Multichannel Speech Separation and Enhancement Using the Convolutive Transfer Function

Xiaofei Li, Laurent Girin, Sharon Gannot, *Senior Member, IEEE*, and Radu Horaud

***Abstract*—This paper addresses the problem of speech separation and enhancement from multichannel convolutive and noisy mixtures, *assuming known mixing filters*. We propose to perform speech separation and enhancement in the short-time Fourier transform domain using the convolutive transfer function (CTF) approximation. Compared to time-domain filters, the CTF has much less taps. Consequently, it requires less computational cost and sometimes is more robust against the filter perturbations. We propose three methods: 1) for the multisource case, the multichannel inverse filtering method, i.e., the multiple input/output inverse theorem (MINT), is exploited in the CTF domain; 2) a beamforming-like multichannel inverse filtering method applying the single-source MINT and using power minimization, which is suitable whenever the source CTFs are not all known; and 3) a basis pursuit method, where the sources are recovered by minimizing their ℓ_1-norm to impose spectral sparsity, while the ℓ_2-norm fitting cost between microphone signals and mixing model is constrained to be lower than a tolerance. The noise can be reduced by setting this tolerance at the noise power level. Experiments under various acoustic conditions are carried out to evaluate and compare the three proposed methods. Comparison with four baseline methods—beamforming-based, two time-domain inverse filters, and time-domain Lasso—shows the applicability of the proposed methods.**

***Index Terms*—Audio source separation, speech enhancement, short-time Fourier transform, convolutive transfer function, MINT, Lasso optimization.**

I. Introduction

SPEECH recordings in the real world consist of the sum of convolutive images of multiple audio sources plus some additive noise. A convolutive image is the convolution between the source signal and the room impulse response (RIR), which is also called the mixing filter in the multisource context. Interfering speakers, reverberation and additive noise heavily deteriorate the intelligibility of a target speech source for both human listening and machine recognition. This work aims to suppress these distortions, i.e., recover speech source signals from multichannel recordings. In general, suppressing interfering speakers, reverberation and noise are respectively referred to as source separation, dereverberation and noise reduction. Each of these difficult tasks has attracted a lot of research attention. In the microphone recordings, there are three unknown terms, i.e., source signals, mixing filters, and noise. Thence, the problem is often split into two subproblems i) identification of the mixing filters and noise statistics, and ii) estimation of the source signals. This work focuses on the problem of speech source estimation assuming that the mixing filters, and possibly the noise statistics, are known or their estimates are available.

Most convolutive source separation and speech enhancement techniques are designed in the short-time Fourier transform (STFT) domain. In this domain, the convolutive process is usually approximated at each time-frequency (TF) bin by a product between the source STFT coefficient and the Fourier transform of the mixing filter, called the acoustic transfer function (ATF). This assumption is called the multiplicative transfer function (MTF) approximation [1], or the narrowband approximation. Assuming known (or accurately estimated) ATFs or relative transfer functions (RTFs) [2], [3], beamforming techniques are widely used for multichannel source separation and speech enhancement. Popular beamformers include minimum variance/power distortionless response (MVDR/MPDR), and linearly constrained minimum variance/power (LCMV/LCMP) [2], [4]. Alternately, methods based on binary masking [5], [6] and ℓ_1-norm minimization [7] exploit the natural sparsity of audio signals in the TF domain. Many examples of MTF-based techniques can be found in [8] and references therein.

The narrowband assumption is theoretically valid only if the length of the mixing filters is small compared to the length of the STFT window. In practice, this is very rarely the case, even for moderately reverberant environments, since the STFT window length is limited to assume local stationarity of audio signals. Hence the narrowband assumption fundamentally hampers the speech enhancement/separation performance, and this becomes critical for strongly reverberant environments. To avoid this limitation, several source separation methods based on the time-domain representation of mixing filters have been proposed. In the wide-band Lasso method [9], the source signals are estimated by minimizing the ℓ_2-norm fitting cost between microphone signals and mixing model, in which the exact time-domain source-filter convolution is used. Importantly, the ℓ_1-norm of the vector representing the source signals in the

Manuscript received February 27, 2018; revised August 10, 2018, October 23, 2018, and December 11, 2018; accepted December 27, 2018. Date of publication January 11, 2019; date of current version January 25, 2019. This work was supported by the European Research Council Advanced Grant 340113 (project: Vision and Hearing in Action). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Simon Doclo. *(Corresponding author: Xiaofei Li.)*

X. Li and R. Horaud are with the INRIA Grenoble Rhône-Alpes, 38330 Montbonnot-Saint-Martin, France (e-mail: xiaofei.li@inria.fr; radu.horaud@inria.fr).

L. Girin is with the Univ. Grenoble Alpes, Grenoble-INP, GIPSA-lab, 38400 Saint-Martin-d'Hères, France, and also with INRIA Grenoble Rhône-Alpes, 38330 Montbonnot-Saint-Martin, France (e-mail: laurent.girin@gipsa-lab.grenoble-inp.fr).

S. Gannot is with the Faculty of Engineering, Bar Ilan University, Ramat Gan 5290002, Israel (e-mail: Sharon.Gannot@biu.ac.il).

Digital Object Identifier 10.1109/TASLP.2019.2892412

STFT-domain is added to the fitting cost as a regularization term to impose sparsity of the source spectra. In the presence of additive noise, the ℓ_1-norm regularization is able to reduce the noise in the recovered source signals. However, the regularization factor is difficult to set even if the noise power is known. To overcome this, a more flexible scheme is proposed in [10] that relaxes the ℓ_2-norm fitting cost to the noise level and minimizes the ℓ_1-norm of the STFT-domain source signals.

In the family of multichannel inverse filtering or multichannel equalization methods, an inverse filter is estimated with respect to the known mixing filters, and applied to the microphone signals, preserving the desired source and suppressing the interfering sources. The multiple-input/output inverse theorem (MINT) method was proposed in [11] for this aim. However it is sensitive to RIR perturbations (misalignment and estimation error) and to additive noise. To improve the robustness of MINT to RIR perturbations, many techniques have been proposed, preserving not only the direct-path impulse response but also the early reflections, such as channel shortening [12], infinity- and p-norm optimization-based channel shortening/reshaping [13], or partial MINT [14], [15]. In addition, the energy of the inverse filter was used in [16] as a regularization term to avoid the amplification of filter perturbations and noise. In the context of speech enhancement the above-mentioned improved MINT methods are proposed for single source dereverberation, while the multisource case has been rarely studied for source separation. In [17], a two-stage method was proposed, that first converts a multiple-input multiple-output (MIMO) system into a set of single-input multiple-output (SIMO) systems for source separation, and then applies inverse filtering for dereverberation. The multiple-input/output inverse filtering method has also been applied to virtual sound reproduction [18]: the inverse filters of room acoustics are applied as prefilters to the sound clips/files played by loudspeakers, so as to ensure that the respective designated signal reaches the corresponding microphone (listener). For each microphone, the compensation of listening room has an identical formulation with the dereverberation problem. Furthermore, the cancellation of the undesired crosstalk from the loudspeakers by the prefilters has an identical formulation with the multisource separation problem.

The wide-band models mentioned above are all performed in the time domain. The time-domain convolution problem can be transformed into the subband domain, which provides several benefits: i) The original problem is split into subproblems, and each subproblem has a smaller data size and thus a smaller computational complexity; ii) For multichannel equalization, it is demonstrated in [19]–[21] that, compared to the time-domain method, the subband method is less sensitive to filter perturbations and thus achieves better performance measures. As stated in [19], [21], the reason for this phenomenon is associated with numerical problems that appear as a result of inverting large and poorly conditioned matrices in the time domain; this can be mitigated by using shorter subband filters; and iii) In the TF domain, the sparsity of speech signals can be more easily exploited. Several variants of subband MINT were proposed based on filter banks [19]–[23]. In this context, the key issues in the filter-bank design are: i) The time-domain RIRs should be well approximated in the subband domain, and ii) The frequency response of each filter should be quasi-flat, namely it should not contain frequency components with magnitude close to zero. Otherwise, these components are common to all channels, and are problematic in the MINT application. To satisfy the second condition, the filter-bank is either critically sampled [19], [22], which suffers from frequency aliasing, or has a flat-top frequency response [20], [21], [23], which may suffer from time aliasing. Generally speaking, the STFT is the most widely-used subband decomposition because most of the audio processing algorithms are performed in this domain. To represent the time-domain convolution in the STFT domain, especially for the long filter case, cross-band filters were introduced in [24]. To simplify the analysis, the convolutive transfer function (CTF) approximation is further adopted in [25], [26] only using the band-to-band convolution and ignoring the cross-band filters. In [26], the CTF is integrated into the generalized sidelobe canceler (GSC) beamformer. In [27], a CTF-domain Lasso method was proposed, following the spirit of the wide-band Lasso [9].

Several probabilistic techniques have also been proposed for wide-band source separation via maximizing the likelihood of a generative model. Variational Expectation-Maximization (EM) algorithms are proposed in [28] and [29] based on the time-domain convolution and in [30] based on cross-band filters. CTF-based EM algorithms are proposed in [31] and [32] for single-source dereverberation and source separation, respectively. These EM algorithms iteratively estimate the mixing filters and the sources, and intrinsically require a fairly good initialization for both filters and sources.

In the present paper, we propose the following three source recovery methods, all working in the standard (oversampled) STFT domain and using the CTF approximation:

- We propose a CTF-based multisource MINT method for both source separation and dereverberation. The oversampled STFT does not suffer from both frequency aliasing and time aliasing. However, the STFT window is not flat-top, namely the subband signals and filters have a frequency region with magnitude close to zero, which is common to all channels. To overcome this problem, instead of using the conventional impulse function as the target of inverse filtering, we propose a new target, which has a frequency response corresponding to the STFT window. In addition, a filter energy regularization is adopted following [16] to improve the robustness of inverse filtering. This method is an extension of our previous work presented in [33].
- For situations where the CTFs of the sources are not all available, we propose a beamforming-like inverse filtering method. The inverse filters are designed i) to preserve one source with known CTF based on single-source MINT, and ii) to minimize the overall power of the inverse filtering output, and thus suppress interfering sources and noise. This method shares a similar spirit with the MPDR beamformer.
- To overcome the drawback of the CTF-Lasso method [27], namely that the regularization factor is difficult to set with

respect to the noise level, following the spirit of [10], we propose a CTF-based basis pursuit method: recovering the source signals by minimizing the ℓ_1-norm of the STFT-domain source vector with the constraint that the ℓ_2-norm fitting cost is lower than a tolerance. The effect of the tolerance setting is studied. In addition, a complex-valued *proximal splitting* algorithm [34], [35] is investigated to solve the optimization problem.

The remainder of this paper is organized as follows. The problem is formulated based on CTF in Section II. The two multichannel inverse filtering methods are proposed in Section III. The CTF-based basis pursuit method is proposed in Section IV. Experiments are presented in Section V. Section VI concludes the work.

II. CTF-Based Problem Formulation

In this section, we first present the STFT-domain CTF convolution as an approximation of the time-domain convolution. Then we formulate the multichannel source separation problem, first in the time domain, and then in the STFT domain based on the CTF approximation.

