Doubly sparse models for multiple filter estimation in sparse echoic environments

Prasad Sudhakar, Simon Arberet, Rémi Gribonval, Pierre Vandergheynst

Doubly sparse models for multiple filter estimation in sparse echoic environments

Prasad Sudhakar[a,*], Simon Arberet[b], Rémi Gribonval[a], Pierre Vandergheynst[b]

[a] *INRIA Rennes - Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France.*
[b] *LTS2 Signal Processing Lab., Institute of Electrical Engineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland.*

Abstract

We consider the estimation of multiple time-domain sparse filters from echoic mixtures of several unknown sources, when the sources are sparse in the time-frequency domain. We propose a sparse filter estimation framework consisting of two steps: a) a clustering step to group the time-frequency points of mixtures where only one source is active, for each source; b) a convex optimisation step to estimate the filters based on a time-frequency domain *cross-relation*. We propose a new *wideband formulation* of a frequency domain cross-relation, besides the one based on classical *narrowband approximation*. The solutions of the convex optimisation problem, formed using the cross-relation, are characterised. Numerical evaluation shows the benefit of using the wideband cross-relation for sparse echoic filter estimation. Further, the potential of the proposed framework for *blind* estimation of sparse echoic filters is demonstrated in a controlled experimental setting where in the proposed approach outperforms the state of the art blind filter estimation techniques, when the filters are sufficiently sparse.

Keywords:
sparse filter estimation, ℓ^1 minimisation, sparse echoic blind source separation, convex optimisation

1. Introduction

Blind source separation (BSS) finds many applications in speech processing, music transcription, biomedical signal processing, etc. Its aim is to estimate unknown source signals from a set

*Corresponding author.
Email addresses: prasad.sudhakar@gmail.com (Prasad Sudhakar), simon.arberet@epfl.ch (Simon Arberet), remi.gribonval@inria.fr (Rémi Gribonval), pierre.vandergheynst@epfl.ch (Pierre Vandergheynst)

of observed mixtures, without the explicit knowledge of the mixing system associated to *mixing filters*. A standard BSS architecture consists of two stages: the estimation of mixing filters and the estimation of sources, given the estimated mixing filters. This paper focusses on the filter estimation problem.

The filter estimation problem is intrinsically ill-posed: it cannot be solved without further enabling hypotheses on the filters and/or the sources. Independent Component Analysis (ICA) [1, 2] stems from the classical assumption of *statistical independence* of the sources. In recent years, Sparse Component Analysis (SCA) has been successfully applied to separate instantaneous and anechoic mixtures (where the mixing filters are simply scalars or time-delayed scalars). SCA typically exploits *source sparsity in the time-frequency domain* [3, 4, 5, 6], a property satisfied by many acoustic signals.

Beyond the anechoic case, a standard approach to deal with convolutive mixtures is to use the Short-Time-Fourier-Transform (STFT) to convert them (using the so-called *narrowband approximation*) into multiple complex-valued instantaneous mixtures that can be separated using ICA or SCA in each frequency bin. However, this results in *permutation and scaling ambiguity* of each frequency band [7]. To resolve these, prior information must be used such as the location of the sources (e.g. the direction of arrival) [8, 9, 10, 11], or a consistency property of the filters (e.g.: spectral smoothness) [12, 13] or of the sources (e.g.: correlated energy profiles in different subbands) [14, 8, 15]. In the absence of scaling, the permutations of the filters can be resolved using their *time-domain sparsity* [16]. In this paper, the time-domain sparsity of filters is exploited together with the time-frequency domain sparsity of the sources to resolve all ambiguities (permutation and scaling). For that, we build upon the *cross-relation* [17], a tool widely used in communications engineering [18] for blind filter estimation from several filtered versions of a *single source*. This formulation is free of the explicit source term [19, 20, 21] and yields a constrained optimisation problem [22], as recalled in Sec. 2.

Narrowband vs. wideband. In the context of BSS, most existing work is based on the *narrowband approximation* of the mixing process (even if it is not always assumed to be an approximation), including DUET [23] and convolutive ICA methods [24, 25, 26], approaches for filter estimation [4, 5, 6, 16] and other methods for convolutive source separation [27, 28]. Existing

work pointing out the limitations of the narrowband approximation includes the statistical models of time-frequency source images introduced by Duong et al [29, 30], and the convex optimization approaches for source estimation with *known* mixing filters of Kowalski et al [31]. Kowalski et al. show theoretically and experimentally that a *wideband* approach outperforms a narrowband approximation of the optimized criteria for audio problems with reverberant filters ($RT_{60} > 50$ ms). From this perspective, the current paper complements the work of Kowalski et al. by demonstrating and documenting the importance of a wideband formulation of convolution in the context of *blind sparse echoic source separation*.

Relevance of the sparse filter assumption. The framework proposed and investigated in this paper exploits the time-domain sparsity of the filters. Time-domain sparsity is exhibited by channels which have a few reflection paths compared to its length. Typical examples of such channels are encountered in underwater acoustics [32, 33], wideband wireless communications [34, 35] and seismic signal processing [36].
In acoustics, room impulse responses cannot, properly speaking, be considered as sparse: while the early echoes generate relatively sparse filters, reverberation induces non-sparse tails. However, in moderately reverberant rooms, sparse echoic models can be considered as natural extensions of the widely used and coarser anechoic model.

Contributions and organisation of the paper

In a nutshell, SCA exploits the time-frequency sparsity of the sources to estimate mixing filters, but suffers from permutation and scaling ambiguities. Cross-relation techniques to estimate sparse filters are essentially [37] restricted to the single source case. In this context, the contributions of this paper are:

1. The demonstration that single source cross-relation techniques can be extended to handle multichannel mixtures of multiple sources using *partial time-frequency cross-relations* (Sec. 2). The *partial* nature of such cross-relations comes with a price, in that we have to assume certain sparsity assumptions on the sources (in the time-frequency domain) and the filters (in the time domain). The relevance of these assumptions is discussed below.
2. A convex formulation of the sparse filter estimation problem from partial time-frequency cross-relations, together with a characterisation of the ambiguities of the solutions of the

convex optimisation problem (Sec. 3).

3. A two-stage framework for multiple mixing filter estimation, consisting of a time-frequency clustering stage and a convex optimisation stage (Sec. 4).
4. The numerical demonstration that, while the de facto standard in convolutive source separation relies on a narrowband approximation to express filtering in the time-frequency domain, a precise wideband formulation can dramatically improve the ability to estimate sparse echoic filters using the proposed partial time-frequency cross-relations (Sec. 5).

To conclude the paper (Sec. 6), an experimental illustration of the potential of the proposed framework is provided, in a controlled audio scenario where the clustering step can be addressed blindly (all but one of the sources are mixed with linear instantaneous filters). While limited in scope, the experiment confirms the ability of the proposed framework to provide significant accuracy improvements compared to state of the art methods (GCC-PHAT, JADE), over a range of sparsities of the estimated echoic filter, provided the clustering step is conducted with sufficient accuracy. This should motivate further work at the junction between signal processing and machine learning to design clustering techniques adapted to this problem.

2. Partial time-frequency cross-relations

The simplest of the filter estimation problems is the so-called single-input-two-output (SITO) problem: two signals $x_i(t)$, $i = 1, 2$ are observed, which are filtered versions of the same (unknown) source signal $s(t)$: $x_i = a_i \star s$, $i = 1, 2$, where a_i is a filter of length L associated to the path between the source and the i-th sensor.

In this case the *cross-relation* $x_2 \star a_1 = x_1 \star a_2$ holds [17]. To express it in matrix form, let us associate the filter a_i to the column vector $\mathbf{a}_i = [a_i(t)]_{t=0}^{L-1}$ and likewise s to $\mathbf{s}$ and x_i to $\mathbf{x}_i$. The convolution $x_i \star a_j$ is associated to the multiplication between a Toeplitz matrix[1] $\mathcal{T}[\mathbf{x}_i]$ and the vector $\mathbf{a}_j$. Denoting $\mathbf{B} = \mathcal{B}[\mathbf{x}_1, \mathbf{x}_2] := \big[\mathcal{T}[\mathbf{x}_2], -\mathcal{T}[\mathbf{x}_1]\big]$, the cross-relation becomes

$$\mathbf{B} \cdot \mathbf{a} = \mathbf{0}, \text{ where } \quad \mathbf{a} = \begin{bmatrix} \mathbf{a}_1^{\mathbf{T}} & \mathbf{a}_2^{\mathbf{T}} \end{bmatrix}^{\mathbf{T}}. \tag{1}$$

[1] Calligraphic letters denote matrices built from a vector, e.g. $\mathcal{T}[\mathbf{x}_i]$.

A traditional exploitation to estimate $\mathbf{a}$ given the cross-relation is to minimise the ℓ^2 norm of the cross-relation term (under a normalisation constraint [19] to avoid the trivial solution):

$$\text{minimize} \quad \|\mathbf{B} \cdot \mathbf{a}\|_2 \quad \text{s.t.} \quad \|\mathbf{a}\|_2 = 1. \tag{2}$$

The solution is found by solving an eigen-vector problem.

