Sparse Reverberant Audio Source Separation via Reweighted Analysis

Simon Arberet, Pierre Vandergheynst, Rafael E. Carrillo, Jean-Philippe Thiran and Yves Wiaux

***Abstract*— We propose a novel algorithm for source signals estimation from an underdetermined convolutive mixture assuming known mixing filters.**

Most of the state-of-the-art methods are dealing with anechoic or short reverberant mixture, assuming a synthesis sparse prior in the time-frequency domain and a narrowband approximation of the convolutive mixing process. In this paper, we address the source estimation of convolutive mixtures with a new algorithm based on i) an analysis sparse prior, ii) a reweighting scheme so as to increase the sparsity, iii) a wideband data-fidelity term in a constrained from. We show, through theoretical discussions and simulations, that this algorithm is particularly well suited for source separation of realistic reverberation mixtures. Particularly, the proposed algorithm outperforms state-of-the-art methods on reverberant mixtures of audio sources by more than 2 dB of signal-to-distortion ratio.

I. Introduction

Most audio recordings can be viewed as mixtures of several audio signals (e.g., musical instruments or speech), called *source signals* or *sources*, that are usually active simultaneously. The sources may have been mixed synthetically with a mixing console or by recording a real audio scene using microphones.

The mixing of N audio sources on M channels is often formulated as the following convolutive mixing model:

$$x_m(t) = \sum_{n=1}^{N} (a_{mn} \star s_n)(t) + e_m(t), \quad 1 \leq m \leq M, \qquad (1)$$

where $s_n(t) \in \mathbb{R}$ and $x_m(t) \in \mathbb{R}$ denote sampled time signals of respectively the n-th source and the m-th mixture (t being a discrete time index), $a_{mn}(t) \in \mathbb{R}$ denote the finite (sampled) impulse response of some causal filter, and $\star$ denotes convolution.

The goal of the *convolutive Blind Source Separation (BSS)* problem is to estimate the N source signals $s_n(t)$ ($1 \leq n \leq N$), given the M mixture signals $x_m(t)$ ($1 \leq m \leq M$).

When the number of sources is larger than the number of mixture channels ($N > M$), the BSS problem is said to be *underdetermined* and is often addressed by sparsity-based approaches [1]–[4] consisting in the following two steps: i) at the first step the mixing parameters are estimated as in [3], [5]–[7], and ii) at the second step, the source are estimated e.g. using a minimum mean squared error (MMSE) or a Maximum A Posteriori (MAP) estimator given a sparse source prior and the mixing parameters.

The authors are with the Signal Processing Laboratory LTS2, Electrical Engineering Department, École Polytechnique Fédérale de Lausanne (EPFL), Station 11, CH-1015 Lausanne, Switzerland. This work was supported in part by the European Union through the project SMALL (Sparse Models, Algorithms and Learning for Large-Scale data). The project SMALL acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 225913.
E-mail:simon.arberet@epfl.ch.

Since audio signals are usually not sparse in the time domain, the estimation of the source coefficients is done in some time-frequency (TF) domain, where they are sparse, by using for example the short time Fourier transform (STFT). The mixing equation (1) is then approximated by the so-called narrowband approximation [8], as follows [9]:

$$\tilde{\mathbf{x}}(t,f) \approx \tilde{\mathbf{A}}(f)\tilde{\mathbf{s}}(t,f) + \tilde{\mathbf{e}}(t,f) \qquad (2)$$

where $\tilde{\mathbf{x}}(t,f) \in \mathbb{C}^M$, $\tilde{\mathbf{s}}(t,f) \in \mathbb{C}^N$ and $\tilde{\mathbf{e}}(t,f) \in \mathbb{C}^M$ are the vectors of mixture, source and noise STFT coefficients in TF bin (t,f), and $\tilde{\mathbf{A}}(f) = [\tilde{a}_{mn}(f)]_{m,n=1}^{M,N}$ is an $M \times N$ complex-valued mixing matrix with elements $\tilde{a}_{mn}(f)$ being the discrete Fourier transforms of the $m \times n$ filters $a_{mn}(t), \forall t$.

More recently, the limitation of the narrowband approximation for reverberant source separation has been pointed out, and some approaches have been proposed to overcome it for the mixing filter estimation [7] and the sources estimation [10]–[12]. In [10] and [11], the narrowband approximation has been circumvented via a statistical modeling of the mixing process, while in [12], Kowalski et al. proposed a convex optimization framework based on a wideband ℓ_2 mixture fitting cost.

While *synthesis* sparse priors have been widely used for the source modeling including ℓ_0 cost (i.e. binary masking) [3] ℓ_1 cost [2], [12], [13], ℓ_p cost [14], and mixed norm $\ell_{1,2}$ cost [12], *analysis* sparse prior has to our knowledge never been used in audio source separation.

In this article, we focus on addressing the source estimation task, assuming that the mixing filters a_{mn} are known. We propose a novel algorithm, for convolutive source separation which is based on a reweighted scheme of an *analysis* sparse prior. This algorithm introduces three important contributions with respect to the state-of-the-art, and which will be carefully evaluated in the experimental section: i) The algorithm is based on an analysis sparsity prior, which is fundamentally different than the synthesis prior when the analysis operator is a redundant frame (such as a redundant STFT), ii) the algorithm is based on a reweighting scheme that mimics the ℓ_0 minimization behaviour and thus promotes a stronger sparsity assumption than the ℓ_1 cost, and iii) the algorithm is based on a wideband mixture fitting constraint, and thus first avoid the narrowband approximation, and secondly offers a strong fidelity term (in a new constrained formulation) without the need to fix a regularization parameter (in a standard regularized formulation).

The organization of the remainder of the paper is the following. In section II, we introduce our notations and the state-of-the-art approaches. In section III, we discuss convex optimization approaches for sparse inverse problems.

In section IV, we introduce our algorithm for audio source separation, and in section V, we provide numerical results of our algorithm compared to the state-of-the-art methods.

II. State of the Art

A. The convolutive mixture model in operator and matrix form

The mixture model (1) can be written as:

$$\mathbf{x} = \mathcal{A}(\mathbf{s}) + \mathbf{e}. \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^{M\times T}$ is the matrix of the mixture composed of $x_m(t)$ entries, $\mathbf{s} \in \mathbb{R}^{N\times T}$ is the matrix of sources composed of the $s_n(t)$ entries, $\mathbf{e} \in \mathbb{R}^{M\times T}$ is the matrix of the noise composed of $e_m(t)$ entries, and $\mathcal{A} : \mathbb{R}^{N\times T} \to \mathbb{R}^{M\times T}$ is the discrete linear operator defined by

$$[\mathcal{A}(\mathbf{s})]_{m,t} = \sum_{n=1}^{N} (a_{mn} \star s_n)(t).$$

The adjoint operator $\mathcal{A}^* : \mathbb{R}^{M\times T} \to \mathbb{R}^{N\times T}$ of $\mathcal{A}$ is obtained by applying the convolution mixing process with the ajoint filters $a^*_{nm}(t) \triangleq a_{mn}(-t), \forall t$ instead of a_{mn}, that is: $[\mathcal{A}^*(\mathbf{x})]_{n,t} = \sum_{m=1}^{M} (a^*_{nm} \star x_m)(t)$.

