interferences within the group. Finally, for groups where the desired source is not one of the dominant sources, we estimate the contribution of this source in those TF bins using a soft TF masking approach.

To illustrate this technique, let's consider the simplest underdetermined case with $N = 3$ sources and $M = 2$ mixtures in this paragraph. Furthermore, we focus on estimating the spatial image $s_{i1}^{\text{img}}(t)$ of source $s_1(t)$ on the sensor with index "i". It is important to note that the spatial images of the other sources ($s_{i2}^{\text{img}}(t)$ and $s_{i3}^{\text{img}}(t)$) are estimated using the same methodology. To begin, we classify the TF bins into $D = \frac{(N-1)!}{(M-1)!(N-M)!} + 1 = 3$ groups based on the following classification scheme:

$$(n,f) \in \begin{cases} E_{12}(f) \;\; if \;\; \min\{M_1(n,f), M_2(n,f)\} > M_3(n,f) \\ E_{13}(f) \;\; if \;\; \min\{M_1(n,f), M_3(n,f)\} \geq M_2(n,f) \\ E_{23}(f) \;\; if \;\; \min\{M_2(n,f), M_3(n,f)\} > M_1(n,f) \end{cases} \tag{28}$$

where $E_{jp}(f)$ gathers the TF bins where both sources $s_j(t)$ and $s_p(t)$ are dominant. Then, at each frequency band, we generate two beamformers, $w_{12}(f)$ and $w_{13}(f)$. The first beamformer filters the TF bins in the set $E_{12}(f)$, while the second beamformer filters the TF bins in the set $E_{13}(f)$. The expressions for these beamformers are as follows [1]:

$$\begin{cases} w_{12}(f) = R_{12}^{-1}(f)C_{12}(f)(C_{12}^H(f)R_{12}^{-1}(f)C_{12}(f))^{-1}g \\ w_{13}(f) = R_{13}^{-1}(f)C_{13}(f)(C_{13}^H(f)R_{13}^{-1}(f)C_{13}(f))^{-1}g \end{cases} \tag{29}$$

where $g = [1, 0]^T$ and the constraint matrices $C_{12}(f)$ and $C_{13}(f)$ are given by:

$$\begin{cases} C_{12}(f) = [F_{i1}(f), F_{i2}(f)] \\ C_{13}(f) = [F_{i1}(f), F_{i3}(f)] \end{cases} \tag{30}$$

The covariance matrices $R_{12}(f)$ and $R_{13}(f)$ are defined by[7] [1]:

$$\begin{cases} R_{12}(f) = E[Z_{12}(n,f)Z_{12}^H(n,f)], \text{ for } (n,f) \in E_{12}(f), \\ R_{13}(f) = E[Z_{13}(n,f)Z_{13}^H(n,f)], \text{ for } (n,f) \in E_{13}(f), \end{cases} \tag{31}$$

where the mixtures $Z_{12}(n,f)$ and $Z_{13}(n,f)$ are defined by [1]:

$$\begin{cases} Z_{12}(n,f) = (M_1(n,f) + M_2(n,f))\, X(n,f), \; (n,f) \in E_{12}(f) \\ Z_{13}(n,f) = (M_1(n,f) + M_3(n,f))\, X(n,f), \; (n,f) \in E_{13}(f) \end{cases} \tag{32}$$

[7] In these equations, the statistical expectations are approximated by averages over the time frames of the two sets $E_{12}(f)$ and $E_{13}(f)$.

The final estimate of the TF spatial images $S_{i1}^{\text{img}}(n,f)$ is expressed as [1]:

$$\hat{S}_{i1}^{\text{img}}(n,f) = \begin{cases} w_{12}^H(f) Z_{12}(n,f) & \text{if } (n,f) \in E_{12}(f) \\ w_{13}^H(f) Z_{13}(n,f) & \text{if } (n,f) \in E_{13}(f) \\ M_1(n,f) X_i(n,f) & \text{if } (n,f) \in E_{23}(f) \end{cases} \tag{33}$$

The temporal versions of the estimated spatial images are then obtained by applying the inverse STFT to the signals $\hat{S}_{i1}^{\text{img}}(n,f)$ as follows:

$$\hat{s}_{i1}^{\text{img}}(t) = \text{ISTFT}\{\hat{S}_{i1}^{\text{img}}(n,f)\}. \tag{34}$$

It is important to note that Eq. (33) can be used to estimate the spatial image of any source (with index j) in any sensor (with index i). Furthermore, it should be mentioned that our method can be readily generalized to cases where $N > 3$.

