Beyond the Narrowband Approximation: Wideband Convex Methods for Under-Determined Reverberant Audio Source Separation

Matthieu Kowalski, Emmanuel Vincent, Rémi Gribonval

Beyond the Narrowband Approximation: Wideband Convex Methods for Under-Determined Reverberant Audio Source Separation

Matthieu Kowalski, Emmanuel Vincent *Senior Member, IEEE* and Rémi Gribonval *Senior Member, IEEE*

***Abstract*—We consider the problem of extracting the source signals from an under-determined convolutive mixture assuming known mixing filters. State-of-the-art methods operate in the time-frequency domain and rely on narrowband approximation of the convolutive mixing process by complex-valued multiplication in each frequency bin. The source signals are then estimated by minimizing either a mixture fitting cost or a ℓ_1 source sparsity cost, under possible constraints on the number of active sources. In this article, we define a wideband ℓ_2 mixture fitting cost circumventing the above approximation and investigate the use of a $\ell_{1,2}$ mixed-norm cost promoting disjointness of the source time-frequency representations. We design a family of convex functionals combining these costs and derive suitable optimization algorithms. Experiments indicate that the proposed wideband methods result in a signal-to-distortion ratio improvement of 2 to 4 dB compared to the state-of-the-art on reverberant speech mixtures.**

***Index Terms*—Source separation, convolutive mixture, narrowband approximation, mixed norms, convex optimization**

I. Introduction

In many situations, such as a concert or a cocktail party, the recorded sound signals are mixtures of several sound sources. The mth mixture channel $x_m(t)$ is then given by

$$x_m(t) = \sum_{n=1}^{N} A_{mn} \star s_n(t) + e_m(t) \ , \qquad (1)$$

where $s_n(t)$ is the nth source signal, the filters $A_{mn}(t)$ are called mixing filters, $\star$ denotes convolution, and $e_m(t)$ is the background noise. Blind source separation is the task of estimating the source signals from the mixture.

In this work, we consider the so-called under-determined setting, where the number of sources is larger than the number of mixture channels (ie $N > M$). State-of-the-art under-determined source separation methods operate in the time-frequency domain and rely on narrowband approximation of the convolutive mixing process by complex-valued multiplication in each frequency bin. The separation task is split into two successive subtasks. First, frequency-dependent mixing matrices are estimated by clustering the mixture time-frequency coefficients based on the associated sound directions. The source time-frequency coefficients are then separately estimated in each time-frequency bin, typically either by minimizing some mixture fitting cost under the constraint that at most one source be active [1], [2], a method known as binary masking, or by minimizing a ℓ_1 source sparsity cost [3], [4]. These costs implement the assumption that the source time-frequency representations are disjoint or sparse, respectively. According to a recent evaluation [5], these methods achieve limited separation performance in realistic reverberant environments.

In this article, we focus on addressing the second subtask, namely the estimation of the source signals *assuming that the mixing filters A_{mn} are known*. We investigate two possible reasons for the limited performance of state-of-the-art methods. Firstly, while the narrowband approximation is valid when the length of the mixing filters is short compared to that of the time-frequency analysis window, this condition does not hold in reverberant environments. Significant performance improvements have been observed in the determined setting using wideband methods that jointly process all frequency bins [6]. Yet, these methods do not apply to the under-determined setting. Secondly, while maximum disjointness of the source time-frequency representations appears to be a reasonable assumption, the additional constraint exploited by binary masking that at most one source be active in each time-frequency bin does not hold in practice.

This article provides four contributions in light of the above issues. Firstly, we define a wideband ℓ_2 mixture fitting cost that circumvents the narrowband approximation. It is the first time, to our knowledge, that a way of avoiding this approximation is proposed in the under-determined setting. Secondly, motivated by recent theoretical results about the so-called mixed norms [7], [8], we investigate the use of a ℓ_{12} mixed-norm cost promoting disjointness of the source time-frequency representations without constraining the number of active sources per time-frequency bin. Thirdly, we design a family of convex functionals combining these costs as well as state-of-the-art costs and exploit recent advances in the area of convex optimization to derive suitable optimization algorithms. Finally, we compare the proposed methods with state-of-the-art methods on a set of speech mixtures with different numbers of sources, reverberation times and microphone spacings. We thereby extend and improve our preliminary paper [9], which presented a single functional and a less efficient algorithm

M. Kowalski is with Laboratoire des Signaux et Systèmes, UMR 8506 CNRS - SUPELEC - Univ. Paris-Sud, 91192 Gif-sur-Yvette Cedex, France (e-mail: matthieu.kowalski@lss.supelec.fr).

E. Vincent and R. Gribonval are with INRIA, Centre Inria Rennes - Bretagne Atlantique, 35042 Rennes Cedex, France (e-mail: emmanuel.vincent@inria.fr; remi.gribonval@inria.fr).

This work was supported in part by the french Agence National de la Recherche project ECHANGE (ANR-08-EMER-006) and in part by the European Union through the project SMALL (Sparse Models, Algorithms and Learning for Large-Scale data).

evaluated on two mixtures only.

The structure of the rest of the article is as follows. In Section II, we introduce some notations and recall the principles of state-of-the-art methods. In Section III, we explain how the source separation problem can be recast as that of minimizing a convex functional and present a family of narrowband and wideband functionals. In Section IV, we summarize relevant theoretical results in the area of convex optimization and provide the details of the resulting source separation algorithms. Finally, we compare the proposed methods with state-of-the-art methods in Section V and conclude in Section VI.

II. State of the art

We start by introducing the notations used in the rest of the article and presenting state-of-the-art methods used to separate under-determined convolutive mixtures.

A. Matrix notation

The problem under consideration is the following: N source signals $s_n(t)$ of duration T are recorded by $M < N$ microphones, yielding M mixture channels $x_m(t)$. The effect of acoustic propagation between the sources and the microphones is modeled by a set of mixing filters $A_{mn}(t)$ of length P. Denoting by $\mathbf{x} \in \mathbb{R}^{M\times T}$ and $\mathbf{s} \in \mathbb{R}^{N\times T}$ the matrices of mixture channels and source signals and by $\mathbf{A} \in \mathbb{R}^{M\times N\times P}$ the three-way array of mixing filters, the mixing process (1) can be rewritten more concisely in matrix form as

$$\mathbf{x} = \mathbf{A} \star \mathbf{s} + \mathbf{e} \ , \tag{2}$$

where $\mathbf{e} \in \mathbb{R}^{M\times T}$ models the background noise. Since $M < N$, $\mathbf{A}$ is not invertible, hence suitable approaches must be found to estimate $\mathbf{s}$ given $\mathbf{x}$ and $\mathbf{A}$.

B. Blind vs non-blind source separation

As we suppose the mixing system $\mathbf{A}$ known, the framework under consideration is the *non blind* source separation. In many practical applications, one has to solve the general problem of *blind* source separation: the mixing system $\mathbf{A}$ and the sources $\mathbf{s}$ must be estimated in the same time. In the convolutive underdetermined case considered here, both part are challenging tasks. Estimating the mixing system is arguably the most difficult part, but, even if the mixing system is known, the source separation quality is still far from the best one could expected [10]. We choose to concentrate here on this second part. Indeed, in order to compare the different models described in this work, it seems natural to evaluate them on the ground-truth model with the true mixing system. In addition, we provide in Section V-G some results to evaluate the robustness of the different models to error in evaluation of the mixing system. These results indicate that the design of wideband under-determined blind source separation algorithms is a relevant long-term research goal. Few wideband filter estimation algorithms exist today, e.g. the subspace-based channel identification algorithm in [11], yet this algorithm is not suitable for long filters such as those arising in the context of audio due to large memory requirements. The design of such algorithms will be the subject of future work.

C. Time-frequency transform

A popular approach is to rely on the assumption that the sources admit disjoint or sparse representations in the time-frequency domain. Under this assumption, only a few sources contribute significantly to the mixture in each time-frequency bin so that the mixing process becomes "locally invertible" and estimates of the source time-frequency coefficients can be obtained [12].

