HAL
open science

Underdetermined Reverberant Blind Source Separation: Sparse Approaches for Multiplicative and Convolutive Narrowband Approximation

Fangchen Feng, Matthieu Kowalski

▶ **To cite this version:**

Fangchen Feng, Matthieu Kowalski. Underdetermined Reverberant Blind Source Separation: Sparse Approaches for Multiplicative and Convolutive Narrowband Approximation. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2019, 27 (2), pp.442-456. 10.1109/taslp.2018.2881925 . hal-01760968

HAL Id: hal-01760968

https://hal.science/hal-01760968v1

Submitted on 6 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Underdetermined Reverberant Blind Source Separation: Sparse Approaches for Multiplicative and Convolutive Narrowband Approximation

Fangchen Feng and Matthieu Kowalski

***Abstract*—We consider the problem of blind source separation for underdetermined convolutive mixtures. Based on the multiplicative narrowband approximation in the time-frequency domain with the help of Short-Time-Fourier-Transform (STFT) and the sparse representation of the source signals, we formulate the separation problem into an optimization framework. This framework is then generalized based on the recently investigated convolutive narrowband approximation and the statistics of the room impulse response. Algorithms with convergence proof are then employed to solve the proposed optimization problems. The evaluation of the proposed frameworks and algorithms for synthesized and live recorded mixtures are illustrated. The proposed approaches are also tested for mixtures with input noise. Numerical evaluations show the advantages of the proposed methods.**

I. Introduction

A. Time model

Blind source separation (BSS) recovers source signals from a number of observed mixtures without knowing the mixing system. Separation of the mixed sounds has several applications in the analysis, editing, and manipulation of audio data [1]. In the real-world scenario, convolutive mixture model is considered to take the room echo and the reverberation effect into account:

$$x_m(t) = \sum_{n=1}^{N} a_{mn}(t) * s_n(t) + n_m(t), \tag{1}$$

where s_n is the n-th source and x_m is the m-th mixture. N and M are the number of sources and microphones respectively. $a_{mn}(t)$ is the room impulse response (RIR) from the n-th source to the m-th microphone. $n_m(t)$ is the additive white Gaussian noise at the m-th microphone. We denote also $s_{mn}^{\text{img}}(t) = a_{mn}(t) * s_n(t)$ the image of the n-th source at the m-th microphone.

B. Multiplicative narrowband approximation

The source separation for convolutive mixtures is usually tackled in the time-frequency domain with the help of STFT (Short-Time-Fourier-Transform) [2], [3], [4]. With the narrowband assumption, the separation can be performed in each frequency band [5]. Because of the permutation ambiguity in each frequency band, the separation is then followed by a permutation alignment step to regroup the estimated frequency components that belong to the same source [6]. In this paper, we concentrate on the separation step.

The multiplicative narrowband approximation [2], [3] deals with the convolutive mixtures in each frequency using the complex-valued multiplication in the following vector form:

$$\tilde{\mathbf{x}}(f,\tau) = \sum_{n=1}^{N} \tilde{\mathbf{a}}_n(f)\tilde{s}_n(f,\tau) + \tilde{\mathbf{n}}(f,\tau), \tag{2}$$

where $\tilde{\mathbf{x}}(f,\tau) = [\tilde{x}_1(f,\tau),\ldots,\tilde{x}_M(f,\tau)]^T$ and $\tilde{s}_n(f,\tau)$ are respectively the analysis STFT coeffcients of the observations and the n-th source signal. $\tilde{\mathbf{a}}_n(f) = [\tilde{a}_{1n}(f),\ldots,\tilde{a}_{Mn}(f)]^T$ is a vector that contains the Fourier transform of the RIR associated with the n-th source. $\tilde{\mathbf{n}}(f,\tau) = [\tilde{n}_1(f,\tau),\ldots,\tilde{n}_M(f,\tau)]^T$ consistes not only the analysis STFT coefficients of the noise, but also the error term due to the approximation. The formulation (2) approximates the convolutive mixtures by using instantaneous mixture in each frequency. This approximation therefore largely reduces the complexity of the problem and is valid when the RIR length is less than the STFT window length.

The sparsity assumption is largely utilized for source separation problem [2], [3], [7], [8]. Based on the model (2) and by supposing that only one source is active or dominant in each time-frequency bin (f,τ), the authors of [2] proposed to estimate the mixing matrix in each frequency by clustering, and then estimate the source in a *maximum a posteriori* (MAP) sense. This method is further improved by [3] where the authors proposed to use a soft masking technique to perform the separation. The idea is to classify each time-frequency bin of the observation $\tilde{\mathbf{x}}(f,\tau)$ into N class, where N is the number of sources. Based on a complex-valued Gaussian generative model for source signals, they inferred a bin-wise *a posteriori* probability $P(\mathcal{C}_n|\tilde{\mathbf{x}}(f,\tau))$ which represents the probability that the vector $\tilde{\mathbf{x}}(f,\tau)$ belongs to the n-th class $\mathcal{C}_n$. This method obtains good separation results for speech signals, however only in low reverberation scenario [3]. The performance of these methods is limited by the multiplicative approximation whose approximation error increases rapidely as the reverberation time becomes long [9]. Moreover, the disjointness of the soures in the time-frequency domain is not realistic [8].

Fangchen Feng is with Laboratoire Astroparticule et Cosmologie, Université Paris Diderot, CNRS/IN2P3, Sorbonne Paris Cité, 75205, Paris, France (email: fangchen.feng@apc.in2p3.fr)

Matthieu Kowalski is with Laboratoire des signaux et systèmes, CNRS, Centralesupélec, Université Paris-Sud, Université Paris-Saclay, 91192, Gif-sur-Yvette, France (email: matthieu.kowalski@l2s.centralesupelec.fr)

C. Beyond the multiplicative narrowband model

A generalization of the multiplicative approximation is proposed in [4] by considering the spatial covariance matrix of the source signals. By modeling the sources STFT coefficients as a phase-invariant multivariate distribution, the authors inferred that the covariance matrix of the STFT coefficients of the n-th source images $\mathbf{s}_n^{\text{img}} = \left[s_{1n}^{\text{img}}, s_{2n}^{\text{img}}, \ldots, s_{Mn}^{\text{img}}\right]^T$ can be factorized as:

$$\mathbf{R}_{\tilde{\mathbf{s}}_n^{\text{img}}}(f,\tau) = v_n(f,\tau)\mathbf{R}_n(f), \tag{3}$$

where $v_n(f,\tau)$ are scalar time-varying variances of the n-th source at different frequencies and $\mathbf{R}_n(f)$ are time-invariant spatial covariance matrices encoding the source spatial position and spatial spread [4]. The multiplicative approximation forces the spatial covariance matrix to be of rank-1 and the authors of [4] exploited a generalization by assuming that the spatial covariance matrix is of full-rank and showed that the new assumption models better the mixing process because of the increased flexibility. However, as we show in this paper by experiments, the separation performance of this full-rank model is still limited in strong reverberation scenarios.

Moreover, as both the bin-wise method [3] and the full-rank approach [4] do not take the error term $\tilde{\mathbf{n}}$ into consideration, they are therefore sensitive to additional noise.

Recently, the authors of [10] investigated the convolutive narrowband approximation for oracle source separation of convolutive mixtures (the mixing systems is known). They showed that the convolutive approximation suits better the original mixing process especially in strong reverberation scenarios. In this paper, we investigate the convolutive narrowband approximation as the generalization of the multiplicative approximation in the full blind setting (the mixing system if unknown).

The contribution of the paper is three-folds: first based on the multiplicative narrowband approximation, we formulate the separation in each frequency as an optimization problem with ℓ_1 norm penalty to exploit sparsity. The proposed optimization formulation is then generalized based on the statistics of the RIR [11] and the convolutive narrowband approximation model [10]. At last, we propose to solve the obtained optimizations with PALM (Proximal alternating linearized minimization) [12] and BC-VMFB (Block coordinate-variable metric forward backward) [13] algorithms which have convergence guarantee.

The rest of the article is organized as follows. We propose the optimization framework based on multiplicative approximation with ℓ_1 norm penalty and present the corresponding algorithm in Section II. The optimization framework is then generalized in Section III based on the statistics of the RIR and the convolutive approximation. The associated algorithm is also presented. We compare the separation performance achieved by the proposed approaches to that of the state-of-the-art in various experimental settings in Section IV. Finally, Section V concludes the paper.

II. The multiplicative narrowband approximation

We first rewrite the formulation (2) with matrix notations by concatenating the time samples and source indexes. In each frequency f, we have:

$$\tilde{\mathbf{X}}_f = \tilde{\mathbf{A}}_f\tilde{\mathbf{S}}_f + \tilde{\mathbf{N}}_f, \tag{4}$$

where $\tilde{\mathbf{X}}_f \in \mathbb{C}^{M\times L_T}$ is the matrix of the analysis STFT coefficients of the observations at the given frequency f. $\tilde{\mathbf{A}}_f \in \mathbb{C}^{M\times N}$ is the mixing matrix at frequency f. $\tilde{\mathbf{S}}_f \in \mathbb{C}^{N\times L_T}$ is the matrix of the analysis STFT coefficients of the sources at frequency f. $\tilde{\mathbf{N}}_f \in \mathbb{C}^{M\times L_T}$ is the noise term which also contains the approximation error. In the above notations, L_T is the number of time samples in the time-frequency domain.

The target of the separation is to estimate $\tilde{\mathbf{A}}_f$ and $\tilde{\mathbf{S}}_f$ from the observations $\tilde{\mathbf{X}}_f$. However, according to the definition of the analysis STFT coefficients , the estimated $\tilde{\mathbf{S}}_f$ has to be in the image of the STFT operator (see in [14] for more details). To avoid this additional constraint, we propose to replace the analysis STFT coefficients $\tilde{\mathbf{S}}_f$ by the synthesis STFT coefficients $\boldsymbol{\alpha}_f \in \mathbb{C}^{N\times L_T}$, which leads to:

$$\tilde{\mathbf{X}}_f = \tilde{\mathbf{A}}_f\boldsymbol{\alpha}_f + \tilde{\mathbf{N}}_f. \tag{5}$$

In the following, we denote also $\boldsymbol{\alpha}_{f,n}$ the n-th source componant (row) of $\boldsymbol{\alpha}_f$ and $\alpha_{f,n}(\tau)$ the scalar element at position τ in $\boldsymbol{\alpha}_{f,n}$.