SITO approaches such as described above were extended to N sources [37] by assuming that one can identify time segments where only one source contributes to the mixtures. However in general, the sources may overlap almost everywhere in the time-domain, limiting the applicability of this approach. Instead of sources with disjoint time supports, it is common to consider sources disjoint in the time-frequency domain. We thus develop time-frequency cross-relations associated to other matrices $\mathbf{B}$ such that $\mathbf{B} \cdot \mathbf{a} \approx \mathbf{0}$, the rows of which are indexed by time-frequency points. The matrices are built in the single source setting but will be later exploited to address filter estimation in the context of multiple sources, where time-frequency disjointness will allow us to select few rows of these matrices associated to time-frequency regions where the cross-relations are valid.

2.1. The Short Time Fourier Transform

We begin with a short recollection of the Short-Time Fourier Transform (STFT). Consider a source signal $s(t)$ with $0 \leq t < T$, and let $w(t)$ be a discrete window function[2]. The STFT of $s(t)$ is defined as

$$\widehat{s}(\tau, f) = \sum_{t=0}^{T-1} s(t) w(t - \tau) e^{-2\mathbf{i}\pi f t}.$$

The STFT coefficients are computed on a discrete grid: $\tau \in \{qF/2 : q_{\min} \leq q \leq q_{\max}\}$, $0 \leq f < F$, where $q, q_{\min}, q_{\max}$ and f are integers, and F is the window length. The STFT can be interpreted as projections of s on a collection of Gabor time-frequency atoms $\psi_{\tau,f}(t) := w(t - \tau) e^{2i\pi f t}$ [38].

2.2. A narrowband approximation *of the cross-relation*

By the narrowband approximation [24], we have

$$\widehat{x}_i(\tau, f) \approx \widehat{a}_i(f) \cdot \widehat{s}(\tau, f),\ i = 1, 2 \tag{3}$$

[2]For experiments we used a Blackman-Harris window [38].

where $\widehat{a}_i(f)$ is the Discrete Fourier Transform (DFT) coefficient of the filter $a_i(t)$ at frequency index f. A narrowband (approximate) time-frequency cross-relation [39] follows:

$$\widehat{a}_2(f) \cdot \widehat{x}_1(\tau, f) - \widehat{a}_1(f) \cdot \widehat{x}_2(\tau, f) \approx 0, \, \forall \tau, f. \tag{4}$$

By collecting such relations for all the time-frequency points, we can write the cross-relation in matrix form $\mathbf{B}_{\mathtt{nb}} \cdot \mathbf{a} \approx \mathbf{0}$, where $\mathbf{B}_{\mathtt{nb}} := \big[\mathbf{B}_{\mathtt{nb}}[\mathbf{x}_2], -\mathbf{B}_{\mathtt{nb}}[\mathbf{x}_1]\big]$. The rows of $\mathbf{B}_{\mathtt{nb}}[\mathbf{x}_i]$ are indexed by (τ, f), and the matrix $\mathbf{B}_{\mathtt{nb}}$ is of size $(Q \cdot F) \times 2L$, with Q the number of STFT frames. The details of the structure of $\mathbf{B}_{\mathtt{nb}}$ are given in Appendix Appendix A.

As it can be noticed from Eq. (4), the narrowband cross-relation is intrinsically approximate, as it relies on the narrowband *approximation*. This is not desirable and more so in the case of multiple sources. Hence, we develop a wideband version of the cross-relation and interpret its meaning.

2.3. A wideband expression *of the cross-relation*

An accurate wideband formulation of the cross-relation can be obtained by first formulating it in the time domain, and then taking a time-frequency transformation. This transformation is interpreted as a projection of the time-domain cross-relation onto a suitable time-frequency atom. For this purpose, we first propose the following standard lemma, which is proved using Fubini's Theorem and a change of variable.

Lemma 1. *Let $x(t)$ be a bounded signal, $a(t)$ and $\psi(t)$ two finite support signals, and let $\langle f, g\rangle = \sum_{t=-\infty}^{+\infty} f(t)\bar{g}(t)$ be the Hermitian inner product between two complex-valued signals f and g, where $\bar{\cdot}$ denotes complex conjugation. Then: $\langle x \star a, \psi\rangle = \langle a, \tilde{x} \star \psi\rangle$ with $\tilde{x}(t) = \bar{x}(-t)$.*

The projection of the cross-relation $x_2 \star a_1 = x_1 \star a_2$ on an atom ψ, that is $\langle x_2 \star a_1 - x_1 \star a_2, \psi\rangle = 0$ can be written as $\langle \tilde{x}_2 \star \psi, a_1\rangle = \langle \tilde{x}_1 \star \psi, a_2\rangle$. Note that since the filters a_i are real-valued with support $[\![0, L-1]\!]$, we have

$$\langle \tilde{x}_i \star \psi, a_j\rangle = \sum_{\ell=0}^{L-1} \langle x_i, \psi_{-\ell}\rangle a_j(\ell)$$

where $\psi_{-\ell}$ is the atom ψ shifted in time by ℓ samples. For Gabor atoms[3] $\psi = \psi_{\tau,f}$ this yields the

[3]More generally it is possible to define cross-relations in any domain associated to a dictionary $\mathcal{D}$ of atoms, such as a multi-scale Gabor dictionary or a union of a wavelet basis and a local Fourier basis.

wideband cross-relation

$$\sum_{\ell=0}^{L-1} \widehat{x}_2(\tau-\ell,f)a_1(\ell) - \sum_{\ell=0}^{L-1} \widehat{x}_1(\tau-\ell,f)a_2(\ell) = 0. \tag{5}$$

Note that unlike the narrowband *approximate* cross-relation (4), the wideband cross-relation (5) is a perfect equality if the time domain cross-relation $x_2 \star a_1 = x_1 \star a_2$ holds.

Wideband cross-relations for all Gabor atoms can be expressed in matrix form $\mathbf{B}_{\mathtt{wb}} \cdot \mathbf{a} = \mathbf{0}$, where $\mathbf{B}_{\mathtt{wb}} := \left[\mathbf{B}_{\mathtt{wb}}[\mathbf{x}_2], -\mathbf{B}_{\mathtt{wb}}[\mathbf{x}_1]\right]$. Each row of $\mathbf{B}_{\mathtt{wb}}[\mathbf{x}_i]$ is of the form

$$\mathbf{b}_{\mathtt{wb}}^T[x_i](\tau,f) := \left[\widehat{x}_i(\tau,f), \widehat{x}_i(\tau-1,f), \ldots, \widehat{x}_i(\tau-L+1,f)\right].$$

2.4. Cross-relations for multiple sources

Equipped with time-frequency cross-relations when the mixture is generated by a single source, we now propose *partial* time-frequency cross-relations to deal with multiple sources. We consider $M = 2$ noiseless mixtures $x_i(t), i = 1, 2$ of N unknown source signals $s_j(t), j = 1 \ldots N$, related by the convolutive model

$$x_i(t) = \sum_{j=1}^{N} (a_{ij} \star s_j)(t), \quad \forall t.$$

Each filter $a_{ij}(t)$ is of length L and models the impulse response between the j^{th} source and the i^{th} sensor. For brevity, we denote the sources, filters and mixtures by s_j, a_{ij} and x_i, by dropping the time index. To obtain the time-frequency cross-relation in multiple source scenario, we can either look at the narrowband formulation or the wideband formulation.

Under the narrowband approximation (3), we have

$$\widehat{x}_i(\tau,f) \approx \sum_{j=1}^{N} \widehat{a}_{ij}(f) \cdot \widehat{s}_j(\tau,f), \quad i = 1, 2, \ \forall(\tau,f).$$

and we have the narrowband cross-relation (4) taking the form, for any source j

$$\begin{aligned} &\widehat{a}_{2j}(f)\cdot\widehat{x}_1(\tau,f) - \widehat{a}_{1j}(f) \cdot \widehat{x}_2(\tau,f) \\ &\qquad = \sum_k \widehat{s}_k(\tau,f)\Big[\widehat{a}_{2j}(f)\widehat{a}_{1k}(f) - \widehat{a}_{1j}(f)\widehat{a}_{2k}(f)\Big]. \end{aligned} \tag{6}$$

This expression does not a priori yield a value close to zero. This is due to the interference of multiple sources at the considered time-frequency point. However, even if the interference from other sources are absent, there is no way to clearly say that the cross-relation has been satisfied

because of the approximate nature of the formulation itself. Therefore, it is desirable to have an accurate formulation of the cross-relation.