More precisely, let us denote by $\Phi \in \mathbb{C}^{T\times B}$ the matrix representing an energy-preserving Short-Time Fourier Transform (STFT) operator. This operator transforms a signal of length T into a set of $B \geq T$ time-frequency coefficients. The STFT coefficients $\tilde{\mathbf{x}} \in \mathbb{C}^{M\times B}$ of the mixture $\mathbf{x}$ are given by

$$\tilde{\mathbf{x}} = \mathbf{x}\Phi \tag{3}$$

while the sources $\mathbf{s}$ can be resynthesized from their estimated STFT coefficients $\tilde{\mathbf{s}} \in \mathbb{C}^{N\times B}$ by

$$\mathbf{s} = \tilde{\mathbf{s}}\Phi^* \tag{4}$$

where $\Phi^* \in \mathbb{C}^{B\times T}$ is the adjoint operator of Φ, that is its Hermitian transpose. Note that, strictly speaking, (3) defines analysis STFT coefficients, while (4) defines synthesis STFT coefficients. Due to the absence of possible confusion between these two notions, we omit the terms "analysis" or "synthesis" in the following.

D. Narrowband approximation

Besides its desirable effect on the sparsity of the sources, the STFT offers a convenient way of dealing with the convolutive mixing process. Indeed, after applying the STFT, the mixing model (2) becomes

$$\tilde{\mathbf{x}} = (\mathbf{A} \star \mathbf{s} + \mathbf{e})\Phi \ . \tag{5}$$

Considering each frequency bin f individually, the above convolution can be approximated by the complex-valued matrix product [2], [4]

$$\tilde{\mathbf{x}}(t,f) \simeq \tilde{\mathbf{A}}(f)\tilde{\mathbf{s}}(t,f) + \tilde{\mathbf{e}}(t,f) \ , \tag{6}$$

where $\tilde{\mathbf{x}}(t,f)$, $\tilde{\mathbf{s}}(t,f)$ and $\tilde{\mathbf{e}}(t,f)$ are the vectors of mixture and source STFT coefficients in time-frequency bin $b = (t,f)$ and $\tilde{\mathbf{A}}(f)$ is a mixing matrix equal to the Fourier transform of the mixing system. Denoting by $\times_f$ the frequency-wise matrix product operator [13], this can also be written in matrix form as

$$\tilde{\mathbf{x}} \simeq \tilde{\mathbf{A}} \times_f \tilde{\mathbf{s}} + \tilde{\mathbf{e}} \ . \tag{7}$$

This ubiquitous narrowband approximation stems from a first-order Taylor expansion of $\tilde{\mathbf{x}}$. It is generally assumed to be valid when the mixing filters are short compared to the STFT window. However, no mathematical quantification of the approximation error can be found in the literature to our knowledge. Lemma 1 in the Appendix shows that the approximation error is small when the Fourier transform of the STFT window is concentrated at low frequencies and the derivative of $\tilde{\mathbf{A}}$ is small. This is equivalent to the respective conditions that the window be smooth enough, which is usually the case, and that the mixing filters be concentrated around the null

delay. Consequently, the narrowband approximation holds for anechoic mixtures with small delays, but not for reverberant mixtures or anechoic mixtures with larger delays.

We provide in section V-A some numerical results to illustrate the approximation error of the narrowband approximation (6) compared with model (2) with no additional noise ($\mathbf{e} = 0$). The results illustrate the conditions highlighted by Lemma 1 in Appendix. Note however that no direct relationship exists between the approximation error on the mixture and the resulting error on the sources

E. DUET

The well-known binary masking method for source separation exploits the assumption that the sources are disjoint in the time-frequency domain, so that a single source is active in each time-frequency bin [1], [2]. The active source and its STFT coefficients are estimated by minimizing some cost of fitting the observed mixture STFT coefficients by the product of the mixing matrix and the source STFT coefficients. The Degenerate Unmixing Estimation Technique (DUET) relies on the ℓ_2 mixture fitting cost [1]

$$\min_{\tilde{\mathbf{s}}(t,f) \text{ s.t. } \|\tilde{\mathbf{s}}(t,f)\|_0=1} \|\tilde{\mathbf{x}}(t,f) - \tilde{\mathbf{A}}(f)\tilde{\mathbf{s}}(t,f)\|_2^2 \quad \text{(N-DUET)}$$

where the ℓ_p norm of a vector or a matrix $\mathbf{z}$ with I entries indexed by i is defined for $p > 0$ by

$$\|\mathbf{z}\|_p = \left(\sum_{i=1}^{I} |z_i|^p \right)^{1/p} \tag{8}$$

and $\|\mathbf{z}\|_0$ denotes the number of nonzero entries of $\mathbf{z}$. This cost can be minimized by computing the cost associated with each possible active source by least-squares projections and selecting the source leading to the lowest cost [14]. This approach can be generalized to up to $M-1$ active sources per time-frequency bin, resulting in a combinatorial optimization problem [14].

F. ℓ_1 norm minimization

An alternative method is to rely on the assumption that the sources are sparse in the time-frequency domain, *i.e.* only a few STFT coefficients significantly differ from zero for each source. This assumption can be exploited by minimizing the ℓ_1 norm of the coefficients subject to an exact mixture fitting constraint [15], [4]

$$\min_{\tilde{\mathbf{s}}(t,f) \text{ s.t. } \tilde{\mathbf{x}}(t,f)=\tilde{\mathbf{A}}(f)\tilde{\mathbf{s}}(t,f)} \|\tilde{\mathbf{s}}(t,f)\|_1. \quad \text{(N-}\ell_1\text{)}$$

This sparsity cost generally results in M or little more than M active sources per time-frequency bin [4] and can be optimized by *e.g.* FOCUSS [16], second-order cone programming [4] or gradient-based algorithms [17]. Note that the ℓ_1 norm can be replaced by an ℓ_p quasi-norm with $p < 1$, but the problem is not convex anymore. The additional constraint that exactly M sources are active in each time-frequency bin is sometimes assumed, resulting in a combinatorial optimization problem [4].

III. Source separation by minimization of narrowband or wideband functionals

In this section, we recast the source separation problem into a more general convex optimization framework and construct a family of convex functionals that generalize those underlying DUET or ℓ_1 norm minimization. This approach will allow us to re-use and adapt efficient algorithms proposed in the convex optimization community.

A. Convex optimization framework

The general form of convex optimization problems we shall consider reads

$$\min_{\tilde{\mathbf{s}} \in \mathbb{C}^{N \times B}} \mathcal{L}(\mathbf{x}, \mathbf{A}, \tilde{\mathbf{s}}) + \lambda \mathcal{P}(\tilde{\mathbf{s}}) \tag{9}$$

where the different components of the functional to be minimized are

- a convex loss or *data term* $\mathcal{L}(\mathbf{x}, \mathbf{A}, \tilde{\mathbf{s}})$ measuring the fit between the observed mixture $\mathbf{x}$ and the source STFT coefficients $\tilde{\mathbf{s}}$ given the mixing system $\mathbf{A}$,
- a convex penalty or *regularization term* $\mathcal{P}(\tilde{\mathbf{s}})$ modeling the sparsity or disjointness assumptions about the source STFT coefficients $\tilde{\mathbf{s}}$,
- an hyperparameter $\lambda \in \mathbb{R}_+$ governing the balance between the data term and the regularization term.

In the following, we propose two possible data terms and two possible regularization terms, yielding four distinct functionals.

B. Mixed norms

While the assumption that the source STFT coefficients are sparse translates into the convex cost (N-ℓ_1), the alternative assumption underlying DUET that a single source be active in each time-frequency bin does not translate into a convex cost. Moreover, this assumption holds only for mixtures with a sufficiently small number of sources. We design a convex cost promoting disjointness of the source time-frequency representations through the use of a mixed norm defined hereafter.

Definition 1 (Mixed norm): Let $p \geq 1$ and $q \geq 1$. Let $\mathbf{z} \in \mathbb{C}^{I \times J}$ be a matrix with row index i and column index j. The $\ell_{p,q}$ norm of $\mathbf{z}$ is called mixed norm and defined by [7], [8]

$$\|\mathbf{z}\|_{p,q} = \left(\sum_{j=1}^{J} \left(\sum_{i=1}^{I} |z_{ij}|^p \right)^{q/p} \right)^{1/q}.$$

The classical ℓ_p norm in (8) is a particular instance of the $\ell_{p,q}$ norm with $p = q$.