A. Formulation of the optimization

Based on the model (5), we propose to formulate the separation as an optimization problem as follow:

$$\min_{\tilde{\mathbf{A}}_f,\boldsymbol{\alpha}_f} \frac{1}{2}\|\tilde{\mathbf{X}}_f - \tilde{\mathbf{A}}_f\boldsymbol{\alpha}\|_F^2 + \lambda\|\boldsymbol{\alpha}_f\|_1 + \imath_{\mathcal{C}}(\tilde{\mathbf{A}}_f), \tag{6}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|_1$ is the ℓ_1 norm of the matrix which is the sum of the absolute value of all the elements. $\imath_{\mathcal{C}}(\tilde{\mathbf{A}}_f)$ is an indicator function to avoid the trivial solution caused by the scaling ambiguity between $\tilde{\mathbf{A}}_f$ and $\boldsymbol{\alpha}_f$:

$$\imath_{\mathcal{C}}(\tilde{\mathbf{A}}_f) = \begin{cases} 0, & \text{if } \|\tilde{\mathbf{a}}_{f,n}\| = 1,\ n = 1,2,\ldots,N \\ +\infty, & \text{otherwise} \end{cases} \tag{7}$$

with $\tilde{\mathbf{a}}_{f,n}$ the n-th column of $\tilde{\mathbf{A}}_f$. λ is the hyperparameter which balances between the data term $\frac{1}{2}\|\tilde{\mathbf{X}}_f - \tilde{\mathbf{A}}_f\boldsymbol{\alpha}_f\|_F^2$ and the penalty term $\|\boldsymbol{\alpha}_f\|_1$.

For instantaneous mixtures, the formulation (6) has been firstly proposed in [7] and recently investigated in [15]. Compared to the masking technique of separation [3], the ℓ_1 norm term exploits only sparsity which is more realistic than the disjointness assumption for speech signals. Moreover, the Lagrangian form with the data term $\frac{1}{2}\|\tilde{\mathbf{X}}_f - \tilde{\mathbf{A}}_f\boldsymbol{\alpha}_f\|_F^2$ allows us to take the noise/approximation error into consideration.

B. Algorithm: N-Regu

The optimization problem (6) is non-convex with a non-differentiable term. In this paper, we propose to solve the problem by applying the BC-VMFB (block coordinate variable metric forward-backward) [16] algorithm. This algorithm relies on the proximal operator [17] given in the next definition.

Definition 1. *Let ψ be a proper lower semicontinuous function, the proximal operator associated with ψ is defined as:*

$$\mathrm{prox}_{\psi} := \underset{\mathbf{y}}{\mathrm{argmin}}\, \psi(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_F^2. \tag{8}$$

When the function $\psi(\mathbf{y}) = \lambda\|\mathbf{y}\|_1$, the proximal operator becomes the entry-wise soft thresholding presented in the next proposition.

Proposition 2. *Let $\boldsymbol{\alpha} \in \mathbb{C}^{N\times L_T}$. Then, $\hat{\boldsymbol{\alpha}} = \mathrm{prox}_{\lambda\|\cdot\|_1}(\boldsymbol{\alpha}) := \mathcal{S}_\lambda(\boldsymbol{\alpha})$ is given entrywise by soft-thresholding:*

$$\hat{\alpha}_i = \frac{\alpha_i}{|\alpha_i|}\left(|\alpha_i| - \lambda\right)^+, \tag{9}$$

where $(\alpha)^+ = \max(0,\alpha)$.

When the function ψ in Definition 1 is the indicator function $\imath_{\mathcal{C}}$, the proximal operator reduces to the projection operator presented in Proposition 3.

Proposition 3. *Let $\tilde{\mathbf{A}} \in \mathbb{C}^{M\times N}$. Then $\hat{\mathbf{A}} = \mathrm{prox}_{\imath_{\mathcal{C}}}(\tilde{\mathbf{A}}) := \mathcal{P}_{\mathcal{C}}(\tilde{\mathbf{A}})$ is given by the column-wise normalization projection:*

$$\hat{\mathbf{a}}_n = \frac{\tilde{\mathbf{a}}_n}{\|\tilde{\mathbf{a}}_n\|},\ n = 1,2,\ldots,N \tag{10}$$

With the above proximal operators, we present the algorithm derived from BC-VMFB in Algorithm 1. We denote the data term by $Q(\boldsymbol{\alpha}_f, \mathbf{A}_f) = \frac{1}{2}\|\tilde{\mathbf{X}}_f - \tilde{\mathbf{A}}_f\boldsymbol{\alpha}_f\|_F^2$. $L^{(j)} = \|\tilde{\mathbf{A}}_f^{(j+1)^H}\tilde{\mathbf{A}}_f^{(j+1)}\|_2$ is the Lipschitz constant of $\nabla_{\boldsymbol{\alpha}_f} Q(\boldsymbol{\alpha}_f^{(j)}, \tilde{\mathbf{A}}_f^{(j)})$ with $\|\cdot\|_2$ denoting the spectral norm of matrix. Details of the derivation of this algorithm and the convergence study are given in Appendix VI-A. In the following, this algorithm is referred as N-Regu (Narrowband optimization with regularization).

Algorithm 1: N-Regu

Initialisation : $\boldsymbol{\alpha}_f^{(1)} \in \mathbb{C}^{N\times L_T}$, $\tilde{\mathbf{A}}_f^{(1)} \in \mathbb{C}^{M\times N}$,
$L^{(1)} = \|\tilde{\mathbf{A}}_f^{(1)^H}\tilde{\mathbf{A}}_f^{(1)}\|_2$, $j = 1$;
repeat
- $\nabla_{\boldsymbol{\alpha}_f} Q\left(\boldsymbol{\alpha}_f^{(j)}, \tilde{\mathbf{A}}_f^{(j)}\right) = -\tilde{\mathbf{A}}_f^{(j)^H}\left(\tilde{\mathbf{X}}_f - \tilde{\mathbf{A}}_f^{(j)}\boldsymbol{\alpha}_f^{(j)}\right)$;
- $\boldsymbol{\alpha}_f^{(j+1)} = \mathcal{S}_{\lambda/L^{(j)}}(\boldsymbol{\alpha}_f^{(j)} - \frac{1}{L^{(j)}}\nabla_{\boldsymbol{\alpha}_f} Q(\boldsymbol{\alpha}_f^{(j)}, \tilde{\mathbf{A}}_f^{(j)})$;
- $\tilde{\mathbf{A}}_f^{(j+1)} = \mathcal{P}_{\mathcal{C}}(\tilde{\mathbf{X}}_f\boldsymbol{\alpha}_f^{(j+1)^H})$;
- $L^{(j+1)} = \|\tilde{\mathbf{A}}_f^{(j+1)^H}\tilde{\mathbf{A}}_f^{(j+1)}\|_2$;
- $j = j+1$;

until *convergence*;

III. The convolutive narrowband approximation

A. Convolutive approximation

Theoretically, the multiplicative narrowband approximation (2) is valid only when the RIR length is less than the STFT window length. In practice, this condition is rarely varified because the STFT window length is limited to ensure the local stationarity of audio signals [10]. To avoid such limitation, the convolutive narrowband approximation was proposed in [18], [19]:

$$\tilde{\mathbf{x}}(f,\tau) = \sum_{n=1}^{N}\sum_{l=1}^{\mathcal{L}} \tilde{\mathbf{h}}_n(f,l)\tilde{s}_n(f,\tau-l), \tag{11}$$

where $\tilde{\mathbf{h}}_n = \left[\tilde{h}_{1n},\ldots,\tilde{h}_{Mn}\right]^T$ is the vector that contains the impulse responses in the time-frequency domain associated with the n-th source. $\mathcal{L}$ is the length of the convolution kernel in the time-frequency domain.

The convolutive approximation (11) is a generalization of the multiplicative approximation (2) as it considers the information diffusion along the time index. When the kernel length $\mathcal{L} = 1$, it reduces to the multiplicative approximation. The convolution kernel in the time-frequency domain $\tilde{h}_{mn}(f,\tau)$ is linked to the RIR in the time domain $a_{mn}(t)$ by [10]:

$$\tilde{h}_{mn}(f,\tau) = \left[a_{mn}(t) * \zeta_f(t)\right]|_{t=\tau k_0}, \tag{12}$$

which represents the convolution with respect to the time index t evaluated with a resolution of the STFT frame step k_0 with:

$$\zeta_f(t) = e^{2\pi i f t/L_F}\sum_j \varphi(j)\tilde{\varphi}(t+j), \tag{13}$$

where L_F is the number of frequency bands. $\varphi(j)$ et $\tilde{\varphi}(j)$ denote respectively the analysis and synthesis STFT window.

With matrix notations, for each frequency f, the convolutive approximation (11) can be written as:

$$\tilde{\mathbf{X}}_f = \tilde{\mathbf{H}}_f \star \tilde{\mathbf{S}}_f + \tilde{\mathbf{N}}_f, \tag{14}$$

where $\tilde{\mathbf{H}}_f \in \mathbb{C}^{M\times N\times\mathcal{L}}$ is the mixing system formed by concatenating the impulse responses of length $\mathcal{L}$. In the following, we denote also $\tilde{\mathbf{h}}_{f,mn}$ the vector that represents the impulse response at position (m,n) in $\tilde{\mathbf{H}}_f$ and $\tilde{h}_{f,mn}(\tau)$ the scalar element at position (m,n,τ). The operator $\star$ denotes the convolutive mixing process (11).

Compared to the original mixing process in the time domain (1), the convolutive approximation (14) largely reduces the length of the convolution kernel, thus makes the estimation of both the mixing system and the source signals practically possible.