A. Convolutive Transfer Function

In a room, the microphone signal can be modeled as: $x(n) = a(n) \star s(n)$, where n is the time index, $x(n)$, $s(n)$ and $a(n)$ are the microphone signal, the source signal and the RIR, respectively, and $\star$ denotes convolution. The STFT representation of the microphone signal $x(n)$ is

$$x_{p,k} = \sum_{n=-\infty}^{+\infty} x(n)\tilde{w}(n-pD)e^{-j\frac{2\pi}{N}k(n-pD)}, \tag{1}$$

where p and k denote the frame index and the frequency index, respectively. $\tilde{w}(n)$ is the STFT analysis window, and N and D denote the frame (window) length and the frame step, respectively. In the filter bank interpretation, the analysis window is considered as the low-pass filter, and D as the decimation factor.

The cross-band filter model [24] consists in representing the STFT coefficient $x_{p,k}$ as a summation over multiple convolutions (between the STFT-domain source signal $s_{p,k}$ and filter $a_{p,k,k'}$) across frequency bins. Formally, the linear time-invariant system can be written in the STFT domain as

$$x_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'} a_{p',k,k'}\ s_{p-p',k'}. \tag{2}$$

If $D < N$, then $a_{p',k,k'}$ is non-causal, with $\lceil N/D \rceil - 1$ non-causal coefficients, where $\lceil \cdot \rceil$ denotes the ceiling function. The number of causal filter coefficients is related to the reverberation time. For notational simplicity, let the frame index of the filters be in $[0, L_a - 1]$, with L_a being the filter length. The non-causal coefficients are shifted to the causal part, which only leads to a constant shift of the frame index of the source signal. Let $w(n)$ denote the STFT synthesis window. The STFT-domain impulse response $a_{p',k,k'}$ is related to the time-domain impulse response $a(n)$ by:

$$a_{p',k,k'} = (a(n) \star \zeta_{k,k'}(n))|_{n=p'D}, \tag{3}$$

which represents the convolution with respect to the time index n evaluated at frame steps, with

$$\zeta_{k,k'}(n) = e^{j\frac{2\pi}{N}k'n} \sum_{m=-\infty}^{+\infty} \tilde{w}(m)\, w(n+m)\, e^{-j\frac{2\pi}{N}m(k-k')}.$$

To simplify the analysis, we consider the CTF approximation, i.e., only band-to-band filters with $k = k'$ are considered, and simplify $a_{p,k,k'}$ as $a_{p,k}$:

$$x_{p,k} \approx \sum_{p'=0}^{L_a-1} a_{p',k} s_{p-p',k} = a_{p,k} \star s_{p,k}. \tag{4}$$

B. Problem Formulation

In the time domain, a multichannel convolutive mixture with J sources and I microphones is expressed as:

$$x^i(n) = \sum_{j=1}^{J} a^{i,j}(n) \star s^j(n) + e^i(n), \tag{5}$$

where $i = 1, \ldots, I,\ I \geq 2$ and $j = 1, \ldots, J,\ J \geq 2$ are the indices of the microphones and of the sources, respectively. The RIR $a^{i,j}(n)$ relates the j-th source to the i-th microphone. The noise signals $e^i(n)$ are uncorrelated with the source signals, and could be spatially uncorrelated, diffuse, or directional. Note that the relation between I and J is not specified here and will be discussed afterwards with respect to the proposed methods. The goal of this work is to recover the multiple source signals $s^j(n)$, given the microphone signals, the RIRs and the noise statistics. The RIRs and noise statistics can be blindly estimated from the microphone signals, which is not trivial but beyond the scope of this work, and the estimated values generally suffer from distortions.

Based on the CTF approximation, we obtain the STFT-domain model corresponding to the time-domain model (5):

$$x_p^i \approx \sum_{j=1}^{J} a_p^{i,j} \star s_p^j + e_p^i. \tag{6}$$

Note that here (and hereafter) the frequency index k is omitted, unless it is necessary, since the proposed methods are applied frequency-wise. Accordingly, the goal is transferred to recover the STFT coefficients of the source signals, i.e., s_p^j, and then applying the inverse STFT to obtain an estimation of the time-domain source signals.

III. CTF-Based Multichannel Inverse Filtering

The multichannel inverse filtering method is based on the MINT method. In this section, we propose two MINT-based methods in the CTF domain for the multisource case.

A. Problem Formulation for Inverse Filtering

Define the CTF-domain inverse filters as h_p^i with $i = 1, \ldots, I$, and with a length of L_h. The output of inverse filtering is:

$$y_p = \sum_{i=1}^{I} h_p^i \star x_p^i = \sum_{j=1}^{J} s_p^j \star \left(\sum_{i=1}^{I} h_p^i \star a_p^{i,j} \right) + \sum_{i=1}^{I} h_p^i \star e_p^i, \tag{7}$$

which comprises the mixture of the inverse filtered sources and the inverse filtered noise. To facilitate the analysis, we denote the convolution in vector form. We define the convolution matrix for the microphone signal $x_p^i,\ p \in [1, P]$ as:

$$\mathbf{X}^i = \begin{bmatrix} x_1^i & 0 & \cdots & 0 \\ x_2^i & x_1^i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ x_P^i & \vdots & \ddots & x_1^i \\ 0 & x_P^i & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & x_P^i \end{bmatrix} \in \mathbb{C}^{(P+L_h-1)\times L_h}, \tag{8}$$

and the vector of filter coefficients h_p^i as:

$$\mathbf{h}^i = [h_0^i, \ldots, h_p^i, \ldots, h_{L_h-1}^i]^\top \in \mathbb{C}^{L_h \times 1},$$

where $^\top$ denotes vector or matrix transpose. Then the convolution $h_p^i \star x_p^i$ can be written as $\mathbf{X}^i \mathbf{h}^i$. The inverse filtering process (7) can be written as:

$$\mathbf{y} = \mathbf{X}\mathbf{h}, \tag{9}$$

with:

$$\mathbf{y} = [y_1, \ldots, y_p, \ldots, y_{P+L_h-1}]^\top \in \mathbb{C}^{(P+L_h-1)\times 1},$$
$$\mathbf{X} = [\mathbf{X}^1, \ldots, \mathbf{X}^i, \ldots, \mathbf{X}^I] \in \mathbb{C}^{(P+L_h-1)\times IL_h},$$
$$\mathbf{h} = [\mathbf{h}^{1\top}, \ldots, \mathbf{h}^{i\top}, \ldots, \mathbf{h}^{I\top}]^\top \in \mathbb{C}^{IL_h \times 1}.$$

Similarly, we define the convolution matrix for the CTF $a_p^{i,j}$ as $\mathbf{A}^{i,j} \in \mathbb{C}^{(L_a+L_h-1)\times L_h}$, and write $h_p^i \star a_p^{i,j}$ as $\mathbf{A}^{i,j}\mathbf{h}^i$. Moreover, we define $\mathbf{A}^j = [\mathbf{A}^{1,j}, \ldots, \mathbf{A}^{i,j}, \ldots, \mathbf{A}^{I,j}] \in \mathbb{C}^{(L_a+L_h-1)\times IL_h}$, and write $\sum_{i=1}^{I} h_p^i \star a_p^{i,j}$ as $\mathbf{A}^j\mathbf{h}$.

B. The CTF-MINT Method

To preserve a desired source, e.g., the j_d-th source, inverse filtering of the CTF filters, i.e., $\sum_{i=1}^{I} h_p^i \star a_p^{i,j_d}$, should target an impulse function d_p with length $L_a + L_h - 1$. To suppress the interfering sources, inverse filtering of the CTF filters of the other sources, i.e., $\sum_{i=1}^{I} h_p^i \star a_p^{i,j\neq j_d}$, should target a zero signal. Let $\mathbf{d}$ denote the vector form of d_p, and $\mathbf{0}$ denote a $(L_a + L_h - 1)$-dimensional zero vector. We define the following I-input J-output MINT equation:

$$\begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{d} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{1,1} & \cdots & \mathbf{A}^{I,1} \\ \vdots & \ddots & \vdots \\ \mathbf{A}^{1,j_d-1} & \cdots & \mathbf{A}^{I,j_d-1} \\ \mathbf{A}^{1,j_d} & \cdots & \mathbf{A}^{I,j_d} \\ \mathbf{A}^{1,j_d+1} & \cdots & \mathbf{A}^{I,j_d+1} \\ \vdots & \ddots & \vdots \\ \mathbf{A}^{1,J} & \cdots & \mathbf{A}^{I,J} \end{bmatrix} \begin{bmatrix} \mathbf{h}^1 \\ \vdots \\ \mathbf{h}^I \end{bmatrix} = \begin{bmatrix} \mathbf{A}^1 \\ \vdots \\ \mathbf{A}^{j_d-1} \\ \mathbf{A}^{j_d} \\ \mathbf{A}^{j_d+1} \\ \vdots \\ \mathbf{A}^J \end{bmatrix} \mathbf{h},$$

which can be rewritten in a compact form as

$$\mathbf{g} = \mathbf{A}\mathbf{h}. \tag{10}$$

From [11], the condition for this equation to have a solution is that the CTFs of the desired source, i.e., $a_p^{i,j_d}, i = 1, \ldots, I$, do not have any common zeros. Under this condition, when the matrix $\mathbf{A} \in \mathbb{C}^{J(L_a+L_h-1)\times IL_h}$ is either square or wide, namely $IL_h \geq J(L_a + L_h - 1)$ and thus $L_h \geq \frac{J(L_a-1)}{I-J}$, (10) has an exact solution, which means an exact inverse filtering can be achieved. This condition implies an over-determined recording system, i.e., $I > J$.

