Multichannel Source Separation and Speech Enhancement Using the Convolutive Transfer Function

Xiaofei Li, Laurent Girin, Sharon Gannot, Radu Horaud

Multichannel Source Separation and Speech Enhancement Using the Convolutive Transfer Function

Xiaofei Li, Laurent Girin, Sharon Gannot and Radu Horaud

***Abstract*—This paper addresses the problem of audio source recovery from multichannel noisy convolutive mixture for source separation and speech enhancement, *assuming known mixing filters*. We propose to conduct the source recovery in the short-time Fourier transform domain, and based on the convolutive transfer function (CTF) approximation. Compared to the time domain filters, CTF has much less taps, and thus less near-common zeros among channels and less computational complexity. This work proposes three source recovery methods, i) the multichannel inverse filtering method, i.e. multiple input/output inverse theorem (MINT), is exploited in the CTF domain, and for the multisource case, ii) a beamforming-like multichannel inverse filtering method is proposed appling the single source MINT and power minimization, which is suitable for the case that not the CTFs of all the sources are known, iii) a constrained Lasso method. The sources are recovered by minimizing their ℓ_1-norm to impose the spectral sparsity, with the constraint that the ℓ_2-norm fitting cost between the microphone signals and the mixture model involving the unknown source signals is less than a tolerance. The noise can be reduced by setting the tolerance to the noise power. Experiments under various acoustic conditions are conducted to evaluate the three proposed methods. The comparison among them and with the baseline methods are presented.**

***Index Terms*—Source separation, Speech enhancement, CTF, MINT, Lasso**

I. Introduction

Speech recordings in the real world consist of the convolutive images of multiple audio sources and some additive noise. A convolutive image is the convolution between the source signal and the room impulse response (RIR), which is also called mixing filter in the multisource context. Correspondingly, the distortions on the source signals, i.e. interfering speakers, reverberations and additive noise, heavily deteriorate the speech intelligibility for both human listening and machine recognition. This work aims to suppress these distortions, in other words, to recover the respective source signals from the multichannel recordings. In generall, suppressing interfering speakers, reverberations and noise are respectively refered to source separation, dereverberation and noise reduction. Each individual of them is difficult, and attracts lots of research attentions. In the microphone recordings, there are three unknown terms, i.e. source signals, mixing filters, and noise. Thence, the problem is often divided into two subproblems i) identification of mixing filters and noise statistics, ii) and estimate of source signals. This work focuses on the problem of speech source estimate assuming that the mixing filters and possibly the noise statistics are either known or their estimations are available.

X. Li and R. Horaud are with INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France.

L. Girin is with GIPSA-lab and with Univ. Grenoble Alpes, Saint-Martin d'Hères, France.

Sharon Gannot is with Bar Ilan University, Faculty of Engineering, Israel.

This work was supported by the ERC Advanced Grant VHIA #340113.

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 approximated at each time-frequency (TF) bin by a product between the source STFT coefficient and the Fourier transform of the mixing filter. This assumption is called the multiplicative transfer function (MTF) approximation [1], or the narrowband approximation, and the frequency domain mixing filter is called the acoustic transfer function (ATF). Based on the known ATFs, or the relative transfer functions (RTFs) [2], [3], the beamforming technique is widely used for multichannel source separation and speech enhancement, such as the minimum variance/power distortionless response (MVDR/MPDR) beamformer, and the linearly constrained minimum variance/power (LCMV/LCMP) beamformer [2], [4]. Moreover, the sparsity of the audio signals in the TF domain is a desirable effect. Based on this property, the binary masking [5], [6] and the ℓ_1-norm minimization [7] approaches could be applied for source separtion. For more multichannel source separation and speech enhancement techniques based on the narrowband assumption, please refer to a comprehensive overview [8] and references therein.

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

Importantly, the ℓ_1-norm of the STFT domain source signals is added to the mixture fitting cost as a regularization term to impose the spectral 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 knwon. To overcome this, a more flexible scheme is proposed in [10] that relaxes the ℓ_2-norm mixture fitting cost to the noise level and minimizes the ℓ_1-norm. In addition, a reweighting approach is also proposed in [10] to approximate the ℓ_0-norm. In the family of multichannel inverse filtering or multichannel equalization, 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 [11] was first proposed to this aim. It is known that MINT is sensitive to RIR perturbations (misalignment / estimation error) and to microphone noise. To improve the robustness of MINT to the 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], partial MINT [14], [15], etc. In addition, the energy of the inverse filter was used in [16] as a regularization term to avoid the amplification of filter perturbations and microphone noise.

The wide-band models mentioned above are all performed in the time domain. The time domain convolution problem can be somehow transfromed to the subband domain, which provides several benefits i) the original problem is splitted into subproblems, and each subproblem has a smaller data size and thus a smaller computational complexity, ii) the subband mixing filters are shorter than the time-domain filters, thence are likely to have less near-common zeros among microphones, which benefits both the filter identification and the multichannel equalization, even if the former is beyond the scope of this work, iii) in the TF domain, the sparsity of the speech signal can be more easily exploited. Several variant of subband MINT were proposed based on filter bank [17], [18], [19], [20], [21]. The key issues in the filter-bank design are 1) the time-domain RIRs should be well approximated in the subband domain, 2) the frequency response of each filter-bank should be fully excited, i.e. should not involve the frequency components with the 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 either is critically sampled [17], [18], which suffers from frequency aliasing, or has a flat-top frequency response[19], [20], [21], which may suffer from time aliasing. Generally speaking, the STFT transform is more preferable in the sense that most of the acoustic algorithms in the current literature are preformed 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 [22]. To simplify the analysis, the convolutive transfer function (CTF) approximation is further adopted in [23], [24] only using the band-to-band convolution and ignoring the cross-band filters. In [24], CTF is integrated into the generalized sidelobe canceler beamformer. In our previous work [25], a CTF-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 the generative model. A variational Expectation-Maximization (EM) algorithm is proposed in [26], [27] based on the time domain convolution and in [28] based on cross-band filters. CTF-based EM algorithms are proposed in [29] and [30] for single source dereverberation and source separation, respectively. These (variational) EM algorithms iteratively estimate the mixing filters and the sources, and intrinsically require a fairly good initialization for both filters and sources. The output of the methods of this work could be a good initialization for them.

In this work, we propose the following three source recovery methods in the standard oversampled STFT domain using the CTF approximation.

- All the above-mentioned improved MINT methods are proposed for single source dereverberation. The multi-source case has been rarely studied, even if the multi-source MINT was presented in the original paper [11]. We exploit the 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 a 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 the 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.
- For the situation that not the CTFs of all the sources are available, we propose a beamforming-like inverse filtering method. The inverse filters are designed to 1) preserve one source with known CTFs based on single source MINT, and 2) minimize the overall power of the inverse filtering output, and thus suppress the interfering sources and noise. This method shares a similar spirit with the MPDR beamformer.
- To overcome the drawback of the CTF-Lasso method [25] that the regularization factor is difficult to set with respect to the noise level, following the spirit of [10], we propose to recover the source signals by minimizing the ℓ_1-norm of the source spectra with the constraint that the ℓ_2-norm mixture fitting cost are less than a tolerance. The setting of the tolerance is studied. In addition, a complex-valued *proximal splitting* algorithm [31], [32] 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 improved CTF-Lasso method is proposed in Section IV. Experiments are presented in Section V. Section VI concludes the work.