The meaning of a given mixed norm depends on the choice of the exponents p and q and the associated matrix indexes. In the convex optimization framework under study, the entries of the matrix $\tilde{\mathbf{s}} \in \mathbb{C}^{N \times B}$ of source STFT coefficients are each associated with a source n and a time-frequency bin $b = (t, f)$. We define the $\ell_{1,2}$ norm of $\tilde{\mathbf{s}}$ as

$$\|\tilde{\mathbf{s}}\|_{1,2}^2 = \sum_{t,f} \left(\sum_{n=1}^{N} |\tilde{s}_n(t,f)| \right)^2. \tag{10}$$

While minimization of the ℓ_1 norm induces sparsity over the whole considered matrix, minimization of the $\ell_{1,2}$ norm induces sparsity over each column separately [7]. Hence, this norm promotes maximum disjointness of the source time-frequency representations, without constraining the proportion of significant STFT coefficients for each source.

C. Narrowband Lasso and E-Lasso

The constraint of exact fitting of the mixture STFT coefficients in (N-ℓ_1) does not account for the fact that (6) is an approximation of the actual mixing process. Consequently, it appears natural to relax this constraint and replace it by the ℓ_2 mixture fitting cost in (N-ℓ_1), which yields the following convex optimization problem in each time-frequency bin:

$$\min_{\tilde{\mathbf{s}}(t,f)\in\mathbb{C}^N} \frac{1}{2}\|\tilde{\mathbf{x}}(t,f) - \tilde{\mathbf{A}}(f)\tilde{\mathbf{s}}(t,f)\|_2^2 + \lambda\|\tilde{\mathbf{s}}(t,f)\|_1. \quad (11)$$

This functional was introduced in [18] to estimate the number of active sources in each time-frequency bin under the constraint that at most M sources are active, while estimating the STFT coefficients of the active sources either by (N-ℓ_1) with M active sources or by generalization of (N-DUET) with up to $M-1$ active sources. By contrast, we exploit this functional to estimate the STFT coefficients of all sources. This approach is mathematically more consistent and enables the recovery of more than M active sources in theory.

The summation of (11) over all time-frequency bins results in the global optimization problem

$$\min_{\tilde{\mathbf{s}}\in\mathbb{C}^{N\times B}} \quad \frac{1}{2}\|\tilde{\mathbf{x}} - \tilde{\mathbf{A}}\times_f\tilde{\mathbf{s}}\|_2^2 + \lambda\|\tilde{\mathbf{s}}\|_1. \quad \text{(N-Lasso)}$$

This problem is equivalent to (11) and is a particular instance of the so-called Lasso [19] or basis pursuit denoising [15] convex optimization problem. Because it relies on the narrowband approximation, we call it *narrowband Lasso*.

As mentioned above, the choice of the ℓ_1 norm for the regularization term is known to induce sparsity over the time-frequency plane. When λ is large, the influence of the regularization term is stronger hence most estimated source STFT coefficients are set to zero. In particular, high frequencies are typically zeroed out, since the associated STFT coefficients have smaller absolute values. This undesirable behavior can be circumvented by considering $\ell_{1,2}$ norm regularization instead, which leads to a particular instance of the so-called Elitist-Lasso (E-Lasso) optimization problem [7]

$$\min_{\tilde{\mathbf{s}}\in\mathbb{C}^{N\times B}} \quad \frac{1}{2}\|\tilde{\mathbf{x}} - \tilde{\mathbf{A}}\times_f\tilde{\mathbf{s}}\|_2^2 + \frac{\lambda}{2}\|\tilde{\mathbf{s}}\|_{1,2}^2 \quad \text{(N-E-Lasso)}$$

that we call *narrowband E-Lasso*. This regularization promotes a small number of active sources in each time-frequency bin. However, it results in at least one nonzero coefficient over each column of the considered matrix [8], which means that at least one source is estimated to be active in each time-frequency bin.

The state-of-the-art ℓ_1 norm minimization method can be viewed as a particular case of the proposed narrowband Lasso and E-Lasso methods, as justified by the following remark.

Remark 1: There exists $\alpha \geq 0$ such that if $\lambda \leq \alpha$, then any solution of (N-Lasso) or (N-E-Lasso) is a solution of (N-ℓ_1).

Proof: Let $\hat{\tilde{\mathbf{s}}}$ be a solution of (N-Lasso) and $\check{\tilde{\mathbf{s}}}$ a solution of (N-ℓ_1).

A classical theorem of convex optimization theory (see *e.g.* [20, p. 24]) implies that there exists $\alpha \geq 0$ such that for any $\lambda \leq \alpha$, the solution is such that the data term is equal to zero hence $\tilde{\mathbf{A}}(f)\hat{\tilde{\mathbf{s}}}(t,f) = \tilde{\mathbf{x}}(t,f)$ for all (t,f).

Since $\|\tilde{\mathbf{x}} - \tilde{\mathbf{A}}\times_f\hat{\tilde{\mathbf{s}}}\|_2^2 = \|\tilde{\mathbf{x}} - \tilde{\mathbf{A}}\times_f\check{\tilde{\mathbf{s}}}\|_2^2 = 0$ and by definition of $\hat{\tilde{\mathbf{s}}}$, $\|\hat{\tilde{\mathbf{s}}}\|_1 \leq \|\check{\tilde{\mathbf{s}}}\|_1$. If there existed (t,f) such that $\|\hat{\tilde{\mathbf{s}}}(t,f)\|_1 > \|\check{\tilde{\mathbf{s}}}(t,f)\|_1$, then there would exist $(t',f') \neq (t,f)$ such that $\|\hat{\tilde{\mathbf{s}}}(t',f')\|_1 < \|\check{\tilde{\mathbf{s}}}(t',f')\|_1$ which contradicts the definition of $\check{\tilde{\mathbf{s}}}(t',f')$. Hence $\|\hat{\tilde{\mathbf{s}}}(t,f)\|_1 = \|\check{\tilde{\mathbf{s}}}(t,f)\|_1$ for all (t,f), which proves that $\hat{\tilde{\mathbf{s}}}$ is a solution of (N-ℓ_1).

A similar proof applies to the solutions of (N-E-Lasso). ■

D. Wideband Lasso and E-Lasso

Given the form of the narrowband Lasso (N-Lasso) and E-Lasso functionals (N-E-Lasso), a natural way of circumventing the narrowband assumption is now to replace the approximate mixing model (6) within the data term by the true time-domain model (2) where the time-domain source signals are obtained from their STFT coefficients via (4). This leads to the following wideband counterparts of the above functionals that we call *wideband Lasso* and *wideband E-Lasso* respectively:

$$\min_{\tilde{\mathbf{s}}\in\mathbb{C}^{N\times B}} \frac{1}{2}\|\mathbf{x} - \mathbf{A}\star\tilde{\mathbf{s}}\Phi^*\|_2^2 + \lambda\|\tilde{\mathbf{s}}\|_1 \quad \text{(W-Lasso)}$$

$$\min_{\tilde{\mathbf{s}}\in\mathbb{C}^{N\times B}} \frac{1}{2}\|\mathbf{x} - \mathbf{A}\star\tilde{\mathbf{s}}\Phi^*\|_2^2 + \frac{\lambda}{2}\|\tilde{\mathbf{s}}\|_{1,2}^2. \quad \text{(W-E-Lasso)}$$

Ideally, since the mixing model (2) is assumed to be exact, one would expect that the ℓ_1 or $\ell_{1,2}$ norm of the source STFT coefficients should be minimized under an exact time-domain mixture fitting constraint:

$$\min_{\tilde{\mathbf{s}}\ \text{s.t.}\ \mathbf{x}=\mathbf{A}\star\tilde{\mathbf{s}}\Phi^*} \|\tilde{\mathbf{s}}\|_1 \quad \text{(W-}\ell_1\text{)}$$

$$\min_{\tilde{\mathbf{s}}\ \text{s.t.}\ \mathbf{x}=\mathbf{A}\star\tilde{\mathbf{s}}\Phi^*} \|\tilde{\mathbf{s}}\|_{1,2}^2 \quad \text{(W-}\ell_{1,2}\text{)}$$

However, as we will see in Section IV, relaxed optimization problems (W-Lasso) and (W-E-Lasso) can be optimized quite efficiently compared to their constrained counterparts (W-ℓ_1) and (W-$\ell_{1,2}$). Moreover, as in the narrowband case, one can make the following remark, which admits a similar proof.