Note that Eq. (3) can be written in the following a matrix form:

$$\mathbf{x}_{\text{vec}} = \mathbf{A}\mathbf{s}_{\text{vec}} + \mathbf{e}_{\text{vec}},$$

where $\mathbf{x}_{\text{vec}} \in \mathbb{R}^{MT}$, $\mathbf{s}_{\text{vec}} \in \mathbb{R}^{NT}$ and $\mathbf{e}_{\text{vec}} \in \mathbb{R}^{MT}$ are the unfolded vectors of the matrices $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{e}$, respectively, and $\mathbf{A}$ is a matrix of size $MT \times NT$ composed of $M \times N$ Toeplitz blocs A_{mn} of size $T \times T$.

B. Time-frequency transform

Underdetermined source separation is an ill-posed inverse problem, which needs additional assumptions to be solved. As stated in the introduction, a powerful assumption is the sparsity of the sources in some representation. Audio signals are known to be sparse in the time-frequency (TF) domain, and a popular TF representation is obtained via the Short Time Fourier Transform (STFT).

The STFT operator $\Psi \in \mathbb{C}^{T\times B}$ transforms a multichannel signal $\mathbf{s}$ of length T, into a matrix $\tilde{\mathbf{s}} \in \mathbb{C}^{N\times B}$ of B time-frequency coefficients per channel:

$$\tilde{\mathbf{s}} = \mathbf{s}\,\Psi, \tag{4}$$

and the ISTFT is obtained by applying the adjoint operator $\Psi^* \in \mathbb{C}^{B\times T}$ on the STFT coefficients $\tilde{\mathbf{s}}$:

$$\mathbf{s} = \tilde{\mathbf{s}}\,\Psi^*. \tag{5}$$

C. Narrowband methods

a) Binary masking: The DUET method [3], as most of the convolutive source separation methods [2], [13], [14] is based on the narrowband approximation (2). DUET exploits the assumption that at each TF point, there is approximately one active source, meaning that the ℓ_0 quasi-norm of the source vector $\tilde{\mathbf{s}}(t,f)$, i.e. the number of nonzero entries of $\tilde{\mathbf{s}}(t,f)$, is equal to one: $\|\tilde{\mathbf{s}}(t,f)\|_0 = 1$. The DUET method minimizes, for each TF point, a narrowband data fidelity term under a ℓ_0 constraint on the source vector:

$$\underset{\tilde{\mathbf{s}}(t,f)\in\mathbb{C} \text{ s.t. } \|\tilde{\mathbf{s}}(t,f)\|_0=1}{\operatorname{argmin}} \|\tilde{\mathbf{x}}(t,f) - \tilde{\mathbf{A}}(f)\tilde{\mathbf{s}}(t,f)\|_2^2. \tag{6}$$

As there is only one possible active source, this problem can be efficiently minimized with a combinatorial optimization strategy, where the data fidelity cost associated with each possible active source is computed, and the selected source is the one leading to the lowest cost.

b) ℓ_p norm minimization: Other methods relax the ℓ_0 cost with an ℓ_p norm, with $p > 0$, which is defined for a vector or a matrix $\mathbf{z}$ with I entries as $\|\mathbf{z}\|_p = (\sum_{i=1}^{I} |z_i|^p)^{1/p}$. The ℓ_p norm promotes sparsity when $p \leq 1$ but is not convex when $p < 1$. As a consequence the ℓ_1 norm is often preferred. Experiments [12] show that the DUET method is more robust than the ℓ_1 norm minimization in reverberant situations.

D. Wideband Lasso

Kowalski et al. [12] proposed a variational formulation of the source estimation problem where the data fidelity term used in (6) which is based on the narrowband model (2) is replaced with a wideband data fidelity term according to the time mixing model (1) (equivalently (3)). This leads to the following *Wideband Lasso* (WB Lasso):

$$\underset{\tilde{\mathbf{s}}\in\mathbb{C}^{N\times B}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{x} - \mathcal{A}(\tilde{\mathbf{s}}\,\Psi^*)\|_2^2 + \lambda\mathcal{P}(\tilde{\mathbf{s}}). \tag{7}$$

The first term is the wideband data fidelity term measuring the fit between the observed mixture $\mathbf{x}$ and the mixing model (3) obtained with the source STFT coefficients $\tilde{\mathbf{s}}$ given the mixing system $\mathcal{A}$, and the second term $\mathcal{P}(\tilde{\mathbf{s}})$ is a sparse synthesis regularization term. The parameter $\lambda \in \mathbb{R}_+$ governs the balance between the data term and the regularization term. Kowalski et al. [12] proposed to minimize problem (7) for an ℓ_1 cost $\mathcal{P}(\tilde{\mathbf{s}}) = \|\tilde{\mathbf{s}}\|_1$ and a mixed norm cost $\mathcal{P}(\tilde{\mathbf{s}}) = \|\tilde{\mathbf{s}}\|_{1,2}^2$ (in order to promote disjointness of the source in the TF domain without constraining the number of active sources per TF bin) using a forward backward scheme (ISTA) [15], [16] and its accelerated versions [17] (FISTA and the Nesterov scheme).

Experimental results in [12] showed that, in reverberant situations, the Wideband Lasso (WB Lasso) was significantly better than the Narrowband Lasso (i.e. problem (7) with the narrowband data fidelity term of (6)) and that the $\ell_{1,2}$ mixed norm regularization on the sources was not performing better than the ℓ_1 norm regularization.

III. Optimization

A. Constrained vs unconstrained problems

The Lasso (also called Basis pursuit denoising (BPDN)) problem (7) with $\mathcal{P}(\tilde{\mathbf{s}}) = \|\tilde{\mathbf{s}}\|_1$ has an alternative constrained formulation:

$$\begin{aligned} &\underset{\tilde{\mathbf{s}}\in\mathbb{C}^{N\times B}}{\operatorname{argmin}} \|\tilde{\mathbf{s}}\|_1 \\ &\text{subject to } \|\mathbf{x} - \mathcal{A}(\tilde{\mathbf{s}}\,\Psi^*)\|_2 \leq \epsilon, \end{aligned} \tag{8}$$

which is equivalent to the unconstrained formulation (7) for some (unknown) value of λ and ϵ. It follows that determining

the proper value of λ in (7) is akin to determining the power limit of the noise [18]. However, there is no optimal strategy to fix the regularization parameter λ even if the noise level is known, therefore constrained problems, such as (8), offer a stronger fidelity term when the noise power is known, or can be estimated a priori.

Moreover, as stated in [12], [19], algorithms to optimize (8) such as FISTA, have some convergence issue for small values of λ, which is the case in our problem where the noise is very small or null. Indeed, for small λ, FISTA requires a larger number of iterations to reach convergence, and secondly, its convergence speed strongly depends on the chosen initialization. So as to let the algorithm converge in practice, it is possible to use a *continuation trick* [12], [20] also known as *warm start*, which consists in running the algorithm multiple times, first with a large value of λ, and then iteratively decrease the value of λ and initialize the algorithm with the result of the previous run. This *continuation trick* is quite efficient in practice, but requires a significant computational effort specially when λ is small (i.e. when the noise is small).