II. CTF-based Problem Formulation

In time domain, we consider a multichannel convolutive mixture with J sources and I microphones,

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

where n is the time index, and $i = 1, \dots, I,\ I \geq 2$ and $j = 1, \dots, J,\ J \geq 2$ are respectively the indices of microphone and source. The signals $x^i(n)$, $s^j(n)$ and $e^i(n)$ are microphone signal, source signal, and noise signal, respectively. Here $\star$ denotes convolution, and $a^{i,j}(n)$ is the RIR relating the j-th source to the i-th microphone. Note that the relation between I and J is not specified here, and this will be discussed afterwards with respect to the respective proposed methods. The noise signal $e^i(n)$ is uncorrelated to the source signals, and could be spatially uncorrelated or diffuse, or directional.

The goal of this paper is to recover the mutiple source signals from the microphone signals, given the RIRs and the noise PSDs. The RIRs and noise PSDs should be blindly estimated from the microphone signals and suffer from some disturbance, which are not trivial but beyond the scope of this work. Overall, the multi-source recovery simultaneously conducts source separation, dereverberation, and noise reduction.

A. Convolutive Transfer Function

In this subsection, the time domain convolution would be transformed to the CTF convolution in the STFT domain. To simplify the analysis, we consider the noise free situation with only one microphone and one source: $x(n) = a(n) \star s(n)$, in which the source index and microphone index are omitted.

The STFT representation of the microphone signal $x(n)$ is

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

where p and k denote the frame index and the frequency index, respectively. $\tilde{w}(n)$ is an analysis window, and N and D denote the frame (window) length, and the frame step, respectively. Equation (2) gives the overlap-add view of STFT. 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 [22] consists in representing the STFT coefficient $x_{p,k}$ as a summation over multiple convolutions (between the STFT-domain source signal and filter) across frequency bins. Mathmatically, the linear time invariant system can be written in the STFT domain as

$$x_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'} s_{p-p',k'}\ a_{p',k,k'}, \tag{3}$$

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 ceiling function. The number of causal filter coefficients is related to the reverberation time. For notational simplicity, let the filter index p' be in $[0, L_a - 1]$, with L_a being the filter length, i.e. 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:

$$a_{p',k,k'} = (a(n) \star \zeta_{k,k'}(n))|_{n=p'D}, \tag{4}$$

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:

$$x_{p,k} \approx \sum\nolimits_{p'=0}^{L_a-1} s_{p-p',k}a_{p',k} = s_{p,k} \star a_{p,k}. \tag{5}$$

B. STFT Domain Mixture Model

Based on the CTF approximation, we can obtain the STFT domain mixture model corresponding to the time domain model (1),

$$x^i_p = \sum_{j=1}^{J} a^{i,j}_p \star s^j_p + e^i_p, \tag{6}$$

Note that here (and hereafter) the frequency index k is omitted, since the proposed methods are applied frequency-wise until the final inverse STFT. Let $p \in [1, P]$ and $p \in [0, L_a - 1]$ respectively denote the frame indices of the microphone signals and the CTFs.

III. 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^i_p with $i = 1, \dots, I$ and $p = 0, \dots, L_h - 1$, where L_h denotes the length of the inverse filters. The output of the inverse filtering is

$$y_p = \sum_{i=1}^{I} h^i_p \star x^i_p = \sum_{j=1}^{J} s^j_p \star (\sum_{i=1}^{I} h^i_p \star a^{i,j}_p) + \sum_{i=1}^{I} h^i_p \star e^i_p, \tag{7}$$

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. Define the convolution matrix for the microphone signal x_p^i as

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

and the vector of filter 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 the transpose of vector or matrix. Then the convolution $h_p^i \star x_p^i$ can be written as $\mathbf{X}^i\mathbf{h}^i$. The inverse filtering (7) can be further written as

$$\mathbf{y} = \mathbf{X}\mathbf{h}, \quad (9)$$

where

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

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

B. CTF-MINT

To preserve a desired source, e.g. the j_d-th source, the inverse filtering of the CTF filters, i.e. $\sum_{i=1}^I h_p^i \star a_p^{i,j_d}$, generally should target to an impulse function function d_p with the length of L_a+L_h-1. To suppress the interfering sources, the 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 to 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

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

or

$$\mathbf{g} = \mathbf{A}\mathbf{h}. \quad (10)$$

When the matrix $\mathbf{A} \in \mathbb{C}^{J(L_a+L_h-1)\times IL_h}$ is 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 overdetermined recording system, i.e. $I > J$.

From [11], the solvable condition of (10) is that the CTFs of the desired source $a_p^{i,j_d}, i = 1, \ldots, I$ do not have any common zero. The subband filters, i.e. CTFs, are much shorter than the time domain filters, and are thus likely to have much less near-common zeros, which is a major benefit. Unfortunately, the filter banks induced from the short-time windows will lead to some structured common zeros. From (4), for any RIR $a^{i,j}(n)$, its CTF (with $k' = k$) is computed as

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

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 $e^{j\frac{2\pi}{N}kn}$. This cross-correlation has a similar frequency response as the windows $\tilde{w}(n)$ and $w(n)$ in the sense that it is also a low-pass filter with the same bandwidth denoted by $\bar{\omega}$. 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 $\zeta_{p,k} = \zeta_k(n)|_{n=pD}$. One can see $\zeta_{p,k}$ is obtained by downsampling $\zeta_k(n)$ by the decimation factor D. The downsampling operation folds the frequency response with the period of $2\pi/D$. To avoid the frequency aliasing, the period should not be smaller than the bandwidth $\bar{\omega}$. 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 $D \leq N/4$. 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. This phenomenon that the frequency response of $\zeta_{p,k}$ and thus of $a_{p,k}^{i,j}$ are not fully excited is common to all microphones, which is problematic for solving (10). Fortunately, it is trivially known that the common zeros are introduced by the frequency response of $\zeta_{p,k}$. To make (10) solvable, we propose to determine the desired target $\mathbf{d}$ to have the same frequency response as $\zeta_{p,k}$, instead of the impulse function that has a full-band frequency response. To this end, the target $\mathbf{d}$ is designed as

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

where $\boldsymbol{\zeta}$ denotes the vector form of $\zeta_{p,k}$. 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.