B. Proposed optimization approach

a) Basic extension of the multiplicative model: Once again, to circumvent the additional constraint brought by the analysis coefficients of the sources, we replace the analysis STFT coefficients $\tilde{\mathbf{S}}_f$ by the synthesis coefficients $\boldsymbol{\alpha}_f$, which leads to:

$$\tilde{\mathbf{X}}_f = \tilde{\mathbf{H}}_f \star \boldsymbol{\alpha}_f + \tilde{\mathbf{N}}_f. \tag{15}$$

Based on (15), we generalize (6) by replacing the multplicative operator by the convolutive mixing operator:

$$\min_{\tilde{\mathbf{H}}_f,\boldsymbol{\alpha}_f} \frac{1}{2}\|\tilde{\mathbf{X}}_f - \tilde{\mathbf{H}}_f \star \boldsymbol{\alpha}_f\|_F^2 + \lambda\|\boldsymbol{\alpha}_f\|_1 + \imath_{\mathcal{C}}^{\mathrm{Conv}}(\tilde{\mathbf{H}}_f), \tag{16}$$

where $\imath_{\mathcal{C}}^{\text{Conv}}(\tilde{\mathbf{H}}_f)$ is the normalisation constraint to avoid trivial solutions:

$$\imath_{\mathcal{C}}^{\text{Conv}}(\tilde{\mathbf{H}}_f) = \begin{cases} 0, \text{ if } \sqrt{\sum_{m,\tau} |\tilde{h}_{f,mn}(\tau)|^2} = 1, \ n = 1, \dots, N \\ +\infty, \text{ otherwise.} \end{cases} \tag{17}$$

b) Regularization for the convolution kernel: In [11], the authors consider the problem of estimating the RIR supposing that the mixtures and the sources are known. They formulated the estimation problem as an optimization problem and proposed a differentiable penalty for the mixing system in the time domain:

$$\sum_{m,n,t} \frac{|a_{mn}(t)|^2}{2\rho^2(t)}, \tag{18}$$

where $\rho(t)$ denotes the amplitude envelope of RIR which depends on the reverberation time RT_{60}:

$$\rho(t) = \sigma 10^{-3t/\text{RT}_{60}}, \tag{19}$$

with σ being a scaling factor. The penalty (18) is designed to force an exponential decrease of the RIR which satisfaits the acoustic statistics of the RIR [20].

As the convolutive kernel in the time-frequency domain is linked to the RIR in time domain by (12), in this paper, we consider the penalty in the time-frequency domain in the same form:

$$\mathcal{P}(\tilde{\mathbf{H}}_f) = \sum_{m,n,\tau} \frac{|\tilde{h}_{f,mn}(\tau)|^2}{2\tilde{\rho}^2(\tau)}, \tag{20}$$

where $\tilde{\rho}(\tau)$ is the decreasing coefficients in the time-frequency domain which depends on $\rho(t)$ and the STFT transform.

Other forms of penalty are also proposed in [11]. However, their adaption in the time-frequency domain is not straightforward.

c) Final optimization problem: With the above penalty term, the formulation (16) can be improved as:

$$\min_{\tilde{\mathbf{H}}_f, \boldsymbol{\alpha}_f} \frac{1}{2}\|\tilde{\mathbf{X}}_f - \tilde{\mathbf{H}}_f \star \boldsymbol{\alpha}_f\|_F^2 + \lambda\|\boldsymbol{\alpha}_f\|_1 + \mathcal{P}(\tilde{\mathbf{H}}_f) + \imath_{\mathcal{C}}^{\text{Conv}}(\tilde{\mathbf{H}}_f). \tag{21}$$

C. Algorithm: C-PALM

We propose to use the Proximal Alternating Linearized Minimization (PALM) algorithm [12] to solve the problem. The derived algorithm is presented in Algorithm 2, and one can refer to Appendix VI-C for details on the derivation and the convergence study. We refer to this algorithm as C-PALM (Convolutive PALM) in the following. We denote:

$$Q(\tilde{\boldsymbol{\alpha}}_f, \tilde{\mathbf{H}}_f) = \frac{1}{2}\|\tilde{\mathbf{X}} - \tilde{\mathbf{H}}_f \star \boldsymbol{\alpha}_f\|_F^2 + \mathcal{P}(\tilde{\mathbf{H}}_f)$$

and the gradient of $\mathcal{P}(\tilde{\mathbf{H}}_f)$ is given coordinate-wise by:

$$\left[\nabla_{\tilde{\mathbf{H}}_f} \mathcal{P}(\tilde{\mathbf{H}}_f)\right]_{f,mn\tau} = \frac{\tilde{h}_{f,mn}(\tau)}{\tilde{\rho}^4(\tau)} \ . \tag{22}$$

In Algorithm 2, $\tilde{\mathbf{H}}_f^H$ and $\boldsymbol{\alpha}_f^H$ are respectively the adjoint operators of the convolutive mixtures with respect to the convolution kernel and the sources. Details of derivation of these adjoint operators are given in Appendix VI-B. $L_{\boldsymbol{\alpha}_f}^{(j)}$ and $L_{\tilde{\mathbf{H}}_f}^{(j)}$ are respectively the Lipschitz constant of $\nabla_{\boldsymbol{\alpha}_f} Q(\boldsymbol{\alpha}_f^{(j)}, \tilde{\mathbf{H}}_f^{(j)})$ and $\nabla_{\tilde{\mathbf{H}}_f} Q(\boldsymbol{\alpha}_f^{(j+1)}, \tilde{\mathbf{H}}_f^{(j)})$. $L_{\boldsymbol{\alpha}_f}^{(j)}$ can be calculated with the power iteration algorithm [9] shown in Algorithm 3. $L_{\tilde{\mathbf{H}}_f}^{(j)}$ can be approximately estimated thanks to the next proposition.

Algorithm 2: C-PALM

Initialisation : $\boldsymbol{\alpha}_f^{(1)} \in \mathbb{C}^{N \times L_T}$, $\tilde{\mathbf{H}}_f^{(1)} \in \mathbb{C}^{M \times N}$, $j = 1$;
repeat
- $\nabla_{\boldsymbol{\alpha}_f} Q\left(\boldsymbol{\alpha}_f^{(j)}, \tilde{\mathbf{H}}_f^{(j)}\right) =$
- $-\tilde{\mathbf{H}}_f^{(j)^H} \star \left(\tilde{\mathbf{X}}_f - \tilde{\mathbf{H}}_f^{(j)} \star \boldsymbol{\alpha}^{(j)}\right)$;
- $\boldsymbol{\alpha}_f^{(j+1)} = \mathcal{S}_{\lambda/L_{\boldsymbol{\alpha}_f}^{(j)}}\left(\boldsymbol{\alpha}_f^{(j)} - \frac{1}{L_{\boldsymbol{\alpha}_f}^{(j)}} \nabla_{\boldsymbol{\alpha}_f} Q(\boldsymbol{\alpha}_f^{(j)}, \tilde{\mathbf{H}}_f^{(j)})\right)$;
- $\nabla_{\tilde{\mathbf{H}}_f} Q(\boldsymbol{\alpha}_f^{(j+1)}, \tilde{\mathbf{H}}_f^{(j)}) =$
- $-(\tilde{\mathbf{X}}_f - \tilde{\mathbf{H}}_f^{(j)} \star \boldsymbol{\alpha}_f^{(j+1)}) \star \boldsymbol{\alpha}_f^{(j+1)^H} + \nabla_{\tilde{\mathbf{H}}_f} \mathcal{P}(\tilde{\mathbf{H}}_f^{(j)})$;
- $\tilde{\mathbf{H}}_f^{(j+1)} =$
- $\mathcal{P}_{\imath_{\mathcal{C}}}^{\text{Conv}}\left(\tilde{\mathbf{H}}_f^{(j)} - \frac{1}{L_{\tilde{\mathbf{H}}_f}^{(j)}} \nabla_{\tilde{\mathbf{H}}_f} Q(\boldsymbol{\alpha}_f^{(j+1)}, \tilde{\mathbf{H}}_f^{(j)})\right)$;
- Update $L_{\boldsymbol{\alpha}_f}^{(j)}$ et $L_{\tilde{\mathbf{H}}_f}^{(j)}$; $j = j + 1$;

until *convergence*;

Algorithm 3: Power iteration for the calculation of $L_{\boldsymbol{\alpha}_f}$

Initialisation : $\mathbf{v}_f \in \mathbb{C}^{N \times L_T}$;
repeat
- $\mathbf{W} = \tilde{\mathbf{H}}_f^H \star \tilde{\mathbf{H}}_f \star \mathbf{v}_f$;
- $L_{\boldsymbol{\alpha}_f} = \|\mathbf{W}\|_\infty$;
- $\mathbf{v}_f = \frac{\mathbf{W}}{L_{\boldsymbol{\alpha}_f}}$;

until *convergence*;

Proposition 4. *If we suppose that the source componants $\boldsymbol{\alpha}_{f,1}, \boldsymbol{\alpha}_{f,2}, \dots, \boldsymbol{\alpha}_{f,N}$ are mutually independant and $\mathcal{L} \ll L_T$, then $L_{\tilde{\mathbf{H}}_f}$, the Lipschitz constant of $\nabla_{\tilde{\mathbf{H}}_f} Q(\boldsymbol{\alpha}_f, \tilde{\mathbf{H}}_f)$ can be calculated as:*

$$L_{\tilde{\mathbf{H}}_f} = \max_n (L_{f,n}) + \sqrt{\max_\tau (\frac{1}{\tilde{\rho}^8(\tau)})}, \tag{23}$$

where $L_{f,n} = \|\boldsymbol{\Gamma}_{f,n}\|$, with

$$\boldsymbol{\Gamma}_{f,n} = \begin{pmatrix} \gamma_{f,n}(0) & \gamma_{f,n}(1) & \dots & \gamma_{f,n}(\mathcal{L}-1) \\ \gamma_{f,n}(-1) & \gamma_{f,n}(0) & \dots & \gamma_{f,n}(\mathcal{L}-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{f,n}(1-\mathcal{L}) & \gamma_{f,n}(2-\mathcal{L}) & \dots & \gamma_{f,n}(0) \end{pmatrix}, \tag{24}$$

and $\gamma_{f,n}(\tau)$ is the empirical autocorrelation function of $\boldsymbol{\alpha}_{f,n}$:

$$\gamma_{f,n} = \sum_{\ell=1}^{L_T-1} \alpha_{f,n}(\ell + \tau)\alpha_{f,n}^*(\ell). \tag{25}$$

Proof. The proof is postponed in Appendix VI-D. □

If the independance assumption mentioned in Proposition 4 appears to be strong, it is well adapted for audio signals as it is the basic hypothesis of the FDICA (frequency domain independant component analysis) [21] used for source separation of determined convolutive mixtures. Although we do not have any guarantee of independence in the proposed algorithm, the experiments show that good performances are obtained.