In contrast to the narrowband cross-relation, the proposed wideband cross-relation in a time-frequency activity region Ω_j suffers from only one type of inaccuracy, and hence it is more likely to be more effective for filter estimation. For a given source j and time-frequency atom $\psi_{\tau,f}$ we have by Lemma 1:

$$\begin{aligned}
&\langle x_2 \star a_{1j}, \psi_{\tau,f}\rangle - \langle x_1 \star a_{2j}, \psi_{\tau,f}\rangle \\
&\quad = \sum_k \left[\langle s_k \star a_{2k} \star a_{1j}, \psi_{\tau,f}\rangle - \langle s_k \star a_{1k} \star a_{2j}, \psi_{\tau,f}\rangle\right] \\
&\quad = \sum_{k\neq j} \langle a_{2k} \star a_{1j} - a_{1k} \star a_{2j}, \tilde{s}_k \star \psi_{\tau,f}\rangle. \qquad (7)
\end{aligned}$$

Since the filters a_{ij} are supported on $[\![0, L-1]\!]$, the filters $a_{ik} \star a_{i'j}$ are supported on $[\![0, 2L-1]\!]$, and the interference terms in the right hand side of (7) will vanish if for all $k \neq j$:

$$\widehat{s_k}(\tau - \ell, f) = \langle s_k, \psi_{\tau-\ell,f}\rangle = (\tilde{s}_k \star \psi_{\tau,f})(\ell) = 0,$$

for all $0 \leq \ell \leq 2L-1$. This determines a time-frequency region around the point (τ, f) such that if source j is the only one significantly active in this region, then the wideband cross-relation holds at the point (τ, f). As a consequence of this, one can build matrices $\mathbf{B}^j = \mathbf{B}_{\mathtt{wb}}^{\Omega_j}$ as restrictions of the matrix $\mathbf{B}_{\mathtt{wb}}$ to the rows indexed by time-frequency points in such regions Ω_j.

Thus, unlike the narrowband cross-relation, the wideband cross-relation does not suffer from approximation error, and hence it is more appropriate to be used for filter estimation. This will be demonstrated numerically in section 5.

Before we proceed with presenting our filter estimation framework and experimental results, we shall first characterise the indeterminacies of the cross-relation.

3. Sparse filter estimation from cross-relations

Suppose we know a matrix $\mathbf{B}$ that embodies the cross-relation $\mathbf{B} \cdot \mathbf{a} \approx 0$ satisfied by the unknown pair of filters $\mathbf{a}$. How do we estimate $\mathbf{a}$ given this knowledge? The most traditional approach, which is exploited when a full time-domain cross-relation is at hand, is to solve the optimization problem (2) through an eigen-value problem. However, when exploiting *partial* (time-frequency)

cross-relations, one can expect that $\mathbf{B}$ will carry less information about $\mathbf{a}$, hence additional regularization can be expected to help favor certain types of solutions such as sparse solutions, which involve a smaller number of degrees of freedom. Indeed, even in the case of pure time-domain cross-relations, it has been shown that the time-domain sparsity of the filters can be exploited [22] using the modified approach:

$$\text{minimize} \;\; \|\mathbf{a}\|_1 \quad \text{s.t.} \quad \|\mathbf{B}\cdot\mathbf{a}\|_2 \leq \;\epsilon \;\; \text{and} \;\; \|\mathbf{a}\|_2 = 1. \tag{8}$$

However, this new problem is non-convex, because of the nonconvex normalisation constraint $\|\mathbf{a}\|_2 = 1$. Below we propose a variant of this problem which is convex, and discuss the shift ambiguities of the solution of the various formulations.

3.1. Proposed convex formulation

To obtain a convex problem, we replace the normalisation $\|\mathbf{a}\|_2 = 1$ with the constraint $a_1(t_0) = 1$, where t_0 is an arbitrarily chosen time index:

$$\text{minimize} \;\; \|\mathbf{a}\|_1 \quad \text{s.t.} \quad \|\mathbf{B}\cdot\mathbf{a}\|_2 \leq \;\epsilon \;\; \text{and} \;\; \mathbf{a}_1(t_0) = 1. \tag{9}$$

Numerical solver. This convex formulation is central to the rest of the paper. It can be solved using any convex optimisation algorithm. In all the experiments reported in this article we have used the `CVX` toolbox [40].

3.2. Debiasing

It is known that the estimation of sparse vectors using ℓ^1 regularization introduces a "bias" corresponding to a soft-thresholding of the significant coefficients. Therefore, it is a well-established common practice to improve estimation performance through a debiasing (DB) step, performed as a post-processing step.

Given the sparsity $k = \|\mathbf{a}\|_0$ of the filters as side-information, the optimisation problem in Eq. (9) is first solved with a constraint $\mathbf{a}_1(L/2) = 1$ to obtain a first estimate $\widetilde{\mathbf{a}}$. Let $\mathbf{b}_{L/2}$ be the $(L/2)^{th}$ column vector of the matrix $\mathbf{B}$ and let $\mathbf{B}_{\Gamma'}$ be a matrix built using the columns of $\mathbf{B}$ whose indices are in the set $\Gamma' = \Gamma - \{L/2\}$, where Γ is the support of the k largest entries of $\widetilde{\mathbf{a}}$. The

debiased estimate $\widetilde{\mathbf{a}}'$ is obtained as

$$\begin{aligned}\widetilde{\mathbf{a}}' &:= \arg\min \|\mathbf{B}\cdot\mathbf{a}\|_2^2 \text{ s.t. } \mathbf{a}(L/2)=1 \text{ and } \mathrm{support}(\mathbf{a})=\Gamma,\\ &= \mathbf{B}_{\Gamma'}^{\dagger}\cdot\left(-\mathbf{b}_{L/2}\right),\end{aligned}$$

where $\mathbf{B}_{\Gamma'}^{\dagger}$ is the Moore-Penrose pseudoinverse of $\mathbf{B}_{\Gamma'}$.

Unless stated otherwise, debiasing is performed in all experiments reported in this paper. Figure 3 will confirm the strong positive impact of debiasing on the filter estimation results.

3.3. Indeterminacies of the cross-relations

When we neglect the boundary effects and consider an infinite length source s and infinite length observations x_i, the exact cross-relation can at best characterize the filters up to a global scaling and a shift. This is a straightforward consequence of the linearity and shift invariance of the convolution: the equality of the signals $x_2 \star a_1 = x_1 \star a_2$ implies that of the shifted and scaled signals $x_2 \star (\alpha T_\tau a_1) = x_1 \star (\alpha T_\tau a_2)$, where α is an arbitrary scalar factor and T_τ is the shift operator, that is to say the convolution with the shifted Dirac δ_τ. To what extent does this ambiguity apply to the matrix formulation (1) of the cross-relation, i.e. to the restriction to finite signals? The answer requires considering more carefully the boundaries by denoting $1 \le \mathrm{l}_i(\mathbf{a}) \le L$ the index of the first nonzero coefficient of the filter on the i-th channel, $\mathbf{a}_i \in \mathbb{R}^L$, $\mathrm{r}_i(\mathbf{a})$ the index of its last nonzero coefficient, $\mathrm{l}(\mathbf{a}) = \min_i \mathrm{l}_i(\mathbf{a})$ and $\mathrm{r}(\mathbf{a}) = \max_i \mathrm{r}_i(\mathbf{a})$. For $-\mathrm{l}(\mathbf{a}) \le \tau \le L - \mathrm{r}(\mathbf{a})$, we can simultaneously shift both filter vectors $\mathbf{a}_i$ $(i = 1, 2)$ by τ samples without boundary effect: for example, for $\tau \le 0$, we shift the vector to the left by removing the leading τ zero entries, and zero-padding the end of the filter vector. Denoting again T_τ this shifting operation we obtain:

Proposition 1. *Let* $\mathbf{a} = [\mathbf{a}_1^T, \mathbf{a}_2^T]^T$ *be a pair of filter vectors satisfying the cross-relation* $\mathbf{B}\cdot\mathbf{a} = \mathbf{0}$*, with* $\mathbf{B} := \mathcal{B}[\mathbf{x}_1, \mathbf{x}_2]$*. For any scalar* α *and* $-\mathrm{l}(\mathbf{a}) \le \tau \le L - \mathrm{r}(\mathbf{a})$*, the scaled and shifted vector* $\widetilde{\mathbf{a}} := [\alpha(T_\tau \mathbf{a}_1)^T, \alpha(T_\tau \mathbf{a}_2)^T]^T$ *satisfies the cross-relation* $\mathbf{B}\cdot\widetilde{\mathbf{a}} = \mathbf{0}$*.*

Thus, any solution to the cross-relation (1) admitting either leading or trailing zero entries is not unique, even up to normalisation. The scaling ambiguity of the optimisation problems (2)-(8), is fixed (up to a sign) by the normalization $\|\mathbf{a}\|_2 = 1$ (resp. $\|\mathbf{a}\|_1 = 1$), but a shift ambiguity can remain.