The solution of (10) gives an exact recovery of the j_d-th source plus the filtered noise $\sum_{i=1}^I h_p^i \star e_p^i$ as shown in (7). In this method, a directional noise is considered as an interfering source, and is modeled in the MINT formulation. Therefore, noise e_p^i is usually temporally white, and spatially uncorrelated or diffuse. To suppress the noise, a straightforward way is to minimize the power of the filtered noise under the MINT constraint (10). As proposed in [16], an alternative way to suppress the noise is to reduce the energy of the inverse filter $\mathbf{h}$. This strategy is equivalent to minimizing the power of the filtered noise if we approximately assume the noise correlation matrix is identity. In addition, this strategy is also able to suppress the perturbations of the CTFs, if the disturbance noise is also assumed to have an identity correlation matrix. This leads to the following optimization problem

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

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 the regularization factor. 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 unrelated to the power level of the CTF filters. This property is especially useful and necessary for the present frequency-wise algorithm where all the frequencies can share the same regularization factor δ, although the CTFs level may significantly vary along the frequencies. The solution of (13), i.e. the inverse filter, 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. We refer to this method as CTF-MINT.

As mentioned above, to perform the exact inverse filtering, the matrix $\mathbf{A}$ should be square or wide. In (13), the exact match between $\mathbf{A}\mathbf{h}$ and $\mathbf{g}$ is relaxed, which means the exact inverse filtering is abandoned to improve the robustness of the inverse filters estimate. Let ρ denotes the ratio between the number of columns and the number of rows of $\mathbf{A}$, then we have $IL_h = \rho J(L_a + L_h - 1)$. To specify the method, write L_h as L_h^{mint}, then

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

For the overdetermined recording system, i.e. $I > J$, it can be set $\rho \geq 1$ to have a square or wide $\mathbf{A}$. When $I \leq J$, ρ should be less than $\frac{I}{J}$, consequently $\mathbf{A}$ is narrow, however, the optimization problem (13) is still feasible. 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. CTF-MPDR

The above CTF-MINT approach requires the CTFs knowledge of all the considerable sources. In this subsection, we consider the situation that not the CTFs of all the sources can be obtained/estimated. One source is recovered based on only its own CTFs.

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

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

where $\phi_x = \sum_{i=1}^{I} \sum_{p=0}^{P-1} |x_p^i|^2$ is the energy of the microphone signals. Similar to CTF-MINT, the normalization factor $\frac{\phi_a^{j_d}}{\phi_x}$ makes the choice of the regularization factor κ unrelated to the power level of the CTF filters and the power level of the microphone signals. Therefore, all the frequencies can share the same regularization factor κ, even if the level of microphone signals significantly vary along the frequencies. This optimization problem considers any type of noise signal equally by minimizing the overall ouput power.

The solution of (16), i.e. the inverse filter, is

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

This method shares a similar spirit with the MPDR beamformer, more exactly with the speech distortion weighted multichannel wiener filter [33] since the source distortionless is relaxed, we still refer to this method as CTF-MPDR.

Similarly, let ϱ denote the ratio between the number of columns and the number of rows of $\mathbf{A}^{j_d}$, then we have $IL_h = \varrho(L_a + L_h - 1)$. Write 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 filters are constrained by only one source, i.e. the desired source, it can always be set $\varrho \geq 1$ to have a square or wide $\mathbf{A}^{j_d}$.

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

IV. CTF-based Constrained Lasso

Instead of explicitly estimating an inverse filter, the source signals can be directly recovered by matching the microphone signals and the mixture model involving the unknown source signals. To this end, the spectral sparsity of the speech signals could be exploited as the *prior* knowledge.

A. Problem Formulation for the Mixture Model

The mixture model (6) can be 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 sensor signals, source signals and noise signals, respectively, and $\mathcal{A} \in \mathbb{C}^{I\times J\times P}$ denotes the three-way CTF array. In the multichannel inverse filtering methods, because of the use of the ℓ_2-norm of the filtered signal and the system inverse, the convolution between two signals is formulated as the multiplication of the convolution matrix of one signal and the vector form of the other signal. Unlike this way, here the convolution operator $\star$ is preserved. The reason is, in this section, the methodology proposal is in the first-order, namely only the convolution operation itself and its adjoint are used. The convolution operation can be realized by the fast Fourier transform, which is computationally simple.

In our previous work [25], it is proposed to estimate the source signals by sloving a ℓ_1-norm regularized ℓ_2-norm mixture fitting cost minimization problem

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

where λ is the regularization factor. Note that both the ℓ_2 and ℓ_1-norms on matrices are defined here as vector norms. The first term is to minimize the mixture fitting cost, and the second term is to impose the sparsity on the speech source signals. In the presence of the additional noise $\mathbf{e}$, the regularization factor λ can be adjusted to impose the sparsity and thus to remove the noise from the estimated source signals. However, it is difficult to automatically tune λ even though the noise PSD is known. Especially, the source recovery is performed frequency by frequency in this work, and it is common that the noise PSDs are different at different frequencies. This requires an unique λ for each frequency, which further increases the difficulty of the setting of λ.

In this work, we solve this problem by transforming the above problem to a constrained optimization problem.

B. CTF-based Constrained Lasso

The problem (20) is equivalent to the following formulation

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

for some unknown λ and ϵ. In the constraint, the ℓ_2-norm mixture fitting cost is relaxed to at most a tolerance ϵ. This formulation is first proposed in [10] for audio source separation in time domain. We adapted it to the CTF-magnitude domain in our previous work [34] for single source dereverberation. In this work, we further extend it to the complex-valued CTF domain for multisource recovery.

The setting of the tolerance ϵ is critical to the quality of the recovered source signals. The tolerance ϵ is related to the noise power in the microphone signals. The noise signal is assumed to be stationary. Let σ_i^2 denote the noise PSD in the i-th microphone, which can be estimated from the pure noise signal or estimated by a noise PSD estimator, e.g. [35]. Let $\mathbf{e}^i \in \mathbb{C}^{1\times P}$ denotes the noise signal in the i-th microphone in vector form. The squared ℓ_2-norm of the noise signal $\parallel \mathbf{e}^i \parallel^2$, i.e. the energy, follows an Erlang distribution with mean $P\sigma_i^2$ and variance $P\sigma_i^4$. Assume the noise signal in the different microphones are uncorrelated, then for all the microphones, the squared ℓ_2-norm $\parallel \mathbf{e} \parallel^2$ has the mean $\sum_{i=1}^I P\sigma_i^2$ and variance $\sum_{i=1}^I P\sigma_i^4$. To relax the ℓ_2 fitting cost to the noise power, we set the noise relaxing term as

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

Here two times the standard deviation is subtracted, because i) this makes the probability that the ℓ_2 fitting cost being larger than $\parallel \mathbf{e} \parallel^2$ very small. When the ℓ_2 fitting cost is allowed to be larger than $\parallel \mathbf{e} \parallel^2$, the minimization of $|\mathbf{s}|$ will distort the source signal. Here we prefer less source signal distortion on more noise reduction, ii) the minimization of $|\mathbf{s}|$ tends to make the residual noise in the estimated source signals sparse. The sparse noise is perceptually not fine even the noise power is low. As a result, some noise remains in the estimated source signal. Note that this method depends on the spectral sparsity of the source, therefore the possible directional noise is not considered as a source. Even for the directional noise, this method needs only an estimation of the single channel noise auto-PSD, but not the cross-PSD among microphones or frames.