In addition to the near-common zeros originally present in the room impulse responses, the filter banks induced from the STFT windows lead to some structured common zeros. From (3), for any RIR $a^{i,j}(n)$, its CTF (with $k' = k$) is computed as

$$a_{p,k}^{i,j} = (a^{i,j}(n) \star \zeta_k(n))|_{n=pD}, \tag{11}$$

with

$$\zeta_k(n) = e^{j\frac{2\pi}{N}kn} \sum_{m=-\infty}^{+\infty} \tilde{w}(m)\, w(n+m)$$

being the cross-correlation of the analysis window $\tilde{w}(n)$ and the synthesis window $w(n)$ modulated (frequency shifted) by $e^{j\frac{2\pi}{N}kn}$ (note that for the purpose of the present analysis, the frequency index k is reintroduced here). This cross-correlation has a similar frequency response as the windows $\tilde{w}(n)$ and $w(n)$, since it is also a low-pass filter with the same bandwidth denoted by $\bar{\omega}$. The frequency response of $a_{p,k}^{i,j}$ is the frequency response of $a^{i,j}(n)$ multiplied by the frequency response of $\zeta_k(n)$, and then folded by downsampling with a period of $2\pi/D$. To avoid frequency aliasing, the period should not be smaller than the bandwidth $\bar{\omega}$ not to fold the passband of the low-pass filter. For example, in this work, we use the Hamming window, the width of the main lobe is considered as the bandwidth, i.e., $\bar{\omega} = 8\pi/N$. Consequently, we set the constraint $D \leq N/4$. If we consider the magnitude of side lobes to be zero, the frequency response of $a_{p,k}^{i,j}$ can be interpreted as the k-th frequency band of $a^{i,j}(n)$ multiplied by the frequency response of the downsampled $\zeta_k(n)$, i.e., $\zeta_{p,k} = \zeta_k(n)|_{n=pD}$. When $D < N/4$, the frequency response of $\zeta_{p,k}$ involves some side lobes, which have a magnitude close to zero. When $D = N/4$, only the main lobe is involved, and because the magnitude is dramatically decreasing from the center of the main lobe to its margin, the frequency region close to the margin of the main lobe has magnitude close to zero. Fig. 1 depicts an example of CTF (at one STFT frequency bin) and its frequency response (in magnitude). It can

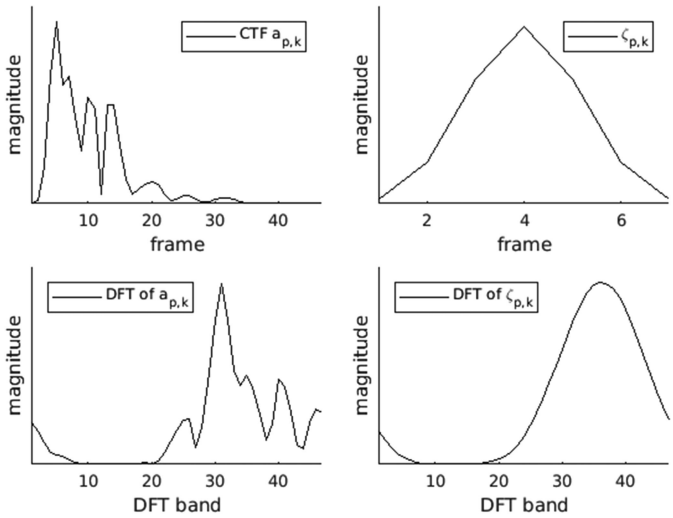

Fig. 1. An example of CTF for a 10,000-sample RIR, at frequency bin $k = 100$, and its frequency response. The STFT window is a 1,024-sample Hamming window, with 256-sample subband filterstep. The length of the CTF is 47 taps. For the detailed dataset and parameter description, please see Section V-A. **Top-left**: CTF magnitude, **bottom-left**: magnitude of the DFT of the CTF, **top-right**: $\zeta_{p,k}$ magnitude, **bottom-right**: magnitude of the DFT of $\zeta_{p,k}$. Note that $\zeta_{p,k}$ is zero-padded when applying DFT, to have the same length as the CTF.

be seen that the frequency response of the CTF is reshaped by the frequency response of $\zeta_{p,k}$, and that they have a same region with magnitude close to zero.

This phenomenon, namely that the frequency response of $\zeta_{p,k}$ and thus of $a_{p,k}^{i,j}$ have a frequency region with magnitude close to zero, is common to all microphones, which is problematic for solving (10). When the impulse function is taken as the target function, the inverse of the CTF frequency region with magnitude close to zero will target a nonzero value, which is obviously an ill-posed problem. Fortunately, we know that the common zeros are introduced by the frequency response of $\zeta_{p,k}$. To make (10) solvable, we propose to set the desired target $\mathbf{d}$ to have the same frequency response as $\zeta_{p,k}$. To this end, $\mathbf{d}$ is designed as:

$$\mathbf{d} = [0, \ldots, 0, \boldsymbol{\zeta}^\top, 0, \ldots, 0]^\top \in \mathbb{C}^{(L_a+L_h-1)\times 1}, \tag{12}$$

where $\boldsymbol{\zeta}$ denotes the vector form of $\zeta_{p,k}$. From the definition of $\zeta_{p,k}$, it is deduced that the length of $\boldsymbol{\zeta}$ is $\lfloor (2N-1)/D \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function, e.g., the length is 7 when $D = N/4$ as shown in Fig. 1. In contrast to the impulse function, with the target function (12), the inverse of the CTF frequency region with magnitude close to zero targets a value close to zero, which is no longer ill-posed. Note that this target function only accounts for the structured common zeros caused by the STFT windows, but not for the near-common zeros originally present in the room impulse responses. The zeros before $\boldsymbol{\zeta}$ introduce a modeling delay. As shown in [16], this delay is important for making the inverse filtering robust to perturbations of the CTF.

As shown in (7), the solution of (10) gives an exact recovery of the j_d-th source plus the filtered noise $\sum_{i=1}^{I} h_p^i \star e_p^i$. In this method, a directional noise can be treated as an interfering source, and can be modeled in the MINT formulation. Therefore, here we only need to consider the case of a spatially uncorrelated or diffuse noise. To suppress the noise, a straightforward way is to minimize the power of the filtered noise under the MINT constraint (10), which requires the knowledge of the noise statistics. As proposed in [16], an alternative way to suppress the noise is to reduce the energy of the inverse filter $\mathbf{h}$. This strategy leads to the minimization of the power of the filtered noise if the noise correlation matrix is assumed to be the identity. In other words, this strategy is the optimal way to suppress a spatially and temporally uncorrelated noise, but is less effective for spatially or temporally correlated noises. As an additional merit, this strategy is also the optimal way to suppress the CTF perturbation noise, e.g., the CTF estimation error, if the perturbation noise is also assumed to be spatially and temporally uncorrelated. This leads to the following optimization problem:

$$\min_{\mathbf{h}} \|\mathbf{A}\mathbf{h} - \mathbf{g}\|^2 + \delta \phi_a^{j_d} \|\mathbf{h}\|^2, \tag{13}$$

where $\phi_a^{j_d} = \sum_{i=1}^{I} \sum_{p=0}^{L_a-1} |a_p^{i,j_d}|^2$ is the CTF energy for the desired source (summed over channels and frames), used as a normalization term, and δ is a regularization factor.[1] Indeed, the power of the inverse filter $\mathbf{h}$ is at the level of $1/\phi_a^{j_d}$, thus $\|\mathbf{h}\|^2$ is somehow normalized by $\phi_a^{j_d}$. As a result, the choice of δ, which controls the trade-off between the two terms in (13), is made independent of the energy level of the CTF filters. This property is especially relevant for the present frequency-wise algorithm since all frequencies can share the same regularization factor δ, although the CTF energy may significantly vary along the frequencies. The solution of (13) is

$$\hat{\mathbf{h}}^{\text{mint}} = (\mathbf{A}^H \mathbf{A} + \delta \phi_a^{j_d} \mathbf{I})^{-1} \mathbf{A}^H \mathbf{g}, \tag{14}$$

where $\mathbf{I}$ is the IL_h-dimensional identity matrix. Due to the regularization term, this solution is no longer an exact inverse filter, and is more like a regularized least squares estimator. Concerning the fact that it is derived from the MINT formulation, we refer to this method as CTF-MINT.

The length of the inverse filters, i.e., L_h, should be set based on the desired shape of matrix $\mathbf{A}$. Let ρ denote the ratio between the number of columns and the number of rows of $\mathbf{A}$, then we have $IL_h = \rho J(L_a + L_h - 1)$. Rename L_h as L_h^{mint}, then:

$$L_h^{\text{mint}} = \frac{L_a - 1}{\frac{I}{\rho J} - 1}, \quad \text{with } \rho < \frac{I}{J}. \tag{15}$$

Based on pilot experiments, the best choice is to set $\mathbf{A}$ to be square, i.e., $\rho = 1$, which is feasible when $I > J$. For non-overdetermined recordings, i.e., $I \leq J$, ρ should be less than $\frac{I}{J}$, and consequently $\mathbf{A}$ is a narrow matrix. In this case, the optimization problem (13) is still feasible, since exact inverse filtering is not required in (13). Note that $L_h^{\text{mint}} \to +\infty$ when $\rho \to \frac{I}{J}$, thence in practice ρ should be sufficiently small to avoid a very large L_h^{mint}.

C. The CTF-MPDR Method

The above CTF-MINT approach requires the knowledge of CTFs for all the sources. In this section, we consider the situation

[1] Note that $\|\cdot\|$ and $|\cdot|$ denote the ℓ_2-norm of a vector and the absolute value of a scalar, respectively.

where the CTFs of the sources are not all known/estimated. One source is recovered based on its own CTFs only.

For the desired source, the inverse filter $\mathbf{h}$ should still satisfy $\mathbf{A}^{j_d}\mathbf{h} = \mathbf{d}$ to achieve a distortionless estimated source. At the same time, the power of the output, i.e., $\|\mathbf{X}\mathbf{h}\|^2$, should be minimized. Again, by relaxing the match between $\mathbf{A}^{j_d}\mathbf{h}$ and $\mathbf{d}$, we define the following optimization problem:

$$\min_{\mathbf{h}} \|\mathbf{A}^{j_d}\mathbf{h} - \mathbf{d}\|^2 + \kappa \frac{\phi_a^{j_d}}{\phi_x}\|\mathbf{X}\mathbf{h}\|^2, \tag{16}$$

where $\phi_x = \sum_{i=1}^{I}\sum_{p=0}^{P-1}|x_p^i|^2$ is the energy of the microphone signals (summed over frames and channels). Similarly to CTF-MINT, the normalization factor $\frac{\phi_a^{j_d}}{\phi_x}$ makes the choice of the regularization factor κ independent of the energy of the CTF filters and of the energy of the microphone signals. Therefore, all frequencies can share the same regularization factor κ, even if the energy of microphone signals significantly varies across frequencies. This optimization problem considers any type of noise signal equally by minimizing the overall output power. This method is similar in spirit with to the MPDR beamformer, more exactly it is similar to the speech distortion weighted multichannel Wiener filter [36] since the source distortionless constraint is relaxed. We still refer to this method as CTF-MPDR. It is known that minimizing the overall output power risks to distort the desired signal, and a better strategy is to minimize the output power of interfering signals, as done by the MVDR beamformer and the partial MINT method proposed in [15]. However, for the present method, interfering signals include not only stationary noise, but also non-stationary speech signals, whose power spectral density (PSD) is very difficult to estimate in practice. In addition, also due to the non-stationarity of speech signals, we present the regularization term in (16) as an instance of the output power, instead of the power expectation.

The solution of (16), i.e., the CTF-based beamforming-like inverse filter, is

$$\hat{\mathbf{h}}^{\text{mpdr}} = \left(\mathbf{A}^{j_d H}\mathbf{A}^{j_d} + \kappa\frac{\phi_a^{j_d}}{\phi_x}\mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{A}^{j_d H}\mathbf{d}. \tag{17}$$

Similarly, let ϱ denote the ratio between the number of columns and the number of rows of $\mathbf{A}^{j_d}$, then we have $IL_h = \varrho(L_a + L_h - 1)$. Rename L_h as L_h^{mpdr}, then:

$$L_h^{\text{mpdr}} = \frac{L_a - 1}{\frac{I}{\varrho} - 1}, \quad \text{with } \varrho < I. \tag{18}$$

Because the inverse filter is constrained by only one source, i.e., the desired source, we can always set $\varrho = 1$ in order to have $\mathbf{A}^{j_d}$ square.