An immediate consequence is that if the true filter (which by definition satisfies the cross-relation) satisfies either $\mathrm{l}(\mathbf{a}) > 1$ or $\mathrm{r}(\mathbf{a}) < L$, then it cannot be the unique solution to any of the optimization problems (2)- (8). This is consistent with known results on the identifiability of filters (see e.g. [41, Theorem 1, Theorem 2]), both in principle and in the context of cross-relations [17]. Yet, this could be problematic in a context where we consider sparse filters, which typically have many zero entries, in particular leading and trailing zeroes. [4]

The optimisation problem (9), which is convex, admits either a unique solution or a convex set of solutions. As shown below, under mild conditions it cannot be subject to the same shift and scaling ambiguities as the more standard formulations (2)-(8).

Lemma 2. *Let* $\mathbf{a} = [\mathbf{a}_1^T, \mathbf{a}_2^T]^T$ *be a pair of filter vectors satisfying the cross-relation* $\mathbf{B} \cdot \mathbf{a} = \mathbf{0}$*, with* $\mathbf{B} := \mathcal{B}[\mathbf{x}_1, \mathbf{x}_2]$*, and the constraint* $\mathbf{a}_1(t_0) = 1$*. Assume that:*

1. $t_0 \in [\![2\mathrm{r}(\mathbf{a}) - L,\ 2\mathrm{l}(\mathbf{a})]\!]$,
2. $\|\mathbf{a}_1\|_\infty > 1$*;*

then, there exist a shifted and scaled version $\tilde{\mathbf{a}}$ *of* $\mathbf{a}$ *that also satisfies the cross-relation* $\mathbf{B} \cdot \tilde{\mathbf{a}} = \mathbf{0}$ *and the constraint* $\tilde{\mathbf{a}}_1(t_0) = 1$*, such that* $\|\tilde{\mathbf{a}}\|_1 < \|\mathbf{a}\|_1$.

Proof. Denote $t_1 \in \arg\max_t |\mathbf{a}_1(t)|$. By assumption 1:

$$-\mathrm{l}(\mathbf{a}) \leq \mathrm{l}(\mathbf{a}) - t_0 \leq t_1 - t_0 \leq \mathrm{r}(\mathbf{a}) - t_0 \leq L - \mathrm{r}(\mathbf{a}).$$

Hence, the shift $\tau = t_1 - t_0$ satisfies the hypothesis of Proposition 1, and the cross-relation $\mathbf{B} \cdot \widetilde{\mathbf{a}} = \mathbf{0}$ must hold for the scaled and shifted filter vector $\widetilde{\mathbf{a}} := [\alpha(T_\tau \mathbf{a}_1)^T, \alpha(T_\tau \mathbf{a}_2)^T]^T$. With $\alpha := 1/\mathbf{a}_1(t_1)$, the the additional constraint $\mathbf{a}_1(t_0) = 1$ also holds, and $\|\tilde{\mathbf{a}}\|_1 = |\alpha| \|\mathbf{a}\|_1 = \|\mathbf{a}\|_1 / \|\mathbf{a}_1\|_\infty < \|\mathbf{a}\|_1$. □

An immediate consequence is that, if the filter vector $\mathbf{a}$ that is solution to (9) satisfies assumption 1 in Lemma 2, then $\mathbf{a}_1(t)$ must reach its maximum magnitude at $t = t_0$. No other shift is allowed, unless the filter reaches its maximum magnitude in at least two different locations.

The above analysis suggests to choose $t_0 = L/2$ (for simplicity we consider the case of an even filterlength L) for the additional constraint $\mathbf{a}(t_0) = 1$. With such a choice, if the original sparse filter

[4] Note that identifiability from cross-relations is only possible if the filters have no common zero, in the sense of common *roots* of associated polynomials. This should not be confused with the notion of *zero entries* of the filters, which is the main source of shift ambiguity.

satisfies $\mathrm{r}(\mathbf{a}) - \mathrm{l}(\mathbf{a}) \leq L/2$, there is a shifted and scaled version $\tilde{\mathbf{a}}$ so that $\mathbf{a}_1(t_0) = \|\mathbf{a}_1(t)\|_\infty = 1$. One then can hope to recover this particular filter vector as the solution to (9), even though we have no formal guarantee of recovery. This is why, in numerical simulations, we chose $t_0 = L/2$ and restrict the support of the ground truth input filters to the set $[\![\frac{L}{4}, \frac{3L}{4}]\!]$.

4. Proposed framework

Recalling that the rows of $\mathbf{B}_{\mathtt{wb}}$ are indexed by time-frequency points, the cross-relation $\mathbf{B}^j \cdot \mathbf{a}^j \approx \mathbf{0}$ actually captures *partial frequency information* about the unknown filters $\mathbf{a}^j$. Reconstructing the filters from this information leads to a potentially ill-posed linear inverse problem, if the number of time-frequency points in Ω_j is small. Sparse regularisation overcomes this difficulty by exploiting sparsity, provided the filters are sparse enough compared to the amount of frequency information actually captured. This leads to the proposed sparse filter estimation framework, in two stages:

1. **time-frequency clustering**: to identify time-frequency activity regions Ω_j for each source j;
2. **convex optimisation**: to estimate the filters by solving an ℓ^1 minimisation problem (with or without debiasing)

$$\min \|\mathbf{a}^j\|_1 \text{ s.t. } \|\mathbf{B}^j \cdot \mathbf{a}^j\|_2 \leq \ \epsilon \text{ and } \mathbf{a}^j_1(t_0) = 1.$$

The success of the convex optimisation stage will of course heavily depend on the amount of available frequency information, i.e., the size of the detected time-frequency activity regions Ω_j and the reliability of the approximation $\mathbf{B}^j \cdot \mathbf{a}^j \approx 0$. In other words, the success of the convex optimisation stage can drastically depend on the success of the time-frequency clustering stage, which is itself a challenging task.

After the filters are estimated as solutions of convex optimisation problems, a post processing step is performed to improve the quality of the solutions. This step, called *debiasing* is a popular post processing in sparse signal processing. We describe this step in Sec. 3.2.

The subsequent two sections correspond to two experiments to assess the performance of the proposed framework.

- *Oracle time-frequency clustering* (Section 5): to assess the performance of *convex optimisation stage alone*, with the clustering step being "ideally" solved using an oracle.

- *Blind time-frequency clustering* (Section 6): to assess the *overall performance* of the proposed framework, in a controlled experimental setting were we are able to propose a *blind* time-frequency clustering technique.

5. Wideband *vs* Narrowband: oracle experiments

To assess the performance of the convex optimisation stage when the time-frequency clustering stage is "ideally" solved, we propose to design an "oracle" that provides the activity regions Ω_j *given the true filters* $\mathbf{a}^j$ *as side information.*

5.1. Principle of the oracle clustering

The true filters $\mathbf{a}^j$ indeed satisfy the cross-relation at all time-frequency locations where only source j is dominant. Therefore, we can identify the time-frequency locations where source j is dominant by examining the points where the cross-relation holds.

To determine whether the cross-relation holds for a given time-frequency point (τ, f), it is natural to consider the magnitude of the entry of $\mathbf{B} \cdot \mathbf{a}^j$ associated to the row $\mathbf{b}^T(\tau, f)$:

$$\mathtt{CR}^j(\tau, f) := \langle \mathbf{b}[\mathbf{x}_2](\tau, f), \mathbf{a}_{1j} \rangle - \langle \mathbf{b}[\mathbf{x}_1](\tau, f), \mathbf{a}_{2j} \rangle.$$

However, the fact that this quantity is close to zero does not guarantee that the j-th source dominates the other sources at time-frequency location τ, f: it could simply happen, for example, that no source is active at this location. To avoid such degenerate conditions, we measure the energy content at the considered time-frequency locations by evaluating

$$\mathtt{EC}^j(\tau, f) := \langle \mathbf{b}[\mathbf{x}_2](\tau, f), \mathbf{a}_{1j} \rangle + \langle \mathbf{b}[\mathbf{x}_1](\tau, f), \mathbf{a}_{2j} \rangle.$$

We propose to classify a time-frequency location as belonging to source j when the cross-relation term is small while the energy content remains significant, given a threshold ν:

$$(\tau, f) \in \Omega_j \iff 20 \cdot \log_{10} \frac{|\mathtt{EC}^j(\tau, f)|}{|\mathtt{CR}^j(\tau, f)|} \geq \nu. \tag{10}$$

5.2. Condensation of matrix $\mathbf{B}_{wb}^{\Omega_j}$

Each row of the matrix $\mathbf{B}_{\mathtt{wb}}$ corresponds to a time-frequency point in the mixture. Hence, these matrices can be very large if the mixtures have long durations, and even $\mathbf{B}_{\mathtt{wb}}^{\Omega_j}$ can be very

large, raising computational challenges without necessarily bringing relevant information about the unknown filters: indeed, several rows of the matrices correspond to identical frequency bins, potentially yielding redundant information. To ease the computations, we chose to condense the matrix $\mathbf{B}_{\mathtt{wb}}^{\Omega_j}$: for each frequency bin, we keep only the row associated to the largest ratio in the right-hand side of (10). With this, $\mathbf{B}_{\mathtt{wb}}^{\Omega_j}$ can have at most F rows, where F is the STFT window size.