Besides, the ℓ_2 fit should also be relaxed with respect to the CTF approximation error and the CTF filters perturbations. The tolerance is akin to the energy of the noise-free signal, which can be estimated by spectral subtraction as

$$\hat{\Gamma}_s = \max(\parallel \mathbf{x} \parallel^2 - \sum\nolimits_{i=1}^I P\sigma_i^2,\ 0). \tag{23}$$

Empirically, the tolerance with respect to the noise-free signal is set to $\epsilon_s = 0.01\hat{\Gamma}_s$. Overall, the tolerance is set to $\epsilon = \epsilon_e + \epsilon_s$.

Thanks to the sparsity imposing, the optimization problem (21) is feasible for any case in terms of the relation between I and J. We refer to this method as CTF-based Constrained Lasso (CTF-C-Lasso).

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 proposed in the complex domain, and thus the optimization algorithm is also complex-valued. The optimization problem consists of a 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 written as the following unconstrained optimization problem

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

where C denotes the convex set of the constraint, i.e. $C = \{\mathbf{s} \mid \parallel \mathcal{A}\star\mathbf{s}-\mathbf{x}\parallel^2\leq \epsilon\}$, 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 (nonsmooth), convex functions. For this problem, the *Douglas-Rachford* splitting approach

[31] is suitable, which is an iterative method, at each iteration, the two functions are splitted, and their proximity operators $\text{Prox}_{\iota_C(\cdot)}$ and $\text{Prox}_{\gamma|\cdot|}$ are individually applied. The *Douglas-Rachford* approach does not require the differentiability of any of the two functions, and is a generalization of the *proximal splitting* method [32]. Algorithm 1 summarizes the *Douglas-Rachford* approach. Here α and γ are set as constant values over iterations, e.g. respectively 1 and 0.01 in the experiments, for all iterations. The initialization of $\mathbf{s}_0$ is set as the matrix composed of J replication of the first microphone signal.

The convergence criteria is set to check if the optimization objective is almost invariant, i.e. $||\mathbf{s}_k| - |\mathbf{s}_{k-1}||/|\mathbf{s}_k| < \eta_1$, where η_1 is a threshold, e.g. 0.01 in the experiments. In addition, the maximum number of iterations is set to 20.

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

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

The proximity operator of the ℓ_1-norm $\gamma|\cdot|$ at point $\mathbf{z}$, aka the shrinkage operator, is given entrywise 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 [36], an iterative method has been derived in [37], and used in [10]. However, this method converges linearly, which is very slow especially for the case that the convex set C (also ϵ) is small. As hinted in [37], it can be accelerated to the squared speed via the Nesterov's scheme [38], [39]. The accelerated method is summarized in Algorithm 2. The acceleration procedure is composed of Step 3 and 4, which is based on the derivation in [39]. 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 $\mathcal{A}^* \circ \mathcal{A}$. The power iteration method is used to compute ν, which is summarized in Algorithm 3. We set μ as a constant value over iterations, e.g. $1/\nu$ in the experiments. In Step 2, the *projection* of a variable $\mathbf{u}$ onto the convex set $\{\mathbf{v} \mid \parallel \mathbf{v} \parallel^2 \leq \epsilon\}$ can be easily obtained as

$$\text{Prox}_{\iota_{\|\cdot\|^2<\epsilon}}(\mathbf{u}) = \min(1, \frac{\sqrt{\epsilon}}{\parallel \mathbf{u} \parallel})\mathbf{u}. \tag{27}$$

In Algorithm 2, the variable $\mathbf{p}_k$ iteratively moves from the initial point $\mathbf{s}$ to its *projection*, thence the convergence criteria is set to check the feasibility of the constraint, namely if $\parallel \mathcal{A} \star \mathbf{p}_k - \mathbf{x} \parallel^2 \leq 1.1 * \epsilon$, then stop. Here 1.1 is a slack factor. In addition, the maximum number of iterations is set to 300.

Algorithm 1 Douglas-Rachford

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

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

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

V. Experiments

Appling the inverse STFT to the frequency-wise output of the three methods, i.e. the outputs of (7) and (21), an estimation of the time domain source signals can be obtained. In this section, we evaluate the quality of the estimated source signals, in terms of the performance of source separation, speech dereverberation and noise reduction.

A. Experimental Configuration

1) Dataset: The multichannel impulse response data [40] is used, which is recorded using a 8-channel linear microphone array in the speech and acoustic lab of Bar-Ilan University, with room size of 6 m × 6 m × 2.4 m. The reverberation time is controlled by 60 panels covering the room facets. In the reported experiments, we used the recordings with $T_{60} = 0.61$ s. The RIRs are truncated to correspond to T_{30}, and has a length of 5600 samples. The speech signals from the TIMIT dataset [41] are taken as the source signals, with the duration of about 3 s. A TIMIT speech is convolved with a RIR as the image of one source. Multiple image sources are summed as a mixture. For one mixture, the source direction and the microphone-to-source distance of each source are randomly selected from $-90°$:$15°$:$90°$ and $\{1, 2\}$ m, respectively. Note that the mutiple sources consist of different TIMIT speech utterances and different impulse responses in terms of the source directions. To generate the

Algorithm 3 Power Iteration

Input: $\mathcal{A}$, $\mathcal{A}^*$
Initialization: $\mathbf{v} \in \mathbb{C}^{I \times P}$
repeat
  $\mathbf{w} = \mathcal{A}^* \star (\mathcal{A} \star \mathbf{v})$
  $\mathbf{v} = \mathbf{w} / \parallel \mathbf{w} \parallel$
until convergence
Output: $\nu = \parallel \mathbf{w} \parallel$

noisy microphone signal, a spatially uncorrelated stationary speech-like noise is added to the noise-free mixture, the noise level is controlled by a wide-band 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 random Gaussian 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 noise level is denoted as the normalized projection misalignment (NPM) [42] in decibels (dB), i.e.

$$\text{NPM} = 10\log_{10}\frac{\sum_n (a^{i,j}(n) - \tilde{a}^{i,j}(n))^2}{\sum_n a^{i,j}(n)^2}.$$

Various acoustic conditions in terms of the number of microphones and sources, SNRs, and NPMs are tested. For each condition, 20 runs are performed, and the averaged performance measures are computed.

2) performance Metrics: The signal-to-distortion ratio (SDR) [43] 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-C-Lasso, are evaluated as the output scores.

The signal-to-interference ratio (SIR) [43] in dB is used to specially evaluate the source separation performance. This metric focuses on the suppression of interfering sources, thence the additive noise would be 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 can simply take the noise-free output, i.e. $\sum_{i=1}^{I} h_p^i \star (\sum_{j=1}^{J} a_p^{i,j} \star s_p^j)$ in (7), for evaluation. However, for CTF-C-Lasso, we have to test the overall outputs, since the noise-free output is not available. Experimental results show that CTF-C-Lasso has low residual noise, thus the SIR measure is assumed to not be significantly influenced, and to be believable.