Remark 2: There exists $\alpha \geq 0$ such that if $\lambda \leq \alpha$, then any solution of (W-Lasso) is a solution of (W-ℓ_1) and any solution of (W-E-Lasso) is a solution of (W-$\ell_{1,2}$).

In particular, Remarks 2 emphases that the solution of the two methods (W-Lasso) and (W-E-Lasso) will be different, in general, as soon as $\lambda \neq 0$.

E. Remark on the ℓ_2 data term

From a purely Bayesian point of view, a ℓ_2 data term corresponds to a Gaussian prior. Consequently, when $\mathbf{e}$ in model (2) is a iid Gaussian white noise, the choice of such a data term in (W-Lasso) and (W-E-Lasso) is sounded. However, after the STFT, the noise $\tilde{\mathbf{e}}$ which appears in approximation (6) is no longer iid. The choice of the ℓ_2 data term in (N-Lasso) and (N-E-Lasso) is justified as constraint on the energy of the sought solution, but does not correspond as a "true" prior on the noise $\tilde{\mathbf{e}}$.

IV. Convex optimization algorithms

We now explain how to minimize the four functionals introduced above through recent developments in convex optimization that we apply to the problems at hand.

Let us synthetically describe the framework used to solve the general problem (9). More specifically, the algorithms we are interested in deal with problems of the form

$$\min_{\tilde{\mathbf{s}}} \mathcal{L}(\tilde{\mathbf{s}}) + \lambda \mathcal{P}(\tilde{\mathbf{s}}) \tag{12}$$

with $\mathcal{L}$ and $\mathcal{P}$ being convex, lower semicontinuous functions, and $\mathcal{L}$ L-Lipschitz differentiable.

Various algorithms were developed and studied in the last few years, in particular:

- the Iterative Shrinkage/Thresholding Algorithm (ISTA) [21], [22] used in our preliminary paper [9],
- Nesterov schemes [20],
- the Fast Iterative Shrinkage/Thresholding Algorithm (FISTA) [23].

Variants of ISTA were proposed to speed it up. However, we restrain ourselves to these three algorithms, which have the advantage of remaining simple, with no other parameter than the input functional. Moreover, important theoretical results are available, in particular on the speed of convergence.

All these algorithms have the advantage to tackle non-differentiable convex functionals of the form (12), when $\mathcal{P}$ is a non-differentiable convex function. Resolving such kind of problems relies on proximity operators introduced by Moreau [24]. In this section we shall summarize general concepts and derive particular algorithms suited to the minimization of functionals of the form (9), for the data and regularization terms described in the previous section.

A. Proximity operators and general algorithms

To start with, let us give the formal definition of proximity operators and the standard derivation of these operators for ℓ_1 and ℓ_{12} norms.

Definition 2 (Proximity operator): Let $\varphi : \mathbb{C}^I \to \mathbb{C}$ be a lower semicontinuous, convex function. The proximity operator associated with φ denoted by $\text{prox}_\varphi : \mathbb{C}^I \to \mathbb{C}^I$ is given by

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

The norms ℓ_1 and ℓ_{12} are convex, lower semicontinuous functions, nondifferentiable in 0, whose proximity operators can be computed in closed form. For the sake of simplicity, if $z = 0$, $\frac{z}{|z|} = 0$ by convention in the following.

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

$$u_i = \frac{z_i}{|z_i|}(|z_i| - \lambda)^+$$

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

Proposition 2 (Prox of $\frac{\lambda}{2}\|.\|_{1,2}^2$): Let $\mathbf{z} \in \mathbb{C}^{I \times J}$. For each column j, let i'_j be the sequence of row indexes such that the entries $z_{i'_j,j}$ in that column are ordered by decreasing absolute value: $|z_{i'_j+1,j}| \leq |z_{i'_j,j}|$ for all i'_j. Then, $\mathbf{u} = \text{prox}_{\frac{\lambda}{2}\|.\|_{1,2}^2}(\mathbf{z})$ is given entrywise by

$$u_{ij} = \frac{z_{ij}}{|z_{ij}|}\left(|z_{ij}| - \frac{\lambda}{1 + \lambda I_j} \sum_{i'_j=1}^{I_j} |z_{i'_j,j}|\right)^+$$

with I_j being the index such that

$$\lambda \sum_{i'_j=1}^{I_j} \left(|z_{i'_j,j}| - |z_{I_j,j}|\right) < |z_{I_j,j}|$$

and

$$|z_{I_j+1,j}| \leq \lambda \sum_{i'_j=1}^{I_j+1} \left(|z_{i'_j,j}| - |z_{I_j+1,j}|\right).$$

One can refer to [8] for the proof of these two propositions. It appears that these proximity operators reduce to a simple shrinkage/thresholding operator, hence the names ISTA and FISTA of the derived algorithms.

For the sake of completeness, we provide the general forms of ISTA, FISTA and Nesterov schemes in Algorithms 1, 2 and 3. All these algorithms have been proved to converge to a solution of (12). However, the convergence rate of ISTA is $\mathcal{O}(\frac{1}{k})$ where k is the number of iterations, while that of the other two algorithms is $\mathcal{O}(\frac{1}{k^2})$. In the context of audio source separation, we found that ISTA did not converge in reasonable time [9]. FISTA is based on Nesterov's ideas [25] and has the same convergence rate as the Nesterov schemes described in [20]. However, as it can be seen in Algorithms 2 and 3, Nesterov schemes rely on the computation of two gradients and two proximity operators at each iteration, instead of one gradient and one proximity operator only for FISTA. Preliminary experiments (not showed here) indicated that FISTA is a little bit more efficient than Nesterov schemes in the context of audio source separation, since computation of the gradient is quite expensive in this case. Therefore, we choose to focus on FISTA in the following. However, that does not mean that FISTA is more efficient than Nesterov schemes in general, from an optimization point of view.

Algorithm 1: ISTA [21], [22]

Initialization: $\tilde{\mathbf{s}}^{(0)} \in \mathbb{C}^{N \times B}$, $k = 1$.
repeat
  $\tilde{\mathbf{s}}^{(k)} = \text{prox}_{\frac{\lambda}{L}\mathcal{P}}(\tilde{\mathbf{s}}^{(k-1)} - \frac{\nabla\mathcal{L}(\tilde{\mathbf{s}}^{(k-1)})}{L})$;
until *convergence* ;

B. Application to the proposed functionals

Let us now derive the details of the application of FISTA to the proposed functionals, starting by the regularization term. The proximity operator of the ℓ_1 norm as applied to STFT coefficients can be readily computed from Proposition 1. A practical implementation of the proximity operator of the $\ell_{1,2}$ mixed norm is given in Algorithm 4.

Algorithm 2: FISTA [23]

Initialization: $\tilde{\mathbf{s}}^{(0)} \in \mathbb{C}^{N\times B}$, $\mathbf{z}^{(0)} = \tilde{\mathbf{s}}^{(0)}$, $t^{(0)} = 1$, $k = 1$.
repeat
$\tilde{\mathbf{s}}^{(k)} = \text{prox}_{\frac{\lambda}{L}\mathcal{P}}\left(\mathbf{z}^{(k-1)} - \frac{\nabla\mathcal{L}(\mathbf{z}^{(k-1)})}{L}\right)$;
$\tau^{(k)} = \frac{1+\sqrt{1+4\tau^{(k-1)^2}}}{2}$;
$\mathbf{z}^{(k)} = \tilde{\mathbf{s}}^{(k)} + \frac{\tau^{(k-1)}-1}{\tau^{(k)}}(\tilde{\mathbf{s}}^{(k)} - \tilde{\mathbf{s}}^{(k-1)})$;
$k = k + 1$
until *convergence* ;

Algorithm 3: Nesterov schemes [20]