For both CTF-MINT and CTF-MPDR, the J source signals are estimated by respectively taking the $1, \ldots, J$-th source as the desired source and applying (7). They both do not require the knowledge of noise statistics.

IV. CTF-Based Basis Pursuit

Instead of explicitly estimating an inverse filter, the source signals can be directly recovered by matching the microphone signals and the mixing model involving the unknown source signals. The spectral sparsity of the speech signals can be exploited as *prior* knowledge. Interfering sources, reverberation and noise all tend to make the signal spectra less sparse, hence it is beneficial to suppress them by imposing spectral sparsity on the estimated source signals.

A. Mixing Model and Initial Problem Formulation

The mixing model (6) is rewritten in vector/matrix form as

$$\mathbf{x} = \mathcal{A} \star \mathbf{s} + \mathbf{e}, \tag{19}$$

where $\mathbf{x} \in \mathbb{C}^{I\times P}$, $\mathbf{s} \in \mathbb{C}^{J\times P}$ and $\mathbf{e} \in \mathbb{C}^{I\times P}$ denote the matrices of microphone signals, source signals and noise signals, respectively, and $\mathcal{A} \in \mathbb{C}^{I\times J\times P}$ denotes the three-dimensional CTF array. The convolution $\star$ is carried out along the time frame. Remember that this equation is defined for each frequency bin k and that we omit the k index for clarity of presentation.

In our previous work [27], we proposed to estimate the source signals by solving an ℓ_2-norm fitting cost minimization problem with an ℓ_1-norm regularization term

$$\min_{\mathbf{s}} \|\mathcal{A} \star \mathbf{s} - \mathbf{x}\|^2 + \lambda\|\mathbf{s}\|_1, \tag{20}$$

where $\|\cdot\|_1$ denotes ℓ_1-norm and λ is the regularization factor. Note that both the ℓ_2- and ℓ_1-norms on matrices are redefined here as vector norms. The first term minimizes the fitting cost, and the second term imposes sparsity on the estimated speech source signals, implicitly contributing to suppress the additive noise $\mathbf{e}$ from the estimated source signals. However, it is difficult to automatically tune λ even when the noise statistics are known. Especially, the source estimation is performed frequency by frequency in this work, and it is common that the noise PSD varies significantly across frequencies. This requires a specific value of λ for each frequency, which further increases the difficulty of choosing λ. In this work, we solve this problem by transforming the above problem to a basis pursuit problem.

B. CTF-Based Basis Pursuit Reformulation

Problem (20) can be reformulated as:

$$\min_{\mathbf{s}} \|\mathbf{s}\|_1, \quad \text{s.t. } \|\mathcal{A} \star \mathbf{s} - \mathbf{x}\|^2 \leq \epsilon, \tag{21}$$

for some unknown λ and ϵ. The ℓ_2-norm fitting cost is relaxed to at most a tolerance ϵ. This formulation was first proposed in [10] for audio source separation in the time domain. We adapted it to the CTF-magnitude domain in our previous work [37] for single-source dereverberation. In the present work, we further extend it to the complex-valued CTF domain for multisource recovery. We refer to this method as CTF-based basis pursuit (CTF-BP).

The setting of the tolerance ϵ is critical to the quality of the recovered source signals. We assume that the noise signal is stationary, and we propose to set the tolerance ϵ as a function of the noise PSD. Let σ_i^2 denote the noise PSD in the i-th microphone. Let $\mathbf{e}^i \in \mathbb{C}^{1\times P}$ denote the noise signal in the i-th microphone in vector form. The squared ℓ_2-norm of the noise signal, i.e., the noise energy $\|\mathbf{e}^i\|^2$, follows an Erlang distribution with mean $P\sigma_i^2$ and variance $P\sigma_i^4$ [38]. If we further assume

that the noise signal is spatially uncorrelated, then for all microphones, the squared ℓ_2-norm $\|\mathbf{e}\|^2$ has mean $\sum_{i=1}^{I} P\sigma_i^2$ and variance $\sum_{i=1}^{I} P\sigma_i^4$. To adjust the ℓ_2-norm fitting cost to the noise power, we define the noise-related tolerance as:

$$\epsilon_e = \sum_{i=1}^{I} P\sigma_i^2 - 2\sqrt{\sum_{i=1}^{I} P\sigma_i^4}. \tag{22}$$

Twice the standard deviation is subtracted in the above equation to guarantee that the probability of the ℓ_2-norm fitting cost being larger than $\|\mathbf{e}\|^2$ is very small. The reason for this setting is the following. If the ℓ_2-norm fitting cost happens to be larger than $\|\mathbf{e}\|^2$, the minimization of $\|\mathbf{s}\|_1$ in (21) will distort the source signal. In other words, by subtracting twice the standard deviation, the estimated source signal will be less distorted at the price of less noise reduction. Note that a directional noise source with non-sparse spectra cannot be well recovered by this method. Therefore, such directional noise is considered as noise but not as a source, i.e., its power is included in $\|\mathbf{e}\|^2$, and thus in ϵ_e. Note also that this method needs only an estimation of the single-channel noise PSD, but not the inter-channel or inter-frame noise cross-PSDs.

Besides, the ℓ_2-norm fitting cost should also be relaxed with respect to the CTF approximation error and the CTF perturbations. The CTF perturbations lead to a mismatch between $\mathcal{A} \star \mathbf{s}$ and $\mathbf{x}$. The level of this mismatch is supposed to be proportional to the energy of the source images, since convolution is a linear operation. Therefore, the tolerance should also be adjusted to the energy of the noise-free signal, which can be estimated by spectral subtraction as:

$$\hat{\Gamma}_s = \max\left(\|\mathbf{x}\|^2 - \sum_{i=1}^{I} P\sigma_i^2,\ 0\right). \tag{23}$$

We define the tolerance with respect to the noise-free signal as $\epsilon_s = \varsigma\hat{\Gamma}_s$, where the factor ς should be set based on the level of CTF perturbations. Overall, the tolerance is set to $\epsilon = \epsilon_e + \epsilon_s$.

C. Convex Optimization Algorithm

The optimization algorithm presented in this section mainly follows the principle proposed in [10]. Unlike [10], the target optimization problem (21) is carried out in the complex domain, and thus the optimization algorithm is also complex-valued. The optimization problem consists of an ℓ_1-norm minimization and a quadratic constraint, which are both convex. The difficulty of this convex optimization problem is that the ℓ_1-norm objective function is not differentiable.

The constrained optimization problem (21) can be recast as the following unconstrained optimization problem:

$$\min_{\mathbf{s}} \|\mathbf{s}\|_1 + \iota_C(\mathbf{s}), \tag{24}$$

where C denotes the convex set of signals verifying the constraint, $C = \{\mathbf{s}\ |\|\mathcal{A} \star \mathbf{s} - \mathbf{x}\|^2 \leq \epsilon\}$, and $\iota_C(\mathbf{s})$ denotes the indicator function of C, namely $\iota_C(\mathbf{s})$ equals 0 if $\mathbf{s} \in C$, and $+\infty$ otherwise. This unconstrained problem consists of two lower semi-continuous, non-differentiable (non-smooth), convex functions. For this problem, the *Douglas-Rachford* splitting method [34] is suitable, which is an iterative method, summarized in Algorithm 1. At each iteration, the two functions are split, and their respective proximity operators $\text{Prox}_{\iota_C(\cdot)}$ and $\text{Prox}_{\gamma|\cdot|}$ (see below) are individually applied. The *Douglas-Rachford* method does not require the differentiability of any of the two functions, and is a generalization of the *proximal splitting* method [35]. In our experiments, α and γ are set to 1 and 0.01 respectively. The initialization of $\mathbf{s}_0$ is set as the matrix composed of J replications of the first microphone signal. The convergence criteria is set to check if the optimization objective is almost invariant from one iteration to the next. The threshold η_1 is set to 0.01 and the number of iterations is limited to 20.

Algorithm 1: Douglas-Rachford.

Initialization: $l = 0$, $\mathbf{s}_0 \in \mathbb{C}^{I\times P}$, $\alpha \in (0,2)$, $\gamma > 0$,
repeat
  $\mathbf{z}_l = \text{Prox}_{\iota_C(\cdot)}(\mathbf{s}_l)$
  $\mathbf{s}_{l+1} = \mathbf{s}_l + \alpha(\text{Prox}_{\gamma\|\cdot\|_1}(2\mathbf{z}_l - \mathbf{s}_l) - \mathbf{z}_l)$
  $l = l + 1$
until $|\|\mathbf{s}_l\|_1 - \|\mathbf{s}_{l-1}\|_1|/\|\mathbf{s}_l\|_1 < \eta_1$

Algorithm 2: $\text{Prox}_{\iota_C(\cdot)}(\mathbf{s})$.

Input: $\mathbf{x}$, $\mathcal{A}$, $\mathcal{A}^*$, $\mathbf{s}$
Initialization: $l = 0$, $\mathbf{u}_0 = \mathbf{x}$, $\mathbf{p}_0 = \mathbf{s}$, $t_0 = 1$, $\mu \in (0, 2/\nu)$
repeat
  1. $l = l + 1$
  2. $\mathbf{u}_l = \mu(\mathbf{I} - \text{Prox}_{\iota_{\|\cdot\|^2\leq\epsilon}})(\mu^{-1}\mathbf{u}_{l-1} + \mathcal{A} \star \mathbf{p}_{l-1} - \mathbf{x})$
  3. $t_l = \left(1 + \sqrt{(1 + 4t_{l-1}^2)}\right)/2$
  4. $\tilde{\mathbf{u}}_l = \mathbf{u}_{l-1} + \frac{t_{l-1}-1}{t_l}(\mathbf{u}_l - \mathbf{u}_{l-1})$
  5. $\mathbf{p}_l = \mathbf{s} - \mathcal{A}^* \star \tilde{\mathbf{u}}_l$
until $\|\mathcal{A} \star \mathbf{p}_k - \mathbf{x}\|^2 \leq 1.1\epsilon$
Output: $\mathbf{p}_l$

The proximity operator plays an important role in the optimization of nonsmooth functions. In Hilbert space, the proximity of a complex-valued function f is

$$\text{Prox}_f(\mathbf{z}) = \underset{\mathbf{y}}{\text{argmin}}\, f(\mathbf{y}) + \|\mathbf{z} - \mathbf{y}\|^2. \tag{25}$$

The proximity operator of the ℓ_1-norm $\gamma\|\cdot\|_1$ at point $\mathbf{z}$, aka the shrinkage operator, is given entry-wise by

$$y_i = \frac{z_i}{|z_i|}\max(0, |z_i| - \gamma). \tag{26}$$

The proximity of the indicator function $\iota_C(\mathbf{s})$ is the *projection* of $\mathbf{s}$ onto C. To compute this proximity, based on the *proximal splitting* method and the Fenchel-Rockafellar duality [39], an iterative method was derived in [40], and used in [10]. However, this method converges linearly, which is very slow especially when the convex set C (also ϵ) is small. As hinted in [40], it can be accelerated to the squared speed via the Nesterov's scheme [41], [42]. The accelerated method is summarized in Algorithm 2. The acceleration procedure is composed of Step 3 and 4, which are based on the derivation in [42]. Here $\mathcal{A}^*$ is the adjoint matrix of $\mathcal{A}$, and is obtained by conjugate transposing the source and channel indices, and then temporally reversing the filters. Here ν is the tightest frame bound of the quadratic

operation in the indicator function, and thus is the largest spectral value of the frame operator $\mathcal{A}^* \circ \mathcal{A}$. It is computed using the power iteration method, for which please refer to [27]. In our experiments, we set $\mu = 1/\nu$. In Step 2, the *projection* of a variable $\mathbf{u}$ onto the convex set $\{\mathbf{v} \; | \|\mathbf{v}\|^2 \leq \epsilon\}$ can be easily obtained as

$$\mathrm{Prox}_{\iota_{\|\cdot\|^2 \leq \epsilon}}(\mathbf{u}) = \min\left(1, \frac{\sqrt{\epsilon}}{\|\mathbf{u}\|}\right)\mathbf{u}. \tag{27}$$

In Algorithm 2, the variable $\mathbf{p}_k$ iteratively moves from the initial point $\mathbf{s}$ to its *projection*, thence a convergence criteria is set to check the feasibility of the constraint. The slack factor 1.1 is set to avoid a too long convergence, which however leads to a possible small bias of the ℓ_2-norm constraint. In addition, the maximum number of iterations is set to 300.