III. TEST RESULTS

A. TEST CONDITIONS

In this section, we will evaluate the separation performance of the proposed method through three different experiments. The first and second experiments will evaluate the separation performance on determined and underdetermined artificial mixtures of speech signals, respectively. In contrast, the third experiment will evaluate these performances on real underdetermined mixtures from the Signal Separation Evaluation Campaign (SiSEC) database [48]. For comparison, we selected methods known for their effectiveness and compatibility with our working hypotheses and test protocol. These methods are *TFS* [42], *UCBSS* [14], *TFLC* [43] and those proposed by Sawada et al. [7] (referred to as *Sawada*) and Ito et al. [20] (referred to as *Ito*). These methods have been chosen for their good performance and applicability in the determined and underdetermined cases. The experiments were conducted using multiple sets of mixtures of speech signals. Each set consisted of two mixtures of speech sources,[8] which were sampled at a rate of 16 kHz and had a duration of 10 s each. The mixing filters were generated using the toolbox described in [49], which simulates an acoustic room with dimensions 4.45 m × 3.55 m × 2.5 m and characterized by a varying Reverberation Time (RT_{60}). The coefficients $h_{ij}(t)$ of these mixing filters depend on the inter-microphone distance d, the angular distance $\delta\varphi$[9] between the source signals, and the microphone-source distance D.

For the computation of the STFT, an analysis window of length 2048 was used in tests where $RT_{60} < 250$ ms and a length of 4096 was used when $RT_{60} \geq 250$ ms.

In all tests, the Hanning window was employed as the analysis window, with an overlap of 75% and the threshold η introduced in (21) was set to 0.95.

[8]The sources used were taken from the database of the 2011 Signal Separation Evaluation Campaign (SiSEC 2011).

[9]$\delta\varphi$ represents the absolute value of the difference between directions of arrival of the sources.

B. PERFORMANCE MEASURES

We selected three commonly used metrics in the BSS community to evaluate performance of the tested methods: Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), and Signal to Artifact Ratio (SAR). The SDR metric measures the overall performance of any BSS method, while the SIR metric assesses the method's performance in terms of interference reduction. The SAR metric provides information on the method's performance regarding the presence of artifacts in the separated signals. All three criteria are provided by the *BSSeval* toolbox [50] and are expressed in decibels (dB). Computation of these criteria requires knowledge of the true spatial images of the sources $s_{ij}^{\text{img}}(t)$, which is necessary to decompose each estimated spatial image $\hat{s}_{ij}^{\text{img}}(t)$, as follows [50]:

$$\hat{s}_{ij}^{\text{img}}(t) = s_{ij}^{\text{img}}(t) + e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{int}}(t) + e_{ij}^{\text{artif}}(t), \tag{35}$$

where e_{ij}^{spat}, e_{ij}^{int} and e_{ij}^{artif} are error terms indicating spatial distortion, interference, and artifacts caused by the separation process, respectively. These performance metrics are defined for each source with index j based on the decomposition (35) of $\hat{s}_{ij}^{\text{img}}(t)$ by calculating the energy ratios as follows [50]:

$$\begin{cases} \text{SDR}_j = 10\log_{10} \dfrac{\sum_{i=1}^{M}\sum_t s_{ij}^{\text{img}}(t)^2}{\sum_{i=1}^{M}\sum_t \left(e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{int}}(t) + e_{ij}^{\text{artif}}(t)\right)^2} \\ \text{SIR}_j = 10\log_{10} \dfrac{\sum_{i=1}^{M}\sum_t \left(s_{ij}^{\text{img}}(t) + e_{ij}^{\text{spat}}(t)\right)^2}{\sum_{i=1}^{M}\sum_t e_{ij}^{\text{int}}(t)^2} \\ \text{SAR}_j = 10\log_{10} \dfrac{\sum_{i=1}^{M}\sum_t \left(s_{ij}^{\text{img}}(t) + e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{int}}(t)\right)^2}{\sum_{i=1}^{M}\sum_t e_{ij}^{\text{artif}}(t)^2} \end{cases} \tag{36}$$

Then the average SDR, SIR, and SAR are computed as follows:

$$\begin{cases} \text{SDR} = \dfrac{1}{N}\sum_{j=1}^{N} \text{SDR}_j \\ \text{SIR} = \dfrac{1}{N}\sum_{j=1}^{N} \text{SIR}_j \\ \text{SAR} = \dfrac{1}{N}\sum_{j=1}^{N} \text{SAR}_j \end{cases} \tag{37}$$

C. SEPARATION RESULTS IN THE DETERMINED CASE

In the first experiment, we were interested in the measurement of the performance of our method as a function of RT_{60}, and this for $RT_{60} \in \{50, 100, 150, 250, 500\}$ms, in the determined case $M = N = 2$. In this experiment we used two different microphone-source distances $D \in \{1, 2\}$ m and three different inter-source angles $\delta\varphi \in \{80°, 45°, 15°\}$. Fig. 3 groups the performance obtained as a function of RT_{60} for $d = 1$ m. Since we have 60 realizations,[10] this figure presents each metric's mean and standard deviation (SDR, SIR, and SAR).

[10]We averaged over the three values of $\delta\varphi$, the two values of D, and ten different realizations of the source signals, i.e. 60 mixtures in total.