Initialization: $\mathbf{s}^{(0)} \in \mathbb{R}^{N\times B}$, $\mathbf{g}^{(0)} = \mathbf{0}$, $\gamma = \frac{2}{L}$, $\kappa^{(0)} = 0$.
repeat
$\tau^{(k)} = \frac{\gamma+\sqrt{\gamma^2+4\gamma\kappa^{(k-1)}}}{2}$;
$\mathbf{v}^{(k)} = \text{prox}_{\kappa^{(k-1)}\lambda\mathcal{P}}(\tilde{\mathbf{s}}^{(0)} - \mathbf{g}^{(k-1)})$;
$\mathbf{z}^{(k)} = \frac{\kappa^{(k-1)}\tilde{\mathbf{s}}^{(k-1)}+\tau^{(k)}\mathbf{v}^{(k)}}{\kappa^{(k-1)}+\tau^{(k)}}$;
$\tilde{\mathbf{s}}^{(k)} = \text{prox}_{\frac{\lambda}{L}\mathcal{P}}(\mathbf{z}^{(k)} - \frac{\nabla\mathcal{L}(\mathbf{z}^{(k)})}{L})$;
$\mathbf{g}^{(k)} = \mathbf{g}^{(k-1)} + \tau^{(k)}\nabla\mathcal{L}(\tilde{\mathbf{s}}^{(k)})$;
$\kappa^{(k)} = \kappa^{(k-1)} + \tau^{(k-1)}$;
$k = k + 1$
until *convergence* ;

Algorithm 4: Computation of $\mathbf{u} = \text{prox}_{\frac{\lambda}{2}\|.\|_{1,2}^2}(\mathbf{z})$

for *each* (t, f) **do**
$\mathbf{y}(t, f) =$ sort $\mathbf{z}(t, f)$ in order of decreasing absolute value;
for $n = 1 : N$ **do**
$\|\mathbf{y}_{1:n}(t, f)\|_1 = \sum_{i=1}^{n} |y_i(t, f)|$;
find I such that:
$|y_I(t, f)| > \frac{\lambda}{1+\lambda I}\|\mathbf{y}_{1:I}(t, f)\|_1$
and
$|y_{I+1}(t, f)| \leq \frac{\lambda}{1+\lambda(I+1)}\|\mathbf{y}_{1:I+1}(t, f)\|_1$;
for $n = 1 : N$ **do**
$u_n(t, f) = \frac{z_n(t,f)}{|z_n(t,f)|}\left(|z_n(t, f)| - \frac{\lambda\|\mathbf{y}_{1:I}(t,f)\|_1}{1+\lambda I}\right)^+$;

Algorithm 5: Computation of L via power iteration

Initialization: $\mathbf{v} \in \mathbb{R}^{N\times B}$.
repeat
$\mathbf{w} = (\mathbf{A}^* \star \mathbf{A} \star \mathbf{v}\Phi^*)\Phi$;
$L = \|\mathbf{w}\|_\infty$;
$\mathbf{v} = \frac{\mathbf{w}}{L}$;
until *convergence* ;

Under the narrowband assumption, the data term $\mathcal{L}(\tilde{\mathbf{s}}) = \frac{1}{2}\|\tilde{\mathbf{x}} - \tilde{\mathbf{A}}\times_f\tilde{\mathbf{s}}\|_2^2$ is L-Lipschitz differentiable with gradient

$$\nabla\mathcal{L}(\tilde{\mathbf{s}}) = -\tilde{\mathbf{A}}^*\times_f(\tilde{\mathbf{x}} - \tilde{\mathbf{A}}\times_f\tilde{\mathbf{s}}) \tag{14}$$

where the adjoint $\tilde{\mathbf{A}}^*$ of $\tilde{\mathbf{A}}$ is such that $\tilde{\mathbf{A}}^*(f)$ is the Hermitian transpose of $\tilde{\mathbf{A}}(f)$ for each f. The Lipschitz constant L is given by

$$L = \max_f \|\tilde{\mathbf{A}}^*(f)\tilde{\mathbf{A}}(f)\|_{2_{op}} \tag{15}$$

with $\|.\|_{2_{op}}$ denoting the operator norm associated with the ℓ_2 vector norm. For a matrix, this norm is equal to the maximum absolute singular value.

Under the wideband assumption, the data term $\mathcal{L}(\tilde{\mathbf{s}}) = \frac{1}{2}\|\mathbf{x} - \mathbf{A} \star \tilde{\mathbf{s}}\Phi^*\|_2^2$ is L-Lipschitz differentiable with gradient

$$\nabla\mathcal{L}(\tilde{\mathbf{s}}) = [\mathbf{A}^* \star (\mathbf{x} - \mathbf{A} \star \tilde{\mathbf{s}}\Phi^*)]\Phi \tag{16}$$

where that the adjoint $\mathbf{A}^*$ of $\mathbf{A}$ is obtained by transposition of source and channel indexes and time reversal of the filters. Introducing the linear operator $\mathcal{T} : \mathbb{C}^{N\times B} \to \mathbb{R}^{M\times T}$ defined by

$$\mathcal{T}(\tilde{\mathbf{s}}) = \mathbf{A} \star (\tilde{\mathbf{s}}\Phi^*) \tag{17}$$

and its adjoint operator $\mathcal{T}^* : \mathbb{R}^{M\times T} \to \mathbb{C}^{N\times B}$ defined by

$$\mathcal{T}^*(\mathbf{x}) = (\mathbf{A}^* \star \mathbf{x})\Phi, \tag{18}$$

the Lipschitz constant L is given by

$$L = \|\mathcal{T}^*\mathcal{T}\|_{2_{op}}. \tag{19}$$

This operator norm can be computed using the well-known power iteration algorithm. Algorithm 5 recalls this procedure as applied to $\mathcal{T}^*\mathcal{T}$.

where $\|\mathbf{w}\|_\infty$ norm denotes the maximum absolute value of the vector (or matrix) $\mathbf{w}$.

By combining the above results for the data term and the regularization term, the application of FISTA to the proposed functionals leads to Algorithm 6.

While FISTA seems to be an efficient algorithm to optimize our functionals, some practical issues still remain. Firstly, the convergence criterion is not easy to choose. Popular choices are criteria like $\|\tilde{\mathbf{s}}^{(k+1)} - \tilde{\mathbf{s}}^{(k)}\| < \varepsilon$ or $\frac{\|\tilde{\mathbf{s}}^{(k+1)} - \tilde{\mathbf{s}}^{(k)}\|}{\tilde{\mathbf{s}}^{(k)}} < \varepsilon$. However, these quantities do not decrease monotonically and the choice of ε clearly depends on the mixture $\mathbf{x}$ and the hyper-parameter λ. More precisely, for small λ, FISTA requires a larger number of iterations to reach convergence. Secondly, the convergence speed of FISTA strongly depends on the chosen initialization. One can refer to [26] for more practical details on the convergence of these different algorithms.

Consequently, we chose to fix *a priori* the number of iterations to 20000 in practice, and we used the *continuation trick*, also known as *warm start* or *fixed point continuation* [27]: we first run FISTA with a large value of λ, then iteratively decrease the value of λ and initialize FISTA with the result of the previous run.

V. Experimental evaluation

We evaluated the narrowband and wideband source separation methods proposed in Section III over convolutive mixtures of speech sources and compared them to the state-of-the-art methods described in Section II. For all experiments, the test signals were sampled at 11 kHz and the STFT was computed with half-overlapping sine windows of 512 samples ($\simeq$ 46 ms).