5.3. Experimental protocol

The experiments reported below, with three audio sources ($N = 3$), are designed to evaluate the wideband cross-relation method and to calibrate two parameters that drive the performance of the approaches: a) the size F of the STFT window; b) the value of the clustering threshold parameter ν.

For comparison, experiments are also carried out using the cross-relation that relies on the narrowband approximation. The narrowband cross-relation for the multiple sources setting is developed in appendix Appendix A. For both experiments the data are generated in the following way.

5.3.1. Audio source signals

The three sources used in all experiments are real audio recordings: 1) a flute sound, 2) a guitar sound and 3) a vocal recording. All sources are of length $T = 80,000$ samples, corresponding to approximately 8 seconds of recording, sampled at 8KHz.

5.3.2. Sparse filter generation

The sources are mixed with sparse filters to obtain the mixtures. Each time-domain sparse filter $\mathbf{a}_i$, $i = 1, 2$ of length $L = 256$ is generated to have $k/2$ non-zero coefficients, for various even integer values of k. That is, $\|\mathbf{a}_i\|_0 = k/2$, $i = 1, 2$. The $k/2$ support indices on each channel are chosen uniformly at random in the set $(\frac{L}{4}, \frac{3L}{4})$ (See Sec. 3.3 for an explanation). The filter coefficients are generated i.i.d. Gaussian with zero mean, unit variance and sorted to have decreasing magnitudes along the time axis. The first channel filter $a_1(t)$ is then normalised and shifted to have $a_1(L/2) = 1$. Note that the filter coefficient at the index $L/2$ need not be the largest coefficient in magnitude. The vector $\mathbf{a}$ defined in Eq. (1) has totally k non-zero coefficients.

Remark: for $k = 2$ there is a single peak in each filter: this corresponds to anechoic filters.

5.3.3. Parameters of the convex optimisation stage

The solution to the optimisation problem critically depends on the error threshold ϵ, which is hard to tune. We observed that a slight change in the value of ϵ can affect the recovery performance drastically. After repeated experimentation, we fixed $\epsilon = 10^{-5}$ for all experiments.

5.3.4. Performance measure

The output SNR of the estimated filters is measured in decibel (dB) scale, accounting for the possible scaling and shift ambiguity of the obtained solution:

$$\texttt{SNR}_{out} = 10 \log_{10} \frac{\sum_j \|\mathbf{a}^j\|_2^2}{\sum_j \min_{\mu_j, t_j} \|\mathbf{a}^j - \mu_j \cdot T_{t_j} \tilde{\mathbf{a}}^j\|_2^2}$$

where $\mathbf{a}^j$ is the vector associated to the true filter of the j^{th} source and T_τ is the operator that shifts both channels of the estimated vector $\tilde{\mathbf{a}}^j$ from τ samples, cf Sec. 3.3. The overall recovery performance is computed by averaging the output SNR of 20 independent trials for each configuration.

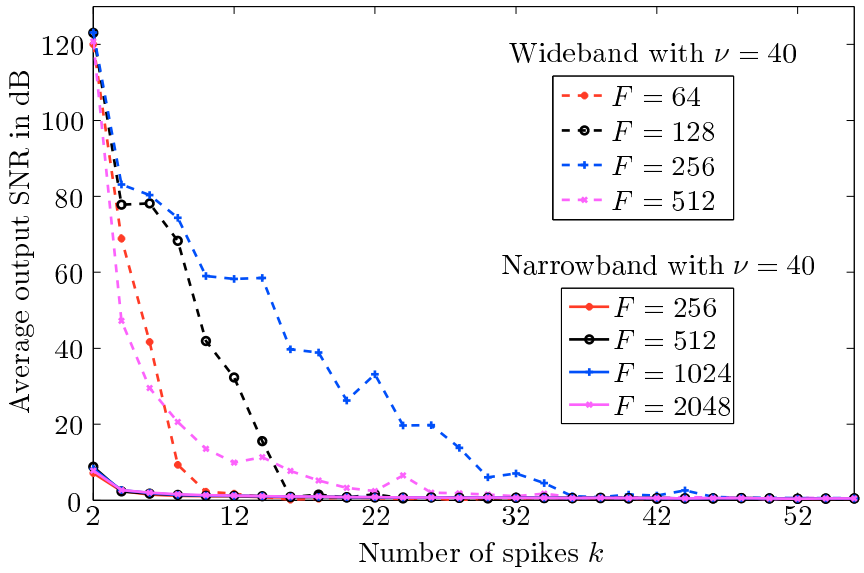
5.4. Results

Fig. 1(a) shows the filter recovery performance of both the wideband and narrowband approaches as a function of the filter sparsity k, for different settings of the STFT window length F. The clustering threshold ν is set to 40 dB in both the cases. The dashed and solid lines correspond to the wideband and the narrowband methods respectively.

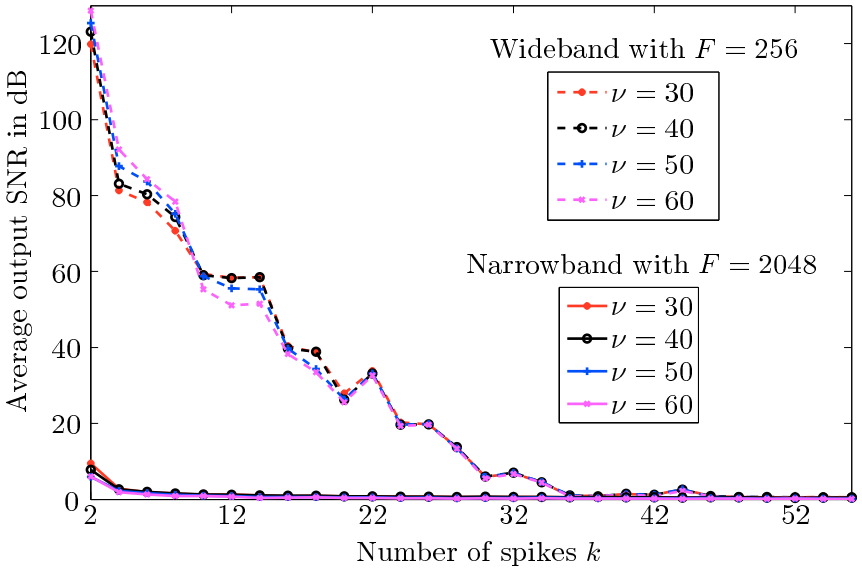
For the narrowband method, STFT window lengths much longer than the filter size provide better approximation of the cross-relation in Eq. (4) than shorter window lengths, and hence we expect it to aid filter recovery. However, in spite of long STFT window lengths $F = 1024$ and 2048, the output SNR does not exceed 8 dB even for filters with sparsity $k = 2$.

The wideband method has its best performance when the STFT window size equals the individual filter length, $F = L = 256$, with the output SNR being more than 20 dB for sparsities $k \leq 26$. For filter sparsity $k = 2$, the output SNR for all the STFT window sizes are comparable, whereas the SNR falls rapidly for higher values of k when $F = 64, 128$ and 512.

The effect of the clustering threshold ν on the filter estimation performance is shown in Fig. 1(b). Here again, the dashed and solid lines correspond to the wideband and the narrowband methods respectively. The STFT window length for the wideband method is set to $F = 256$, because the



(a) Wideband and narrowband performance for a fixed clustering threshold $\nu = 40$ dB.



(b) Wideband and narrowband performance for fixed STFT window sizes $F = 256$ and $F = 2048$ respectively.

Figure 1: Oracle performance of filter estimation using the wideband and narrowband cross-relations.

best performance is obtained for this value of F as seen from Fig. 1(a). For the narrowband method, the window size is set to $F = 2048$ because it provides the best approximation of the cross-relation amongst the window sizes we have chosen to experiment with.

The narrowband method displays a very poor performance for filter sparsities as low as $k = 2$

(anechoic filters) even when the STFT window size is eight times the size of filters to be recovered. We see from the results that in spite of using the *oracle* clustering step, the narrowband method is ineffective in recovering the filters. This is due to the inherent approximate nature of the narrowband cross-relation.

The wideband approach, on the other hand, recovers the filters with more than 20 dB output SNR for sparsity up to $k \leq 26$, irrespective of the STFT window length, and the SNR falls off when $k \geq 36$.

For higher clustering thresholds ν, the time-frequency points which satisfy the cross-relation poorly are rejected by the clustering stage. Therefore, as ν increases, lesser number of time-frequency points are likely to be selected. As a result, the information about the unknown filters, contained in the clustered time-frequency points, is potentially incomplete. However, thanks to the accurate nature of the wideband cross-relation, the method can still successfully recover the sparse filters from partial information.