B. Analysis vs synthesis problems

The BPDN (constrained (8) or unconstrained (7)) defines the optimization in the sparse representation domain finding the optimal representation vector $\tilde{\mathbf{s}}$ and then recovering the true signal trough the synthesis relation $\mathbf{s} = \tilde{\mathbf{s}}\,\Psi^*$. These methods are known as *synthesis* based methods in the literature. Synthesis-based problems may also be substituted by *analysis* based problems, where instead of estimating a sparse representation of the signal, the methods recover the signal $\mathbf{s}$ itself [21]:

$$\begin{aligned} &\underset{\mathbf{s}\in\mathbb{R}^{N\times T}}{\operatorname{argmin}} \ \|\mathbf{s}\,\Psi\,\|_1 \\ &\text{subject to } \|\mathbf{x}-\mathcal{A}(\mathbf{s})\|_2 \leq \epsilon, \end{aligned} \tag{9}$$

In the case of orthonormal bases, the two approaches are equivalent. However, when Ψ is a redundant frame or an overcomplete dictionary, the two problems are no longer equivalent. The analysis of the geometry of the two problems, studied in [21], [22], show that there is a large gap between the two formulations. One remark to make is that the analysis problem does not increase the dimensionality of the problem (relative to the signal dimension) when overcomplete dictionaries are used. Empirical studies have shown very promising results for the analysis approach [21]. [23] provides a theoretical analysis of the ℓ_1 analysis problem coupled with redundant dictionaries in the context of compressed sensing.

C. Reweighted ℓ_1 vs ℓ_1 minimization

As discussed above, the ℓ_1 minimization problem is equivalent to ℓ_0 minimization when the *forward* operator $\mathcal{A}$ (in Eq. (3)) satisfies certain conditions defined in the context of compressed sensing. The main difference between these two problems, is that the ℓ_0 minimization does not depend on the magnitudes of the coefficients, while the ℓ_1 does. One way to mimic the minimization behavior of the ℓ_0 cost is to replace the ℓ_1 norm in (8) by a weighted ℓ_1 norm [24], which is defined, for a vector or a matrix $\mathbf{z}$ with I entries z_i, as $\sum_{i=1}^{I} w_i|z_i|$. The idea behind the weighted ℓ_1 norm, is that large weights will encourage small components while small weights will encourage larger components. Moreover, if the non-zero components have very small weights, the weighted norm will be independent of the precise value of the non-zero components. The appropriate weights can be computed by solving a sequence of weighted ℓ_1 problems, each using as weights essentially the inverse of the values of the solution of the previous problem. However, in order to avoid infinite weights, a small parameter has to be added to the signal values when computing the inverse. This procedure has been observed to be very effective in reducing the number of measurements needed for recovery, and to outperform standard ℓ_1-minimization in many situations, see e.g. [23], [24].

D. Source recovery guarantees

The literature on Compressed Sensing (CS) and sparse recovery gives some insight about the theoretical guarantees to recover the sources from a mixture by solving problem (9).

The sufficient recovery condition for the analysis problem (9) depends [23] both on the analysis sparsity of the sources ($k = \|\mathbf{s}\,\Psi\,\|_0$) and on properties of the *forward* operator $\mathcal{A}$, more precisely its matrix form $\mathbf{A}$. The analysis dictionary Ψ is supposed to be a tight frame and can be highly coherent. A sufficient condition for accurate recovery, is that the matrix $\mathbf{A}$ satisfies the so-called D-RIP and that the signal has a sparse representation in Ψ. We do not have any proof of the validity of the D-RIP for matrix $\mathbf{A}$, but there are some clues about the conditions of its validity.

Indeed, $\mathbf{A}$ is a matrix composed of $M \times N$ Toeplitz blocks, each of them corresponding to the convolution with the filter a_{mn}. It has been shown [25] that Toeplitz matrices satisfy, with an overwhelming probability, the restricted isometry property (RIP), for filters of length P having i.i.d. Gaussian (or zero mean bounded distribution) entries, when the sparsity k of the signal of length T is such that $k < c\sqrt{P/\log(T)}$, for a constant c. Other results [23] show that a lot of matrices that satisfy the RIP (e.g. matrices with Gaussian, subgaussian, or Bernoulli entries) also satisfy the D-RIP. If we assume it is also the case for random Toeplitz matrices, we see that, in the case of a single-channel, single-source "mixture" ($M = N = 1$), the longer the filters, the larger the sparsity of the sources can be. This discussion suggests that the source estimation, solving problem (9), should be better in reverberant conditions, where the filters are long (large P) and where the coefficients of the filters can be relatively well modeled by an i.i.d. Gaussian distribution or zero mean bounded distribution. This trend will be confirmed in the experimental section V-D.

IV. Reweighted Analysis Algorithm

Our proposed algorithm is based on a reweighted ℓ_1 analysis method.

Let us define the weighted ℓ_1 problem

$$\begin{aligned} &\underset{\mathbf{s}\in\mathbb{R}^{M\times T}}{\operatorname{argmin}} \ \|\mathbf{s}\,\Psi\,\|_{\mathrm{W},1} \\ &\text{subject to } \|\mathbf{x}-\mathcal{A}(\mathbf{s})\|_2 \leq \epsilon, \end{aligned} \tag{10}$$

where $\mathrm{W} \in \mathbb{R}_+^{N \times B}$ is a matrix with positive entries w_{ij}, and $\|\mathbf{z}\|_{\mathrm{W},1} \triangleq \sum_{i,j} w_{ij}|z_{ij}|$ is the weighted ℓ_1 norm and ϵ is a bound on the ℓ_2 norm of the noise $\mathbf{e}$.

Assuming i.i.d. real Gaussian noise with variance σ_x^2, the ℓ_2 norm term in (10) follows a χ^2 distribution with MT degrees of freedom. Thus we can set $\epsilon^2 = (MT + 2\sqrt{2MT})\sigma_x^2$, where σ_x^2 is the variance of the noise. This choice provides a likely bound for $\|\mathbf{e}\|_2$, since the probability that $\|\mathbf{e}\|_2^2$ exceeds ϵ^2 is the probability that a χ^2 with MT degrees of freedom exceeds its mean, MT, by at least two times the standard deviation $\sqrt{2MT}$, which is very small.

In the noise-free case, we can choose a very small value of ϵ ($\epsilon \to 0$), or replace the ℓ_2 constraint by the linear equality constraint $\mathbf{x} = \mathcal{A}(\mathbf{s})$. The solution to (10) is denoted as $\Delta(\mathbf{x}, \mathcal{A}, \mathrm{W}, \epsilon)$, which is a function of the data vector $\mathbf{x}$, the mixing operator $\mathcal{A}$, the weights matrice W, and the bound ϵ on the noise level estimator.

Recall that in the reweighting approach a sequence of weighted ℓ_1 problems is solved, each using as weights essentially the inverse of the values of the solution of the previous problem. In practice, we update the weights at each iteration, i.e. after solving a complete weighted ℓ_1 problem, by the function

$$f(\delta, \cdot) = \frac{\delta}{\delta + |\cdot|} \tag{11}$$

applied entrywise on the weights $w_{ij}, \forall i, j$.