V. Experiments

In this section, we evaluate the quality of the source signals estimated by the proposed methods. The quality is measured by the performance of source separation, speech dereverberation and noise reduction.

A. Experimental Configuration

1) Dataset: The multichannel impulse response dataset [43] is used, which was recorded using an 8-channel linear microphone array in the speech and acoustic lab of Bar-Ilan University, with room size of 6 m $\times$ 6 m $\times$ 2.4 m. The reverberation time was controlled by 60 panels covering the room facets. In the reported experiments, we use the recordings with $T_{60} = 0.61$ s. The RIRs are truncated to correspond to T_{60}, and have a length of 10,000 samples. Monophonic speech signals from the TIMIT dataset [44] are used as source signals, with a duration of about 3 s. A multichannel source image is obtained by convolving a source signal with the RIRs corresponding to the different source-to-microphone paths. Multiple source images corresponding to different speakers, phonetic content, directions and microphone-to-source distances are summed up to obtain a mixture signal. The source directions and the microphone-to-source distances are randomly selected from $-90°$:$15°$:$90°$ and $\{1\,\mathrm{m}, 2\,\mathrm{m}\}$, respectively. To generate noisy microphone signals, a spatially uncorrelated stationary speech-like noise is added to the noise-free mixture. The noise level is controlled by a wideband input signal-to-noise ratio (SNR). Note that SNR refers to the averaged single source-to-noise ratio over multiple sources. To evaluate the robustness of the methods to the perturbations of the RIRs/CTFs, a proportional Gaussian random noise is added to the original filters $a^{i,j}(n)$ in the time domain to generate the perturbed filters denoted as $\tilde{a}^{i,j}(n)$. The perturbation level is quantified by the normalized projection misalignment (NPM) measure [45] in decibels (dB).

2) Performance Metrics: The signal-to-distortion ratio (SDR) [46] in dB is used to evaluate the overall quality of the outputs. The unprocessed microphone signals are evaluated as the baseline scores. The overall outputs, i.e., (7) for CTF-MINT and CTF-MPDR, and (21) for CTF-BP, are evaluated as the output scores.[2]

The signal-to-interference ratio (SIR) [46] in dB is used to evaluate the source separation performance. This metric focuses on the suppression of interfering sources, thence SIR is calculated with the additive noise being eliminated. The unprocessed noise-free mixtures, i.e., $\sum_{j=1}^{J} a_p^{i,j} \star s_p^j$, are evaluated as the baseline scores. For CTF-MINT and CTF-MPDR, we take the noise-free output for SIR evaluation. The noise-free output and output noise are obtained by applying the inverse filtering to the noise-free mixture and to the noise-only signal, i.e., $\sum_{i=1}^{I} h_p^i \star (\sum_{j=1}^{J} a_p^{i,j} \star s_p^j)$ and $\sum_{i=1}^{I} h_p^i \star e_p^i$ in (7), respectively. However, for CTF-BP, the noise-free output is not available, since speech and noise components are mixed together in the estimated source signal, and cannot be separated. Therefore, we test the overall outputs. Experimental results show that CTF-BP has low residual noise, thus SIR measures are not significantly influenced by the output noise.

The perceptual evaluation of speech quality (PESQ) measure [47] is specially used to evaluate the dereverberation performance. PESQ is calculated with interfering sources and noise being eliminated. The unprocessed source images, i.e., $a_p^{i,j} \star s_p^j$ are evaluated as the baseline scores. For CTF-MINT and CTF-MPDR, the noise-free single-source output, i.e., $\sum_{i=1}^{I} h_p^i \star (a_p^{i,j} \star s_p^j)$ is evaluated. For CTF-BP, again we have to test the overall outputs. However, the residual interfering sources and noise affect the PESQ measure to a large extent. Therefore, we should note that the dereverberation performance of CTF-BP cannot be precisely reflected and is actually underestimated by the PESQ measure.

The output SNR in dB is used to evaluate the noise reduction performance. The input SNR is taken as the baseline score. For CTF-MINT and CTF-MPDR, the output SNR is computed as the power ratio between the noise-free outputs and the output noise. For CTF-BP, the noise PSD in the output signals is first blindly estimated using the method proposed in [48], where the estimation error was shown to be around 1 dB. The power of the noise-free outputs are estimated by spectral subtraction following the principle in (23), and then the output SNR is obtained by taking the ratio of them. Because the noise PSD estimation error is around 1 dB, the estimated output SNR also deviates from the ground truth output SNR by about 1 dB. For all the four metrics, the higher the better.

3) Parameter Settings: The sampling rate is 16 kHz. The STFT is calculated using a Hamming window, with window length and frame step of $N = 1,024$ (64 ms) and $D = N/4 = 256$, respectively. The CTFs are computed from the time-domain filters using (11). The CTF length L_a is 47. For the overdetermined recording system, i.e., $I > J$, the length of the inverse filter of CTF-MINT, i.e., L_h^{mint}, is computed via (15) with $\rho = 1$, which makes $\mathbf{A}$ square. Pilot experiments show that a longer inverse filter (or a larger ρ) does not noticeably improve the performance measures, while leading to a larger

[2]All metrics are actually evaluated on time-domain signals, obtained using inverse STFT. Here, we refer to the STFT-domain processed signals by convenience.

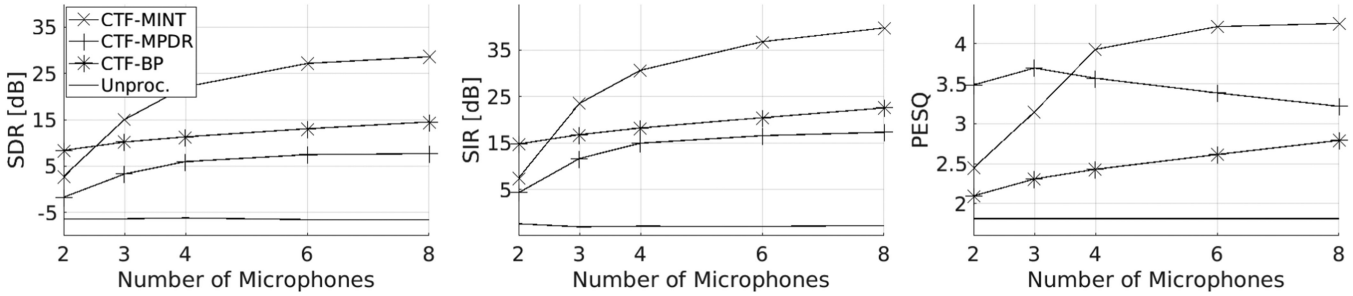

Fig. 2. Performance of the proposed methods as a function of the number of microphones I ($J = 3$, noise-free mixtures). Note that the legends in this figure are common to all the following figures.

computational cost. For the case of $I \leq J$, ρ is set to be less than and close to $\frac{I}{J}$. The exact values of ρ will be given in the following experiments depending on the specific values of I and J. The length of the inverse filter of CTF-MPDR, i.e., L_h^{mpdr}, is computed via (18) with $\varrho = 1$, thus $\mathbf{A}^{j_d}$ is square. The optimal setting of the modeling delay in $\mathbf{d}$ is related to the length of the inverse filters. In the experiments, it is set to 6 and 3 taps for CTF-MINT and CTF-MPDR, respectively, as a good tradeoff for the different inverse filter lengths under various acoustic conditions. The tolerance factor with respect to the CTF perturbations, i.e., ς, is set to 0.01.

Thanks to the normalization factors in (13) and (16), a unique regularization factor δ and a unique regularization factor κ are suitable for all frequencies. Moreover, they are robust to large variations in the numerical range of both filters and signals in different datasets. For CTF-MINT, a large δ value makes the inverse filtering inaccurate and thus deteriorates the performance of source separation and dereverberation, but it also reduces the energy of the inverse filters and thus reduces the influence of noise and CTF perturbations. In the following experiments, we consider two representative choices of δ: i) a small one, i.e., 10^{-5}, is used for the cases where both noise and CTF perturbations are absent, and ii) the small one and a large one, i.e., 10^{-1}, are respectively tested and compared for the cases where noise or CTF perturbations are present. For CTF-MPDR, the minimization of the output power suppresses both the interfering sources and the noise, but there is a risk to suppress the desired source. Based on pilot experiments, we set κ to 10^{-1}, which achieves a good tradeoff between distortion of desired source and suppression of interfering signals.

In the following sections, we first evaluate the three proposed methods under various acoustic conditions, in terms of number of microphones and sources, SNRs, and NPMs. For each condition, 20 runs are executed, and the averaged performance measures over these 20 runs are reported. Then, the comparison with four baseline methods will be conducted, in terms of both performance and computational complexity.

B. Influence of the Number of Microphones

Fig. 2 shows the results as a function of the number of microphones, for 3-source mixtures. In this experiment, the microphone signals are noise-free, thus the output SNR is not reported. For CTF-MINT, ρ is set to 0.55 and 0.8 for the cases of two and three microphones, respectively. Consequently the length of the inverse filters are about five times the CTF length.