The perceptual evaluation of speech quality (PESQ) [44] is used to specially evaluate the dereverberation performance. The interfering sources and noise would be eliminated. For each source, its unprocessed image sources, 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-C-Lasso, 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 PESQ scores of CTF-C-Lasso are highly underestimated.

The output SNR in dB is used to evaluate the noise reduction performance. The input SNR is taken as the baseline scores. For CTF-MINT and CTF-MPDR, the output SNR is computed as the power ratio between the noise-free outputs and the output noise, i.e. $\sum_{i=1}^{I} h_p^i \star e_p^i$. For CTF-C-Lasso, the noise PSDs in the output signals are first blindly estimated using the method proposed in [35]. 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. It is shown in [35] that the estimation error of noise PSD is around 1 dB, thence the estimated output SNRs are reliable.

SDR, SIR and PESQ are evaluated in the time domain, thence the signals mentioned above are actually their corresponding time domain signals. The output SNR for CTF-MINT and CTF-MPDR are computed either in the time domain or in the STFT-domain, while the output SNR for CTF-C-Lasso is clearly computed in the STFT domain.

3) Parameters Setting: The sampling rate is 16 kHz. STFT takes the Hamming window, with the 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 29. 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. The pilot experiments show that a longer inverse filter (or a larger ρ) does not noticeably improve the preformance measures, while leads to a larger computational cost. For the case of $I \leq J$, ρ is set to be less than and close to $\frac{I}{J}$, and ρ should be small to avoid an unreasonabe long inverse filter. The exact values of ρ will be given in the following experiments concerning the specific value 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 $\boldsymbol{d}$ is related to the length of the inverse filters. In the experiments, it is respectively set to 6 and 3 taps for CTF-MINT and CTF-MPDR as a good tradeoff for the different inverse filter lengths in various acoustic conditions.

Thanks to the normalization factors in (13) and (16), the same regularization factors δ and κ are suitable for all frequencies, moreover are robust to any possible numerical scales of the filters and the signals in different datasets. Fig. 1 shows the performance measures of CTF-MINT and CTF-MPDR as a function of δ and κ, respectively. For CTF-MINT, with the increase of δ, the inaccuracy of inverse filtering increases, while the energy of the inverse filters decreases. From the *left* figure, it is observed that the output SNR gets larger with the increase of δ, which verifies that the additive noise can be suppressed by decreasing the energy of the inverse filter. However, SIR and PESQ scores become smaller with the increase of δ due to the larger inaccuracy of inverse filtering, which leads to more residual interfering source and reverberation. Integrating these effects, SDR first increases then decreases with the increase of δ. In a similar way, the energy of the inverse filters also affects the robustness of the inverse filtering to the CTF perturbations. In summary, we consider two representative choices of δ i) a relative small one, i.e. 10^{-5}, leads to an accurate inverse filtering but a large energy of the inverse filter. This is suitable for the case that both the microphone noise and the CTF perturbations are small, ii) a large one, i.e. 0.1, achieves an output SNR being slightly larger than the input SNR to avoid the amplification of the additive noise. In the following experiments, the former is used for the noise-free case, and the latter is used for the noisy case. This partially oracle configuration is a bit unreal,

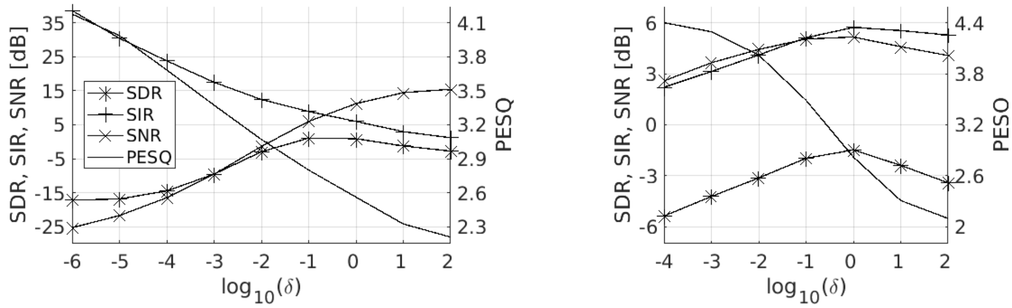

Fig. 1: The performance measures as a function of *left* δ for CTF-MINT and *right* κ for CTF-MPDR. $I = 4$ and $J = 3$. The input SNR is 10 dB. SDR, SIR and PESQ of the unprocessed signals are -6.9 dB, -3.0 dB and 1.85, respectively. Two vertical axes are used due to the different scales and units of the performance measures.

but will show the full potentials of CTF-MINT. See [14] for further discussion on the optimal setting of δ.

For CTF-MPDR, κ controls the tradeoff between the distortionless of the desired source and the power of the output. The minimization of the power of the output will suppress both the interfering sources and the noise. From the *right* figure, we observe that PESQ decreases along with the increasing of κ, due to the increased distortions of the desired source. SIR and output SNR can be increased by increasing κ until $\kappa = 1$. A larger κ, e.g. 10^2, leads to a smaller SIR and output SNR although the power of the output is smaller, since the desired signal is also heavily distorted and suppressed. Overall, κ is set to 0.1, which achieves a high PESQ score and good other measures.

B. Influence of the Number of Microphones

Fig. 2 shows the results as a function of the number of microphones. The source number is fixed to 3. The microphones 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 2- and 3-microphone, respectively. Consequently the length of the inverse filters are about 5 times the CTF length.

For CTF-MINT, the scores of all the three metrics dramatically decrease when the microphone number goes from 4 to 3 and to 2, namely from the overdetermined case to the non-overdetermined case. This indicates that the inaccuracy of the inverse filtering is large for the non-overdetermined case, due to the insufficient spatial freedom of the inverse filters. CTF-MPDR suppresses the interfering sources by minimizing the power of the output, and implicitly also by the inverse filtering with a target of zero signal. Therefore, as for CTF-MPDR, the metrics to measure the interfering sources suppression performance, i.e. SDR and SIR, also significantly degrade for the non-overdetermined case. Along with the increase of microphone number, the PESQ score slightly varies, which means 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-C-Lasso almost linearly increases with the increase of microphone number, no matter it is underdetermined or overdetermined, thanks to the spectral sparsity exploiting. For the overdetermined case, SDR and SIR of the three methods slowly increase with the increase of microphone number, and CTF-MINT has a reletive larger changing rate. CTF-C-Lasso achieves the worst PESQ score due to the influence of the residual interfering sources. By listening to the outputs of CTF-C-Lasso, it is not perceived more reverberation.

Overall, without considering the noise reduction, CTF-MINT performs the best for the overdetermined case. For instance, CTF-MINT achieves a SDR of 21.9 dB by using 4 microphones, which is close to a perfect source recovery. CTF-C-Lasso performs the best for the underdetermined case. For instance, CTF-C-Lasso achieves a SDR of 8.4 dB by using only two microphones. By using only the mixing filters of one source, the source separation performance of CTF-MPDR is worse than the other two.