Algorithm 6: Source separation via FISTA

Initialization: $\tilde{\mathbf{s}}^{(0)} \in \mathbb{C}^{N\times B}$, $\mathbf{z}^{(0)} = \tilde{\mathbf{s}}^{(0)}$, $\tau^{(0)} = 1$, $k = 1$.
repeat
  switch *data term* **do**
    case *narrowband*
      $\mathbf{a}^{(k)} = \mathbf{z}^{(k-1)} - \frac{1}{L}\tilde{\mathbf{A}}^* \times_f (\tilde{\mathbf{x}} - \tilde{\mathbf{A}} \times_f \tilde{\mathbf{s}}^{(k-1)})$;
    case *wideband*
      $\mathbf{a}^{(k)} = \mathbf{z}^{(k-1)} - \frac{1}{L}[\mathbf{A}^* \star (\mathbf{x} - \mathbf{A} \star \tilde{\mathbf{s}}^{(k-1)}\Phi^*)]\Phi$;
  switch *regularization term* **do**
    case ℓ_1
      $\tilde{\mathbf{s}}^{(k)} = \mathrm{prox}_{\frac{\lambda}{L}\|\cdot\|_1}(\mathbf{a}^{(k)})$ (see Proposition 1);
    case $\ell_{1,2}$
      $\tilde{\mathbf{s}}^{(k)} = \mathrm{prox}_{\frac{\lambda}{2L}\|\cdot\|_{1,2}^2}(\mathbf{a}^{(k)})$ (see Algorithm 4);
  $\tau^{(k)} = \frac{1+\sqrt{1+4\tau^{(k-1)^2}}}{2}$;
  $\mathbf{z}^{(k)} = \tilde{\mathbf{s}}^{(k)} + \frac{\tau^{(k-1)}-1}{\tau^{(k)}}(\tilde{\mathbf{s}}^{(k)} - \tilde{\mathbf{s}}^{(k-1)})$;
  $k = k + 1$
until *convergence* ;

A. Numerical comparison between wideband and narrowband models

First of all, we report in Table I the numerical approximation error of the narrowband model (6) compared with the wideband model (2) for $N = 4$ sources. These relative approximation errors are computed as:

$$E_a = \frac{\|\mathbf{x}_{\mathrm{narrow}} - \mathbf{x}_{\mathrm{wide}}\|_2}{\|\mathbf{x}_{\mathrm{wide}}\|_2} \ ,$$

where $\mathbf{x}_{\mathrm{narrow}}$ is the mixture obtained from the narrowband model and $\mathbf{x}_{\mathrm{wide}}$ the mixture from the wideband model. Errors for $N = 3$, 5 or 6 sources are of the same order of magnitude. One can remark that these numerical results confirm that the narrowband model collapses for filters with long delay, as expected from Lemma 1. Moreover, Table I provides the value of the corresponding infinity norm of the Fourier transform of the mixing system $\tilde{\mathbf{A}}'$. Again, these numerical results are in adequacy with Lemma 1: $\|\tilde{\mathbf{A}}\|_\infty$ reaches the highest value for $RT_{60} = 250$ ms.

TABLE I
NUMERICAL ILLUSTRATION OF LEMMA 1: APPROXIMATION ERROR OF THE NARROWBAND APPROXIMATION COMPARED WITH THE WIDEBAND MODEL AND VALUE OF $\|\tilde{\mathbf{A}}'\|_\infty$.

RT_{60}	dist	E_a	$\|\tilde{\mathbf{A}}'\|_\infty$
Anechoic	5 cm	0.018	0.80
	1 m	0.018	0.84
50 ms	5 cm	0.035	1.95
	1 m	0.039	1.97
250 ms	5 cm	0.188	4.87
	1 m	0.217	4.87

B. Experimental protocol

For the following, the experimental protocol is described hereafter. The mixing filters were room impulse responses simulated via the image technique [28] using the Roomsim software[1] with the same room size as in [5]. The number of microphones was set to $M = 2$ and the number of sources was varied in the range $3 \leq N \leq 6$. For each number of sources N, six different sets of mixing filters were generated corresponding to three different reverberation times RT_{60} (anechoic, $RT_{60} = 50$ ms and $RT_{60} = 250$ ms) and two different microphone spacings d ($d = 5$ cm and $d = 1$ m). Each set of mixing filters was convolved with ten different sets of male and/or female speech sources from various nationalities, yielding ten mixtures per mixing condition.

In order to only evaluate the different methods in the light of the source separation efficiency, we choose not to add any artificial noise. Such a choice was made to do not measure the denoising abilities of the algorithms.

Each mixture was separated with the proposed narrowband and wideband Lasso and E-Lasso methods in (N-Lasso), (N-E-Lasso), (W-Lasso) and (W-E-Lasso) for different values of the hyper-parameter λ. The separation performance was then assessed for each λ using the Signal-to-Distortion Ratio (SDR), the Signal-to-Interference Ratio (SIR) and the Signal-to-Artifacts Ratio (SAR) in decibels (dB) as defined in [29]. The SDR indicates the overall quality of each estimated source compared to the target, while the SIR reveals the amount of residual crosstalk from the other sources and the SAR is related to the amount of musical noise. These measures were subsequently averaged over all sources and all mixtures for each mixing condition. The state-of-the-art narrowband DUET and ℓ_1 norm minimization methods in (N-DUET) and (N-ℓ_1), with the same known mixing system $\mathbf{A}$, were also performed as a baseline.

C. Performance analysis as a function of λ

Figure 1 illustrates the variation of the SDR as a function of λ with $N = 4$, $\mathrm{RT}_{60} = 250$ ms and $d = 1$ m. The curves in this figure are quite different from those in our preliminary paper [9]. Indeed, this work relied on the Modified Discrete Cosine Transform (MDCT) instead of the STFT and on ISTA without the continuation trick, which did not experimentally converge, contrary to FISTA here.

Several interesting trends can be seen. Firstly, the performance of the narrowband Lasso and E-Lasso methods becomes equal to that of the narrowband ℓ_1 norm minimization method when $\lambda \to 0$, which is consistent with Remark 1. Furthermore, the performance of the wideband Lasso and the wideband E-Lasso are different for $\lambda \to 0$ (*c.f. Remark 2)*. Secondly, the performance of the narrowband Lasso and E-Lasso methods is maximum for $\lambda > 0$, while that of the wideband Lasso and E-Lasso methods is maximum for $\lambda \to 0$. As we choose not to add any noise ($\mathbf{e} = 0$ in Eq. (2)), observing a better behavior of the wideband methods when $\lambda \to 0$ is consistent with the mixing model. Furthermore, since narrowband methods are an approximation of model (2), it is expected that relaxing the equality constraint yields a better estimation, according to SDR, when $\lambda > 0$. Thought we observe an improvement using this relaxation in practice, it remains small. This disappointing

[1] http://media.paisley.ac.uk/~dccampbell/Roomsim/

observation could come from the bad ℓ_2 prior made on the data term, as explained in Section III-E.

Similar trends were observed for other mixing conditions. In the following, we compare all methods after selecting the best λ for each mixing condition as that maximizing the average SDR over the ten corresponding mixtures.

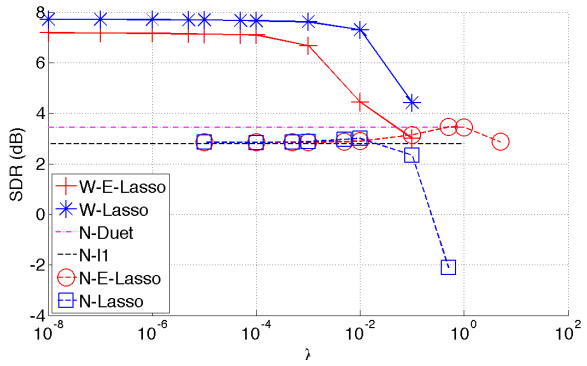

Fig. 1. Variation of the average SDR as a function of λ over speech mixtures with $N = 4$ sources, $\mathrm{RT}_{60} = 250$ ms and $d = 1$ m.

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

The resulting SDR figures are shown in Table II for different values of RT_{60} and d. Among state-of-the-art methods, DUET outperforms ℓ_1 norm minimization in all cases but the anechoic case. In order to be able to decide if a method performs better than an another one, we have applied an ANOVA at 1% on the results. If the null hypothesis of the ANOVA is rejected (i.e. some methods can be considered statistically better than some another), we applied the Fisher protocol (i.e. several paired Fisher-Student T-test) to determine which methods perform best. Algorithms which perform statistically best are in bold.

The proposed narrowband Lasso and E-Lasso methods consistently improve performance over ℓ_1 norm minimization. This confirms again that introducing a data term into the model helps dealing with the error resulting from the narrowband approximation. Moreover, under this approximation, E-Lasso regularization appears always more appropriate than the Lasso. This corroborates our analysis of the limitations of the Lasso with large λ in Section III-B and our claim that promoting disjointness instead of sparsity of the source time-frequency representations is a better strategy in this context. Note however that the observed improvements remain inferior to 1 dB in all cases but one. This is not enough to justify the use of the narrowband E-Lasso method in practice since it performs at most as well as DUET except in the anechoic case, despite its larger computational cost and the use of an additional hyper-parameter λ.