Finally, we must stress that the BC-VMFB algorithm is not suitable for (21) as it relies on the second derivative of $Q(\tilde{\boldsymbol{\alpha}}_f, \tilde{\mathbf{H}}_f)$ w.r.t $\tilde{\mathbf{H}}_f$, which does not necessarily simplify the algorithm.

IV. Experiments

A. Permutation alignment methods

For the proposed approaches, we use the existing permutation alignment methods. For N-Regu, we compare the approach based on TDOA (Time Difference Of Arrival) used in Full-rank method [4] and the approach based on inter-frequency correlation used in the Bin-wise approach [3]. For the inter-frequency correlation permutation, we use the power ratio [6] of the estimated source to present the source activity. For C-PALM, as the TDOA permutation is not adapted, we use only the correlation permutation.

For the proposed approaches (N-Regu and C-PALM) and the reference algorithms (Bin-wise and Full-rank), we also designed an oracle permuation alignment method. In each frequency, we look for the permutation that maximizes the correlation between the estimated and the original sources. Such permutation alignment is designed to show the best permutation possible in order to have a fair comparison of the separation approaches instead of the choice made for solving the permutation problem.

B. Experimental setting

We first evaluated the proposed approaches with 10 sets of synthesized stereo mixtures ($M = 2$) containing three speech sources ($N = 3$) of male/female with different nationalities. The mixtures are sampled at 11 kHz and truncated to 6 s. The room impulse response were simulated via the toolbox [22]. The distance between the two microphone is 4 cm. The reverberation time is defined as 50 ms, 130 ms, 250 ms and 400 ms. The Fig. 1 illustrates the room configuration. For each mixing situation, the mean values of the evaluation results over the 10 sets of mixtures are shown.

We then evaluated the algorithm C-PALM with the live recorded speech mixtures from the dataset SiSEC2011 [23]. Music mixtures are avoided because the instrumental sources are often synchronized to each other and this situation is difficult for the permutation alignment based on inter-frequency correlation [3]. An effective alternative way is to employ nonnegative matrix factorization [24]. The parameters of STFT for the synthesized and live recorded mixtures are summarized in Table I. The STFT window length (and window shift) for synthesized mixtures are chosen to preserve local stationarity of audio sources without bringing too much computational costs. The parameters for the live recorded mixtures are the same as the reported reference algorithm Bin-wise [3].

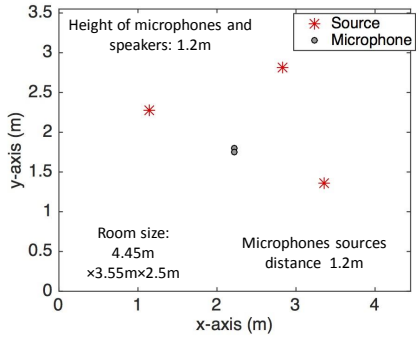

Fig. 1. Room configuration for synthesized mixtures

TABLE I
Experimental conditions

	synthesized	live recorded
Number of microphones	$M = 2$	$M = 2$
Number of sources	$N = 3$	$N = 3, 4$
Duration of signals	6 s	10 s
Reverberation time (RT_{60})	$50, 130, 250, 400$ ms	$130, 250$ ms
Sample rate	11 KHz	16 kHz
Microphone distance	4 cm	5 cm, 1 m
STFT window type	Hann	Hann
STFT window length	512 (46.5 ms)	2048 (128 ms)
STFT window shift	256 (23.3 ms)	512 (32 ms)

The separation performance is evaluated with the signal to distortion ratio (SDR), signal to interference ratio (SIR), source image to spatial distortion ratio (ISR) and signal to artifact ratio (SAR) [25]. The SDR reveals the overall quality of each estimated source. SIR indicates the crosstalk from other sources. ISR measures the amount of spatial distortion and SAR is related to the amount of musical noise.

N-Regu is initialized with Gaussian random signals. C-PALM is initialized with the results of N-Regu with 1000 iterations. This choice of initialization for C-PALM compensates the flexibility of the convolutive model (and then the number of local minima in (21)) without bringing too much computational cost. We use the stopping criteria $\|\boldsymbol{\alpha}_f^{(j+1)} - \boldsymbol{\alpha}_f^j\|_F < 10^{-4}$ for both algorithms.

C. Tuning the parameters

For the proposed methods, we chose several pre-defined hyperparameter λ and select the λ which corresponds to the best SDR. Even though such a way of choosing this hyper-parameter is not possible for real applications, such evaluation offers a "fair" comparison with the state-of-the-art approaches and gives some empirical suggestions of choosing this parameter in practice. We implement the continuation trick, also known as *warm start* or *fixed point continuation* [26] for a fixed value of λ: we start the algorithms with a large value of λ and iteratively decrease λ to the desired value.

It is also important to mention that the hyperparameter λ should be theoretically different for each frequency since the sparsity level of the source signals in each frequency can be very different (for speech signals, the high frequency

componants are usually sparser than the low frequency componants). Therefore, different λ should be determined for each frequency. However, in this paper, we used a single λ for all the frequencies and the experiments show that this simplified choice can achieve acceptable results if we perform a whitening pre-processing for each frequency.

For C-PALM, as the reverberation time is unknown in the blind setting, we pre-define the length of the convolution kernel in the time-frequency domain $\mathcal{L} = 3$ as well as the penalty parameter $\tilde{\rho}(\tau) = [1.75, 1.73, 1.72]^T$. Although these parameters should vary with the reverberation time, we show in the following that the proposed pre-defined parameters work well in different strong reverberation conditions.

D. Synthesized mixtures without noise

We first evaluate the algorithms with synthesized mixtures in the noiseless case as a function of the reverberation time RT_{60}. The results are shown in Fig. 2.

For $\text{RT}_{60} = 50$ ms, it is clear that the Full-rank method performs the best in terms of all four indicators. Its good performance is due to the fact that the full-rank spatial covariance model suits better the convolutive mixtures than the multiplicative approximation and the fact that the TDOA permutation alignement has relatively good performance in low reverberation scenario. N-Regu outperforms Bin-wise only in SDR and SAR. It is because that N-Regu has better data fit than the masking-based Bin-wise method while Bin-wise obtains time-frequency domain disjoint sources which have lower inter-source interference. C-PALM is dominated by other methods in SDR, SIR and ISR. We believe it is because that the pre-defined penalty parameter $\tilde{\rho}(\tau)$ does not fit the low reverberation scenario. The advantages of C-PALM can be seen in relatively stronger reverberation scenarios (especially $\text{RT}_{60} = 130, 250$ ms) where C-PALM outperforms other methods in SDR and SIR. For $\text{RT}_{60} = 400$ ms, all the presented algorithms have similar performance while C-PALM performs slightly better in SIR. To compare the two permutation methods used for N-Regu, TDOA permutation performs better than inter-frequency correlation permutation in SDR, SIR and SAR.

Fig. 3 compares the presented algorithms with oracle permutation alignment. For $\text{RT}_{60} = 50$ ms, once again, Full-rank has the best performance in all four indicators. This confirms the advantages of the full rank spatial covariance model. In high reverberation conditions, C-PALM performs better than others in SDR and SIR. In particular, for $\text{RT}_{60} = 130, 250$ ms, C-PALM outperforms Full-rank by more than 1 dB in SDR and outperforms Bin-wise by about 1.2 dB in SIR. N-Regu performs slightly better than Bin-wise in SDR for all reverberation conditions.

The above observations show the better data fit brought by the optimization framework used in N-Regu (and C-PALM) and confirm the advantages of convolutive narrowband approximation used in C-PALM for high reverberation conditions (especially $\text{RT}_{60} = 130, 250$ ms).

Fig. 4 illustrates the performance of the presented algorithm as a function of the sparsity level [1] of the estimated synthesis coeffcients of the sources for $\text{RT}_{60} = 130$ ms. As the sparsity level is directly linked to the hyperparameter λ in the proposed algorithms, this comparison offers some suggestions of choosing this hyperparameter. Full-rank method does not exploits sparsity, thus has 0% as sparsity level. As the number of sources $N = 3$, the sparsity level of the masking-based Bin-wise method is 66.6%.

C-PALM performs better than N-Regu in terms of SDR, SIR and SAR when the sparsity level is less than 60% and its best performance is achieved when the sparsity level is around 40%. For N-Regu, in terms of SDR, SAR and ISR, the best performance is achieved with the least sparse result.

E. Synthesized mixtures with noise

In this subsection, we evaluate the proposed methods with synthesized mixtures with additive white Gaussian noise. The noise of different energy is added which leads to different input SNR. Fig. 5 reports the separation performance as a function of input SNR with the reverberation time fixed to $\text{RT}_{60} = 130$ ms.

It is clear that N-Regu with TDOA permutation outperforms other methods in terms of SDR and SIR. In particular, it performs better than others by about 1 dB in SIR for all the input SNR tested. C-PALM outperforms the state-of-the-art approaches only in SDR. We believe that it is due to the fact that the freedom degree of the convolutive narrowband approximation used in C-PALM could be sensitive to input noise. Another reason is that the inter-frequency correlation based permutation could be sensitive to input noise. The latter conjecture is supported by the observation that, in terms of SDR and SIR, the gap between N-Regu with TDOA permutation and with correlation permutation increases as the input noise becomes stronger. Further evidence can be found by the comparisons between the presented algorithm with oracle permutation alignment in Fig. 6.

In Fig. 6, in terms of SDR and SIR, it is clear that the gap between N-Regu and C-PALM decreases as the input noise gets stronger. This remark confirms that the separation step of C-PALM is sensitive to input noise. Moreover, in terms of SIR, C-PALM with oracle permutation performs consistently better than N-Regu with oracle permutation, while C-PALM with correlation permutation is dominated by N-Regu with TDOA permutation by about 1 dB (Fig. 5). This observation shows that the performance of C-PALM can be largely improved for noisy mixtures if better permutation alignment method is developped.

Fig. 7 reports the separation performance as a function of the sparsity level of the estimated synthesis coefficients of the sources. $\text{RT}_{60} = 130$ ms and the input SNR is 15 dB. The results of Full-rank and Bin-wise method are also shown.

In terms of SDR and SIR, N-Regu with TDOA permutation consistently outperforms the other methods and achieves its best performance when the sparsity level is about 78%.

[1] In this paper, the sparsity level is the percentage of zero elements in a vector or matrix. A higher sparsity level means a sparser vector or matrix.

