Authors' final version of a paper published in "Signal Processing"

Paper reference:
Y. Deville, M. Puigt, "Temporal and time-frequency correlation-based blind source separation methods. Part I: determined and underdetermined linear instantaneous mixtures", Signal Processing, vol. 87, no. 3, pp. 374-407, March 2007.

Elsevier on-line version:
http://dx.doi.org/10.1016/j.sigpro.2006.05.012

Temporal and time-frequency correlation-based blind source separation methods. Part I: determined and underdetermined linear instantaneous mixtures

Yannick DEVILLE[1] and Matthieu PUIGT

Laboratoire d'Astrophysique de Toulouse-Tarbes
Observatoire Midi-Pyrénées - Université Paul Sabatier Toulouse 3
14 Av. Edouard Belin, 31400 Toulouse, France

ydeville@ast.obs-mip.fr , mpuigt@ast.obs-mip.fr

We propose two types of correlation-based blind source separation (BSS) methods, i.e. a time-domain approach and extensions which use time-frequency (TF) signal representations and thus apply to much more general conditions. Our basic TF methods only require each source to be isolated in a tiny TF area, i.e. they set very limited constraints on the source sparsity and overlap, unlike various previously reported TF-BSS methods. Our approaches consist in identifying the columns of the (scaled permuted) mixing matrix in TF areas where these methods detect that a source is isolated. Both the detection and identification stages of these approaches use local correlation parameters of the TF transforms of the observed signals. Two such Linear Instantaneous TIme-Frequency CORRelation-based BSS methods are proposed, using Centered or Non-Centered TF transforms. These methods, which are resp. called LI-TIFCORR-C and LI-TIFCORR-NC, are especially suited to non-stationary sources. We derive their performance from many tests performed with mixtures of speech signals. This demonstrates that their output SIRs have a low sensitivity to the values of their TF parameters and are quite high, i.e. typically 60 to 80 dB, while the SIRs of all tested classical methods range about from 0 to 40 dB. We also extend these approaches to achieve partial BSS for underdetermined mixtures and to operate when some sources are not isolated in any TF area.

Keywords:
blind source separation,
correlation,
linear instantaneous mixtures,
non-stationarity,
partial source separation,
short-time Fourier transform,
sparsity,
speech,
time-frequency analysis,
underdetermined mixtures.

[1]Corresponding author.

1 Introduction

Blind source separation (BSS) methods aim at restoring a set of unknown source signals from a set of observed signals which are mixtures of these source signals [1]-[3]. Most of the approaches that have been developed to this end concern linear instantaneous mixtures and are based on Independent Component Analysis (ICA). They assume the sources to be random stationary statistically independent signals, and they recombine the observed signals so as to obtain statistically independent output signals. The latter signals are then equal to the sources, up to some indeterminacies and under some conditions (especially, at most one source may be Gaussian for such methods to be applicable if no additional constraints are set on the sources).

In addition to ICA, a few other general concepts have been used for achieving BSS. This includes the class of approaches based on time-frequency (TF) analysis, which is the main framework considered in this paper. TF tools have been used in various ways in the BSS methods reported so far, as may be seen e.g. in [4]-[11]. Among the trends which emerge from these papers, the following ones should especially be mentioned. A first set of methods is composed of approaches based on ratios of TF transforms of observed signals [4]-[6]. Some of these methods require the sources to have no overlap in the TF domain [4], which is quite restrictive. On the contrary, only slight differences in the TF representations of the sources are required by the type of methods that we introduced and extended in [5]-[6]. Another general concept that has been proposed for achieving BSS consists in exploiting the sparsity of the sources in an adequate representation of these signals. The representation used in some of these approaches is based on a TF transform of the signals. This yields a second set of time-frequency BSS methods, which especially contains the approaches proposed in [7]. This second set of methods has some relationships with the above-mentioned first set, in the sense that the different constraints on the TF overlap between sources in the first set of methods may be considered as various conditions on the degree of sparsity of these transformed sources. All the approaches presented in the papers [4]- [7] which compose the above two sets use the same TF transform, i.e. the Short-Time Fourier Transform (STFT), which is a linear transform. On the contrary, other approaches use quadratic TF transforms (see e.g. [8]-[11]), thus forming a third set of methods. This set especially includes time-frequency BSS methods which are significantly related to classical BSS approaches, as they consist of TF adaptations of previously developed joint-diagonalization methods, with subsequent modifications. It should be noted that, unlike classical ICA-based BSS methods, TF-based BSS approaches are intrinsically well-suited to non-stationary signals (and set no restrictions on the gaussianity of the sources). They are therefore e.g. especially attractive for speech sources.