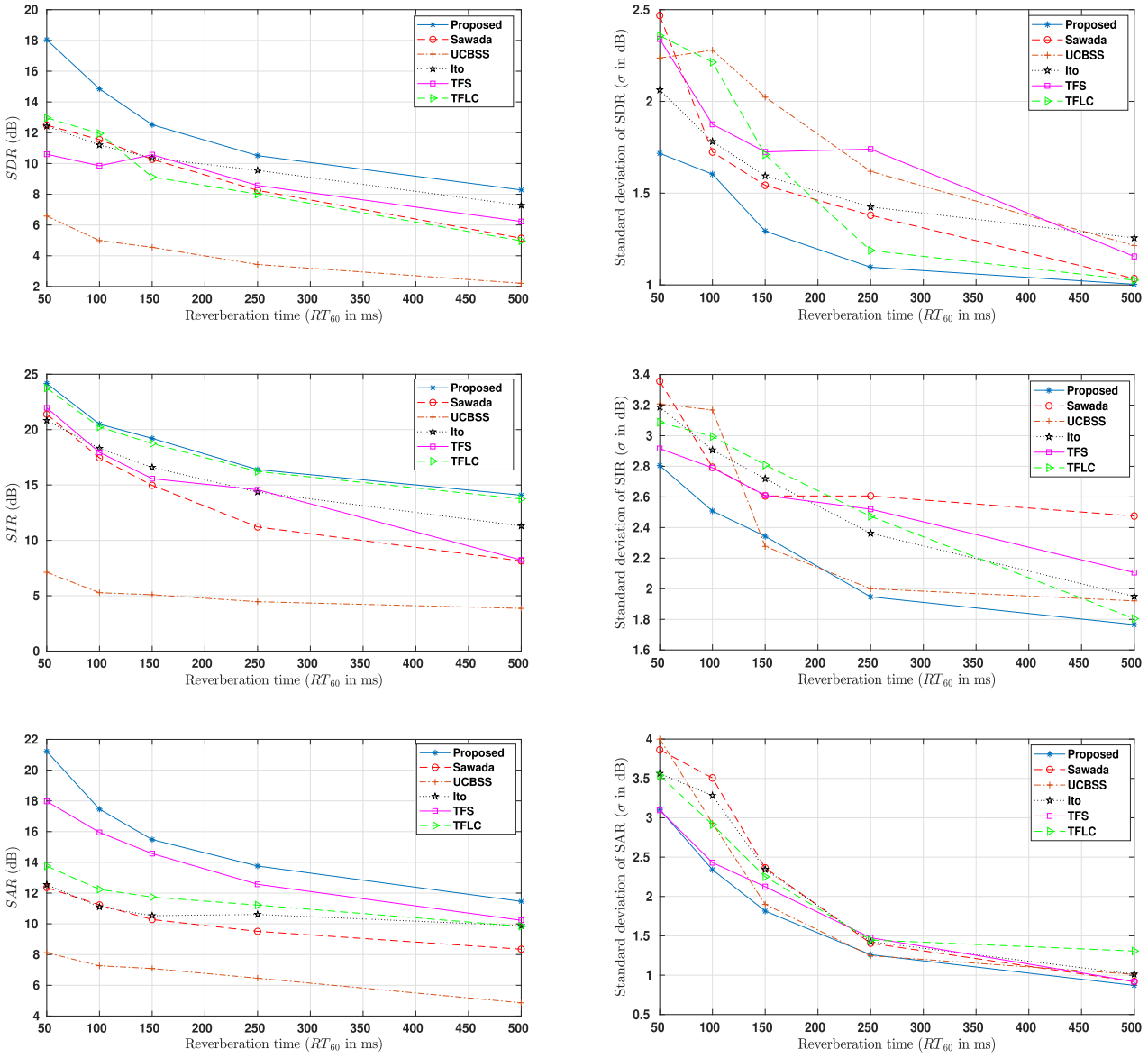


FIGURE 3. Mean (on the left) and standard deviation (on the right) of SDR, SIR and SAR in the determined case as a function of RT_{60}.

From Fig. 3, it can be observed that our method outperforms the other five reference methods (*TFLC*, *Sawada*, *UCBSS*, *Ito*, and *TFS*) in the determined case, regardless of the reverberation time. Indeed, for example, for $RT_{60} = 50$ ms, our method surpasses all the reference methods by more than 5.07 dB, 0.41 dB and 3.25 dB in terms of $\overline{\text{SDR}}$, $\overline{\text{SIR}}$ and $\overline{\text{SAR}}$, respectively. Furthermore, the standard deviation of the proposed method is generally lower than those of the reference methods. We observe, in particular the superiority of our method in terms of $\overline{\text{SAR}}$, which shows that our method is the most effective in terms of artifact reduction. We note that the performances of all reference methods degrade as RT_{60} increases, particularly when it exceeds 250 ms. In fact, the separation performance of all reference methods becomes poor, while our method continues to perform well. Finally, our experiments demonstrate that our method outperforms the *TFLC* method, which requires prior knowledge of the RTFs of the sources. This confirms the superiority of our method, particularly since it continues to provide good performance even without prior information about these RTFs.

D. SEPARATION RESULTS IN THE UNDERDETERMINED CASE

In this second experiment, we are interested at first in the behavior of our method as a function of RT_{60} in the underdetermined case, where $M = 2$ and $N = 3$. The inter-microphone and microphone-source distances are respectively fixed at $d = 1$ m and $D = 1$ m. The other parameters are similar to those of the first experiment.