So as to approximate the ℓ_0 norm, we used the reweighted ℓ_1 algorithm with a homotopy strategy [20] which consists in solving a sequence of weighted ℓ_1 problems with a decreasing sequence $\{\delta^{(k)}\}$ (k denoting the iteration variable) and warm start initialization. This process is then repeated until a stationary solution is reached [20].

A. The SSRA algorithm

The resulting algorithm defined in Algorithm 1 is similar to the Sparsity Averaging Reweighted Analysis (SARA) algorithm recently proposed by part of the authors for compressive imaging [26], [27]. The main difference is that our redundant sparsity operator Ψ is not built as concatenation of orthonormal bases and that the *forward* operator $\mathcal{A}$ involved in (10) to compute $\Delta(\mathbf{x}, \mathcal{A}, \mathrm{W}, \epsilon)$ are different.

A rate parameter β, with $0 < \beta < 1$, controls the decrease of the sequence $\delta^{(k)} = \beta\delta^{(k-1)} = \beta^k\delta_0$ such that $\delta^{(k)} \to 0$ as $k \to \infty$. However, if there is noise, we set a lower bound as $\delta^{(k)} > \sigma_{\tilde{s}}$, where $\sigma_{\tilde{s}}$ is the standard deviation of the noise in the representation domain and is computed as $\sigma_{\tilde{s}} = \sigma_x\sqrt{MT/2NB}$, which gives a rough estimate for a baseline above which significant signal components could be identified.

As a starting point we set $\mathbf{s}^{(0)}$ as the solution of the ℓ_1 problem and $\delta^{(0)} = \mathrm{std}(\mathbf{s}^{(0)}\,\Psi)$, where $\mathrm{std}(\cdot)$ stands for the empirical standard deviation of the signal, fixing the signal scale. The reweighting process ideally stops when the relative variation between successive solutions $\|\mathbf{s}^{(k)} - \mathbf{s}^{(k-1)}\|_2/\|\mathbf{s}^{(k-1)}\|_2$ is smaller than some bound η, with $0 < \eta < 1$, or after the maximum number of iterations allowed, $K_{\max}$, is reached. In our implementation, we fixed $\eta = 10^{-3}$ and $\beta = 10^{-1}$.

Algorithm 1: SSRA algorithm for source estimation

Input: $\mathbf{x}$, $\mathcal{A}$, Ψ, ϵ.
Initialize:
$k := 1$, $\mathrm{W}^{(0)} := \mathbf{1}_{N \times B}$, $\rho := 1$.
Compute the solution of Problem (10):
$\mathbf{s}^{(0)} := \Delta(\mathbf{x}, \mathcal{A}, \mathrm{W}^{(0)}, \epsilon)$,
$\delta^{(0)} := \mathrm{std}(\mathbf{s}^{(0)}\,\Psi)$.
while $\rho > \eta$ and $k < K_{max}$ **do**
  Update the weight matrix:
  $\mathrm{W}_{ij}^{(k)} := f\left(\delta^{(k-1)}, \tilde{s}_{ij}^{(k-1)}\right)$, for
  $i = 1, \ldots, N,\ j = 1, \ldots, B$,
  with $\tilde{\mathbf{s}}^{(k-1)} = \mathbf{s}^{(k-1)}\,\Psi$.
  Compute the solution of Problem (10):
  $\mathbf{s}^{(k)} := \Delta(\mathbf{x}, \mathcal{A}, \mathrm{W}^{(k)}, \epsilon)$.
  Update $\delta^{(k)} := \max(\beta\delta^{(k-1)}, \sigma_{\tilde{s}})$.
  Update $\rho := \|\mathbf{s}^{(k)} - \mathbf{s}^{(k-1)}\|_2/\|\mathbf{s}^{(k-1)}\|_2$
  $k := k + 1$
end
return $\mathbf{s}^{(k-1)}$

B. Convex optimization algorithms

At each iteration of Algorithm 1, the solution $\Delta(\mathbf{x}, \mathcal{A}, \mathrm{W}, \epsilon)$ of problem (10) has to be computed. Problem (10) consists of minimizing a non-smooth convex function under an ℓ_2-ball constraint. Hence, it is not possible to use conventional smooth optimization techniques based on the gradient. However we can use proximal optimization methods [28] that are efficient convex optimization algorithms that can deal with non-smooth functions and which are particularly well suited for large scale problems.

We first introduce the general framework of *proximal splitting methods* for solving convex problems. We then derive the proximity operators involved in our optimization problem (10), which defined the elementary operations that are required to fit problem (10) into the general *proximal splitting* framework, and we finally describe the Douglas-Rachford (DR) algorithm which is a well adapted algorithm to solve convex optimization problems involving two non-smooth functions.

1) Proximal splitting methods: Proximal splitting methods solve optimization problems of the form:

$$\underset{\mathbf{z} \in \mathbb{R}^I}{\operatorname{argmin}}\ f_1(\mathbf{z}) + f_2(\mathbf{z}) \tag{12}$$

where $f_1(\mathbf{z})$, $f_2(\mathbf{z})$, are convex functions from $\mathbb{R}^I$ to $]-\infty, +\infty]$. Note that any convex constrained problem can be formulated as an unconstrained problem by using the indicator function $i_C(\cdot)$ of the convex constraint set C as one of the functions in (12), e.g. $f_2(\mathbf{z}) = i_C(\mathbf{z})$ where C represents the constraint set, and $i_C(\mathbf{z}) = 0$ if $\mathbf{z} \in C$, and $+\infty$ otherwise. Problem (10) can be seen as a particular instance of problem (12), with $f_1(\mathbf{s}) = \|\mathbf{s}\,\Psi\,\|_{\mathrm{W},1}$ and $f_2(\mathbf{s}) = i_{\mathcal{B}_{\ell_2}^{\epsilon}}(\mathbf{s})$, where $\mathcal{B}_{\ell_2}^{\epsilon} = \{\mathbf{s} \in \mathbb{R}^{N \times T} \mid \|\,\mathcal{A}(\mathbf{s}) - \mathbf{x}\|_2 \leq \epsilon\}$ is the set of matrices $\mathbf{s}$ that are satisfying the fidelity constraint $\|\mathbf{x} - \mathcal{A}(\mathbf{s})\|_2 \leq \epsilon$.

The key concept in proximal splitting methods is the use of the proximity operator of a convex function, which is a natural extension of the notion of a projection operator onto a

Algorithm 2: Douglas-Rachford algorithm

Initialize: $k = 0$, $\mathbf{z}^{(0)} \in \mathbb{R}^{N \times T}$, $\alpha_k \in (0, 2)$, $\gamma > 0$.
repeat
  $\mathbf{s}^{(k)} = \mathrm{prox}_{\gamma f_2}(\mathbf{z}^{(k)})$
  $\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + \alpha_k(\mathrm{prox}_{\gamma f_1}(2\mathbf{s}^{(k)} - \mathbf{z}^{(k)}) - \mathbf{s}^{(k)})$
  $k = k + 1$.
until convergence;
return $\mathbf{s}^{(k)}$

convex set. For example, the proximal operator of the ℓ_1 norm is the soft-thresholding operator, and the proximal operator of the indicator function of a constraint is simply the projection operator onto the constraint set. Solution to (12) is reached iteratively by successive application of the proximity operator associated with each function f_1 and f_2. See [28] for a review of proximal splitting methods and their applications in signal and image processing.