This first part of our paper mainly describes original linear time-frequency BSS approaches applicable to linear instantaneous mixtures. These approaches use STFTs, like the above-mentioned two sets of linear methods, but they rely on other types of parameters, which are based on the local correlations of the observed signals in the TF domain. Before describing these time-frequency BSS methods, we present a purely temporal version of such approaches, which only applies to more restrictive conditions, as shown below.

The remainder of this first part of our paper is therefore organized as follows. In Section 2 we define the first configuration, based on determined linear instantaneous mixtures, that we consider and the resulting goal of our investigation. We then present the associated temporal BSS method in Section 3 and we introduce its TF extensions in Section 4. Section 5 is devoted to the versions of these approaches intended for more complex configurations,

especially underdetermined mixtures. Section 6 reports on a detailed analysis of the experimental performance achieved by all our temporal and TF approaches for artificial mixtures of real speech sources. It also contains a comparison to the performance of various BSS methods from the literature and of a somewhat related TF approach that we proposed in a previous paper. Section 7 contains a discussion of the features of the proposed methods, as compared to classical BSS approaches. This section also presents the conclusions drawn from this first part of our overall investigation and outline extensions of the proposed methods. In addition, specific topics are detailed in the appendices.

2 Problem statement

Let us first introduce the features of the considered configuration which apply to the BSS approaches proposed in Sections 3 and 4. We assume that N unknown, possibly complex-valued, source signals $s_j(t)$ are mixed in a linear instantaneous way, thus providing a set of N observed signals $x_i(t)$. In other words, as in most papers dealing with BSS, we consider determined mixtures, i.e. the number of sources is here assumed to be known and equal to the number of available observations (the case when they are different will then be addressed in Section 5). In various applications, these N observations $x_i(t)$ are respectively provided by N sensors. The source-observation relationship reads in matrix form

$$x(t) = As(t), \tag{1}$$

where $s(t) = [s_1(t) \dots s_N(t)]^T$ and $x(t) = [x_1(t) \dots x_N(t)]^T$ and where A is a NxN unknown, supposedly constant and invertible, mixing matrix. Its coefficients a_{ij} may be complex-valued and are assumed to be nonzero.

BSS would then ideally consist in deriving an estimate $\hat{A}$ of A, so as to then determine the output vector

$$\begin{aligned} y(t) &= \hat{A}^{-1}x(t) &(2)\\ &= \hat{A}^{-1}As(t). &(3)\end{aligned}$$

Each component $y_j(t)$ of this vector $y(t)$ would then be equal to the source signal having the same index, i.e. to $s_j(t)$ (up to estimation errors). It is well known however that this can only be achieved up to two types of indeterminacies, which resp. concern the order and scale factors with which the source signals appear in the output vector $y(t)$. We now provide a specific description of this phenomenon, which is intended for the BSS methods that we will introduce in the next sections. Any of the mixed signals corresponding to the matrix form (1) reads explicitly

$$x_i(t) = \sum_{j=1}^{N} a_{ij}s_j(t) \qquad\qquad i = 1 \dots N. \tag{4}$$

However, it may also be expressed in a different way, by applying two transforms to it. The first one consists in changing the order in which the terms, resp. associated to sources $s_1(t) \dots s_N(t)$, appear in the sum in (4). This is achieved by applying an arbitrary

permutation $\sigma(.)$ to the indices j in this sum[2]. The above mixed signal then reads

$$x_i(t) = \sum_{j=1}^{N} a_{i,\sigma(j)} s_{\sigma(j)}(t) \qquad\qquad i = 1 \dots N. \tag{5}$$