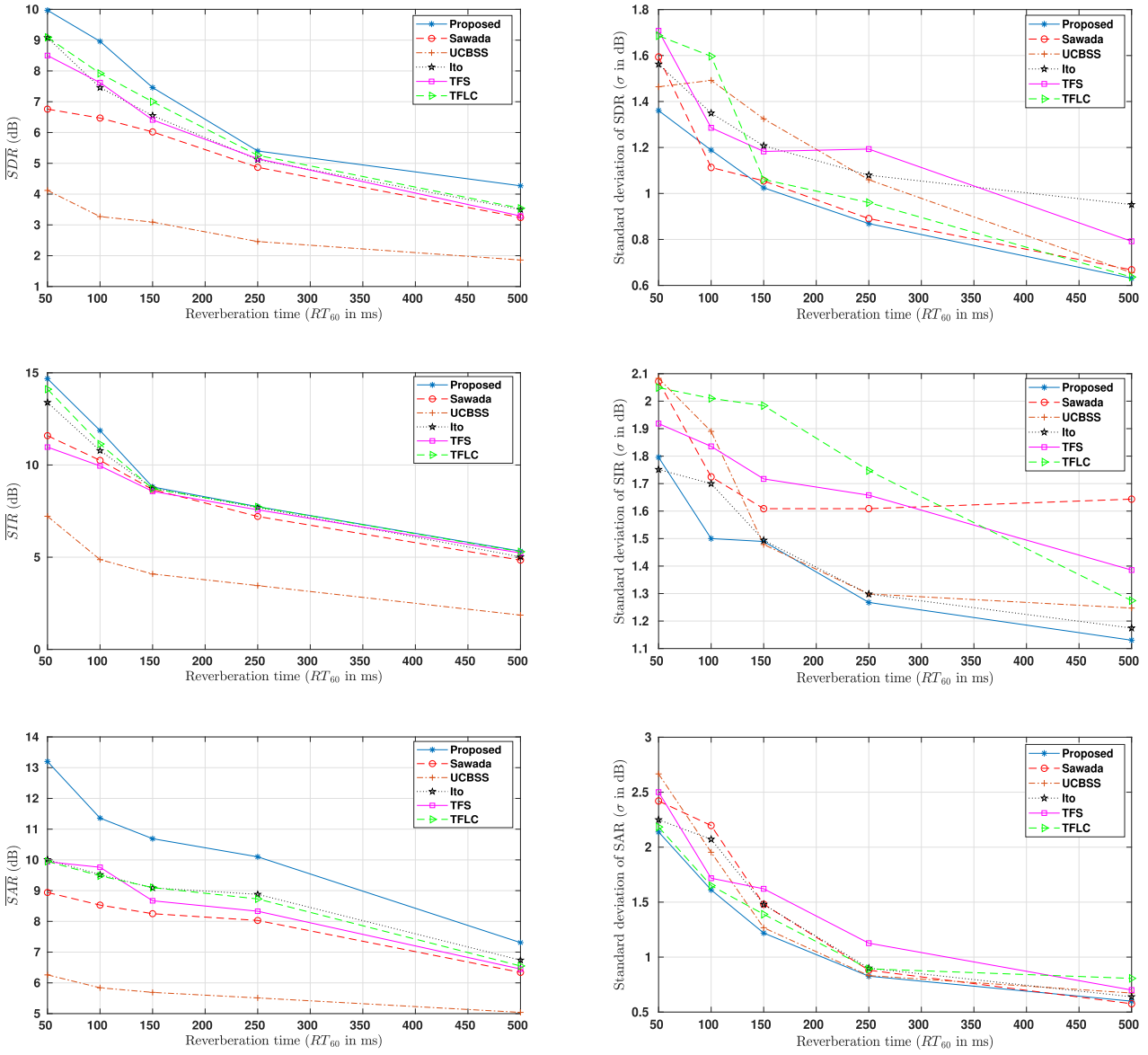


FIGURE 4. Mean (on the left) and standard deviation (on the right) of SDR, SIR and SAR in the underdetermined case as a function of RT_{60}.

Fig. 4 represents the performance obtained in this second experiment. From this figure, we can see again that in the underdetermined case, the best performances are obtained using our method regardless of the RT_{60} value. Indeed, our method shows improvements of more than 0.88 dB, 0.57 dB and 3.19 dB in terms of $\overline{\text{SDR}}$, $\overline{\text{SIR}}$ and $\overline{\text{SAR}}$, respectively, compared to the reference methods, for $RT_{60} = 50$ ms. Furthermore, the standard deviation of our method is usually lower than that of the reference methods. However, it is important to note that all methods show a decrease in performance compared to the determined case. This observation is expected and can be explained by the fact that as the number of sources increases, the sparsity assumption made by these methods on the sources becomes less accurate. We note again that there is a large difference in terms of $\overline{\text{SAR}}$ between our method and the reference methods, confirming that the artifacts introduced by our method are far less significant than those introduced by the reference methods. Furthermore, our method significantly outperforms the *TFS* method, even though both methods combine TF masking and beamforming. This difference in performance can mostly be attributed to the fact that our method uses probabilistic (soft) TF masks, whereas the *TFS* method uses binary TF masks that are known to generate more artifacts in the separated sources.