We give in the appendix, the definition of the proximity operator and then derive these operators for the functions $f_1(\mathbf{s}) = \|\mathbf{s}\,\Psi\,\|_{\mathrm{W},1}$ and $f_2(\mathbf{s}) = i_{\mathcal{B}^{\epsilon}_{\ell_2}}(\mathbf{s})$ of our optimization problem (10).

2) Douglas-Rachford Algorithm: The Douglas-Rachford (DR) algorithm [29] solves problem (12) by splitting, i.e. by performing a sequence of calculations involving separately the individual proximity operators $\mathrm{prox}_{\gamma f_1}$ and $\mathrm{prox}_{\gamma f_2}$. Moreover It does not require Lipschitz-differentiability of any of the functions f_i. The general form of the DR algorithm to solve problem (12) is given in Algorithm 2. This algorithm has been proved to converge to a solution of Problem (12). In practice, we used the value $\alpha_k = 1, \forall k$, and $\gamma = 0.1$.

While the DR algorithm converges when the number of iterations tends to infinity, we have to choose a stopping criterion. We chose to stop the algorithm when the relative change of the objective value between two successive estimates is less than a given value η_{dr}, i.e. $\left|f_1(\mathbf{s}^{(k)}) - f_1(\mathbf{s}^{(k-1)})\right| / f_1(\mathbf{s}^{(k)}) < \eta_{\mathrm{dr}}$, or when the number of iterations is greater than a given value M_{iter}. In our experiments, we fixed $\eta_{\mathrm{dr}} = 0.01$, and $M_{\mathrm{iter}} = 200$.

V. Experiments

We evaluated our algorithm with state-of-the-art methods over convolutive mixtures of speech sources in different mixing conditions. For all the experiments, the test signals are sampled at 11 kHz and we use a STFT with cosine windows.

A. Experimental protocol

We used the same experimental protocol as in [12]. The mixing filters were room impulse responses simulated via the Roomsim toolbox [30], with a room size of dimension 3.55 m $\times$ 4.45 m, and with the same microphones and sources configuration as in [12], [31]. The number of microphones was $M = 2$, and the number of sources was varied in the range $3 \leq N \leq 6$. The different sets of mixing filters were generated corresponding to three different reverberation times RT_{60} (anechoïc, $RT_{60} = 50$ ms, $RT_{60} = 250$ ms) and two different microphone spacings d ($d = 5$ cm and $d = 1$ m). Each set of mixing filters was convolved with 10 different sets of male and female speech sources yielding 10 mixtures per mixing condition. We choose to not add additional noise to the mixture in order to only evaluate the source separation performance of the algorithms.

TABLE I
Source separation methods based on the wideband data-fidelity term.

Methods	Reweighting	Analysis	Constrained
SSRA	✓	✓	✓
BPDN A	×	✓	✓
BPDN S	×	×	✓
WB Lasso	×	×	×

In order to evaluate each feature of our SSRA algorithm (Algorithm 1), we evaluate different variations of it: (i) with and without reweighting, (ii) with synthesis instead of analysis, (iii) using the Lagrangian unconstrained formulation instead of the constrained one. These different variations are summarized in table I. For all the constrained methods, we set $\epsilon = 10^{-4}$. Note that ideally we would have set $\epsilon = 0$ or used the proximity operator (24) in the noise-free case, but both approaches take an infinite number of iterations to reach convergence, and thus we need anyway to specify a tolerance.

Note that the unconstrained synthesis approach corresponds to the WB Lasso method of [12], and that the reweighting approach cannot be easily used in the unconstrained case, because of the difficulty to adjust the λ parameter. Indeed, changing the weights of the weighted ℓ_1 norm would automatically change the balance between the data-fidelity term and the regularization term and thus a new value of λ should be set to compensate this unbalance.

As stated before, the WB Lasso method needs the continuation trick to converge but at the cost of additional computation. In the following experiment, we used the continuation trick (CT) with the sequence $\lambda^{(k)} = 10^{-k}, k = 1, \ldots, 8$. We also give the results without the continuation trick, that is when λ is set directly to $\lambda = 10^{-8}$. The WB Lasso method with the continuation trick is denoted "WB Lasso CT" in the following experiments, while the WB Lasso method without the continuation trick is simply denoted "WB Lasso".

We also compared our algorithm with the classical DUET method [3] for source estimation (i.e. the clustering step of DUET for mixing filters estimation is skipped and the source estimation step of DUET is initialized with the known mixing system $\mathcal{A}$).

The performance is evaluated for each source using the Signal-to-Distortion Ratio (SDR) in decibel (dB) as defined in [32], which indicates the overall quality of each estimated source compared to the target. We then average this measure over all the sources and all the mixtures for each mixing condition.

B. Performance analysis as a function of the window size

In order to setup the good STFT window size L for each of the method, we made a first experiment where, for a given mixing configuration where $N = 4$ sources, $RT_{60} = 250$ ms

and $d = 1$ m, we compute the source separation performance in term of SDR as a function of L.

Figure 1 illustrates the results of this experiment. We can

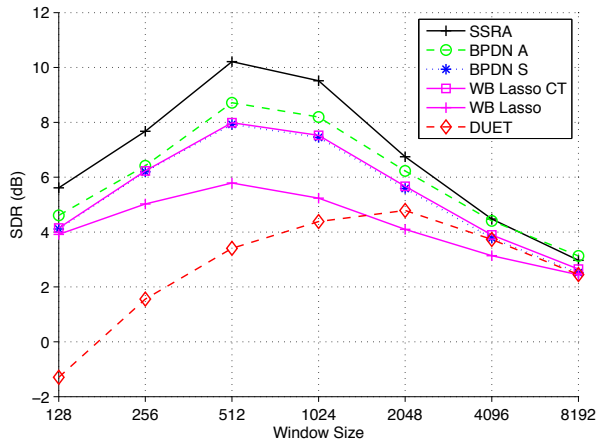
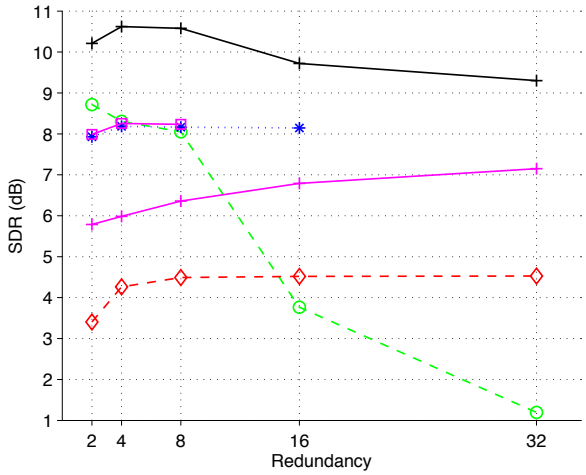


Fig. 1. Variation of the average SDR as a function of the STFT window length L over speech mixtures with $N = 4$ sources, $RT_{60} = 250$ ms and $d = 1$ m.