The second transform concerns the scales of the considered signals. It consists in expressing the contributions of each permuted source signal $s_{\sigma(j)}(t)$ with respect to the contribution of this signal in the first mixed signal[3]. The latter contribution is equal to $a_{1,\sigma(j)} s_{\sigma(j)}(t)$, so that we rewrite (5) as

$$x_i(t) = \sum_{j=1}^{N} \frac{a_{i,\sigma(j)}}{a_{1,\sigma(j)}} a_{1,\sigma(j)} s_{\sigma(j)}(t) \qquad i = 1 \dots N. \tag{6}$$

We therefore introduce the notations

$$s'_j(t) = a_{1,\sigma(j)} s_{\sigma(j)}(t) \qquad j = 1 \dots N \tag{7}$$

$$b_{ij} = \frac{a_{i,\sigma(j)}}{a_{1,\sigma(j)}} \qquad i, j = 1 \dots N, \tag{8}$$

where $s'_j(t)$ are the scaled permuted source signals and b_{ij} are the corresponding scaled permuted mixing coefficients. Eq. (6) then reads

$$x_i(t) = \sum_{j=1}^{N} b_{ij} s'_j(t) \qquad\qquad i = 1 \dots N, \tag{9}$$

or in matrix form

$$x(t) = B s'(t), \tag{10}$$

where $s'(t) = [s'_1(t) \dots s'_N(t)]^T$ and the scaled permuted mixing matrix B consists of the above coefficients b_{ij}. Note that all the coefficients b_{1j}, with $j = 1 \dots N$, are equal to 1 due to (8). These coefficients form the first row of B.

The mixing equation (10) thus obtained is the same as the initial mixture expression (1), except that the mixed signals are now expressed with respect to the scaled permuted source signals $s'_j(t)$. The discussion at the beginning of this section may then be reinterpreted as follows: assume that we succeed in deriving an estimate $\hat{B}$ of B. Then, by computing the output vector

$$y'(t) = \hat{B}^{-1} x(t) \tag{11}$$

$$= \hat{B}^{-1} B s'(t), \tag{12}$$

all components $y'_j(t)$ of this vector are resp. equal to $s'_j(t)$ (up to estimation errors), i.e. to the contributions of the permuted sources in the first mixed signal.

The BSS problem may therefore be solved by first estimating the above matrix B (with an arbitrary permutation function $\sigma(.)$) and then computing the corresponding vector $y'(t)$ of separated source signals. Two types of BSS methods based on this principle are resp. proposed in Sections 3 and 4.

[2]This includes, as a specific case, the situation when $\sigma(.)$ is the identity function, i.e. when no permutation is actually performed.

[3]The same principle may of course be applied to any other mixed signal instead.

3 Proposed standard temporal approach

3.1 Assumptions and definitions

In this section, we consider random signals and introduce a statistical temporal BSS approach. We first present the only assumptions that we make with respect to the sources in this approach, and the associated definitions.

Definition 1: a source is said to be "isolated" in a time area if only this source (among all considered mixed sources) has a nonzero variance in this time area.

This definition corresponds to the theoretical point of view. From a practical point of view, this means that the variances of all other sources are negligible with respect to the variance of the source that is isolated. Each considered time area may be restricted to a single time t from a theoretical point of view, as we use a statistical approach in this section. In practice, the statistical parameters of the signals are estimated over time windows, and each considered time area then consists of such a window.

Definition 2: a source is said to be "accessible" in the time domain if there exist at least one time area where it is isolated.

Assumption 1: each source is accessible in the time domain.