For CTF-MINT, the scores of all three reported metrics dramatically decrease when the number of microphones goes from four to three and to two, namely from the over-determined case to the determined case and to the under-determined case. This indicates that the inaccuracy of the inverse filtering is large for the non over-determined case, due to the insufficient degrees of freedom of the inverse filters as spatial parameters. CTF-MPDR suppresses the interfering sources by minimizing the power of the output, and also implicitly by applying inverse filtering with a zero signal target. Therefore, as for CTF-MPDR, the metrics that measure the suppression of interfering sources, i.e., SDR and SIR, also significantly degrade for the non over-determined case. The PESQ score varies slightly with the increase of number of microphones, which means that the inverse filtering of the desired source is not considerably affected, due to the small variation of the output power. The performance measures of CTF-BP increase almost linearly with the number of microphones, no matter whether it is under-determined or over-determined, thanks to exploiting the spectral sparsity. For the over-determined case, i.e., four microphones or more, SDR and SIR for the three methods slowly increase with the number of microphones, and CTF-MINT has the larger increase rate. CTF-BP achieves the worst PESQ score due to the influence of the residual interfering sources. Informal listening tests show that the outputs of CTF-BP are not perceived as more reverberant than the outputs of CTF-MINT or CTF-MPDR.

Overall, without considering the noise reduction, CTF-MINT performs the best for the over-determined case. For instance, CTF-MINT achieves an SDR of 22.1 dB by using four microphones, which is a very good source recovery SDR score. This indicates that the inverse filtering is very accurate. CTF-BP does not perform as well as CTF-MINT for the over-determined case, since the spectral sparsity of source signals cannot be accurately quantified. CTF-BP performs the best for the under-determined case. For instance, CTF-BP achieves an SDR of 8.4 dB by using only two microphones. By only using the mixing filters of one source, the source separation performance of CTF-MPDR is worse than the other two methods.

C. Performance for Various Number of Sources

Fig. 3 shows the results as a function of the number of sources. In this experiment, the number of microphones is fixed to six.

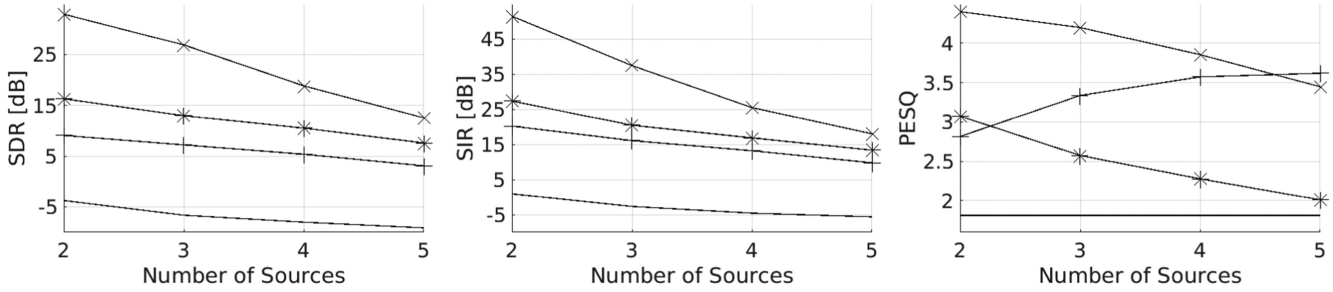

Fig. 3. Performance measures of the proposed methods as a function of the number of sources J ($I = 6$, noise-free mixtures).

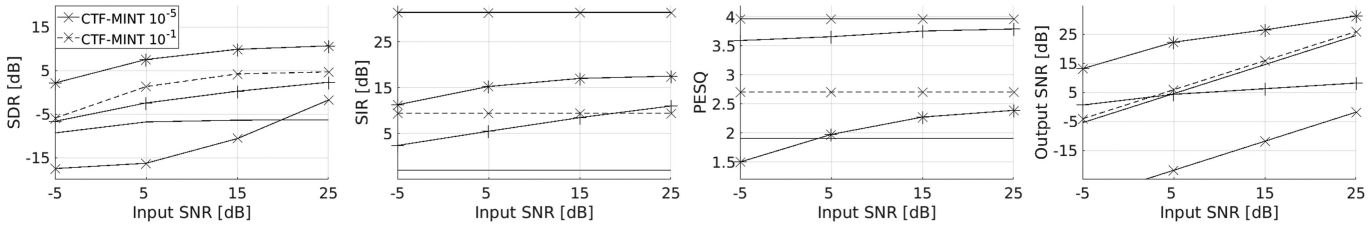

Fig. 4. Performance measures of the proposed methods as a function of input SNR ($I = 4$ and $J = 3$).

The microphone signals are noise-free, thus the output SNR is not reported. From this figure, we can see that the performance measures of the three methods degrade with the increase of the number of sources, except for the PESQ score of CTF-MPDR. CTF-MINT achieves the best performance, even if it exhibits the largest performance degradation. This is somehow consistent with the experiments with various number of microphones: Good performance requires a large ratio between the number of microphones and the number of sources. Both CTF-MPDR and CTF-BP have smaller performance degradation. At first sight, it is surprising that CTF-MPDR achieves a larger PESQ score when more sources are present in the mixture. The reason is that the normalized output power, i.e., $\frac{\phi_a^{jd}}{\phi_x}\|\mathbf{X}\mathbf{h}\|^2$, becomes smaller with the increase of the number of sources due to a larger ϕ_x. Correspondingly, the inverse filtering inaccuracy of the desired source, i.e., $\|\mathbf{A}^{jd}\mathbf{h} - \mathbf{d}\|^2$, becomes smaller.

D. Influence of Additive Noise

Fig. 4 shows the results as a function of the input SNR, for 4-microphone 3-source mixtures. As mentioned above, for the noisy case, two δ settings, i.e., 10^{-5} and 10^{-1}, are tested. The inverse filters of CTF-MINT are invariant for various input SNRs, since they depend only on the CTF filters, but not on the microphone signals. As a result, the SIR and PESQ scores are constant, and the scores for $\delta = 10^{-5}$ are much higher than the scores for $\delta = 10^{-1}$, since the inverse filtering is more accurate with $\delta = 10^{-5}$. However, with $\delta = 10^{-5}$, the noise is amplified by about 30 dB, and consequently the SDR scores are very low. In contrast, with $\delta = 10^{-1}$, the noise is slightly suppressed. For CTF-MPDR, SIR and PESQ scores are smaller when the input SNR is low, since a larger input noise leads to a larger output noise, thus degrades the suppression of the interfering sources, and distorts the inverse filtering of the desired source. The output SNR increases with the increase of the input SNR, but the SNR improvement decreases. It even becomes negative when the input SNR is larger than 5 dB, which means the noise is amplified. The output SNR of CTF-BP is always larger than the input SNR, which means that the noise is efficiently reduced. SDR and SIR of CTF-BP slightly degrade for the low SNRs.

For CTF-MINT and CTF-MPDR, the residual noise is significant, which indicates that the temporal and spatial (inverse) filtering is not able to efficiently suppress the white noise. A single-channel noise reduction process can be used as a postprocessing, as in [49], [50]. One should choose to either apply or not to apply the postprocessing depending on the specific application. For example, for human hearing aids, postprocessing is necessary, since the residual noise is very annoying for human hearing. On the contrary, if the enhanced signals are fed to an automatic speech recognition system, postprocessing may deteriorate the speech recognition performance [51]. The development of a postprocessing method is beyond the scope of this work.

E. Influence of CTF Perturbations

Fig. 5 shows the results as a function of NPM, for 4-microphone 3-source mixtures. As for the previous experiment, both δ settings are tested. As expected, all metrics become worse with the increase of NPM, thus we only analyze the SDR scores. Note that, when NPM is -65 dB, the three methods achieve almost the same performance measures as for the perturbation-free case. Along with the increase of NPM, the performance of CTF-MINT with $\delta = 10^{-5}$ dramatically degrades from a large score to a very small score, which indicates its high sensitivity to CTF perturbations. In contrast, CTF-MINT with $\delta = 10^{-1}$ has a small performance degradation rate, but the performance is poor even for the low NPM case. The performance measures of CTF-MPDR almost linearly decreases with a relatively large degradation rate. The performance of CTF-BP is stable until NPM $= -35$ dB, and quickly degrades when NPM is larger than -25 dB.

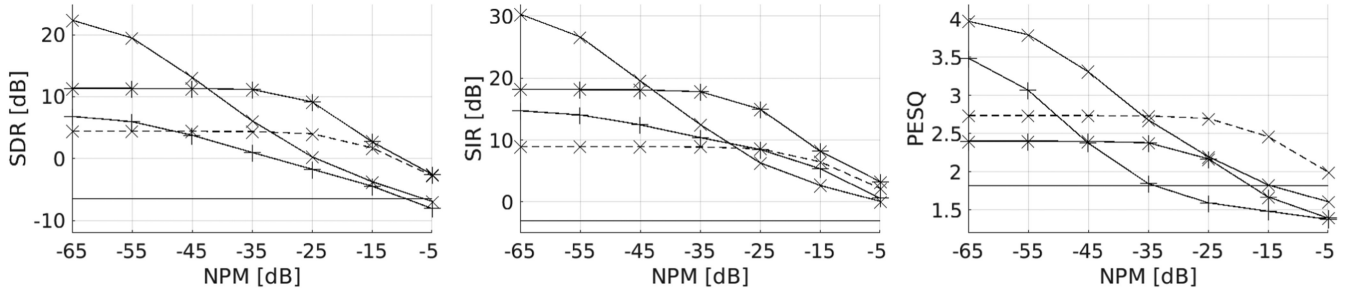

Fig. 5. Performance measures of the proposed methods as a function of NPM ($I = 4$, $J = 3$, noise-free mixtures).

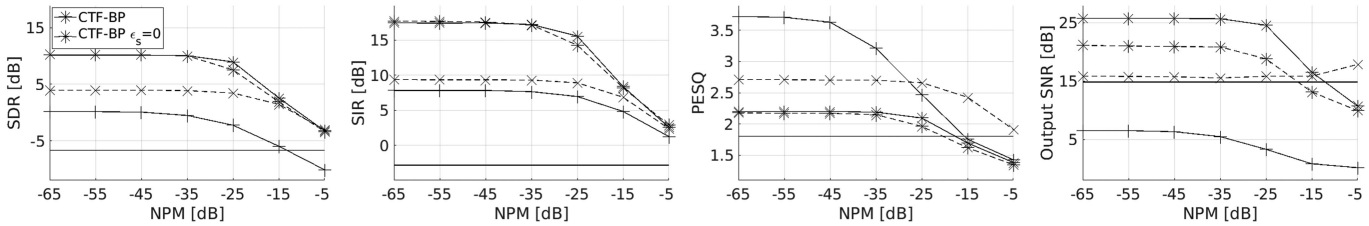

Fig. 6. Performance measures of the proposed methods as a function of NPM ($I = 4$, $J = 3$, input SNR = 15 dB).

In CTF-MINT, the inverse filter is designed to satisfy the targets of both the desired source and interfering sources. Therefore, the CTF perturbations of the desired source will not significantly affect the suppression of interfering sources, and vice versa. Moreover, in CTF-MPDR, the inverse filter is computed depending only on the CTFs of the desired source, thence the CTF perturbations of the interfering sources will not affect the inverse filtering at all. In contrast, in CTF-BP, all sources are simultaneously recovered based on the CTFs of all of them, consequently the CTF perturbations of one source will affect the recovery of all sources. These assertions have been verified by some pilot experiments.