6. Filter estimation by blind clustering

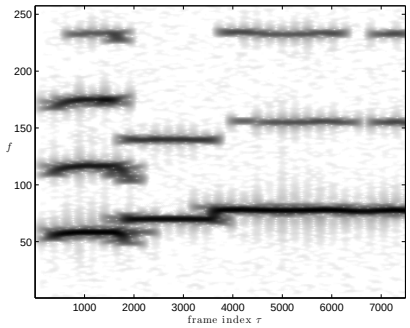
We now wish to investigate the performance of filter estimation while performing *blind* time-frequency clustering.

6.1. Considered simplified setting

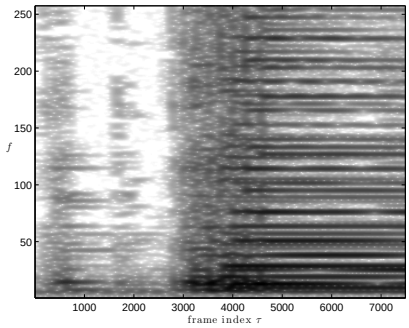
We consider a simplified setting in which *all the sources but one are mixed by linear instantaneous filters*. Let us denote the length of the associated filters by $L_j = 1$. The mixing parameter of each of the instantaneously mixed sources is associated to a corresponding *Intensity Parameter* (IP) defined as $\theta_j = \tan^{-1}(a_{2j}/a_{1j})$ [6]. For convenience let us assume that the sparsely mixed source is the last one $j = N$.

6.2. Blind clustering strategy

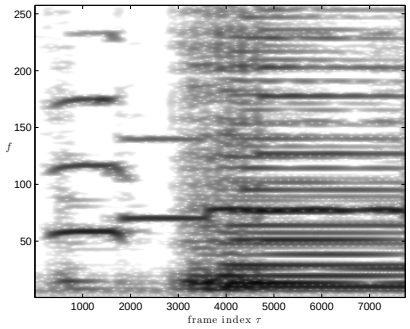
Several existing approaches such as DUET [3] or DEMIX [6] can cluster time-frequency points belonging to the instantaneously mixed sources, and estimate their corresponding intensity parameters $\hat{\theta}_j$, $j < N$. An indicator of where source j is prominently active is the deviation between the inverse tangent of the ratio of the mixtures at each time-frequency point and the estimated



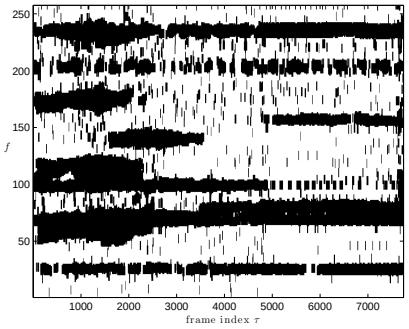
(a) Spectrogram of source s_1: a flute sound



(b) Spectrogram of source s_2: a guitar sound



(c) Spectrogram of mixture x_1



(d) Time-frequency mask. The white pixels correspond to the points in the set Ω_2

Figure 2: Illustration of the proposed blind time-frequency clustering method. See main text for description.

intensity parameter $\hat{\theta}_j$: given a threshold η we thus define

$$(\tau, f) \in \Omega_j \iff \left| \tan^{-1}\left(|\widehat{\mathbf{x}}_2(\tau, f)/\widehat{\mathbf{x}}_1(\tau, f)|\right) - \hat{\theta}_j \right| < \eta. \tag{11}$$

In this way we can find the regions $\Omega_j, j < N$ in the mixtures where sources other than $j = N$ are active. Finally the time-frequency points associated to the instantaneously mixed sources can be "removed" from the full time-frequency plane Ω, yielding a set of points where it is likely that the sparsely mixed source is predominant. This is done by:

1. building regions $\overline{\Omega}_j$ containing all time-frequency points close, either in time or in frequency, to the points in Ω_j; This operation is done by performing a binary dilation of the set Ω_j with

a square kernel $\mathcal{S}_\gamma$ of side γ:

$$\begin{aligned}\overline{\Omega}_j &:= \Omega_j \oplus \mathcal{S}_\gamma \\ &:= \{(\tau+\tau', f+f') | (\tau, f) \in \Omega_j, (\tau', f') \in \mathcal{S}_\gamma\};\end{aligned} \tag{12}$$

2. retaining a set Ω_N as the complement of the set $\cup_{j<N}\overline{\Omega}_j$ in Ω.

This is illustrated on Fig. 2. Figs. 2(a)-2(b) show the STFT of the two sources used in the experimental protocol below: a flute source and a guitar source (black corresponds to high energy, white to low energy). Fig. 2(c) displays the STFT of one of the mixtures x_1. Figure 2(d) illustrates the set $\Omega_2 = \overline{\Omega}_1^c$ obtained with the described approach (white indicates points in Ω_2). We can see that, as expected, Ω_2 only contains time-frequency points where source s_1 is not very active.

6.3. Condensation of matrix $\mathbf{B}_{wb}^{\Omega_j}$

As in the oracle setting, the size of the matrix $\mathbf{B}_{\mathtt{wb}}^{\Omega_2}$ is reduced for computational gain. However, unlike previously, it is impossible to observe which time-frequency point at a given frequency most strongly satisfies the cross-relation. Therefore, instead of selecting a single time-frequency point, we merge all time-frequency points associated to a frequency bin f by computing the first principal component $PC(.)$ of the rows of the $\mathbf{B}_{\mathtt{wb}}^{\Omega_2}$ corresponding to f, that is:

$$\breve{\mathbf{b}}^T(f) := PC\left\{\left[\mathbf{b}^T(\tau, f)\right]_{\tau|(\tau,f)\in\Omega_2}\right\}. \tag{13}$$

6.4. Experimental protocol

The recovery performance was experimentally assessed on mixtures of two ($N = 2$) audio sources.

6.4.1. Generation of the mixtures.

The flute sound is mixed using a pair of linear instantaneous filters with a known intensity parameter $\theta_1 = 0.2$ radians, and the guitar sound is mixed with filters of sparsity k and length $L = 256$. For each sparsity level k, twenty sets of filters are generated according to the procedure described in Sec. 5.3.2.

6.4.2. Tested blind filter estimation algorithms

For the sake of comparison, two existing methods for filter estimation have been tested in addition to the proposed one.

GCC-PHAT. GCC-PHAT [42] is a method primarily used to estimate the delays associated with a pair of anechoic filters. In our experiments, we also estimated the magnitudes of the peaks by averaging the intensity parameter of all the time-frequency points in the set Ω_1.

Joint-diagonalisation with oracle scaling factors. A well-established method for source separation in a convolutive setting is based on joint-diagonalisation [25, 26], without any exploitation of the sparsity of the mixing filters. Though the joint-diagonalisation algorithms described in the references are primarily targeted for source separation, we tweaked an existing implementation available at ICA Central [43] to obtain the associated frequency-domain estimate of the mixing filters.

The frequency-domain estimates of the filters obtained using joint - diagonalisation naturally suffer from a scaling ambiguity in each frequency bin. Before transforming them into the time-domain, we chose to correct this scaling ambiguity by using the true filters as an oracle.

Let $\check{\mathbf{a}}_i = \{\check{a}_i(f)\}_{f=0}^{F-1},\ 1 \le i \le 2$ be the frequency-domain estimate vectors of the filters corresponding to the source that is mixed using a sparse filter, and let $\widehat{\mathbf{a}}_i = \{\hat{a}_i(f)\}_{f=0}^{F-1},\ 1 \le i \le 2$ be the true filters in the frequency domain. Then the oracle scaling coefficients $\{c(f)\}_{f=0}^{F-1}$ are found by solving

$$c(f) := \arg\min_{c} \sum_{i=1,2} (\hat{a}_i(f) - c \cdot \check{a}_i(f))^2$$

Proposed wideband method. The main steps of the proposed wideband method are given below.

Algorithm 1: wideband filter estimation method (N=2)

1. Estimate intensity parameter θ_1 using e.g. DEMIX or DUET[5];
2. Determine the activity region Ω_1 (cf Eq. (11), $\eta = 0.1$) and its closure $\overline{\Omega}_1$ (cf Eq. (12), $\gamma = 8$);
3. Build the matrix $\mathbf{B}_{\mathtt{wb}}^{\Omega_2}$ as described in Sec. 2.3, with $\Omega_2 = \Omega \backslash \overline{\Omega}_1$;
4. Fuse the rows as $\mathbf{B} = [\check{\mathbf{b}}^T(f)]_f$ (cf Eq. (13));
5. Estimate $\tilde{\mathbf{a}}^2$ as the solution of the ℓ^1 minimisation problem (9) with $\epsilon = 0.02$;
6. Debias the estimate(cf Sec. 3.2).

[5] In the following experiments we actually used the true values of θ_1.

6.5. Results

Fig. 3 shows the performance curves for all considered methods: GCC-PHAT; joint-diagonalisation; wideband method with / without debiasing.

The wideband approach with debiasing significantly outperforms the wideband method without debiasing by often more than 10dB, when $k \leq 16$.