Then, we are interested in the behavior of our method as a function of $\delta\varphi$, d, and D in this underdetermined case while fixing $RT_{60} = 100$ ms. Fig. 5 shows the performance in terms of SDR obtained in this experiment. From this figure, it can be observed that the performance of our method remains stable as a function of $\delta\varphi$ and d, indicating that our method

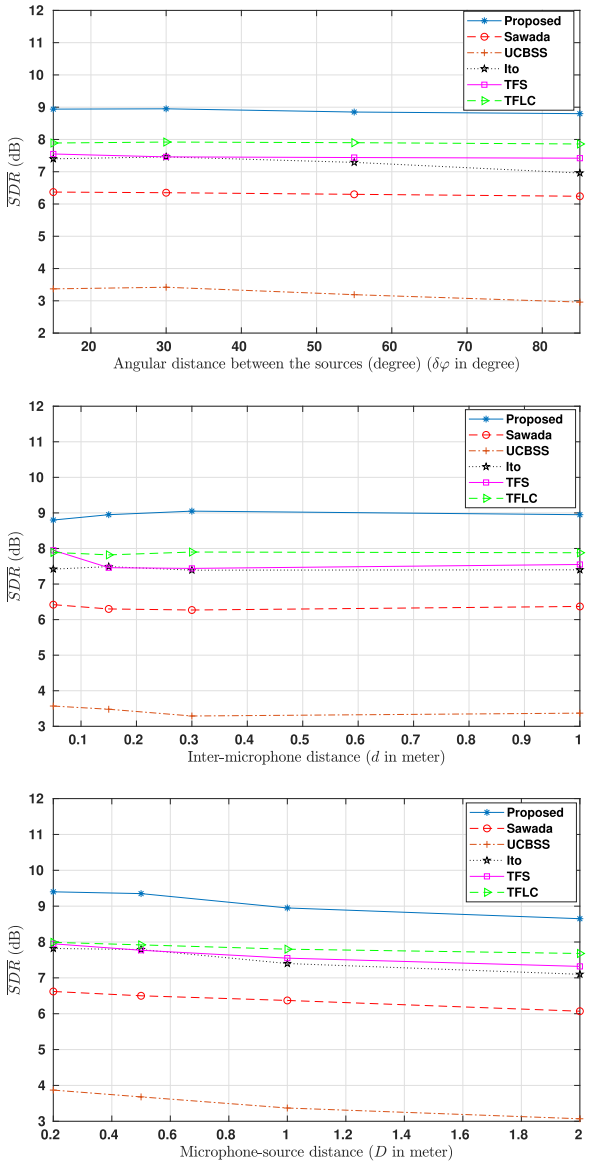


FIGURE 5. Mean of SDR in the underdetermined case as a function of $\delta\varphi$, d and D.

is insensitive to changes in these two parameters. However, we notice a slight degradation in the performance of both the proposed and the reference methods as the microphone-source distance D increases.

Finally, despite the encouraging performance of the proposed method, its computation time is almost comparable to that of the *Ito* method and is at least twice as slow as that of the other reference methods.

E. SEPARATION RESULTS OF THE SISEC DATA

This third experiment focuses on evaluating the performance of our method in the case of real[11] underdetermined mixtures from the SiSEC database [48], where $M = 2$ and $N = 3$, to assess our method's effectiveness in real-world environments. To accomplish this, we utilized some sets from this database that correspond to mixtures of female and male speech signals. In this experiment, we consider two reverberation times $RT_{60} = 130$ ms and $RT_{60} = 250$ ms and two inter-microphone distances $d = 5$ cm and $d = 1$ m. Table 1 presents the average performance obtained in this third experiment in terms of SDR, where each value in this table represents the average value of SDR on six different mixtures.[12]

TABLE 1. $\overline{\text{SDR}}$ in dB of the SiSEC data.

	$RT_{60} = 130$ ms			$RT_{60} = 250$ ms		
Method	Male	Female	SDR	Male	Female	SDR
Sawada [7]	5.28	6.16	5.72	4.76	4.09	4.43
UCBSS [14]	2.23	2.65	2.44	1.43	1.62	1.53
Ito [20]	6.43	6.87	6.64	4.72	5.74	5.23
TFS [42]	4.37	6.20	5.29	3.15	4.36	3.75
TFLC [43]	4.28	6.78	5.53	3.65	5.37	4.51
Proposed	6.91	7.59	**7.25**	5.19	6.10	**5.65**

From Table 1, it can be observed that our method achieves the best performances in terms of $\overline{\text{SDR}}$ for both male and female speech voices. Indeed, our method outperforms the reference methods Sawada [7], UCBSS [14], Ito [20], TFS [42], and TFLC [43] by approximately 1.53 dB, 4.81 dB, 0.61 dB, 1.96 dB, and 1.72 dB respectively, for $RT_{60} = 130$ ms. Similarly, for $RT_{60} = 250$ ms, our method achieves an improvement of approximately 1.22 dB, 4.12 dB, 0.42 dB, 1.90 dB, and 1.14 dB over the same reference methods. These results are consistent with those obtained in the case of underdetermined artificial mixtures.