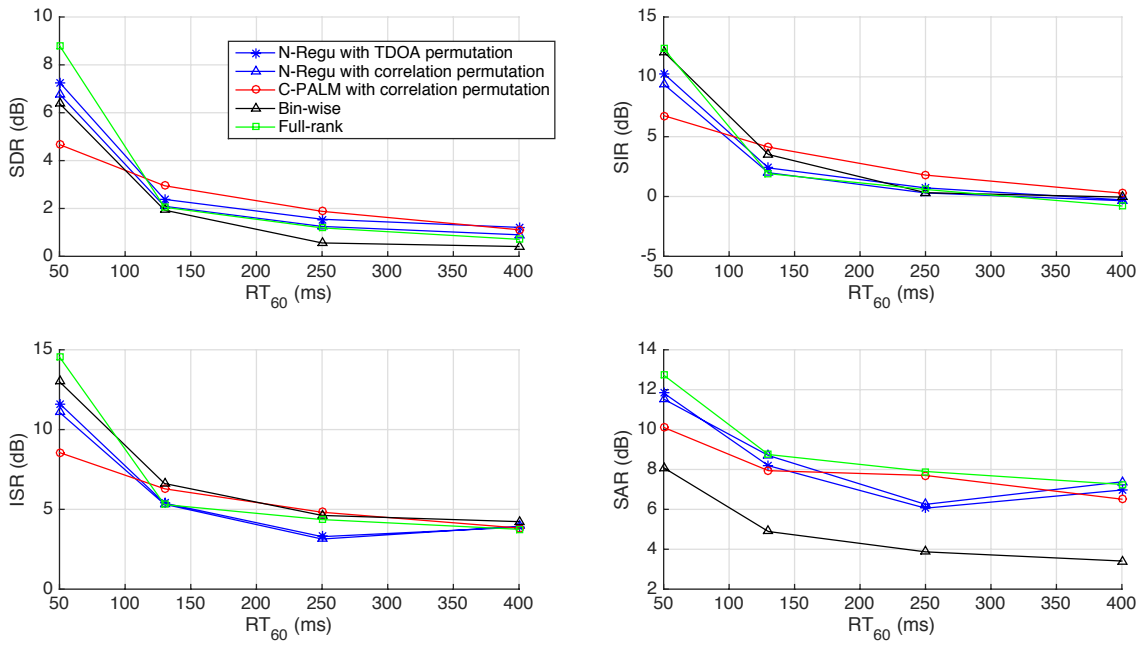

Fig. 2. Separation performance as a function of the reverberation time RT_{60} in noiseless case

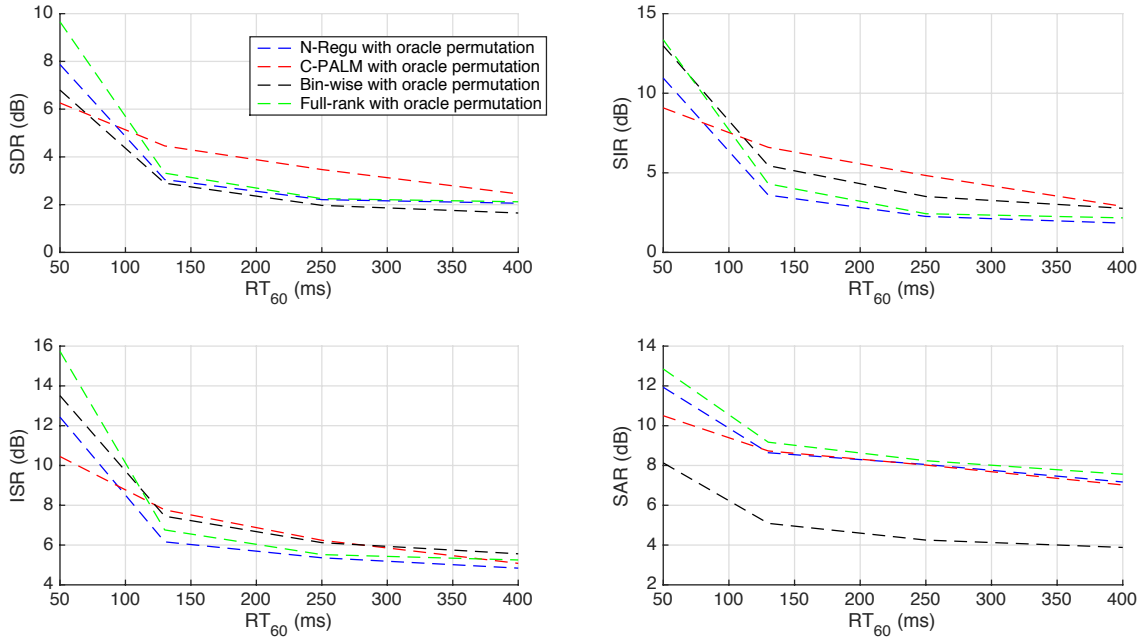

Fig. 3. Separation performance of different algorithms with oracle permutation alignment as a function of the reverberation time RT_{60} in noiseless case

Compared to Bin-wise method, this observation coincides with the intuition that, for noisy mixtures, the coefficients of the noise in the observations should be discarded to achieve better separation. C-PALM achieves its best performance in terms of SDR and SIR when the sparsity level is about 75%.

Fig. 8 illustrates the results of separation as a function of the reverberation time for a fixed input SNR (SNR=15 dB). We can see that N-Regu with TDOA permutation has the best performance in terms of SDR.

F. Synthesized mixtures with different sources positions

In this subsection, we tested the robusteness of the proposed algorithms w.r.t the sources positions. The same room setting as shown in Fig. 1 is used. Fig. 9 illustrates the four tested

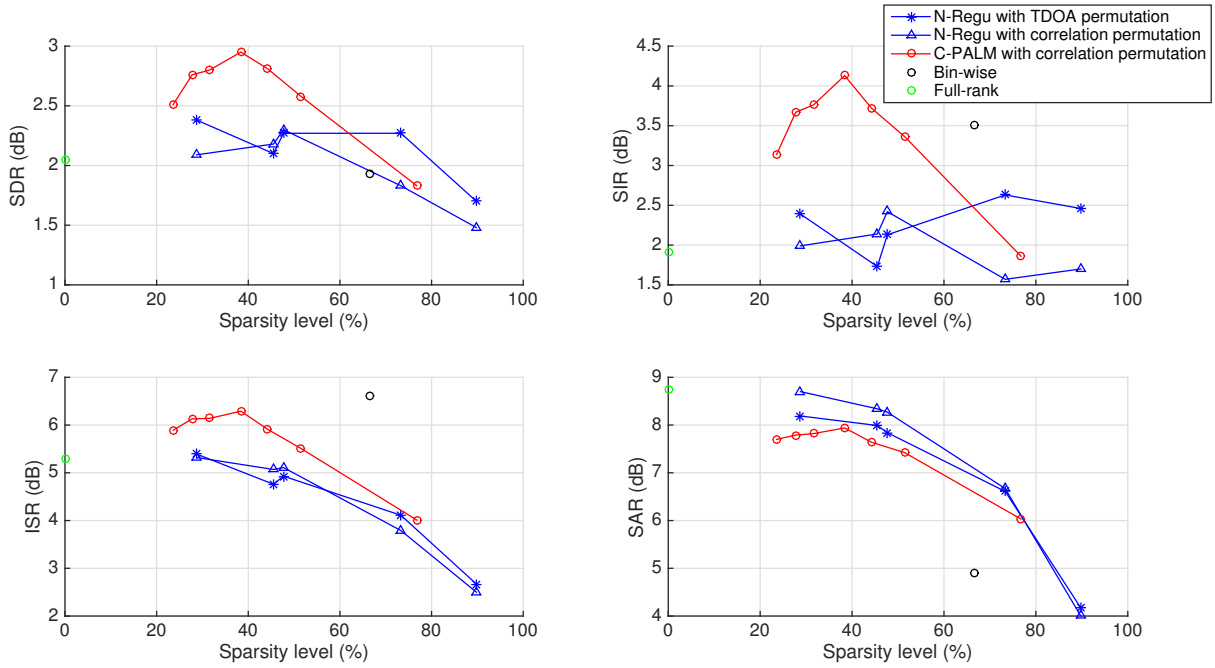

Fig. 4. Separation performance of different algorithms as a function of the sparsity level in noiseless case. $\mathrm{RT}_{60} = 130$ ms

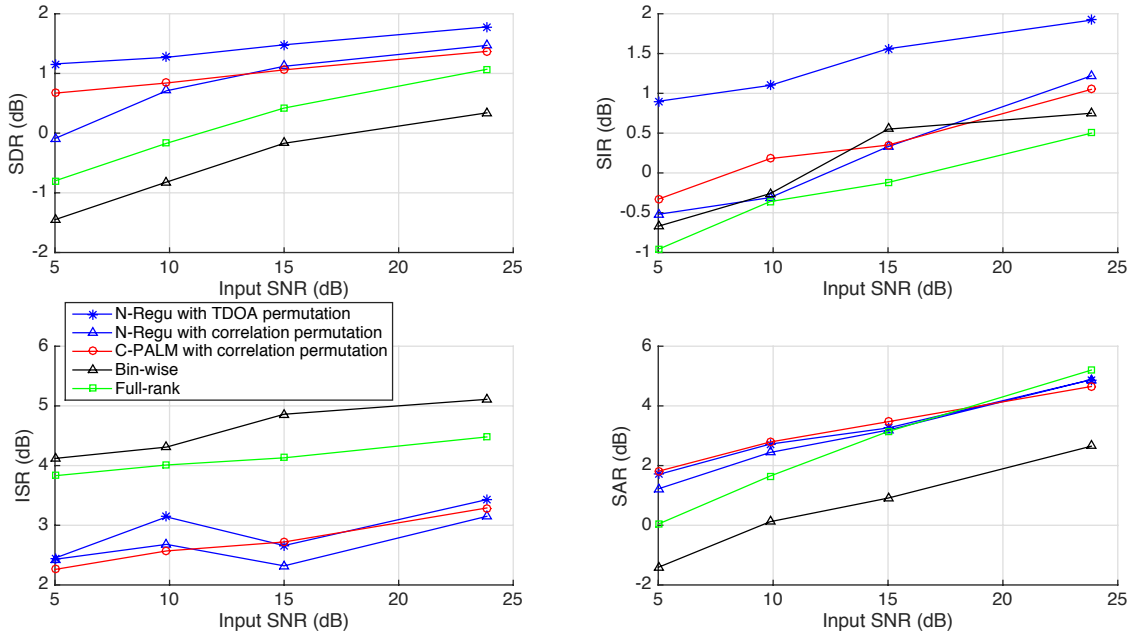

Fig. 5. Separation performance of different algorithms as a function of the input SNR for $\mathrm{RT}_{60} = 130$ ms

sources positions. The same sources positions configuration as in Fig. 1 is also shown as **setting 1** . In these experiments, the reverberation time is fixed to $\mathrm{RT}_{60} = 130$ ms and no noise is added to the mixtures. Fig. 10 shows the separation performance.