F. Results in the Presence of Noise and CTF Perturbations

In practice, both noise and CTF perturbations are present. Fig. 6 displays the results as a function of NPM, for the input SNR set to 15 dB. For CTF-MINT, only the setting with $\delta = 10^{-1}$ is tested. As already mentioned, the inverse filters of CTF-MINT are irrelevant to the noise, thence the SIR and PESQ scores are identical to the noise-free case. The output SNRs for various NPMs are close to each other, which indicates that the energy of the inverse filters do not noticeably vary in the presence of CTF perturbations. For CTF-MPDR, the major influence of the noise is that the performance measures become less sensitive to the CTF perturbations in the low NPM region. For example, SDR and SIR do not significantly decrease with the increase of NPM until -25 dB. The possible reason is that, in order to suppress the noise, the energy of the inverse filters becomes lower, and thus the sensitivity to CTF perturbations becomes lower as well. However, the price of this is that the performance is worse than in the noise-free case.

For CTF-BP, compared with the noise-free case, the noise systematically degrades the performance measures of source separation and dereverberation. Remember that in the proposed CTF-BP scheme, the tolerance for the ℓ_2-norm fitting cost is set to $\epsilon = \epsilon_e + \epsilon_s$, where ϵ_s accounts for the CTF perturbations (see Section IV-B). To validate this scheme, we also tested the tolerance without using ϵ_s, i.e., setting $\epsilon_s = 0$ instead of $\epsilon_s = 0.01\hat{\Gamma}_s$. The results are illustrated in Fig. 6 by the dashed lines with the asterisk marker. We can see that the performance scores for $\epsilon_s = 0$ are lower than for $\epsilon_s = 0.01\hat{\Gamma}_s$, especially when NPM is larger than -35 dB. This indicates that a non-zero value of ϵ_s enables to tackle the CTF perturbations to some extent. However, the performance improvement is not significant when NPM is larger than -15 dB. The reason is that, when NPM is large, setting $\epsilon_s = 0.01\hat{\Gamma}_s$ is not high enough to represent the mismatch between $\mathcal{A} \star \mathbf{s}$ and $\mathbf{x}$. In other words, a factor larger than 0.01 should be used for high NPMs. In practice, the CTF perturbations level is highly related to the specific methods used for CTF estimation, hence ϵ_s should be empirically set with an optimal ς.

G. Comparison With Baseline Methods

To benchmark the proposed methods, we compare them with four baseline methods:

- LCMP beamformer [4] based on the narrowband assumption. Based on the steering vectors and the correlation matrix of the microphone signals, a beamformer is computed to preserve one desired source and zero out the others, and to minimize the power of the output. The RIRs are longer than the STFT window, thus the steering vector should be computed as the Fourier transform of the truncated RIRs. In this experiment, the steering vector is set to the CTF tap with the largest power.
- Time-domain MINT (TD-MINT) [16]. This method is also set to recover the direct-path source signal with an energy regularization. In this experiment, we extend this method to the multisource case. Following the principle of the proposed method, the length of inverse filter is set to $\frac{10000-1}{IJ-1}$, which makes the channel convolution matrix square. We

TABLE I
SDR SCORES AND COMPUTATION TIMES FOR SIX REPRESENTATIVE ACOUSTIC CONDITIONS. SDR SCORES OF THE UNPROCESSED SIGNALS ARE GIVEN IN THE PREVIOUS EXPERIMENTS

Acoustic Condition				SDR [dB]							Computation Time per Mixture [s]						
I	J	SNR	NPM	CTF-MINT	CTF-MPDR	CTF-BP	LCMP	TD-MINT	CTC	W-Lasso	CTF-MINT	CTF-MPDR	CTF-BP	LCMP	TD-MINT	CTC	W-Lasso
4	3	-	-	22.1	6.0	11.3	-3.6	-	17.5	17.8	49.4	5.4	1306	1.1	-	74847	3989
6	2	-	-	33.0	9.1	16.3	-0.3	33.7	32.4	29.1	6.3	4.1	1090	1.1	759	2119	3714
6	3	-	-	26.9	7.2	13.0	-0.6	-	22.3	21.8	16.6	5.9	1861	1.2	-	19902	5679
6	5	-	-	12.5	3.1	7.6	-6.4	-	9.2	15.4	683.3	13.2	3567	1.8	-	214730	9635
4	3	15 dB	-	4.2	0.2	9.9	-14.7	-	5.7	11.2	49.4	5.4	884	1.1	-	74847	3784
6	2	-	-15 dB	9.4	-0.3	10.9	-3.5	7.9	0.8	11.5	6.3	4.1	551	1.1	759	2165	3578
4	3	-	-15 dB	1.7	-5.4	2.7	-4.1	-	-0.3	6.5	49.4	5.4	1294	1.1	-	69594	3722
4	3	15 dB	-15 dB	1.4	-6.0	2.6	-6.7	-	-3.6	6.2	49.4	5.4	873	1.1	-	69594	3713

only test the condition with $I = 6$ and $J = 2$, with the length of inverse filter of $5{,}000$. Other conditions require a too long inverse filter that cannot be implemented within the basic memory resources on a personal computer. The modeling delay is set to $1{,}024$.

- Crosstalk cancellation (CTC) based on RIR shortening/reshaping with p-norm optimization [18]. For the application of virtual sound reproduction, the knowledge of room filters is exploited to tackle the spatial mismatch problem in [18], which however is not suitable for the present speech separation problem. Therefore, in this experiment, we only test *Algorithm A* of [18], which does not exploit the knowledge of room filters. We use the MATLAB implementations, for multichannel gradient computation and line search, provided by the authors of [18].[3] Some pilot experiments show that the inverse filter length of CTC should be set as the one of TD-MINT, i.e., $\lceil \frac{10000-1}{I/J-1} \rceil$, which makes the channel convolution matrix square. This value is optimal in the sense that a smaller value is insufficient and thus noticeably reduces the performance measures, while a larger value largely increases the computational complexity with only a slight performance gain. Unlike TD-MINT that needs to invert the channel convolution matrix, CTC uses a gradient-descent based optimization method, and thus does not have the memory problem caused by the very long inverse filter. However, when the inverse filter is very long, for example the inverse filter for the configuration with $I = 6/J = 5$ would have a length of 49,995, the computational complexity will be extremely large, and the gradient-descent iteration cannot converge in a reasonable time. To avoid this, we limit the inverse filter length to be not larger than 20,000. As a result, for the four representative configurations that will be tested in this section, i.e., $I = 6/J = 2$, $I = 6/J = 3$, $I = 4/J = 3$ and $I = 6/J = 5$, the inverse filter length is set to 5,000, 9,999, 20,000 and 20,000, respectively. Accordingly, the numbers of iterations for these four configurations are set to 2×10^4, 1×10^5, 2×10^5 and 2×10^5, respectively. Note that due to the very large computation time, instead of 20 runs executed for the other methods, 5 runs are executed for this method.
- Wideband Lasso (W-Lasso) [9]. Based on some pilot experiments, the regularization factor is set to 10^{-5} and 10^{-3} for the case where noise and RIR perturbations are both absent, and for the case where noise or RIR perturbations are present, respectively.

Table I presents the SDR scores for eight representative acoustic conditions, as well as the computation times which will be analyzed in the next section. Note that '-' means noise-free and perturbation-free in the columns of SNR and NPM, respectively. LCMP performs poorly for all conditions, which tends to verify that the narrowband assumption is poorly suitable for the long RIR case. CTF-MINT performs similarly to TD-MINT for the filter perturbation-free case, while noticeably outperforms TD-MINT when NPM $= -15$ dB, which demonstrates that the subband MINT is less sensitive to the filter perturbations than the time-domain MINT due to the smaller filter length. W-Lasso noticeably outperforms CTF-BP for the noise-free and perturbation-free cases, due to its exact time-domain convolution. When noise or filter perturbation present, the performance gap between CTF-BP and W-Lasso becomes smaller. For the noise-free and perturbation-free cases, CTC performs worse than CTF-MINT and TD-MINT. However, we should note that our settings for the number of iterations exclude the very long convergence tail, which means more iterations can further improve the performance. The SDR scores listed in Table I can be increased by 1–4 dB with sufficient iterations, and finally are comparable to the scores of CTF-MINT and TD-MINT. However, involving the long convergence tail is very time consuming, which will double or triple the current compuation time. Compared to CTF-MINT, the performance degradation caused by noise is smaller for CTC, while the one caused by filter perturbation is larger for CTC.

H. Analysis of Computational Complexity

Table I also presents the averaged computation time for the processing of a 3-s mixture. All methods were implemented in MATLAB. CTF-MINT and CTF-MPDR computation times comprise both the computation of the inverse filters and their application on the microphone signals, and the former dominates the computation time. The computations include the matrix inversion and multiplication in (14) and (17), thence the complexity is cubic in matrix dimension. We consider square matrices $\mathbf{A}$ in (14) and $\mathbf{A}^{jd}$ in (17), whose dimension is equal to IL_h. From (15) and (18), IL_h is proportional to the filter

[3] https://www.isip.uni-luebeck.de/index.php?id=617&L=ht%20and%201%3D1

length L_a, to $\frac{I-J}{IJ}$ for CTF-MINT, and to $\frac{I}{I-1}$ for CTF-MPDR. The inverse filters are respectively computed for each source and each frequency. Overall, CTF-MINT and CTF-MPDR have a computational complexity of $\mathcal{O}(\frac{KL_a^3I^3J^4}{(I-J)^3})$ and $\mathcal{O}(\frac{KL_a^3I^3J}{(I-1)^3})$, respectively, where $K = N/2+1$ is the number of frequency bins. The complexity of TD-MINT can be derived from the complexity of CTF-MINT by replacing the CTF length with the RIR length and setting K to 1. Since it is proportional to the cube of RIR length, the complexity is prohibitive for most settings. The LCMP beamformer is similar to CTF-MINT, just using an instantaneous steering vector and an instantaneous inverse filter, namely the length of CTF and inverse filter are both 1, thence it has the lowest computation complexity. These methods have a close-form solution and thus low computational complexity. For CTC, let L_t and L_i denote the length of RIR and inverse filter, respectively. The convolution operation employed in the gradient computation is implemented by $(L_{fft} = L_t + L_i - 1)$-point Fast Fourier Transform (FFT), and each gradient computation needs IJ FFT computations. Overall, the computation complexity of CTC is $\mathcal{O}(N_{iter}IJL_{fft}\log(L_{fft}))$, where N_{iter} is the number of iterations.