C. Performance for Various Numbers of Sources

Fig. 3 shows the results as a function of the number of sources. The microphone number is fixed to 6. The microphones signals are noise free, thus the output SNR is not reported. From the figure, we can observe that the performance measures of the three methods degrade with the increase of source number, except for the PESQ score of CTF-MPDR. CTF-MINT achieves the best performance, while also has the largest performance degradation. This is somehow consistent to the experiments with various numbers of microphones that a good performance requires a large microphone number to source number ratio. CTF-MPDR and CTF-C-Lasso have a smaller performance degradation. At first sight, it is surprising that CTF-MPDR achieves a larger PESQ score when more sources exist. The reason is that the normalized output power, i.e. $\frac{\phi_a^{j_d}}{\phi_x} \parallel \mathbf{X}\mathbf{h} \parallel^2$, becomes smaller with the increase of source number due to a larger ϕ_x, correspondingly the inverse filtering inaccuracy of the desired source, i.e. $\parallel \mathbf{A}^{j_d}\mathbf{h} - \mathbf{d} \parallel^2$, becomes smaller as well.

D. Influence of Additive Noise

Fig. 4 shows the results as a function of input SNRs. The number of microphones and sources are respectively fixed to 4

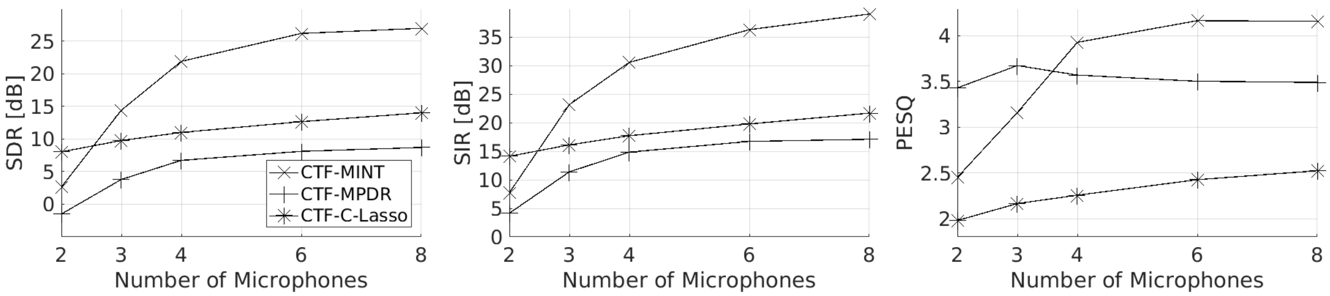

Fig. 2: The performance measures as a function of the number of microphones. $J = 3$. The microphone signals are noise free. SDR, SIR and PESQ of the unprocessed signals are -6.9 dB, -3.0 dB and 1.85, respectively. Note that the legends in this figure are common to all the following figures.

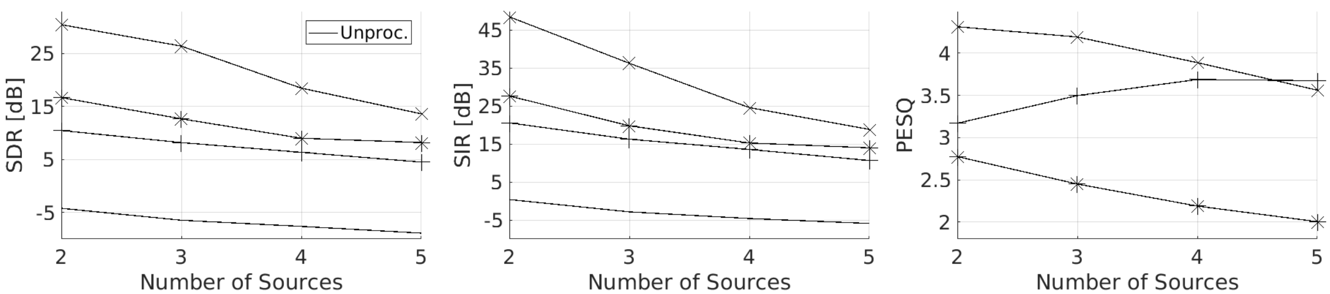

Fig. 3: The performance measures as a function of the number of sources. $I = 6$. The microphone signals are noise free. PESQ of the unprocessed signals is 1.85.

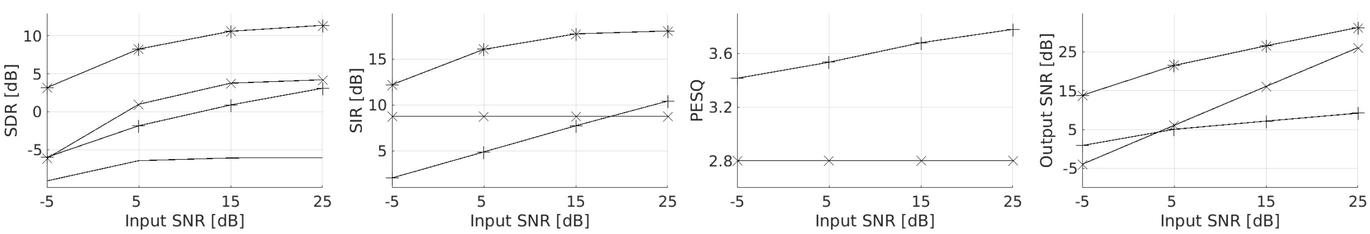

Fig. 4: The performance measures as a function of input SNRs. $I = 4$, $J = 3$. SIR and PESQ of the unprocessed signals are -3.0 dB and 1.85, respectively. PESQ for CTF-C-LASSO is not presented since it is inaccurate due to the residual noise.

TABLE I: The SDR scores and the computation times for 6 representative acoustic conditions. The SDR scores of the unprocessed signals were given in the previous experiments.

Acoustic Condition				SDR [dB]						Computation Time per Mixture [s]					
I	J	SNR	NPM	CTF-MINT	CTF-MPDR	CTF-C-Lasso	LCMP	TD-MINT	W-Lasso	CTF-MINT	CTF-MPDR	CTF-C-Lasso	LCMP	TD-MINT	W-Lasso
4	3	-	-	21.9	6.7	11.0	-3.6	-	18.9	25.4	4.9	1987	1.1	-	4284
6	2	-	-	30.4	10.4	16.6	-0.3	30.0	31.2	5.8	4.2	1688	1.1	142	3843
6	3	-	-	26.3	8.2	12.6	-0.6	-	23.8	12.2	5.9	2827	1.2	-	5961
6	5	-	-	13.6	4.5	8.2	-6.4	-	14.7	229.6	12.4	5679	1.9	-	10134
4	3	15 dB	-	3.8	0.9	10.6	-14.7	-	-	21.9	6.7	1500	1.1	-	-
4	3	-	-15 dB	1.7	-4.3	4.2	-4.1	-	0.5	21.9	6.7	1440	1.1	-	4245

and 3. As mentioned above, for noisy case, the regularization factor δ is set to 0.1. The inverse filter of CTF-MINT is invariant for various input SNRs, since it depend only on the CTF filters, but not on the microphone signals. As a result, the SIR and PESQ score are constant, but are much smaller than the noise-free case with $\delta = 10^{-5}$, see Fig. 2. The SNR improvement is also a constant value about 1 dB. For CTF-MPDR, SIR and PESQ are smaller when the input SNR is lower, since the 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. Along with the increase of the input SNR, the output SNR increases, but the SNR improvement decreases. The SNR improvement is negative when the input SNR is larger than 5 dB, which means the microphone noise is amplified. For CTF-MINT and CTF-MPDR, the residual noise is significant, which indicates that the inverse filtering is not able to efficiently suppress the white noise. Therefore, a single channel noise reduction process is needed as a postprocessing, as in [45], [46]. The output SNRs of CTF-C-Lasso are always larger than the input SNRs, which means the microphone noise is efficiently reduced. SDR and SIR of CTF-C-Lasso degrades for the low SNR case, but not