It is clear that in terms of SDR, SIR and ISR, all the presented algorithms have the worst performance in **setting 3**. This remark shows that having two sources close to each other and one source relatively far (**setting 3**) could be a more difficult situation for blind source separation than having three sources close to each other (**setting 4**). For C-PALM, it has the best performance in terms of SDR, SIR and ISR for all the settings. This observation shows that C-PALM (and the pre-defined penalty parameter) is robust to sources positions,

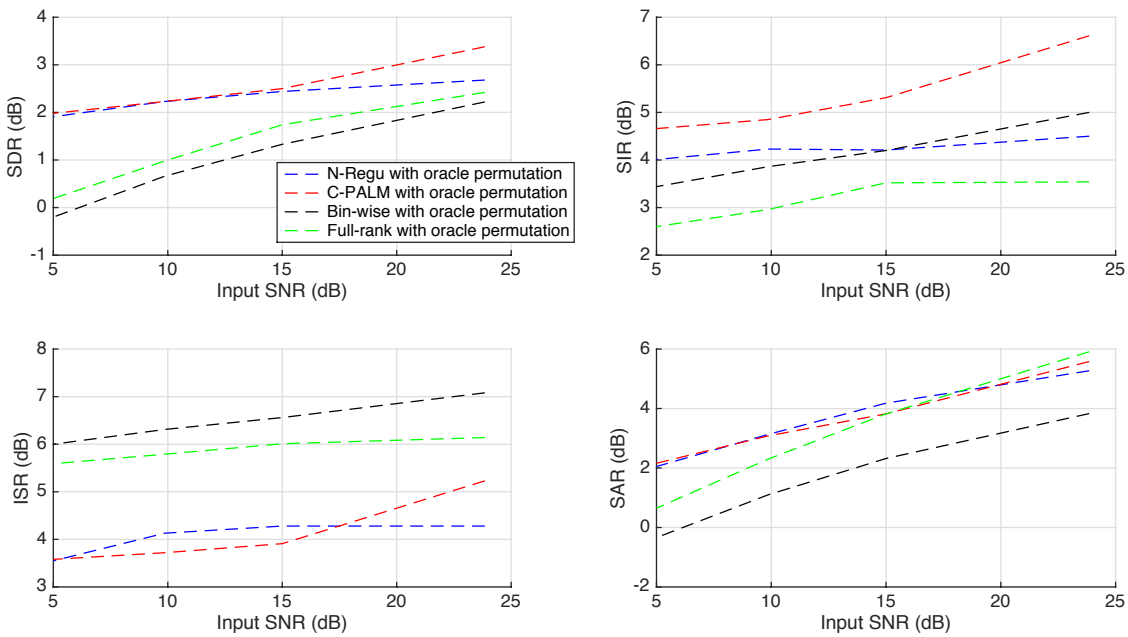

Fig. 6. Separation performance of different algorithms with oracle permutation alignment as a function of the input SNR for $\mathrm{RT}_{60} = 130$ ms

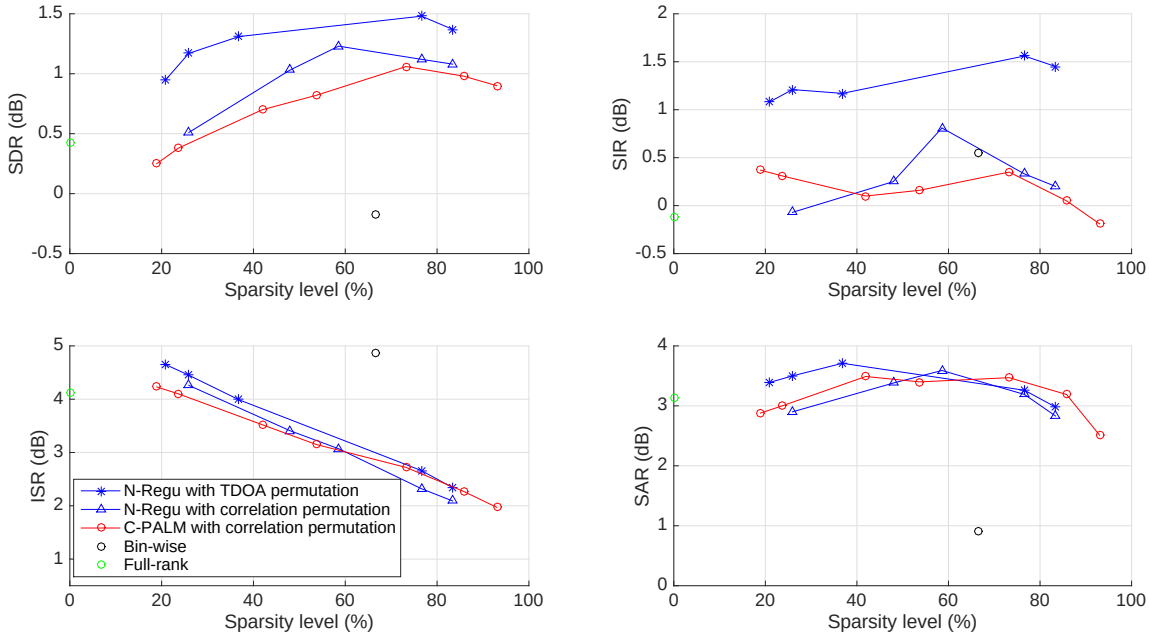

Fig. 7. Separation performance of different algorithms as a function of the sparsity level. $\mathrm{RT}_{60} = 130$ ms. SNR = 15 dB

even with difficult configuration (**setting 3**).

G. Live recorded mixtures without noise

This subsection reports the separation results of C-PALM for publicly avaiable benchmark data in SiSEC2011 [23]. We used the speech signals (*male3, female3, male4* and *female4*) from the first development data (dev1.zip) in "Under-determined speech and music mixtures" data sets. Table II shows the separation results. For C-PALM, we chose the hyperparameter λ such that the sparsity level of the estimated coefficients of the sources is about $20\%, 60\%$ for $\mathrm{RT}_{60} = 130, 250$ ms respectively. Compared to the performances reported in [23], C-PALM obtains relatively good separation results epsecially when the number of sources $N = 3$.

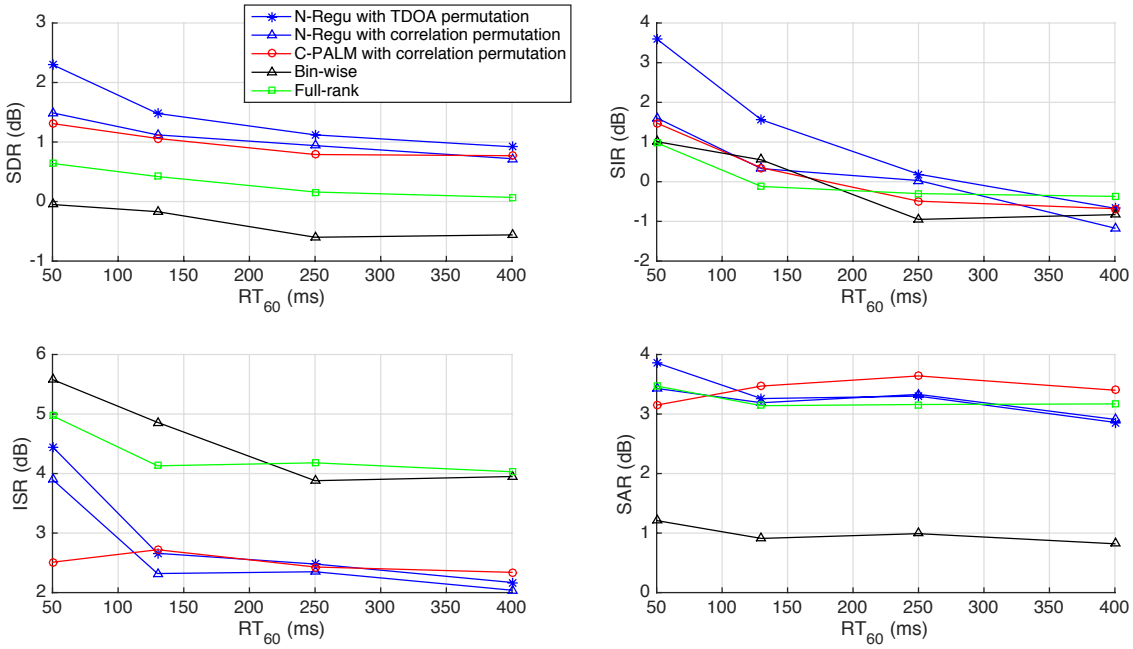

Fig. 8. Separation performance of different algorithms as a function of the reverberation time RT_{60} with input SNR=15 dB

TABLE II
SEPARATION RESULTS OF C-PALM FOR LIVE RECORDED MIXTURES FROM SISEC2011 (SDR/SIR/ISR/SAR IN DB)

	$RT_{60} = 130$ ms		$RT_{60} = 250$ ms	
microphone space	5 cm	1 m	5 cm	1 m
male3	7.65 / 11.38 / 12.10 / 10.65	7.53 / 11.27 / 11.77 / 10.58	5.20 / 7.67 / 9.01 / 8.62	4.98 / 10.62 / 6.67 / 7.04
female3	6.69 / 9.81 / 10.90 / 10.52	9.77 / 14.49 / 14.13 / 13.02	5.29 / 9.16 / 7.77 / 8.75	7.34 / 11.22 / 10.97 / 11.02
male4	3.25 / 4.65 / 6.09 / 6.01	2.34 / 2.15 / 5.16 / 5.47	2.10 / 1.79 / 4.63 / 5.49	3.08 / 4.22 / 6.00 / 6.11
female4	2.36 / 2.05 / 5.37 / 6.53	3.66/ 6.05 / 6.80 / 7.15	2.39 / 2.20 / 5.27 / 6.51	3.12 / 4.51 / 6.07 / 6.84

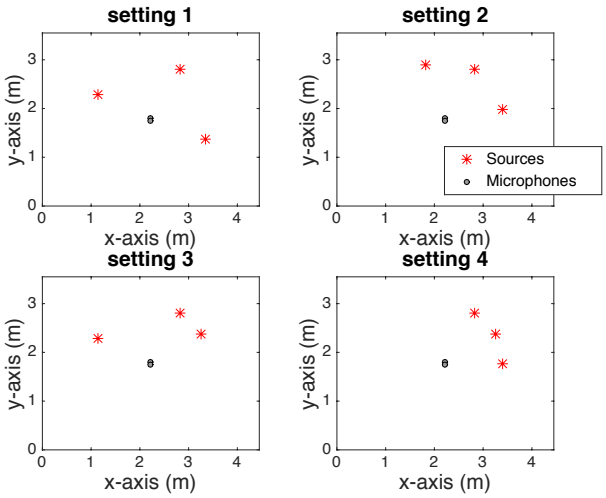

Fig. 9. Different settings of source positions for synthesized mixtures without input noise

H. Computational time

We terminate the expriment section by presenting the computational time of the presented algorithm for the synthesized mixtures in Table III.