The iterative optimization of CTF-BP leads to a high computational complexity. The *Douglas-Rachford* optimization method is a first-order method, thence the complexity is linear with respect to the problem size, specifically the length of microphone signals and filters, and the number of microphones and sources. The most time-consuming procedure in Algorithm 1 is the computation of the proximity of the indicator function (i.e., the projection). To verify this, we can compare the *Douglas-Rachford* method with the optimization algorithm for the Lasso problem (20) that does not have an ℓ_2-norm constraint and thus an indicator function. In [27], we solved the Lasso problem using the fast iterative shrinkage-thresholding algorithm (FISTA) [42], which is also a *proximal splitting* method just without computing the proximity of the indicator function. As reported in [27], FISTA needs only about tens of seconds per mixture, while here *Douglas-Rachford* needs thousands of seconds per mixture, see Table I. As stated in Section IV-C, in Algorithm 2, the variable iteratively moves from the initial point to its projection in the ℓ_2 convex set. Therefore, a large convex set caused by a large noise power (a large ϵ) needs less iterations to reach the projection, and needs less computation time. This is corroborated by the fact that the case with SNR $= 15$ dB needs less computation time than the noise-free case. The CTF convolution at one frequency has a much smaller size than the time-domain convolution. As a result, the CTF-based *Douglas-Rachford* method only requires about ten iterations to converge, while W-Lasso requires tens of thousands iterations to converge. As shown in Table I, W-Lasso needs more computation time than CTF-BP, although it is optimized by FISTA.

VI. Conclusion

Three source recovery methods based on CTF were proposed in this paper. CTF-MINT is an ideal over-determined source recovery method when the noise and mixing filter perturbations are small. It has a relative low computational complexity. However, it is sensitive to noise and filter perturbations. CTF-MPDR is also more suitable for the over-determined case than for the under-determined and determined cases. It achieves the worst performance among the three proposed methods but with the lowest computational cost. The major virtue of CTF-MPDR is that it only requires the mixing filters of the desired source, which makes it more practical. Thanks to exploiting the spectral sparsity, CTF-BP is able to perform well in the under-determined case, and to efficiently reduce the noise. However, it requires the mixing filters of all sources, which are not easy to obtain in practice. In addition, the computational cost is high due to the iterative optimization procedure.

References

[1] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.

[2] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[3] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 320–324.

[4] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. Hoboken, NJ, USA: Wiley, 2004.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[6] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[7] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, 2007, Art. no. 024717.

[8] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[9] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1818–1829, Sep. 2010.

[10] S. Arberet, P. Vandergheynst, J.-P. Carrillo, R. E. Thiran, and Y. Wiaux, "Sparse reverberant audio source separation via reweighted analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1391–1402, Jul. 2013.

[11] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[12] M. Kallinger and A. Mertins, "Multi-channel room impulse response shaping-a study," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2006, vol. 5, pp. V101–V104.

[13] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity- and p-norm optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 249–259, Feb. 2010.

[14] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[15] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 680–693, Apr. 2016.

[16] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, 2007, Art. no. 034013.

[17] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 882–895, Sep. 2005.

[18] J. O. Jungmann, R. Mazur, M. Kallinger, T. Mei, and A. Mertins, "Combined acoustic MIMO channel crosstalk cancellation and room impulse response reshaping," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1829–1842, Aug. 2012.
[19] H. Yamada, H. Wang, and F. Itakura, "Recovering of broadband reverberant speech signal by sub-band MINT method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1991, pp. 969–972.
[20] N. D. Gaubitch and P. A. Naylor, "Equalization of multichannel acoustic systems in oversampled subbands," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1061–1070, Aug. 2009.
[21] F. Lim and P. A. Naylor, "Robust speech dereverberation using subband multichannel least squares with variable relaxation," in *Proc. Eur. Signal Process. Conf.*, 2013, pp. 1–5.
[22] H. Wang and F. Itakura, "Realization of acoustic inverse filtering through multi-microphone sub-band processing," *IEICE Trans. Fundam. Electron., Commun. Comput. Sci.*, vol. 75, no. 11, pp. 1474–1483, 1992.
[23] S. Weiss, G. W. Rice, and R. W. Stewart, "Multichannel equalization in subbands," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 1999, pp. 203–206.
[24] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.
[25] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2009.
[26] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1420–1434, Sep. 2009.
[27] X. Li, L. Girin, and R. Horaud, "Audio source separation based on convolutive transfer function and frequency-domain lasso optimization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 541–545.
[28] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation: Variational inference of time-frequency sources from time-domain observations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 26–30.
[29] S. Leglaive, R. Badeau, and G. Richard, "Separating time-frequency sources from time-domain convolutive mixtures using non-negative matrix factorization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 264–268.
[30] R. Badeau and M. D. Plumbley, "Multichannel high-resolution NMF for modeling convolutive mixtures of non-stationary signals in the time-frequency domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 11, pp. 1670–1680, Nov. 2014.
[31] B. Schwartz, S. Gannot, and E. A. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 2, pp. 394–406, Feb. 2015.
[32] X. Li, L. Girin, and R. Horaud, "An EM algorithm for audio source separation based on the convolutive transfer function," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 56–60.
[33] X. Li, S. Gannot, L. Girin, and R. Horaud, "Multisource MINT using the convolutive transfer function," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proces.*, 2018, pp. 756–760.
[34] P. L. Combettes and J.-C. Pesquet, "A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 564–574, Dec. 2007.
[35] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.
[36] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel Wiener filtering techniques for noise reduction," in *Speech Enhancement*. Berlin, Germany: Springer, 2005, pp. 199–228.
[37] X. Li, S. Gannot, L. Girin, and R. Horaud, "Multichannel identification and nonnegative equalization for dereverberation and noise reduction based on convolutive transfer function," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1755–1768, Oct. 2018.
[38] C. Forbes, M. Evans, N. Hastings, and B. Peacock, "Erlang distribution," in *Statistical Distributions*, 4th ed. Hoboken, NJ, USA: Wiley, 2010, pp. 84–85.
[39] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 2015.
[40] M. J. Fadili and J.-L. Starck, "Monotone operator splitting for optimization problems in sparse recovery," in *Proc. IEEE Int. Conf. Image Process.*, 2009, pp. 1461–1464.
[41] Y. Nesterov, "Gradient methods for minimizing composite objective function," Int. Assoc. Res. Teaching, Center Operations Res. Econometrics (CORE), Catholic Univ. Louvain, Louvain-la-Neuve, Belgium, Tech. Rep. 2007/76, 2007.
[42] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[43] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 313–317.
[44] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *Getting started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*, Nat. Inst. Standards Technol., Gaithersburgh, MD, USA, vol. 107, 1988.
[45] D. R. Morgan, J. Benesty, and M. M. Sondhi, "On the evaluation of estimated impulse responses," *IEEE Signal Process. Lett.*, vol. 5, no. 7, pp. 174–176, Jul. 1998.
[46] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[47] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 749–752.
[48] X. Li, L. Girin, S. Gannot, and R. Horaud, "Non-stationary noise power spectral density estimation based on regional statistics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 181–185.
[49] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and postfiltering system for nonstationary noise environments," *EURASIP J. Appl. Signal Process.*, vol. 2003, pp. 1064–1073, 2003.
[50] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 6, pp. 561–571, Nov. 2004.
[51] A. Moore, P. P. Parada, and P. Naylor, "Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures," *Comput. Speech Lang.*, vol. 46, pp. 574–584, 2017.

Xiaofei Li received the Ph.D. degree in electronics from Peking University, Beijing, China, in 2013. He is currently a Postdoctoral Researcher with the French Computer Science Research Institute, Montbonnot Saint-Martin, France. His research interests include multi-microphone speech processing for sound source localization, separation and dereverberation, single microphone signal processing for noise estimation, voice activity detection, and speech enhancement.

Laurent Girin received the M.Sc. and Ph.D. degrees in signal processing from the Institut National Polytechnique de Grenoble, Grenoble, France, in 1994 and 1997, respectively. In 1999, he joined the Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble, as an Associate Professor. He is currently a Professor with the Physics, Electronics, and Materials Department, Institut Polytechnique de Grenoble, Grenoble, France, where he lectures signal processing theory and applications to audio. His research activity is carried out at the Grenoble Laboratory of Image, Speech, Signal, and Automation, Institut Polytechnique de Grenoble. It deals with speech and audio processing (analysis, modeling, coding, transformation, and synthesis), with a special interest in multimodal speech processing (e.g., audiovisual, articulatory-acoustic, etc.) and speech/audio source separation. He is also a regular collaborator of the French Computer Science Research Institute, as an associate member of the Perception Team.

Sharon Gannot (S'92–M'01–SM'06) received the B.Sc. degree (*summa cum laude*) from the Technion—Israel Institute of Technology, Haifa, Israel, in 1986, and the M.Sc. (*cum laude*) and Ph.D. degrees from Tel-Aviv University, Tel Aviv, Israel, in 1995 and 2000, respectively, all in electrical engineering. In 2001, he was a Postdoctoral Researcher with the Department of Electrical Engineering (ESAT-SISTA), Katholieke Universiteit Leuven, Leuven, Belgium. From 2002 to 2003, he held a research and teaching position with the Faculty of Electrical Engineering, Technion—Israel Institute of Technology. He is currently a Full Professor with the Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel, where he is heading the Speech and Signal Processing Laboratory and the Signal Processing Track. His research interests include multi-microphone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation, dereverberation, single microphone speech enhancement, and speaker localization and tracking.

Prof. Gannot is the recipient of the Bar-Ilan University Outstanding Lecturer Award for 2010 and 2014. He is also a co-recipient of seven best paper awards. He has served as an Associate Editor for the *EURASIP Journal of Advances in Signal Processing* in 2003–2012 and as an Editor for several special issues on Multi-microphone Speech Processing of the same journal. He has also served as a Guest Editor for Elsevier's *Speech Communication* and *Signal Processing* journals. He was an Associate Editor for the IEEE TRANSACTIONS ON SPEECH, AUDIO AND LANGUAGE PROCESSING in 2009–2013. He is a Senior Area Chair of the IEEE TRANSACTIONS ON SPEECH, AUDIO AND LANGUAGE PROCESSING. He also serves as a Reviewer of many IEEE journals and conferences. He has been a member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE since January 2010. Since January 2017, he has been the Committee Chair. He is also a member of the Technical and Steering Committee member of the International Workshop on Acoustic Signal Enhancement (IWAENC) in 2005 and was the General co-Chair of the IWAENC held at Tel Aviv, in August 2010. He has served as the General co-Chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics in October 2013. He was selected (with colleagues) to present a tutorial sessions in the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), the 2012 European Signal Processing Conference (EUSIPCO), ICASSP 2013, and EUSIPCO 2013.

Radu Horaud received the B.Sc. degree in electrical engineering, the M.Sc. degree in control engineering, and the Ph.D. degree in computer science from the Institut National Polytechnique de Grenoble, Grenoble, France. He is currently the Director of research with INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France, where he is the Founder and Head of the PERCEPTION team. His research interests include computer vision, machine learning, audio signal processing, audiovisual analysis, and robotics. He and his collaborators received numerous best paper awards. He was an Area Editor for the *Elsevier Computer Vision and Image Understanding* (1999–2017). He is a member of the Advisory Board of the *Sage International Journal of Robotics Research* and an Associate Editor for the *Kluwer International Journal of Computer Vision*. He was the program co-Chair of the 2001 IEEE International Conference on Computer Vision and the 2015 ACM International Conference on Multimodal Interaction. He was a recipient of an European Research Council (ERC) Advanced Grant for his project *Vision and Hearing in Action* in 2013 and an ERC Proof of Concept Grant in 2017.