The wideband approach with debiasing also consistently outperforms GCC-PHAT by between 15dB and up to more than 35 dB. The comparison with GCC-PHAT for $k > 2$ is not a surprise since GCC-PHAT was designed only to estimate the delays associated with anechoic filters. However, the wideband approach still outperforms GCC-PHAT for anechoic filters $k = 2$.

Regarding the joint diagonalisation approach, which does not exploit the sparsity of the filters, one can observe that its performance is almost constant irrespective of the sparsity k.

For sufficiently sparse filters (i.e., $k \leq 10$), the wideband cross-relation approach with debiasing outperforms the joint diagonalisation approach by up to 40dB. For less sparse filters, the joint diagonalisation method has similar results as the proposed wideband approach.

One should however remember that in the joint diagonalisation method evaluated here, *the scaling problem was solved using an oracle*, in contrast to the blind nature of the proposed wideband approach. In this light, it is remarkable that the wideband approach with debiasing still outperforms joint-diagonalisation for sufficiently sparse filters.

7. Conclusions and perspectives

This paper focusses on the problem of multiple sparse filter estimation from convolutive mixtures. Traditionally, on the one hand, in Sparse Component Analysis for anechoic source separation, the sources are assumed to be sparse in the time-frequency domain, but methods for filter estimation based this suffer from permutation and scaling ambiguities. In addition they also suffer from the narrowband approximation which becomes critical when the filter lengths become realistic. On the other hand, the time-domain sparsity of the filters is exploited by cross-relation based methods for channel estimation problems, but these are only applicable to mixtures where only a single source contributes.

Existing methods to estimate multiple filters in the multi-source scenario is limited by the assumption of time-domain disjointness of the sources. To this end, the method proposed in this

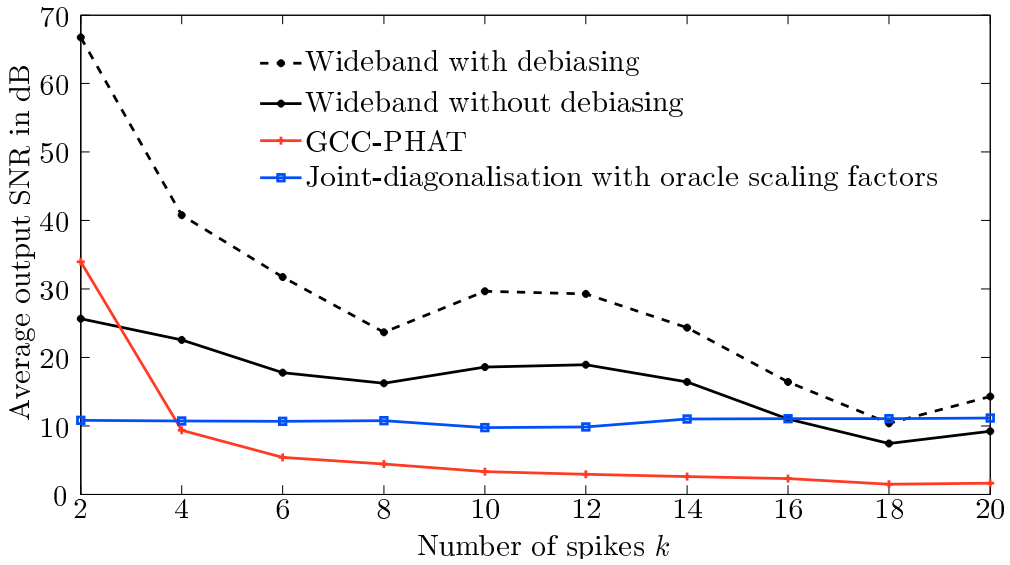


Figure 3: Performance of filter recovery using wideband CR approach for $\theta_1 = 0.2$ radian, in comparison with a joint diagonalisation based approach.

paper is the first attempt to solve the multiple filter estimation problem where sources need not be disjoint in the time domain.

As a main contribution of this paper, we proposed to combine the source sparsity and filter sparsity hypothesis and developed a framework for the blind estimation of multiple sparse filters, based on new wideband time-frequency cross-relations. We proposed to exploit the time-domain sparsity of the filters by formulating a convex optimisation problem based on these cross-relations, after a time-frequency clustering stage which exploits the time-frequency sparsity of the sources. Unlike the classical methods where the filters are estimated in the frequency domain, with scaling and permutation ambiguities, our method estimates the filters directly in the time domain.

We have shown that, under adequate sparsity assumptions, the solution of the newly proposed optimisation problem resolves the ambiguities that arise in similar formulations for blind filter estimation previously proposed. Moreover, while a standard approach in convolutive source separation uses a narrowband approximation to transform the convolutive problem into several complex-valued linear instantaneous problems, we have shown through experiments that, for the considered problem, the proposed wideband formulation of the cross-relation can yield drastic performance improvements.

Our approach was illustrated with experiments in a controlled blind audio source separation

setting, where all but one sources are mixed using instantaneous filters. A truly blind time-frequency clustering method developed to work in this context was shown to outperform existing methods when the time-domain sparsity of the filters is sufficient.

While the work presented in this paper demonstrates the gain that can be achieved by exploiting the time-domain sparsity of the filters as side information, and combining it with the time-frequency disjointness of the sources, it also highlights a number of challenges lying ahead, which will be the object of further work.

A key step of the proposed framework is the blind time-frequency clustering to detect time-frequency regions where the cross-relation associated to a given source is valid. This is a difficult problem in general, though we have been able to solve it blindly so far in a particular setting. An important remark, which can be traced back to the pioneering work of Deville and coworkers [4, 5], may help future work in this direction: we need not find a *partition* of the time-frequency plane into large regions where each source is predominantly active; instead, it is sufficient to find "large enough" time-frequency regions where a given source is "sufficiently visible".

The time-domain filters estimated by our method can be used to estimate the sources by making use of the wideband source estimation method proposed in [31]. This gives a fully wideband framework for source separation.

Further, a better model of temporal sparsity of the filters would take into account echoes that are not aligned with the sampling rate, yielding issues related to subsample precision estimation. For example, one may wish to combine the proposed approach with techniques in the spirit of MUSIC and Finite Rate of Innovation [44] sampling to exploit this type of sparsity. Another possible extension is to seek an overcomplete dictionary in which the filters admit a sparse representation.

Lastly, the success of the approach for sufficiently sparse filters raises several theoretical questions regarding the well-posedness of the filter estimation problem.

Appendix A. Narrowband cross-relation

The formulation of the narrowband cross-relation for multiple sources setting relies on the assumption of the *approximate w-disjoint orthogonality* [3, 45] property of the sources. We assume that at each time-frequency point, there is *at most only one dominant source*

$$\widehat{s}_i(\tau, f)\widehat{s}_j(\tau, f) \approx 0, \quad \forall\, \tau, f,\ i \neq j.$$

Defining Ω_j the time-frequency activity region of source j, i.e., the set of time-frequency locations where source j is dominant, we observe that for any $(\tau, f) \in \Omega_j$ all terms of the right hand side of (6) associated to $k \neq j$ vanish. Moreover, the term associated to $k = j$ also vanishes since $\widehat{a}_{2j}(f)\widehat{a}_{1j}(f) - \widehat{a}_{1j}(f)\widehat{a}_{2j}(f) \approx 0$. Thus, the narrowband cross-relation is still satisfied at certain time-frequency locations due to the time-frequency domain disjointness of the sources. Given the region Ω_j, one builds the matrix $\mathbf{B}^j = \mathbf{B}_{\mathtt{nb}}^{\Omega_j}$ as the restriction of the matrix $\mathbf{B}_{\mathtt{nb}}$ defined in Sec. 2.2 (see also Eq. (A.2)) to the rows indexed by Ω_j.

Let $\widehat{\mathbf{a}}_i = [\widehat{a}_i(f)]_{f=0}^{F-1}$ be the DFT vector corresponding to the time domain filter vector $\mathbf{a}_i$:

$$\widehat{\mathbf{a}}_i = \mathbf{F}^* \cdot \begin{bmatrix} \mathbf{I}_{L\times L} \\ \mathbf{0}_{(F-L)\times L} \end{bmatrix} \cdot \mathbf{a}_i =: \mathbf{F}^*_{F\times L} \cdot \mathbf{a}_i,\ i = 1, 2, \tag{A.1}$$

where $\mathbf{F}^*$ is the forward Fourier matrix of size $F \times F$, $\mathbf{I}_{L\times L}$ is an identity matrix of size $L \times L$ and $\mathbf{0}_{(F-L)\times L}$ is a zero matrix of size $(F - L) \times L$. The matrix $\mathbf{F}^*_{F\times L}$ is associated to zero-padding followed by the forward DFT.