The same remark applies to the proposed wideband Lasso and E-Lasso methods in environments with low reverberation time $\mathrm{RT}_{60} \leq 50$ ms. This was expected since the narrowband assumption is valid when the mixing filters are shorter than the STFT window length, hence circumventing this assumption does not provide significant benefit. However, wideband Lasso and E-Lasso improve the average SDR by 2 to 4 dB compared to DUET in a more realistic environment with $\mathrm{RT}_{60} = 250$ ms. This improvement is huge compared to the small difference of performance between state-of-the-art methods. One possible explaination is that wideband methods are less sensitive to spatial aliasing than narrowband methods [30]. Also, the issue of choosing the hyper-parameter λ does not arise here since $\lambda \to 0$ appears to be always a reasonable choice, so that these methods can be applied in practice. Note also that E-Lasso regularization performs worse than the Lasso in this context.

Table III provides more insight into the performance of state-of-the-art methods and wideband methods when $\mathrm{RT}_{60} = 250$ ms and $d = 1$ m. It can be seen that, besides improving the SDR, the wideband Lasso method also improves both the SIR by 4 dB compared to DUET and the SAR by 3 dB compared to ℓ_1 norm minimization, which are the state-of-the-art methods providing fewer interference and fewer artifacts respectively. Again, algorithms which perform statistically best are in bold.

E. Performance analysis as a function of the number of sources

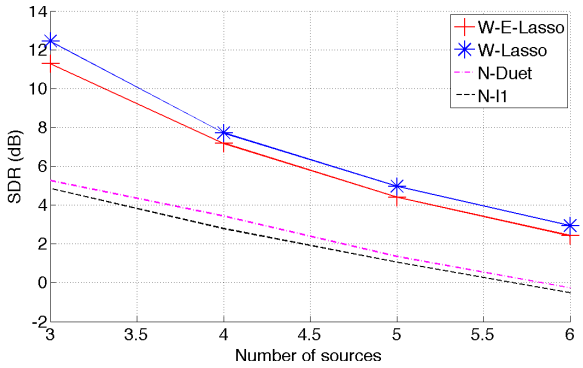

Fig. 2. Variation of the average SDR as a function of N over speech mixtures with $\mathrm{RT}_{60} = 250$ ms and $d = 1$ m.

In addition to the above discussion for $N = 4$ sources, we provide in Figure 2 the results for $3 \leq N \leq 6$ sources with $\mathrm{RT}_{60} = 250$ ms and $d = 1$ m. The performance of all methods as well as the performance improvement brought by wideband methods appears to decrease when the number of sources increases. This is natural since the disjointness or sparsity assumptions underlying all methods hold to a lesser extent with a large number of sources. Yet, wideband E-Lasso still improves the average SDR by 3 dB compared to DUET with $N = 6$ sources.

F. Summary

We resume here the results shown above.

- If $RT_{60} = 250$ ms, the wideband methods outperform the narrowband methods. (W-Lasso) seems to perform a bit better than (W-E-Lasso), but the difference between the two methods is less than 1 dB. The parameter λ should be chosen close to zero in the noise-free case, and one can use the continuation trick to perform this optimization more efficiently.

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

RT_{60}	d	narrowband				wideband	
		DUET	ℓ_1 min.	Lasso	E-Lasso	Lasso	E-Lasso
anechoic	5 cm	5.9	4.5	**6.4**	**6.7**	5.7	**6.5**
	1 m	7.3	7.7	7.8	8.0	7.6	**8.1**
50 ms	5 cm	5.5	4.2	4.3	4.3	4.4	4.5
	1 m	6.4	6.3	6.4	6.4	**7.0**	**7.4**
250 ms	5 cm	2.7	1.8	2.1	2.1	**5.9**	5.0
	1 m	3.4	2.8	3.0	3.4	**7.6**	**7.2**

TABLE III
AVERAGE SDR, SIR AND SAR IN DECIBELS OVER SPEECH MIXTURES WITH $N = 4$ SOURCES, $\mathrm{RT}_{60} = 250$ MS AND $d = 1$ M.

method	narrowband				wideband	
	DUET	ℓ_1 min.	Lasso	E-Lasso	Lasso	E-Lasso
SDR	3.4	2.8	3.0	3.4	**7.6**	**7.2**
SIR	10.0	6.4	6.8	7.7	**14.0**	**13.9**
SAR	5.1	6.5	6.4	6.4	**9.1**	8.5

- If $RT_{60} \leq 50$ ms, even if the wideband methods (in particular (W-E-Lasso)) perform better than the narrowband methods, improvement is not high enough to justify the use of such methods because of the computational cost (*c.f.* Section V-H): (N-ℓ_1) seems the more appropriate method in the anechoic case and (N-DUET) otherwise.

G. Robustness to error in filtering system evaluation

All the previous experiments were made in the non blind case, when the mixing system $\mathbf{A}$ is perfectly known. However, in a practical situation, one would solve the blind problem, and needs to estimate this mixing system. In order to evaluate the robustness of the proposed wideband methods, we made the two following experiments. In these two experiments, we kept the best performance reached by the wideband methods, which implies to choose the appropriate parameter λ.

In the first experiment we cut the true filters at 25, 50, 100 and 150 ms after the direct sound arrival. The aim of this experiment is to show whether the good performance of the wideband methods was due to accurate modeling of the first echos or of the tail of the filters. Figure 3 shows that the longer the filters, the better the SDR of the wideband methods. The tail of the filters seems important to reach satisfactory results. While the best performance can be reached for a specific $\lambda > 0$, one can choose in practice $\lambda \to 0$. Indeed, the difference of performance is very small (< 0.5 dB). On the other hand, the narrowband methods (N-DUET) and (N-ℓ_1) perform best when the tail of the filters is cut, and it would be interesting to investigate a new approach to model error of the narrowband methods. Even when filters are cut to 50 ms, which is the same order of magnitude than the length of the STFT window, wideband methods perform better than narrowband methods.

Figure 4 shows the results of the second experiment, where we add Gaussian noise with exponentially decaying amplitude with the same slope as the reverberation. The input SNR was computed only on the reverberant part of the filters. Wideband methods still perform better than narrowband methods. However, in that case, one must choose the right parameter λ. If the (W-Lasso) performs best, as shown in Figure 4, it is in practice much more sensitive to the choice of λ than the (W-E-Lasso):

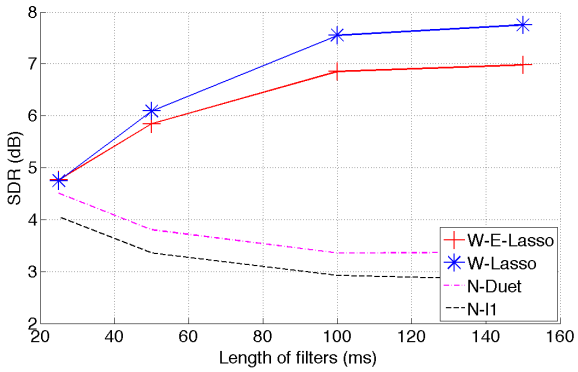

Fig. 3. Variation of the average SDR as a function of the length of the filters over speech mixtures with $\mathrm{RT}_{60} = 250$ ms and $d = 1$ m.

after reaching the maximum SDR, the performance of the (W-Lasso) collapses very quickly contrary to the (W-E-Lasso).

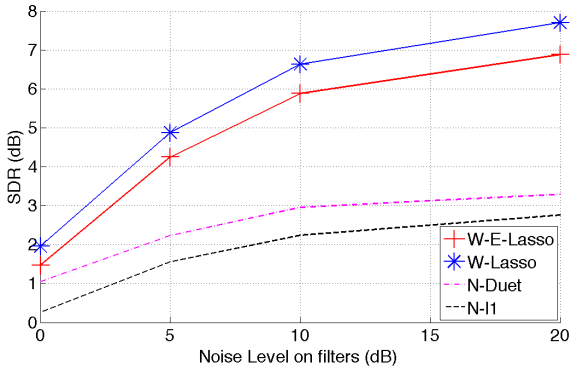

Fig. 4. Variation of the average SDR as a function of the noise level over speech mixtures with $\mathrm{RT}_{60} = 250$ ms and $d = 1$ m.