IV. CONCLUSION AND PERSPECTIVES

In this article, we proposed a new BSS method for convolutive mixtures that can be underdetermined, combining TF masking and beamforming. By combining these two techniques, we aim to minimize the artifacts that impact the signals separated by most BSS methods that use only TF masking. Our method differs from existing methods [42], [43] in that it blindly estimates the RTFs of the sources, eliminating the need for any prior information about them. Additionally, we introduced a new technique that extends the LCMP beamformer, allowing us to obtain undistorted spatial images of the separated sources even in the underdetermined case. According to the results of the tests conducted in this article, the proposed method outperforms the methods proposed in [7], [14], [20], [42], and [43] in terms of SDR, SIR, and SAR, both in the determined case ($M = N = 2$) and the underdetermined case ($M = 2, N = 3$). Furthermore, these results were validated by testing the considered methods on real underdetermined mixtures.

[11] The sources are played through loudspeakers placed in a meeting room. Each source is recorded separately using a pair of omnidirectional microphones, and the resulting recordings are then combined to create the mixture.

[12] We averaged the two values of d and three different realizations of source signals, i.e., six mixtures in total.

Regarding perspectives, it would be interesting to propose a technique for estimating the number of sources N, as our method assumes this information is known, which may only sometimes be the case.

APPENDIX A
DERIVATION OF THE MAXIMIZATION STEP UPDATE EQUATIONS

In this appendix, we present the derivation of the equations for updating the mixing model parameters θ used in the Maximization step in Section II-A1. Indeed, these equations are obtained by maximizing the auxiliary function $Q(\theta, \theta')$ defined in (38) according to each parameter in θ and using the posterior probabilities $\gamma_j(n)$ obtained in the *expectation step* of Section II-A1:

$$\begin{aligned} Q(\theta,\theta') &= \sum_{n=0}^{T-1}\sum_{j=1}^{N}\gamma_j(n)\log(\alpha_j p(X(n)|j,\phi_j(n)B_j)) \\ &\quad + \log(p(\theta)) \\ &= \sum_{j=1}^{N}(\sum_{n=0}^{T-1}(\gamma_j(n)) + \beta - 1)\log(\alpha_j) \\ &\quad - M\sum_{j=1}^{N}\sum_{n=0}^{T-1}\gamma_j(n)\log(\phi_j(n)) \\ &\quad - \sum_{j=1}^{N}\sum_{n=0}^{T-1}(\gamma_j(n))\log(\det(B_j)) \\ &\quad - \sum_{n=0}^{T-1}\sum_{j=1}^{N}\left(\frac{\gamma_j(n)}{\phi_j(n)}X^H(n)B_j^{-1}X(n)\right) + \mathrm{C}, \end{aligned} \tag{38}$$

where C is a constant independent of θ.

The equation for updating α_j is obtained using the Lagrange multiplier method, with the constraint $\sum_{j=1}^{N}\alpha_j = 1$. Consider the function:

$$L(\alpha_j,\lambda) = Q(\theta,\theta') + \lambda(\sum_{j=1}^{N}\alpha_j - 1). \tag{39}$$

We calculate the partial derivative of (39) with respect to α_j, we get:

$$\alpha_j = \frac{\sum_{n=0}^{T-1}(\gamma_j(n)) + \beta - 1}{(-\lambda)}. \tag{40}$$

By summing up Eq. (40) for $j = 1, \ldots, N$, we get:

$$\lambda = -(T + N(\beta - 1)) \tag{41}$$

From (40) and (41), we find:

$$\alpha_j = \frac{\sum_{n=0}^{T-1}\gamma_j(n) + \beta - 1}{T + N(\beta - 1)}. \tag{42}$$

As for the updating equation of $\phi_j(n)$, we compute the partial derivative of (38) with respect to ϕ_j, we then obtain:

$$-M\frac{\gamma_j(n)}{\phi_j(n)} + \frac{\gamma_j(n)}{\phi_j^2(n)}X^H(n)B_j^{-1}X(n) = 0, \tag{43}$$

which gives us the equation for updating $\phi_j(n)$ as follows:

$$\phi_j(n) = \frac{1}{M}X^H(n)B_j^{-1}X(n). \tag{44}$$

As for the update equation of B_j, we compute the partial derivative of (38) with respect to B_j, using some matrix algebra properties we obtain:

$$-(\sum_{n=0}^{T-1}\gamma_j(n))B_j^{-1} + B_j^{-1}(\sum_{n=0}^{T-1}\gamma_j(n)X(n)X^H(n)/\phi_j(n))B_j^{-1} = 0, \tag{45}$$

which gives us the equation for updating B_j as follows:

$$B_j = \frac{\sum_{n=0}^{T-1}\gamma_j(n)X(n)X^H(n)/\phi_j(n)}{\sum_{n=0}^{T-1}\gamma_j(n)}. \tag{46}$$

REFERENCES

[1] M. Bella, H. Saylani, S. Hosseini, and Y. Deville, "Bin-wise combination of time-frequency masking and beamforming for convolutive source separation," in *Proc. Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2022, pp. 741–747.