TABLE III
COMPUTATIONAL TIME OF DIFFERENT ALGORITHMS FOR ONE SYNTHESIZED MIXTURE

C-PALM	N-Regu	Bin-wise	Full Rank
4960.7 s	1388.8 s	152.9 s	3415.4 s

C-PALM is of relative big computational cost mainly because of the convolution operator in each iteration of the algorithm.

V. CONCLUSION

In this paper, we have developped several approaches for blind source separation with underdetermined convolutive mixtures. Based on the sparsity assumption for the source signals and the statistics of the room impulse response, we developed the N-Regu with multiplicative narrowband approximation and C-PALM with convolutive narrowband approximation. The numerical evaluations show the advantages of C-PALM for noiseless mixtures in strong reverberation scenarios. The experiments also show the good performance of N-Regu for noisy mixtures.

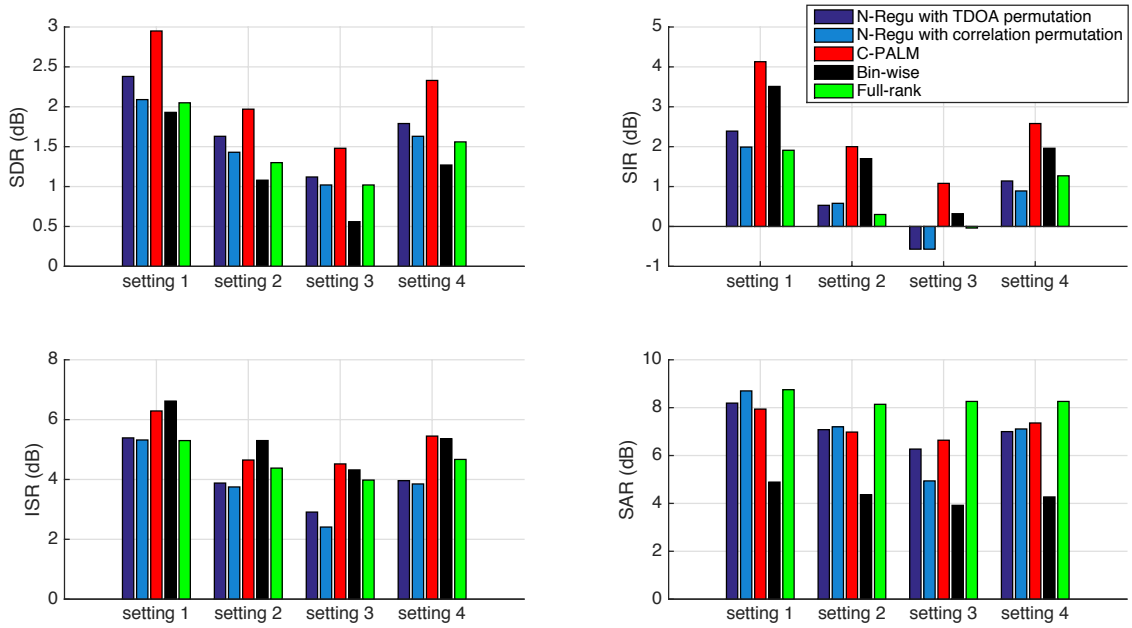

Fig. 10. Separation performance of different algorithms for different sources positions in noiseless case. $\mathrm{RT}_{60} = 130$ ms.

The penalty parameter $\tilde{\rho}(\tau)$ in C-PALM has to be predefined, which makes C-PALM not suitable for low reverberation condition. Future work will concentrate on the estimation of $\tilde{\rho}(\tau)$. In this paper, we used inter-frequency correlation permutation alignment for C-PALM. It would be interesting to exploit TDOA based permutaiton method for convolutive narrowband approximation to improve C-PALM.

VI. Appendix

A. Derivation of N-Regu

We consider the following optimization problem:

$$\min_{\tilde{\mathbf{A}}_f, \boldsymbol{\alpha}_f} \frac{1}{2}\|\tilde{\mathbf{X}}_f - \tilde{\mathbf{A}}_f \boldsymbol{\alpha}_f\|_F^2 + \frac{\mu}{2}\|\tilde{\mathbf{A}}_f\|_F^2 + \lambda\|\boldsymbol{\alpha}_f\|_1 + \imath_{\mathcal{C}}(\tilde{\mathbf{A}}_f). \tag{26}$$

This optimization is equivalent to the problem (6): the indicator function $\imath_{\mathcal{C}}(\tilde{\mathbf{A}}_f)$ forces the normalization on each column of $\tilde{\mathbf{A}}_f$, therefore the term $\frac{\mu}{2}\|\tilde{\mathbf{A}}_f\|_F^2$ is a constant and does not change the minimizer. The reason of adding this term is purely algorithmic. We then solve the optimization (26) with BC-VMFB [13].

Let the general optimization

$$\min_{\mathbf{x},\mathbf{y}} F(\mathbf{x}) + Q(\mathbf{x},\mathbf{y}) + G(\mathbf{y}) \ , \tag{27}$$

where $F(\mathbf{x})$ and $G(\mathbf{y})$ are lower semicontinuous functions, $Q(\mathbf{x},\mathbf{y})$ is a smooth function with Lipschitz gradient on any bounded set. BC-VMFB uses the following update rules to solve (27):

$$\begin{aligned}\mathbf{x}^{(j+1)} =& \operatorname*{argmin}_{\mathbf{x}} F(\mathbf{x}) + \langle \mathbf{x} - \mathbf{x}^{(j)}, \nabla_{\mathbf{x}} Q(\mathbf{x}^{(j)}, \mathbf{y}^{(j)})\rangle \\ &+ \frac{t^{1,(j)}}{2}\|\mathbf{x} - \mathbf{x}^{(j)}\|^2_{\mathbf{U}^{2,(j)}},\end{aligned} \tag{28}$$

$$\begin{aligned}\mathbf{y}^{(j+1)} =& \operatorname*{argmin}_{\mathbf{y}} G(\mathbf{y}) + \langle \mathbf{y} - \mathbf{y}^{(j)}, \nabla_{\mathbf{y}} Q(\mathbf{x}^{(j+1)}, \mathbf{y}^{(j)})\rangle \\ &+ \frac{t^{2,(j)}}{2}\|\mathbf{y} - \mathbf{y}^{(j)}\|^2_{\mathbf{U}^{2,j}},\end{aligned} \tag{29}$$

where $\mathbf{U}^{1,(j)}$ and $\mathbf{U}^{2,(j)}$ are positive definite matrices. $\|\mathbf{x}\|^2_{\mathbf{U}}$ denotes the variable metric norm:

$$\|\mathbf{x}\|^2_{\mathbf{U}} = \langle \mathbf{x}, \mathbf{U}\mathbf{x}\rangle. \tag{30}$$

With the variable metric norm, the proximal operator (8) can be generalized as:

$$\mathrm{prox}_{\mathbf{U},\psi} := \operatorname*{argmin}_{\mathbf{y}} \psi(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2_{\mathbf{U}}. \tag{31}$$

Then (28) and (29) can be rewritten as follow:

$$\begin{aligned}\mathbf{x}^{(j+1)} =& \mathrm{prox}_{\mathbf{U}^{1,(j)}, F/t^{1,(j)}}\Big(\mathbf{x}^{(j)} \\ &- \frac{1}{t^{1,(j)}}\mathbf{U}^{1,(j)^{-1}}\nabla_{\mathbf{x}} Q(\mathbf{x}^{(j)}, \mathbf{y}^{(j)})\Big),\end{aligned} \tag{32}$$

$$\begin{aligned}\mathbf{y}^{(j+1)} =& \mathrm{prox}_{\mathbf{U}^{2,(j)}, G/t^{2,(j)}}\Big(\mathbf{y}^{(j)} \\ &- \frac{1}{t^{2,(j)}}\mathbf{U}^{2,(j)^{-1}}\nabla_{\mathbf{y}} Q(\mathbf{x}^{(j+1)}, \mathbf{y}^{(j)})\Big).\end{aligned} \tag{33}$$

It is shown in [13] that the sequence generated by the above update rules converges to a critical point of the problem (27).