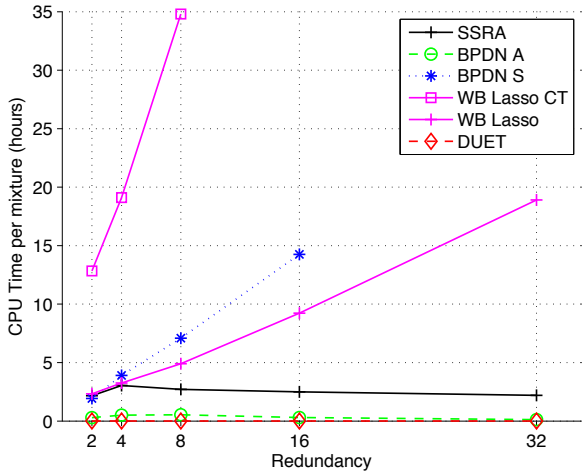
first notice that the proposed SSRA approach outperforms significantly all the other methods whatever the window size. It is interesting to notice that the window length of $L = 512$ samples is the optimal size for all the methods except DUET, for which the optimal window size is $L = 2048$. This is to be expected since the narrowband approximation is better when the window size is large compared to the filter length. Other trends can be observed: the analysis approach (i.e. BPDN A) improves the performance significantly with respect to the synthesis approach (i.e. BPDN S) and the reweighting with analysis approach (i.e. SSRA) improves the performance even more. The performance of the constrained (BPDN S) and the unconstrained (i.e. WB Lasso CT) synthesis approaches, performs very similarly, as predicted by the theory (when $\epsilon \to 0$ and $\lambda \to 0$). Moreover, we will see in section V-C that the computational time of BPDN S is more than 6 time lower than the one of WB Lasso CT. On the other hand, the performance of WB Lasso without the continuation trick, is significantly worse (more than 2 dB of difference with WB Lasso CT at the optimum window size $L = 512$).

C. Performance analysis as a function of redundancy of the STFT

It is known [33] that increasing the redundancy of the STFT of synthesis-based methods can improve the source separation performance by reducing the musical noise. However, it also increases the calculation cost. On the other hand, one of the main advantages of the analysis approach compared to the synthesis approach, is that adding redundancy in the sparse transform (i.e. here the STFT) does not increase the size of the solution. As mentioned by Candès et al. [23], incoherence of the columns of Ψ is not necessary to guarantee the source recovery of the analysis problem. What matters is that the columns of the Gram matrix $\Psi^* \Psi$ are reasonably sparse, which is the case of a redundant STFT. Thus, it is interesting to check if adding redundancy in the STFT, by increasing the overlap ratio between successive windows, can improve the source separation performance. In this experiment we vary the redundancy ratio R by powers of 2 in such a way that Ψ remains a tight frame, which is important, in an algorithm point of view, so as to be able to use Proposition 2 in order to have a fast proximal operator for f_1, and from a theoretical point of view [23] (see section III-D).



(a) Source separation performance



(b) Computational performance

Fig. 2. Average SDR in Decibel and computational time in hours, as a function of the redundancy R over speech mixtures with $N = 4$ sources, $RT_{60} = 250$ ms and $d = 1$ m.

Results depicted in Fig. 2 show that the synthesis approaches improve their performance when the redundancy increases but it stabilizes quickly around R=4 or R=8. Also, as predicted, the time of computation increases quickly with the redundancy. The computation time of the synthesis methods was growing so fast with respect to the redundancy that, for some of these methods, we decided to stop the computation before the end, hence the incomplete curves in Figure 2. Unfortunately, the source separation performance of the analysis approach (BPDN A) decreases when the redundancy

TABLE II
AVERAGE SDR IN DECIBEL AS A FUNCTION OF RT_{60} AND d OVER SPEECH MIXTURES WITH $N = 4$ SOURCES.

RT_{60}	anechoic		50 ms		250 ms	
d	5 cm	1 m	5 cm	1 m	5 cm	1 m
SSRA	1.6	**8.0**	3.9	**8.8**	**6.6**	**10.2**
BPDN A	3.4	5.5	4.2	8.0	5.8	8.7
BPDN S	3.9	7.7	4.3	7.2	6.0	7.9
WB Lasso CT	4.9	7.7	4.4	7.3	6.1	8.0
DUET	**5.9**	7.3	**5.5**	6.4	2.6	3.4

increases. However, for the reweighting analysis approach (SSRA), we can improve the performance by about 0.5 dB with a redundancy of 4 (we called this variant SSRA 4) instead of 2 (called SSRA 2, or simply SSRA as before) without significantly increasing the time of computation.

D. Performance analysis as a function of reverberation time and microphone spacing

We evaluate in Table II the proposed approach and its variants, with respect to the filter length (RT_{60}), and microphone spacing. We also show the results of state-of-the-art methods as a comparison.

According to the results of Table II, the analysis methods seems to work better with realistic (long) RT_{60} while synthesis methods perform better when the RT_{60} is small. The reweighting is working well with realistic (long) RT_{60}, and with shorter RT_{60} when the microphone spacing is large (1 m), but not when the spacing is small (5 cm).

The proposed SSRA method drastically improves performance over the other methods (about 2 dB better than WB Lasso CT) in environments with realistic reverberation ($RT_{60} = 250$ ms).

As for the WB Lasso, the performance of SSRA is less good than DUET in anechoic and low reverberation environment ($RT_{60} = 50$ ms) when the microphone spacing is low ($d = 5$ cm). As shown experimentally in [12], for low reverberant or anechoic environments, it is possible to improve the performance of WB Lasso (and also probably SSRA), by replacing the wideband data fidelity term with the narrowband one, but as discussed in [12], for these mixing conditions, there are a lot of methods based on the narrowband approximation that already work well, and some of them like DUET are moreover very fast. As a consequence, it is not justified to use complex methods like WB Lasso or SSRA for these conditions.

For all the sparse recovery methods based on the wideband formulation, the performance is better when the filter is longer. This trend has a theoretical explanation as discussed in section III-D.

E. Performance with respect to the number of sources

In addition to the previous experiments where we evaluated the different methods in the case of a mixture of 4 sources, we compare now the different methods for mixtures with $3 \leq N \leq 6$ sources and with $RT_{60} = 250$ ms and $d = 1$ m. The results are depicted in Fig. 3.

Whatever the number of sources, the results depicted in Fig. 3 show that the gap of performance between our methods and the state-of-the-art (WB Lasso CT) is nearly constant with respect to the number of sources and that: i) The analysis approach improves the separation performance compared to the synthesis one by between 0.5 dB and 1 dB of SDR. ii) The reweighting combined with analysis, improves the performance by between 2 and 3 dB of SDR. iii) The constrained approach leads to similar results as the non-constrained one, which is predicted by the theory, however, a) it has been necessary to use the costly continuation trick so as to make the non-constrained approach converge at a cost of slowing down the computation by a factor of 8 (i.e. the number of $\lambda^{(k)}$ steps of the continuation trick), b) we cannot simply use (because of the λ setting) the very efficient reweighting scheme, with this approach, iii) In the noise-free case we know that the best setting for λ is zero, however in the noisy case, there is no obvious way to setup λ as opposed to ϵ in the constraint approach.