[2] S. Makino, T. Lee, and H. Sawada, *Blind Speech Separation*. Dordrecht, The Netherlands: Springer, 2007.

[3] M. Bella and H. Saylani, "A new sparse blind source separation method for determined linear convolutive mixtures in time-frequency domain," in *Proc. Int. Conf. Image Signal Process.*, in Lecture Notes in Computer Science, A. E. Moataz, D. Mammass, A. Mansouri, and F. Nouboud, Eds. Cham, Switzerland: Springer, Jun. 2020, pp. 357–366.

[4] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, Jun. 2000, pp. 2985–2988.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, Aug. 2007.

[6] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, *A Survey of Convolutive Blind Source Separation Methods* (Springer Handbook of Speech Processing). Cham, Switzerland: Springer, Nov. 2007.

[7] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[8] Y.-Q. Fu, H. Guo, D.-L. Su, and Y. Liu, "Application of underdetermined blind source separation in ultra-wideband communication signals," *J. China Universities Posts Telecommun.*, vol. 20, no. 3, pp. 13–19, Jun. 2013.

[9] S. Hosseini, Y. Deville, and H. Saylani, "Blind separation of linear instantaneous mixtures of non-stationary signals in the frequency domain," *Signal Process.*, vol. 89, no. 5, pp. 819–830, May 2009.

[10] V. Zarzoso and A. K. Nandi, "Noninvasive fetal electrocardiogram extraction: Blind separation versus adaptive noise cancellation," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 1, pp. 12–18, Jan. 2001.

[11] R. Vigario, J. Sarela, V. Jousmiki, M. Hamalainen, and E. Oja, "Independent component approach to the analysis of EEG and MEG recordings," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 5, pp. 589–593, May 2000.

[12] M. Bella, S. Hosseini, H. Saylani, T. Contini, T. Grégoire, A. Guerrero, and Y. Deville, "Decontamination of galaxy spectra using four dispersion directions," in *Proc. 30th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2022, pp. 1981–1985.

[13] S. Hosseini, A. Selloum, T. Contini, and Y. Deville, "Separation of galaxy spectra measured with slitless spectroscopy," *Digit. Signal Process.*, vol. 106, Nov. 2020, Art. no. 102837.

[14] 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.*, vol. 18, no. 1, pp. 101–116, Jan. 2010.

[15] A. Alinaghi, P. J. Jackson, Q. Liu, and W. Wang, "Joint mixing vector and binaural model based stereo source separation," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 22, no. 9, pp. 1434–1448, Sep. 2014.

[16] J. Marcel, "Blind signal separation of convolutive mixtures: A time-domain joint-diagonalization approach," in *Independent Component Analysis and Blind Signal Separation. ICA 2004*, vol. 3195, A. Prieto and C. G. Puntonet, Eds. Berlin, Germany: Springer, 2004, pp. 578–585.

[17] H. Saylani, S. Hosseini, and Y. Deville, "Blind separation of convolutive mixtures of non-stationary and temporally uncorrelated sources based on joint diagonalization," in *Proc. Image Signal Process. 5th Int. Conf. (ICISP)*, Agadir, Morocco, Jun. 2012, pp. 191–199.

[18] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2007, pp. 139–142.

[19] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3238–3242.

[20] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. 14th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2014, pp. 268–272.

[21] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[22] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. 10th Int. Conf. Inf. Sci., Signal Process. Appl. (ISSPA)*, May 2010, pp. 1–4.

[23] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for joint blind source separation and noise reduction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 109–112.

[24] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 241–244.

[25] F. Abrard and Y. Deville, "A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Process.*, vol. 85, no. 7, pp. 1389–1403, 2005.

[26] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 121–133, Jan. 2010.

[27] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.

[28] H. Saylani, S. Hosseini, and Y. Deville, "Blind separation of convolutive mixtures of non-stationary sources using joint block diagonalization in the frequency domain," in *Latent Variable Analysis and Signal Separation*. Berlin, Germany: Springer, 2010, pp. 97–105.

[29] N. Ito, S. Araki, and T. Nakatani, "Modeling audio directional statistics using a complex Bingham mixture model for blind source extraction from diffuse noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 465–468.

[30] N. Fan, J. Du, and L.-R. Dai, "A regression approach to binaural speech segregation via deep neural network," in *Proc. 10th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Oct. 2016, pp. 1–5.

[31] N. Furnon, R. Serizel, S. Essid, and I. Illina, "DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 29, pp. 2310–2323, 2021.

[32] M. Kumar and V. E. Jayanthi, "Underdetermined blind source separation using CapsNet," *Soft Comput.*, vol. 24, no. 12, pp. 9011–9019, Jun. 2020.

[33] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 196–200.