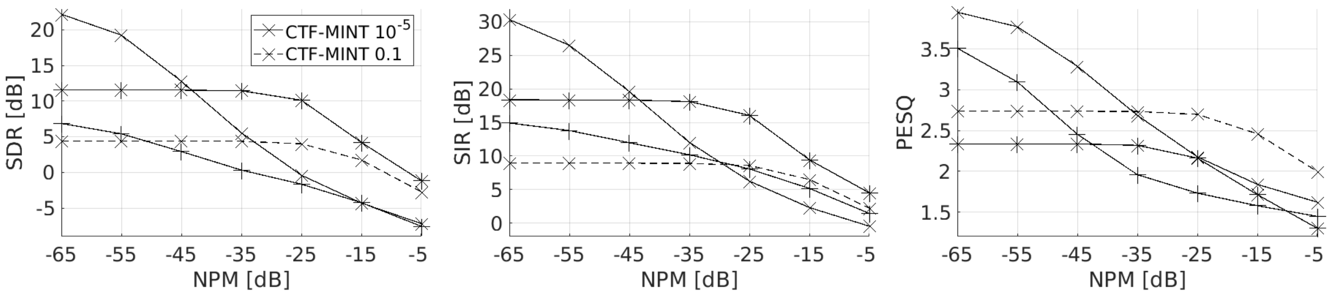

Fig. 5: The performance measures as a function of NPM. $I = 4$, $J = 3$. The microphone signals are noise free. SDR, SIR and PESQ of the unprocessed signals are -6.9 dB, -3.0 dB and 1.85, respectively.

much.

E. Influence of CTF Perturbations

Fig. 5 shows the results as a function of NPMs. For CTF-MINT, two choices of the regularization factor, i.e. 10^{-5} and 0.1, are both tested. As expected, all the metrics become worse with the increase of NPM, thus we only analysis the SDR scores. Note that, when NPM is -65 dB, the three methods achieve almost the same performance measures with the perturbation-free case. Along with the increase of NPMs, 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. In contrast, CTF-MINT with $\delta = 0.1$ 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 relative large degradation rate. The performance of CTF-C-Lasso is stable until NPM equals -35 dB, and quickly degrades when NPM is larger than -25 dB.

In CTF-MINT, the inverse filter is designed to respectively satisfy the targets of 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 depends 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-C-Lasso, all sources are simultaneously recovered based on the CTFs 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. Comparison with Baseline Methods

To benchmark the proposed methods, we compare them with three baseline methods i) LCMP beamformer [4] based on the narrowband assumption. Based on the steering vectors and the correlation matrix of 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 is computed as the Fourier transform of the truncated RIRs, in this experiment is set to the CTF tap with the largest power, ii) 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. We only test the condition with $I = 6$ and $J = 2$, following the principle of the proposed method, the length of inverse filter and the modeling delay are set to 2800 and 1024, respectively. For other conditions with longer inverse filters require a large memory resources that our current computer can not afford, iii) wideband Lasso (W-Lasso) [9]. The regularization factor is set to 10^{-5}, which is empirically suitable for the noise-free case.

Table I presents the SDR scores and the computation time for 6 representative acoustic conditions. The computation time will be analyzed in the next subsection. Note that '-' respectively means noise-free and perturbation-free in the columns of SNR and NPM. LCMP performs poorly for all conditions, which verifies the assertion that the narrowband assumption is not suitable for the long RIR case. CTF-MINT achieves a bit larger SDR score than TD-MINT, despite of the fact that the CTF-based filtering is an approximation of the time-domain filtering. This is mainly due to much shorter filters in the STFT/CTF domain. W-Lasso noticeably outperforms CTF-C-Lasso for the noise-free and perturbation-free cases, due to its exact time domain convolution. W-Lasso has a similar noise reduction capability with CTF-C-Lasso, however the regularization factor is difficult to set for a proper noise reduction, thence the results of W-Lasso for the noisy case is not reported. Compared to CTF-C-Lasso, W-Lasso has a larger performance degradation rate with the increase of the source number and filter perturbations.

G. Analysis of Computational Complexity

Table I also presents the averaged computation time for one mixture with duration of 3 s. All methods were implemented in MATLAB. CTF-MINT and CTF-MPDR comprise the computation of the inverse filters and the inverse filtering on the microphone signals, and the former dominates the computation cost. From (14) and (17), the computations include the multiplication and inverse of the matrices, thence the complexity is cube of the dimensional of matrices. We consider the square matrices $\mathbf{A}$ for (14) and $\mathbf{A}^{j_d}$ for (17), whose dimensionals equal to IL_h. From (15) and (18), IL_h is proportional to the filter length L_a, and to $\frac{I-J}{IJ}$ for CTF-MINT and $\frac{I}{I-1}$ for CTF-MPDR. The inverse filters are respectively computed for each source and each frequency. Overall, CTF-MINT and CTF-MPDR respectively have the computational complexity of $\mathcal{O}(\frac{KL_a^3I^3J^4}{(I-J)^3})$ and $\mathcal{O}(\frac{KL_a^3I^3J}{(I-1)^3})$, where $K = N/2+1$ is the frequency number. 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. The LCMP beamformer is similar to CTF-MINT, just using an instantaneous steering vector and an instantaneous inverse filters, namely the length of CTF and inverse filters are both 1, thence it has the least computation complexity. These methods have a close-formed solution and thus low computational complexity except for TD-MINT due to its long RIRs and inverse filters in the time domain. These can be approximately verified by the computation times shown in Table I.