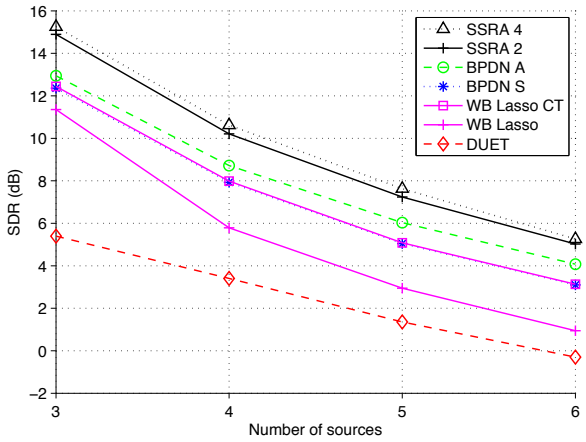


Fig. 3. Variation of the average SDR as a function of N over speech mixtures with $RT_{60} = 250$ ms and $d = 1$ m.

VI. CONCLUSION

We proposed a novel algorithm based on a reweighted analysis sparse prior for reverberant audio source separation, assuming that the mixing filters are known. This algorithm, based on i) an analysis sparse prior, ii) a reweighting scheme so as to increase the sparsity, iii) a wideband data-fidelity term in a constrained form, has been evaluated on convolutive mixtures of sources and compared with state-of-the-art methods. We also evaluated the analysis versus synthesis prior, as well as the reweighted versus non-reweighted scheme, and the constrained versus unconstrained data-fidelity term, on mixtures with different levels of reverberation and different numbers of sources.

Our conclusion is that the reweighted analysis sparse prior with a constrained wideband data-fidelity term works better than any of the tested methods for realistic reverberant mixtures and that the gain of performance is by between 2 and 3 dB of SDR with respect to the state-of-the-art. Another advantage of our algorithm, is that we can easily increase the redundancy of the analysis operator, for example by increasing

the redundancy of the STFT, without significantly increasing the complexity.

A possible extension of this work will be to model the sources with a structured sparsity prior instead of an ℓ_1 cost. Another extension would be, in addition to the sources estimation, to estimate the mixing filters, possibly with an alternating optimization approach. It would also be interesting to study more formally the D-RIP for the narrowband and wideband linear operators so as to have a deeper understanding of the source recovery conditions depending on the mixing conditions.

APPENDIX

A. Proximity operator

We give the definition of the proximity operator and derive these operators for the functions $f_1(\mathbf{s}) = \|\mathbf{s}\,\Psi\,\|_{\mathrm{W},1}$ and $f_2(\mathbf{s}) = i_{\mathcal{B}^{\epsilon}_{\ell_2}}(\mathbf{s})$.

Definition 1. (Proximity operator) *Let f_i be a lower semicontinuous convex function from $\mathbb{C}^I$ to $]-\infty,+\infty]$. The proximity operator of f_i, denoted prox_{f_i} is given by:*

$$\mathrm{prox}_{f_i}(\mathbf{z}) \triangleq \underset{\mathbf{u}\in\mathbb{C}^I}{\mathrm{argmin}}\ f_i(\mathbf{u}) + \frac{1}{2}\|\mathbf{z}-\mathbf{u}\|_2^2. \tag{13}$$

This definition extends naturally for some matrices $\mathbf{z}$ and $\mathbf{u}$, by replacing the ℓ_2 norm with the Frobenius norm.

We recall that $\mathbf{L}$ is a frame if its adjoint satisfies the generalized Parseval relation with bounds ν_1 and ν_2:

$$\nu_1\|\mathbf{z}\|^2 \leq \|\,\mathbf{L}^*\,\mathbf{z}\|^2 \leq \nu_2\|\mathbf{z}\|^2, \tag{14}$$

with $0 < \nu_1 \leq \nu_2 < \infty$. The frame is tight when $\nu_1 = \nu_2 = \nu$ and $\mathbf{L}\,\mathbf{L}^* = \nu\,\mathbf{I}$.

So as to derive the proximity operators of f_1 and f_2, we need the following lemma:

Lemma 1. *If $\mathbf{L}$ is a tight frame, i.e. $\mathbf{L}\,\mathbf{L}^* = \nu\,\mathbf{I}, \nu > 0$, then*

$$\mathrm{prox}_{f(\mathbf{L}\,\cdot)}(\mathbf{z}) = \mathbf{z} + \nu^{-1}\,\mathbf{L}^*(\mathrm{prox}_{\nu f} - \mathbf{I})(\mathbf{L}\,\mathbf{z}) \tag{15}$$

Lemma 2. *If $\mathbf{L}$ is a general frame with bounds ν_1 and ν_2, Let $\mu_k \in (0, 2/\nu_2)$, Define*

$$\mathbf{u}^{(k+1)} = \mu_k(\mathbf{I} - \mathrm{prox}_{\mu_k^{-1} f})(\mu_k^{-1}\mathbf{u}^{(k)} + \mathbf{L}\,\mathbf{p}^{(k)} - \mathbf{y}) \tag{16}$$

$$\mathbf{p}^{(k+1)} = \mathbf{z} - \mathbf{L}^*\,\mathbf{u}^{(k+1)} \tag{17}$$

Then $\mathbf{p}^{(k)} \to \mathrm{prox}_{f(\mathbf{L}\,\cdot - y)}(\mathbf{z})$ linearly.

The proof Lemma 1 can be found in [29] the one of Lemma 2 can be found in [34].

Proposition 1. (Prox of $\lambda\|\cdot\|_1$) *Let $\mathbf{z} \in \mathbb{C}^I$. Then $\mathbf{u} = \mathrm{prox}_{\lambda\|\cdot\|_1}(\mathbf{z}) = (\mathrm{prox}_{\lambda|\cdot|}(z_i))_{1\leq i\leq I}$ is given entrywise by soft thresholding:*

$$u_i = \mathrm{prox}_{\lambda|\cdot|}(z_i) = \frac{z_i}{|z_i|}(|z_i| - \lambda)^+ \tag{18}$$

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

The proof of this proposition can be found in [16].

Applying Lemma 1, we get a closed form solution of the proximal operator of $f_1(\mathbf{s}) = \|\mathbf{s}\,\Psi\,\|_{\mathrm{W},1}$:

Proposition 2. (Prox of $f_1(\cdot) = \|\cdot\,\Psi\,\|_{\mathrm{W},1}$) *Let $\tilde{\mathbf{z}} \in \mathbb{C}^{N\times B}$ and $\mathbf{z} \in \mathbb{R}^{N\times T}$. If $\Psi \in \mathbb{C}^{T\times B}$ is a tight frame, i.e. $\Psi\,\Psi^* = \nu\,\mathbf{I}$, and $\mathrm{W} \in \mathbb{R}_+^{N\times B}$ is a matrix of positive weights w_{ij}, then*

$$\mathrm{prox}_{\|\cdot\Psi\,\|_{\mathrm{W},1}}(\mathbf{z}) = \mathbf{z} + \nu^{-1}(\mathrm{prox}_{\nu\|\cdot\|_{\mathrm{W},1}} - \mathbf{I})(\mathbf{z}\,\Psi)\,\Psi^* \tag{19}$$

with

$$\mathrm{prox}_{\nu\|\cdot\|_{\mathrm{W},1}}(\tilde{\mathbf{z}}) = (\mathrm{prox}_{\nu w_{ij}|\cdot|}(\tilde{z}_{ij}))_{1\leq i\leq N, 1\leq j\leq B} \tag{20}$$

where $\mathrm{prox}_{\nu w_{ij}|\cdot|}$ is the soft thresholding operator given in (18) with $\lambda = \nu w_{ij}$.