[34] 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.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.

[35] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2007, pp. 3247–3250.

[36] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Audio, Speech Lang., Process.*, vol. 15, no. 5, pp. 1592–1604, Jul. 2007.

[37] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 11, Oct. 2003, Art. no. 569270.

[38] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[39] H. L. Van Trees, *Adaptive Beamformers*. Hoboken, NJ, USA: Wiley, 2002, ch. 7, pp. 710–916.

[40] Z. Tian, K. L. Bell, and H. L. Van Trees, "A recursive least squares implementation for LCMP beamforming under quadratic constraint," *IEEE Trans. Signal Process.*, vol. 49, no. 6, pp. 1138–1145, Jun. 2001.

[41] J. Cermak, S. Araki, H. Sawada, and S. Makino, "Blind source separation based on a beamformer array and time frequency binary masking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2007, pp. 145–148.

[42] K. Yamaoka, A. Brendel, N. Ono, S. Makino, M. Buerger, T. Yamada, and W. Kellermann, "Time-frequency-bin-wise beamformer selection and masking for speech enhancement in underdetermined noisy scenarios," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 1582–1586.

[43] K. Yamaoka, N. Ono, and S. Makino, "Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 29, pp. 3461–3475, 2021.

[44] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 25, no. 4, pp. 780–793, Apr. 2017.

[45] N. Ito, S. Araki, and T. Nakatani, "Recent advances in multichannel source separation and denoising based on source sparseness," in *Audio Source Separation*. Cham, Switzerland: Springer, 2018, pp. 279–300.

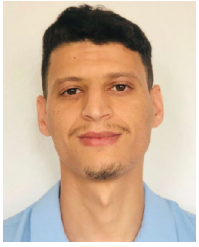
[46] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. USA: Kluwer Academic Publishers, 1981.

[47] Y. Xie, K. Xie, J. Yang, Z. Wu, and S. Xie, "Underdetermined reverberant audio-source separation through improved expectation–maximization algorithm," *Circuits, Syst., Signal Process.*, vol. 38, no. 6, pp. 2877–2889, Jun. 2019.

[48] (2011). *Sisec2011*. [Online]. Available: http://sisec2011.wiki.irisa.fr/tiki-indexbfd7.html?page

[49] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 77–269, 2008.

[50] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech Lang., Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.



MOSTAFA BELLA received the master's degree (Hons.) in electronics (telecommunications) in 2018. He is currently pursuing the Ph.D. degree. He is a member of the IRAP Laboratory, University of Toulouse, France, and the MatSim Laboratory, University Ibnou Zohr, Morocco. His current research interests include signal and image processing, blind source separation, and adaptive signal processing, with a particular focus on their applications in audio processing and astrophysics. He was awarded the prestigious Eiffel Excellence Grant from the Campus France and the Moroccan Excellence Grant from the National Centre for Scientific and Technical Research of Morocco.

HICHAM SAYLANI was born in Kenitra, Morocco, in 1969. He received the Ph.D. degree in signal processing from Ibn Tofail University, Kenitra, in 2009. Since 2011, he has been a Teaching Researcher with the Faculty of Sciences of Agadir, Ibnou Zohr University, Morocco. He is a member of the Laboratory of Materials, Signals, Systems, and Physical Modeling. His current research interests include audio processing, biomedical signal processing, and signal and image processing, with a particular interest in source separation and its various applications.

SHAHRAM HOSSEINI was born in Shiraz, Iran, in 1968. He received the B.Sc. and M.Sc. degrees in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 1991 and 1993, respectively, and the Ph.D. degree in signal processing from Institut National Polytechnique, Grenoble, France, in 2000. He is currently an Associate Professor with the University of Toulouse, France. Since 2004, he has been with the Astrophysics Laboratory, which is part of the University of Toulouse and the French National Center for Scientific Research (CNRS). His current research interests include blind source separation, artificial neural networks, and adaptive signal processing.

YANNICK DEVILLE (Member, IEEE) was born in Lyon, France, in 1964. He received the degree from Ecole Nationale Supérieure des Télécommunications de Bretagne, Brest, France, in 1986, and the D.E.A. and Ph.D. degrees in microelectronics from the University of Grenoble Alpes, France, in 1986 and 1989, respectively. From 1986 to 1997, he was a Research Scientist with Philips Research Labs, Limeil, France. His investigations during that period concerned various fields, including GaAs integrated microwave RC active filters, VLSI cache memory architectures and replacement algorithms, neural network algorithms and applications, and nonlinear systems. Since 1997, he has been a Professor with the University of Toulouse, France. From 1997 to 2004, he was with the Acoustics Laboratory, University of Toulouse. Since 2004, he has been with the Astrophysics Laboratory, which is part of the University of Toulouse and the French National Center for Scientific Research (CNRS). His current research interests include signal and image processing (especially hyperspectral data), higher-order statistics, time-frequency analysis, neural networks, quantum entanglement phenomena, and especially blind source separation and blind identification methods (including independent or sparse component analysis and non-negative matrix factorization) and their applications to remote sensing, astrophysics, and quantum information processing.