H. Computational comparison

To close this experimental section, we give some indications about the computational time of the different methods, to separate $N = 4$ sources with a $RT_{60} = 250$ ms, $d = 1$ m mixing

system. The fastest is obviously (N-DUET), which takes less than one second to proceed the separation. The (N-ℓ_1) method, implemented via the Newton-based optimization technique in [17] takes about three hours. The proposed narrowband methods (N-Lasso) and (N-E-Lasso) take about 1.5 hours for 20000 iterations of FISTA. The (W-E-Lasso) is little bit longer, because of the complexity of the proximity operator associated with the ℓ_{12} norm. Finally, the wideband methods (W-Lasso) and (W-E-Lasso) take respectively about 5 hours and 6 hours to run 20000 iterations of FISTA.

VI. Conclusion

We proposed a general convex optimization framework for under-determined convolutive source separation, which relies on the minimization of a family of narrowband or wideband functionals, assuming that the mixing filters are known or have been estimated using *e.g.* some subspace-based channel identification technique. It is the first time, to our knowledge, that a way of avoiding the narrowband assumption is proposed in the under-determined setting. We translated this framework into a family of algorithms that were carefully compared on a range of speech mixtures. In light of the results, we conclude that the best method is either DUET or ℓ_1 norm minimization depending on the microphone spacing in environments with low reverberation and the proposed wideband Lasso method in more realistic environments whose reverberation time is on the order of a hundred milliseconds or more. In the latter case, wideband Lasso improves the SDR, SIR and SAR measures by several decibels compared to the best state-of-the-art.

There are numerous perspectives to this work. Firstly, the proposed framework could be exploited for the estimation of the mixing filters, possibly using additional sparsity or nonstationarity constraints about filter coefficients. Similarly, this framework could be used to achieve denoising in addition to source separation in case the mixture contains some additive noise by readily modeling such noise by the data term with $\lambda > 0$. Secondly, one must keep a look on the work made in the convex optimization community to find a possibly faster algorithm. Finally, while the mixed norm considered here was not retained for practical separation purposes, different mixed norms could be defined to improve the modeling of the source signals by favoring persistent structures in the source time-frequency representations.

VII. Acknowledgment

We would like to thank Haibin Yang for pointing errors in the implementation of DUET in the published version of this paper.

Appendix

Lemma 1: Let $x \in L_2(\mathbb{R})$, $s \in L_2(\mathbb{R})$ and $A \in C^1(\mathbb{R})$ such that

$$x = A \star s.$$

Then, denoting by S_x (resp. S_s) the STFT of x (resp. s) with an analysis window g in the Sobolev $W^{1,2}(\mathbb{R})$, the STFT of x can be expressed in each time-frequency bin (t, f) by

$$S_x(t,f) = \tilde{A}(f) S_s(t,f) + \mathcal{R}(t,f)$$

where $\tilde{A}$ is the Fourier transform of A and

$$|\mathcal{R}(t,f)| \leq \|\tilde{A}'\|_\infty \|s\|_2 \left(\int_{\mathbb{R}} |\nu \tilde{g}(\nu)|^2 \, \mathrm{d}\nu \right)^{1/2}.$$

Proof: Applying the Fourier transform to (1), we obtain $\tilde{x} = \tilde{A}\tilde{s}$. We can then write the STFT of x as

$$S_x(t,f) = \int_{\mathbb{R}} \tilde{A}(\nu) \tilde{s}(\nu) \tilde{g}(\nu - f) e^{2i\pi\nu t} \, \mathrm{d}\nu. \tag{20}$$

The Taylor expansion of order 1 of $\tilde{A}$ in f is given by

$$\tilde{A}(\nu) = \tilde{A}(f) + (\nu - f) r_f(\nu) \tag{21}$$

where

$$|r_f(\nu)| \leq \|\tilde{A}'\|_\infty \tag{22}$$

and $\tilde{A}'$ denotes the derivative of $\tilde{A}$. By inserting (21) into (20), we get

$$S_x(t,f) = \tilde{A}(f) S_s(t,f) + \mathcal{R}(t,f)$$

with

$$\mathcal{R}(t,f) = \int_{\mathbb{R}} r_f(\nu) \tilde{s}(\nu) (\nu - f) \tilde{g}(\nu - f) e^{2i\pi\nu t} \, \mathrm{d}\nu.$$

Thanks to the Cauchy-Swartz inequality, one can write

$$|\mathcal{R}(t,f)| \leq \|r_f\|_\infty \|s\|_2 \left(\int_{\mathbb{R}} |\nu \tilde{g}(\nu)|^2 \, \mathrm{d}\nu \right)^{1/2},$$

which, combined with (22), yields the desired result. ∎

References

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

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

[3] R. Saab, Ö. Yılmaz, M. McKeown, and R. Abugharbieh, "Blind separation of anechoic under-determined speech mixtures using multiple sensors," in *Proc. IEEE Int. Symposium on Signal Processing and Information Technology (ISSPIT)*, 2006, pp. 642–646.

[4] 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 Advances in Signal Processing*, vol. 2007, 2007, article ID 24717.

[5] E. Vincent, S. Araki, and P. Bofill, "The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation," in *Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 734–741.

[6] W. Kellermann and H. Buchner, "Wideband algorithms versus narrowband algorithms for adaptive filtering in the DFT domain," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, vol. 2, 2003, pp. 1278–1282.

[7] M. Kowalski and B. Torrésani, "Sparsity and persistence: mixed norms provide simple signals models with dependent coefficients," *Signal, Image and Video Processing*, vol. 3, no. 3, pp. 251–264, 2008.

[8] M. Kowalski, "Sparse regression using mixed norms," *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 303–324, 2009.

[9] M. Kowalski, E. Vincent, and R. Gribonval, "Under-determined source separation via mixed-norm regularized minimization," in *Proc. European Signal Processing Conference (Eusipco)*, 2008.

[10] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, August 2007.

[11] A. Aïssa-El-Bey, K. Abed-Meraim, and Y. Grenier, "Blind separation of underdetermined convolutive mixtures using their time-frequency representation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1540–1550, 2007.
[12] R. Gribonval, "Piecewise linear source separation," in *Proc. SPIE '03*, vol. 5207 Wavelets: Applications in Signal and Image Processing X, San Diego, CA, 2003, pp. 297–310.
[13] H. Kiers, "Towards a standardized notation and terminology in multiway analysis," *Journal of Chemometrics*, vol. 14, pp. 105–122, 2000.
[14] J. Rosca, C. Borss, and R. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. III–877–880.
[15] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.
[16] B. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Transactions On Signal Processing*, vol. 47, no. 1, pp. 187–200, January 1999.
[17] 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.
[18] M. Togami, T. Sumiyoshi, and A. Amano, "Sound source separation of overcomplete convolutive mixture using generalized sparseness," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006.
[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Serie B*, vol. 58, no. 1, pp. 267–288, 1996.
[20] P. Weiss, "Fast algorithms for convex optimization. applications to image reconstruction and change detection." Ph.D. dissertation, INRIA Sophia Antipolis / Laboratoire I3S - ARIANA, 2008.
[21] I. Daubechies, M. Defrise, and C. D. 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, August 2004.
[22] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, Nov. 2005.
[23] 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.
[24] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bull. Soc. Math. France*, vol. 93, pp. 273–299, 1965.
[25] Y. E. Nesterov, "method for solving the convex programming problem with convergence rate $o(1/k^2)$," *Dokl. Akad. Nauk SSSR*, vol. 269, no. 3, pp. 543–547, 1983.
[26] I. Loris, "On the performance of algorithms for the minimization of ℓ_1-penalized functionals," *Inverse Problems*, vol. 25, p. 035008 (16pp), 2009.
[27] E. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for l1-minimization: Methodology and convergence," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1107–1130, 2008.
[28] U. Svensson and U. Kristiansen, "Computational modelling and simulation of acoustic spaces," in *Proc. AES 22nd Conf. on Virtual, Synthetic and Entertainment Audio*, 2002, pp. 1–20.
[29] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Speech, Audio and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.
[30] J. Dmochowski, J. Benesty, and S. Affès, "On spatial aliasing in microphone arrays," *IEEE Trans. on Signal Processing*, vol. 57, no. 4, pp. 1383–1395, 2009.