The iterative optimization of CTF-C-Lasso leads to a high computational complexity. Unlike the Newton-style methods employing the second-order derivative, the *Douglas-Rachford* optimization approach is a first-order method, thence the complexity is linearly related 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 evidence this, we can compare the *Douglas-Rachford* approach with the optimization algorithm for the Lasso problem (20) that does not have an ℓ_2-norm constraint and thus an indicator function. In [25], we solve the unconstrained Lasso problem using the fast iterative shrinkage-thresholding algorithm (FISTA) [39], which is also a *proximal splitting* method just without computing the proximity of the indicator function. As reported in [25], FISTA needs only about tens of seconds per mixture, while *Douglas-Rachford* needs thousands of seconds, 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 larger convex set caused by a larger noise power (a larger ϵ) needs less iterations to reach the *projection*, and needs less computation time. This can be verified by the condition with SNR of 15 dB. When the CTF perturbations is large, e.g. NPM is -15 dB, the optimized objective, i.e. $|\mathbf{s}|$, is large, thence less iterations and lower time consumption are needed to converge. The time domain W-Lasso method needs a much higher computation time than CTF-C-Lasso, although it is unconstrained and optimized by FISTA. This indicates that the proposed CTF-based method is advantageous in terms of computational complexity by splitting the time domain problem into the TF domain subproblems. By the way, to converge, the CTF-based *Douglas-Rachford* approach needs only tens of iterations, while the time domain W-Lasso needs ten-thousands of iterations.

VI. Conclusion

Three source recovery methods based on CTF have been proposed in this paper. To conclude, i) CTF-MINT is an ideal overdetermined source recovery method when the microphone noise and filter perturbations are small. It has a relative low computational complexity. However, it is sensitive to the microphone noise and filter perturbations, ii) CTF-MPDR is also more suitable for the overdetermined case. It achieves the worst performance among the three 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, iii) thanks to exploiting the spectral sparsity, CTF-C-Lasso is able to work for the underdetermined case, and to efficiently reduce the microphone noise. However, it requires the mixing filters of all the considerable sources, which are not easy to obtain in practice. In addition, the computational cost is high due to the iterative optimization.

References

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

[2] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 320–324, 2015.

[4] H. L. Van Trees, *Detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

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

[6] M. I. Mandel, R. J. Weiss, and D. P. Ellis, “Model-based expectation-maximization source separation and localization,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 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 Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 81–81, 2007.

[8] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A consolidated perspective on multimicrophone speech enhancement and source separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[9] M. Kowalski, E. Vincent, and R. Gribonval, “Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1818–1829, 2010.

[10] S. Arberet, P. Vandergheynst, J.-P. Carrillo, R. E.Thiran, and Y. Wiaux, “Sparse reverberant audio source separation via reweighted analysis,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1391–1402, 2013.

[11] M. Miyoshi and Y. Kaneda, “Inverse filtering of room acoustics,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 2, pp. 145–152, 1988.

[12] M. Kallinger and A. Mertins, “Multi-channel room impulse response shaping-a study,” in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*, vol. 5, pp. V101–V104, 2006.

[13] A. Mertins, T. Mei, and M. Kallinger, “Room impulse response shortening/reshaping with infinity- and p-norm optimization,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 249–259, 2010.

[14] I. Kodrasi, S. Goetze, and S. Doclo, “Regularization for partial multichannel equalization for speech dereverberation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, 2013.

[15] I. Kodrasi and S. Doclo, “Joint dereverberation and noise reduction based on acoustic multi-channel equalization,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 680–693, 2016.

[16] T. Hikichi, M. Delcroix, and M. Miyoshi, “Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations,” *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–12, 2007.

[17] H. Yamada, H. Wang, and F. Itakura, “Recovering of broadband reverberant speech signal by sub-band MINT method,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 969–972, 1991.

[18] H. Wang and F. Itakura, “Realization of acoustic inverse filtering through multi-microphone sub-band processing,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 75, no. 11, pp. 1474–1483, 1992.

[19] S. Weiss, G. W. Rice, and R. W. Stewart, “Multichannel equalization in subbands,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 203–206, 1999.

[20] N. D. Gaubitch and P. A. Naylor, “Equalization of multichannel acoustic systems in oversampled subbands,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1061–1070, 2009.

[21] F. Lim and P. A. Naylor, “Robust speech dereverberation using subband multichannel least squares with variable relaxation,” in *European Signal Processing Conference (EUSIPCO)*, 2013.

[22] Y. Avargel and I. Cohen, “System identification in the short-time Fourier transform domain with crossband filtering,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[23] R. Talmon, I. Cohen, and S. Gannot, “Relative transfer function identification using convolutive transfer function approximation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[24] R. Talmon, I. Cohen, and S. Gannot, “Convolutive transfer function generalized sidelobe canceler,” *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 7, pp. 1420–1434, 2009.

[25] X. Li, L. Girin, and R. Horaud, “Audio source separation based on convolutive transfer function and frequency-domain lasso optimization,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[26] S. Leglaive, R. Badeau, and G. Richard, “Multichannel audio source separation: variational inference of time-frequency sources from time-domain observations,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[27] S. Leglaive, R. Badeau, and G. Richard, “Separating time-frequency sources from time-domain convolutive mixtures using non-negative matrix factorization,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[28] 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 Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 11, pp. 1670–1680, 2014.

[29] B. Schwartz, S. Gannot, and E. A. Habets, “Online speech dereverberation using kalman filter and EM algorithm,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 2, pp. 394–406, 2015.

[30] X. Li, L. Girin, and R. Horaud, “An EM algorithm for audio source separation based on the convolutive transfer function,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[31] P. L. Combettes and J.-C. Pesquet, “A douglas–rachford splitting approach to nonsmooth convex variational signal recovery,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 564–574, 2007.

[32] P. L. Combettes and V. R. Wajs, “Signal recovery by proximal forward-backward splitting,” *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[33] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, “Speech distortion weighted multichannel wiener filtering techniques for noise reduction,” *Speech enhancement*, pp. 199–228, 2005.

[34] X. Li, R. Horaud, and S. Gannot, “Blind multichannel identification and equalization for dereverberation and noise reduction based on convolutive transfer function,” *CoRR*, vol. abs/1706.03652, 2017.

[35] X. Li, L. Girin, S. Gannot, and R. Horaud, “Non-stationary noise power spectral density estimation based on regional statistics,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 181–185, 2016.

[36] R. T. Rockafellar, *Convex analysis*. Princeton university press, 2015.

[37] M. J. Fadili and J.-L. Starck, “Monotone operator splitting for optimization problems in sparse recovery,” in *IEEE International Conference on Image Processing*, pp. 1461–1464, 2009.

[38] Y. Nesterov, “Gradient methods for minimizing composite objective function,” CORE Discussion Papers, 2007/76, 2007.

[39] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[40] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments,” in *International Workshop on Acoustic Signal Enhancement*, pp. 313–317, 2014.

[41] 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,” *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.

[42] D. R. Morgan, J. Benesty, and M. M. Sondhi, “On the evaluation of estimated impulse responses,” *IEEE Signal processing letters*, vol. 5, no. 7, pp. 174–176, 1998.

[43] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[44] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 749–752, 2001.

[45] I. Cohen, S. Gannot, and B. Berdugo, “An integrated real-time beamforming and postfiltering system for nonstationary noise environments,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1064–1073, 2003.

[46] S. Gannot and I. Cohen, “Speech enhancement based on the general transfer function GSC and postfiltering,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, 2004.