Applying Lemma 2, we get an iterative solution of the proximal operator of $f_2 = i_{\mathcal{B}^{\epsilon}_{\ell_2}}$:

Proposition 3. (Prox of $f_2(\cdot) = i_{\mathcal{B}^{\epsilon}_{\ell_2}}(\cdot)$) *If $\mathcal{A}$ is a general frame with bounds ν_1 and ν_2, Let $\mu_k \in (0, 2/\nu_2)$, Define*

$$\mathbf{u}^{(k+1)} = \mu_k(\mathbf{I} - \mathrm{prox}_{i_{\|\cdot\|_2\leq\epsilon}})(\mu_k^{-1}\mathbf{u}^{(k)} + \mathcal{A}(\mathbf{p}^{(k)}) - \mathbf{x}) \tag{21}$$

$$\mathbf{p}^{(k+1)} = \mathbf{z} - \mathcal{A}^*(\mathbf{u}^{(k+1)}) \tag{22}$$

where :

$$\mathrm{prox}_{i_{\|\cdot\|_2\leq\epsilon}}(\mathbf{u}) = \min(1, \epsilon/\|\mathbf{u}\|_2)\mathbf{u}. \tag{23}$$

Then $\mathbf{p}^{(k)} \to \mathrm{prox}_{i_{\mathcal{B}^{\epsilon}_{\ell_2}}}(\mathbf{z})$ linearly.

The tightest possible frame bound ν_2 is the largest spectral value of the frame operator $\mathcal{A}\,\mathcal{A}^*$, which can be computed using the power iteration algorithm as in [12]. Recursion (21)-(22) is a forward-backward splitting scheme applied to the dual problem [34] which we accelerated with a Nesterov-type update [17].

In the noise-free case, we can also replace the ℓ_2-ball constraint set $\mathcal{B}^{\epsilon}_{\ell_2}$ with the affine constraint set $\mathcal{C}_{eq} = \{\mathbf{s} \in \mathbb{R}^{N\times T} | \mathcal{A}(\mathbf{s}) = \mathbf{x}\}$ and derive a closed form solution of the proximal operator of $f_2(\cdot) = i_{\mathcal{C}_{eq}}(\cdot)$ as in [34]:

Proposition 4. (Prox of $f_2(\cdot) = i_{\mathcal{C}_{eq}}(\cdot)$)

$$\mathrm{prox}_{i_{\mathcal{C}_{eq}}}(\mathbf{z}) = \mathbf{z} + \mathcal{A}^*(\mathcal{A}\,\mathcal{A}^*)^{-1}(\mathbf{x} - \mathcal{A}(\mathbf{z})). \tag{24}$$

In practice, (24) can be solved iteratively with a conjugate-gradient type method such as LSQR [35].

ACKNOWLEDGEMENT

We would like to thank Matthieu Kowalski, for kindly providing his code and data and some details about the implementation of his WB Lasso algorithm.

REFERENCES

[1] A. Belouchrani and M. Amin, "Blind source separation based on time-frequency signal representations," *IEEE Transactions on Signal Processing*, vol. 46, no. 11, pp. 2888–2897, 1998.

[2] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

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

[4] P. O'Grady, B. Pearlmutter, and S. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, pp. 18–33, 2005.

[5] M. Puigt and Y. Deville, "Time–frequency ratio-based blind separation methods for attenuated and time-delayed sources," *Mechanical Systems and Signal Processing*, vol. 19, no. 6, pp. 1348–1379, 2005.

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

[7] S. Arberet, P. Sudhakar, and R. Gribonval, "A wideband doubly-sparse approach for mito sparse filter estimation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011.

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

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

[10] N. 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.

[11] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*, May 2010, pp. 1 –4.

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

[13] M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev, "Blind source separation by sparse decomposition in a signal dictionary," in *Independent Component Analysis : Principles and Practice*. Cambridge Press, 2001, pp. 181–208.

[14] E. Vincent, "Complex nonconvex l_p norm minimization for underdetermined source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2007, pp. 430–437.

[15] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on pure and applied mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[16] P. Combettes, V. Wajs *et al.*, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, 2006.

[17] 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.

[18] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, pp. 129–159, 2001.

[19] I. Loris, "On the performance of algorithms for the minimization of 1-penalized functionals," *Inverse Problems*, vol. 25, p. 035008, 2009.

[20] J. Nocedal and S. Wright, "Numerical optimization. 2006."

[21] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse problems*, vol. 23, p. 947, 2007.

[22] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, "The Cosparse Analysis Model and Algorithms," *Applied and Computational Harmonic Analysis*, 2012.

[23] E. Candes, Y. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Applied and Computational Harmonic Analysis*, vol. 31, no. 1, pp. 59–73, 2011.

[24] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[25] J. Haupt, W. Bajwa, G. Raz, and R. Nowak, "Toeplitz compressed sensing matrices with applications to sparse channel estimation," *Information Theory, IEEE Transactions on*, vol. 56, no. 11, pp. 5862–5875, 2010.

[26] R. E. Carrillo, Y. Wiaux, J. D. McEwen, D. V. D. Ville, and J.-P. Thiran, "Sparsity averaging for compressive imaging," *in preparation*.

[27] R. Carrillo, J. McEwen, and Y. Wiaux, "Sparsity averaging reweighted analysis (sara): a novel algorithm for radio-interferometric imaging," *Arxiv preprint arXiv:1205.3123*, 2012.

[28] P. Combettes and J. Pesquet, "Proximal splitting methods in signal processing," *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212, 2011.

[29] ——, "A douglas–rachford splitting approach to nonsmooth convex variational signal recovery," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 564–574, 2007.

[30] D. Campbell, K. Palomaki, and G. Brown, "Roomsim, a MATLAB simulation of shoebox room acoustics for use in teaching and research," *Computing and Information Systems*, vol. 9, no. 3, pp. 48–51, 2005.

[31] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," *Lecture Notes in Computer Science*, vol. 4666, p. 552, 2007.

[32] R. Gribonval, L. Benaroya, E. Vincent, C. Févotte *et al.*, "Proposals for performance measurement in source separation," 2003.

[33] S. Raki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 3. Ieee, 2005, pp. iii–81.

[34] M. Fadili and J. Starck, "Monotone operator splitting for optimization problems in sparse recovery," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 1461–1464.

[35] C. Paige and M. Saunders, "Lsqr: An algorithm for sparse linear equations and sparse least squares," *ACM Transactions on Mathematical Software (TOMS)*, vol. 8, no. 1, pp. 43–71, 1982.