For the problem (26), we make the following substitutions:

$$\begin{aligned}&F(\boldsymbol{\alpha}_f) = \lambda\|\boldsymbol{\alpha}_f\|_1, \\ &Q(\boldsymbol{\alpha}_f, \tilde{\mathbf{A}}_f) = \frac{1}{2}\|\tilde{\mathbf{X}}_f - \tilde{\mathbf{A}}_f\boldsymbol{\alpha}_f\|_F^2 + \frac{\mu}{2}\|\tilde{\mathbf{A}}_f\|_F^2, \\ &G(\tilde{\mathbf{A}}_f) = \imath_{\mathcal{C}}(\tilde{\mathbf{A}}_f),\end{aligned} \tag{34}$$

Denoting by $L^{(j)}$ the Lipschitz constant of $\nabla_{\boldsymbol{\alpha}_f} Q(\boldsymbol{\alpha}_f^{(j)}, \tilde{\mathbf{A}}_f^{(j)})$, we have chosen:

$$\begin{aligned} &\mathbf{U}^{1,(j)} = L^{(j)}\mathbf{I}, \\ &\mathbf{U}^{2,(j)} = \frac{\partial Q(\tilde{\mathbf{A}}_f, \boldsymbol{\alpha}_f^{(j+1)})^2}{\partial^2 \tilde{\mathbf{A}}_f} = \boldsymbol{\alpha}_f^{(j+1)} \boldsymbol{\alpha}_f^{(j+1)^H} + \mu \mathbf{I}, \\ &t^{1,(j)} = t^{2,(j)} = 1. \end{aligned} \tag{35}$$

The update step of the mixing matrix can be written as:

$$\begin{aligned} &\tilde{\mathbf{A}}_f^{(j+1/2)} = \tilde{\mathbf{X}}_f \boldsymbol{\alpha}_f^{(j+1)^H} (\boldsymbol{\alpha}_f^{(j+1)} \boldsymbol{\alpha}_f^{(j+1)^H} + \mu \mathbf{I})^{-1}, \\ &\tilde{\mathbf{A}}_f^{(j+1)} \in \operatorname{prox}_{\mathbf{U}^{2,(j)}, \imath_{\mathcal{C}}}(\tilde{\mathbf{A}}_f^{(j+1/2)}). \end{aligned} \tag{36}$$

As the choice of the parameter μ does not change the minimizer of (26), by choosing μ sufficiently large, the update step of $\tilde{\mathbf{A}}_f$ becomes:

$$\tilde{\mathbf{A}}_f^{(j+1/2)} = \mathcal{P}_{\mathcal{C}} \left(\tilde{\mathbf{X}}_f \boldsymbol{\alpha}_f^{(j+1)^H} \right). \tag{37}$$

We obtain the N-Regu as shown in Algorithm 1.

B. Convolutive mixing operator and its adjoint operators

Given a signal $\mathbf{s} \in \mathbb{C}^T$, and a convolution kernel $\mathbf{h} \in \mathbb{C}^L$, the convolution can be written under the matrix form:

$$\mathbf{x} = \mathcal{H}\mathbf{s} = \mathcal{S}\mathbf{h} \ , \tag{38}$$

$\mathcal{H} \in \mathbb{C}^{T \times T}$ and $\mathcal{S} \in \mathbb{C}^{T \times L}$ being the corresponding circulant matrices.

The convolutive mixing operator can then be represented by

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_M \end{pmatrix} = \begin{pmatrix} \mathcal{H}_{11} & \mathcal{H}_{12} & \dots & \mathcal{H}_{1N} \\ \mathcal{H}_{21} & \mathcal{H}_{22} & \dots & \mathcal{H}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{H}_{M1} & \mathcal{H}_{M2} & \dots & \mathcal{H}_{MN} \end{pmatrix} \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_N \end{pmatrix}, \tag{39}$$

where $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N \in \mathbb{C}^T$ are N source signals and $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M \in \mathbb{C}^T$ are M observations. $\mathcal{H}_{mn}$ is the convolution matrix from the n-th source to the m-th microphone.

Thanks to these notations, the adjoint operator of convolutive mixing with respect to the mixing system is a linear operator $\mathbb{C}^{M \times T} \to \mathbb{C}^{N \times T}$ and can be represented by the following matrix multiplication:

$$\begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_N \end{pmatrix} = \begin{pmatrix} \mathcal{H}_{11}^H & \mathcal{H}_{21}^H & \dots & \mathcal{H}_{N1}^H \\ \mathcal{H}_{12}^H & \mathcal{H}_{22}^H & \dots & \mathcal{H}_{N2}^H \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{H}_{1M}^H & \mathcal{H}_{2M}^H & \dots & \mathcal{H}_{NM}^H \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_M \end{pmatrix}. \tag{40}$$

In order to coincide with the previous notations in (14), we denote the above formulation as:

$$\mathbf{S} = \mathbf{H}^H \star \mathbf{X}. \tag{41}$$

The adjoint operator of the convolutive mixture with respect to the sources can then be written as:

$$\mathbf{H} = \mathbf{X} \star \mathbf{S}^H, \tag{42}$$

with

$$\mathbf{h}_{mn} = \mathcal{S}_n^H \mathbf{x}_m. \tag{43}$$

C. Derivation of C-PALM

The PALM algorithm [12] is designed to solve the non-convex optimization problem in the general form (27) by the following update rules:

$$\begin{aligned} \mathbf{x}^{(j+1)} = \operatorname*{argmin}_{\mathbf{x}} F(\mathbf{x}) + \langle \mathbf{x} - \mathbf{x}^{(j)}, \nabla_{\mathbf{x}} Q(\mathbf{x}^{(j)}, \mathbf{y}^{(j)}) \rangle \\ + \frac{t^{1,(j)}}{2} \|\mathbf{x} - \mathbf{x}^{(j)}\|_2^2, \end{aligned} \tag{44}$$

$$\begin{aligned} \mathbf{y}^{(j+1)} = \operatorname*{argmin}_{\mathbf{y}} G(\mathbf{y}) + \langle \mathbf{y} - \mathbf{y}^{(j)}, \nabla_{\mathbf{y}} Q(\mathbf{x}^{(j+1)}, \mathbf{y}^{(j)}) \rangle \\ + \frac{t^{2,(j)}}{2} \|\mathbf{y} - \mathbf{y}^{(j)}\|_2^2, \end{aligned} \tag{45}$$

where j is the iteration index and $t^{1,(j)}$ et $t^{2,(j)}$ are two step parameters.

It is shown in [12] that the sequence generated by the above update rules converges to a critical point of the problem (27).

From the general optimization (27), we do the following substitutions:

$$\begin{aligned} F(\boldsymbol{\alpha}_f) &= \lambda \|\boldsymbol{\alpha}_f\|_1, \\ Q(\boldsymbol{\alpha}_f, \tilde{\mathbf{H}}_f) &= \frac{1}{2} \|\tilde{\mathbf{X}}_f - \tilde{\mathbf{H}}_f \star \boldsymbol{\alpha}_f\|_F^2 + \mathcal{P}(\tilde{\mathbf{H}}_f), \\ G(\tilde{\mathbf{H}}_f) &= \imath_{\mathcal{C}}^{\text{Conv}}(\tilde{\mathbf{H}}_f), \end{aligned} \tag{46}$$

and the particular choices:

$$t^{1,(j)} = L^{1,(j)}, \quad t^{2,(j)} = L^{2,(j)},$$

where $L^{1,(j)}$ and $L^{2,(j)}$ are respectively the Lipschitz constant of $\nabla_{\boldsymbol{\alpha}_f} Q(\boldsymbol{\alpha}_f^{(j)}, \tilde{\mathbf{H}}_f^{(j)})$ and $\nabla_{\tilde{\mathbf{H}}_f} Q(\boldsymbol{\alpha}_f^{(j+1)}, \tilde{\mathbf{H}}_f^{(j)})$. We obtain the C-PALM algorithm presented in Algorithm 2.

D. Calculation of the Lipschitz constant in C-PALM

We present the calculation of the Lipschitz constant of the function

$$\begin{aligned} I(\tilde{\mathbf{H}}_f) &:= \tilde{\mathbf{H}}_f \star \boldsymbol{\alpha}_f \star \boldsymbol{\alpha}_f^H + \nabla_{\tilde{\mathbf{H}}_f} \mathcal{P}(\tilde{\mathbf{H}}_f) \\ &= \hat{\mathbf{H}}_f + \nabla_{\tilde{\mathbf{H}}_f} \mathcal{P}(\tilde{\mathbf{H}}_f). \end{aligned} \tag{47}$$

Let $\boldsymbol{\Psi}_n$ denotes the circulant matrix associated with $\boldsymbol{\alpha}_n$. If the synthesis coefficients of different sources are independent, we have

$$\mathrm{E}\left[\boldsymbol{\Psi}_i \boldsymbol{\Psi}_j\right] = \mathbf{0}, \quad \text{for } i \neq j.$$

Then, using similar notations as in Appendix VI-B, one can write $\hat{\mathbf{H}}$ as:

$$\hat{\mathbf{h}}_{mn} = \boldsymbol{\Psi}_n^H \boldsymbol{\Psi}_n \tilde{\mathbf{h}}_m, \tag{48}$$

Finally, Proposition 4 comes from the definition of the Lipschitz constant.

References

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[2] 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.

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

[4] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1830–1840, 2010.

[5] W. Kellermann and H. Buchner, "Wideband algorithms versus narrowband algorithms for adaptive filtering in the DFT domain," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. IEEE, 2003, pp. 1278–1282.

[6] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain bss," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*. IEEE, 2007, pp. 3247–3250.

[7] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural computation*, vol. 13, no. 4, pp. 863–882, 2001.

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

[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, Special Issue on: "Processing Reverberant"*, vol. 17, no. 7, pp. 1818–1829, 2010.

[10] 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*, 2017.

[11] A. Benichoux, L. S. Simon, E. Vincent, and R. Gribonval, "Convex regularizations for the simultaneous recording of room impulse responses," *IEEE Transactions on Signal Processing*, vol. 62, no. 8, pp. 1976–1986, 2014.

[12] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, pp. 1–36, 2013.

[13] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "A block coordinate variable metric forward–backward algorithm," *Journal of Global Optimization*, vol. 66, no. 3, pp. 457–485, 2016.

[14] P. Balazs, M. Doerfler, M. Kowalski, and B. Torresani, "Adapted and adaptive linear time-frequency representations: a synthesis point of view," *IEEE Signal Processing Magazine*, vol. 30, no. 6, pp. 20–31, 2013.

[15] F. Feng and M. Kowalski, "A unified approach for blind source separation using sparsity and decorrelation," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 1736–1740.

[16] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function," *Journal of Optimization Theory and Applications*, pp. 1–26, 2013.

[17] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011, pp. 185–212.

[18] 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.

[19] 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.

[20] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[21] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, pp. 530–538, 2004.

[22] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[23] S. Araki, F. Nesta, E. Vincent, Z. Koldovskỳ, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (sisec2011):-audio source separation," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 414–422.

[24] F. Feng and M. Kowalski, "Sparsity and low-rank amplitude based blind source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 571–575.

[25] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," *Independent Component Analysis and Signal Separation*, pp. 552–559, 2007.

[26] E. T. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for ℓ_1-minimization: Methodology and convergence," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1107–1130, 2008.