Using (4) and (A.1) the time-frequency domain narrowband cross-relation can be written in matrix form as $\mathbf{B}_{\mathtt{nb}} \cdot \mathbf{a} \approx \mathbf{0}$ where $\mathbf{a} = \begin{bmatrix} \mathbf{a}_1^{\mathbf{T}} & \mathbf{a}_2^{\mathbf{T}} \end{bmatrix}^{\mathbf{T}}$, and $\mathbf{B}_{\mathtt{nb}}$ denotes the matrix

$$\begin{bmatrix} \mathrm{diag}\,(\widehat{\mathbf{x}}_2(\tau_1)) & -\mathrm{diag}\,(\widehat{\mathbf{x}}_1(\tau_1)) \\ \vdots & \vdots \\ \mathrm{diag}\,(\widehat{\mathbf{x}}_2(\tau_Q)) & -\mathrm{diag}\,(\widehat{\mathbf{x}}_1(\tau_Q)) \end{bmatrix} \cdot \begin{bmatrix} \mathbf{F}^*_{F\times L} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}^*_{F\times L} \end{bmatrix} \tag{A.2}$$

with $\mathrm{diag}\,(\widehat{\mathbf{x}}_i(\tau))$ the matrix whose diagonal is the vector $\widehat{\mathbf{x}}_i(\tau) := [\widehat{x}_i(\tau, f)]_{f=0}^{F-1}$, corresponding to the STFT coefficients of the i^{th} mixture at frame τ for all frequencies.

Acknowledgment

The authors acknowledge the support of the EU FP7 FET-Open program, SMALL project, under grant no. 225913.

References

[1] P. Comon, Independent component analysis, a new concept?, Signal Processing 36 (1994) 287–314.

[2] P. Comon, C. Jutten (Eds.), Handbook of Blind Source Separation: Independent Component Analysis and Applications, Academic Press, 2010.

[3] A. Jourjine, S. Rickard, O. Yilmaz, Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 5, pp. 2985 –2988.

[4] F. Abrard, Y. Deville, Blind separation of dependent sources using the "time-frequency ratio of mixtures" approach, in: Proceedings of the Seventh International Symposium on Signal Processing and Its Applications, volume 2, pp. 81 – 84.

[5] Y. Deville, M. Puigt, Temporal and time-frequency correlation-based blind source separation methods. part i: Determined and underdetermined linear instantaneous mixtures, Signal Processing 87 (2007) 374–407.

[6] S. Arberet, R. Gribonval, F. Bimbot, A robust method to count and locate audio sources in a multichannel underdetermined mixture, Signal Processing, IEEE Trans. on 58 (2010) 121 –133.

[7] M. S. Pedersen, J. Larsen, U. Kjems, L. C. Parra, A survey of convolutive blind source separation methods, Springer Verlag, 2007.

[8] H. Sawada, R. Mukai, S. Araki, S. Makino, A robust and precise method for solving the permutation problem of frequency-domain blind source separation, IEEE Trans. on Speech and Audio Processing 12 (2004) 530–538.

[9] N. Mitianoudis, M. Davies, Permutation alignment for frequency domain ICA using subspace beamforming methods, in: Proc. of Independent Component Analysis and Blind Signal Separation, pp. 669–676.

[10] M. Z. Ikram, D. R. Morgan, A beamforming approach to permutation alignment for multichannel frequency-domain blind source separation, in: Proc. of ICASSP, pp. 881–884.

[11] M. Ikram, D. Morgan, Permutation inconsistency in blind speech separation: investigation and solutions, IEEE Transactions on Speech and Audio Processing 13 (2005) 1–13.

[12] P. Smaragdis, Blind separation of convolved mixtures in the frequency domain, Neurocomputing 22 (1998) 21 – 34.

[13] L. Parra, C. Spence, Convolutive blind separation of non-stationary sources, IEEE Transactions on Speech and Audio Processing 8 (2000) 320–327.

[14] N. Murata, S. Ikeda, A. Ziehe, An approach to blind source separation based on temporal structure of speech signals, Neurocomputing 41 (2001).

[15] C. Servière, D. Pham, A novel method for permutation correction in frequency-domain in blind separation of speech mixtures, Independent Component Analysis and Blind Signal Separation (2004) 807–815.

[16] P. Sudhakar, R. Gribonval, A sparsity-based method to solve permutation indeterminacy in frequency-domain convolutive blind source separation, in: Proc. of the International Conference on Independent Component Analysis and Signal Separation, pp. 338–345.

[17] H. Liu, G. Xu, L. Tong, A deterministic approach to blind identification of multi-channel FIR systems, in: Proc. of ICASSP, Washington, DC, USA, pp. 581–584.

[18] E. Moulines, P. Duhamel, J.-F. Cardoso, S. Mayrargue, Subspace methods for the blind identification of multichannel FIR filters, Signal Processing, IEEE Trans. on 43 (1995) 516 –525.

[19] G. Xu, H. Liu, L. Tong, T. Kailath, A least-squares approach to blind channel identification, IEEE Transactions on Signal Processing 43 (1995) 2982–2993.

[20] C. Avendano, J. Benesty, D. R. Morgan, A least squares component normalization approach to blind channel identification, in: Proc. of the ICASSP, Washington, DC, USA, pp. 1797–1800.

[21] Y. A. Huang, J. Benesty, Adaptive multi-channel least mean square and newton algorithms for blind channel identification, Signal Processing 82 (2002) 1127–1138.

[22] A. Aïssa-El-Bey, K. Abed-Meraim, Blind SIMO channel identification using a sparsity criterion, in: Proc. of SPAWC, pp. 271 – 275.

[23] O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking, IEEE Trans. on Sig. Proc. 52 (2004) 1830–1847.

[24] H. Sawada, S. Araki, R. Mukai, S. Makino, Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation, EEE Trans. on ALSP 15 (2007) 1592 –1604.

[25] H. Boumaraf, D. T. Pham, C. Servière, Blind separation of convolutive mixture of speech signals, in: Proc. of EUSIPCO.

[26] D. T. Pham, C. Servière, H. Boumaraf, Blind separation of convolutive audio mixtures using nonstationarity, in: Proceedings of ICA, pp. 975–980.

[27] V. Reju, S. N. Koh, I. Y. Soon, Underdetermined convolutive blind source separation via time-frequency masking, IEEE Transactions on ALSP 18 (2010) 101 –116.

[28] H. Sawada, S. Araki, S. Makino, Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment, IEEE Trans. on ALSP 19 (2011) 516–527.

[29] N. Duong, E. Vincent, R. Gribonval, Under-determined reverberant audio source separation using a full-rank spatial covariance model, IEEE Trans. on ALSP 18 (2010) 1830–1840.

[30] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, P. Vandergheynst, Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation, in: Proc. of ISSPA, pp. 1–4.

[31] M. Kowalski, E. Vincent, R. Gribonval, Beyond the narrowband approximation: wideband convex methods for under-determined reverberant audio source separation, IEEE Trans. on ALSP 18 (2010) 1818–1829.

[32] M. Kocic, D. Brady, M. Stojanovic, Sparse equalization for real-time digital underwater acoustic communications, in: OCEANS '95. MTS/IEEE. Challenges of Our Changing Global Environment. Conference Proceedings., volume 3, pp. 1417 –1422 vol.3.

[33] W. Li, J. Preisig, Estimation of rapidly time-varying sparse channels, IEEE Journal of Oceanic Engg. 32 (2007) 927–939.

[34] S. Ariyavisitakul, N. Sollenberger, L. Greenstein, Tap-selectable decision-feedback equalization, Communications, IEEE Transactions on 45 (1997) 1497 –1500.

[35] S. Kim, R. Iltis, A matching-pursuit/gsic-based algorithm for ds-cdma sparse-channel estimation, Signal Processing Letters, IEEE 11 (2004) 12–15.

[36] N. R. Chapman, I. Barrodale, Deconvolution of marine seismic data using the ℓ^1 norm, Geophysical Journal International 72 (1983) 93–100.

[37] A. Aïssa-El-Bey, K. Abed-Meraim, Y. Grenier, Blind separation of underdetermined convolutive mixtures using their time - frequency representation, IEEE Transactions on ALSP 15 (2007) 1540 –1550.

[38] S. Mallat, A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way, Academic Press, 2008.

[39] P. Sudhakar, S. Arberet, R. Gribonval, Double sparsity: Towards blind estimation of multiple channels, in: LVA/ICA, pp. 571–578.

[40] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 1.21, `http://cvxr.com/cvx`, 2011.

[41] Y. Hua, M. Wax, Strict identifiability of multiple fir channels driven by an unknown arbitrary sequence, Signal Processing, IEEE Transactions on 44 (1996) 756 –759.

[42] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. on Acoust., Speech and Signal Proc. 24 (1976) 320 – 327.

[43] ICA Central, http://www.tsi.enst.fr/icacentral/algos.html, 2003.

[44] Y. Barbotin, A. Hormati, S. Rangan, M. Vetterli, Sampling of Sparse Channels with Common Support, in: Proc. of SAMPTA.

[45] O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking, IEEE Transactions on Signal Processing 52 (2004) 1830 – 1847.