Assumption 1 has the following consequence on the statistical properties of the random source signals to be processed by the BSS method introduced below. Consider the (theoretical) variance at time t of any source $s_j(t)$, i.e.

$$\sigma^2_{s_j}(t) = E\{|s_j(t) - E\{s_j(t)\}|^2\} \tag{13}$$

where $E\{.\}$ stands for expectation. Due to Assumption 1, $\sigma^2_{s_j}(t)$ is equal to zero at some times t, i.e. times when *another* source is isolated. Moreover, it is non-zero for other times, otherwise the signal $s_j(t)$ would have zero variance at any time, which is excluded from the considered BSS configurations. Therefore, the statistical parameter $\sigma^2_{s_j}(t)$ of any source signal $s_j(t)$ is time-dependent, so that these source signals are non-stationary[4].

The condition in Assumption 1 and the resulting non-stationarity requirements may be considered as quite restrictive. It should be clear that this only corresponds to the first stage of our overall approach, where this assumption makes it possible to design a simple BSS method in this specific situation. In the TF extension of our approach presented in Section 4, we will show how to replace the above assumption by a much less stringent constraint, thus avoiding non-stationarity requirements.

Assumption 2: the sources are mutually uncorrelated.

Note that we do not require the complete independence of the sources, as the approach introduced below only uses their second-order statistics.

For the sake of simplicity, the notations $s(t)$ and $x(t)$ introduced above directly refer to the centered[5] version of the signals in this section.

[4]The above comments correspond to the theoretical definition of the stationarity of random signals, based on their statistical properties, which are especially defined by the expected values associated to these signals. Now consider practical implementations of the proposed approach, based on a single realization of these random processes. The above discussion then results in the following practical requirements. The source signals are requested to be long-term non-stationary, but they should also be short-term stationary in order to make it possible to estimate correctly the statistical parameters of these signals over short time windows of the available realization. The above constraints on sources should then be interpreted accordingly. Especially, the practical version of Definition 1 means that a source is isolated in a short time window if the *sample* variances of all other sources, obtained by time averaging over this window, are zero (or negligible).

[5]Centering is here handled as usually in BSS investigations, i.e: i) first, in the theoretical analysis

3.2 Temporal BSS method

In this section, we derive a BSS method by taking advantage of the above *Assumption 1*, i.e. of the fact that there exist time areas where each source is isolated. Such areas should first be detected, so as to operate inside them. As all observed signals are proportional in any such area, an appealing approach for detecting these areas consists in checking the cross-correlation coefficients $\rho_{x_1x_i}(t)$ between the observed signals $x_1(t)$ and $x_i(t)$, which are defined as[6]

$$\rho_{x_1x_i}(t) = \frac{E\{x_1(t)x_i^*(t)\}}{\sqrt{E\{x_1(t)x_1^*(t)\}E\{x_i(t)x_i^*(t)\}}} \qquad \forall\ i, \quad 2 \leq i \leq N, \tag{14}$$

where the superscript * denotes the complex conjugate. More precisely, we show in Appendix A that a necessary and sufficient condition for a source to be isolated at time t is

$$|\rho_{x_1x_i}(t)| = 1 \qquad \forall\ i, \quad 2 \leq i \leq N. \tag{15}$$

Now consider such an area where a source is isolated, say $s_k(t)$. The observed signals (4) then become restricted to

$$x_i(t) = a_{ik}s_k(t) \qquad i = 1 \dots N. \tag{16}$$

Other correlation-based parameters associated to these observed signals then make it possible to identify part of the matrix B. More precisely, when (16) is met,

$$\frac{E\{x_i(t)x_1^*(t)\}}{E\{x_1(t)x_1^*(t)\}} = \frac{a_{ik}}{a_{1k}} \qquad i = 2 \dots N. \tag{17}$$

Comparing this expression to (8) shows that the set of such correlation parameters associated to the same single-source time area and to all observations indexed by i identifies one of the columns of B (the first row of B consists of 1, as explained above). By repeatedly performing such column identifications for time areas associated to all sources, we eventually identify the overall matrix B, which completes the proposed approach.

The BSS method thus introduced may be summarized as follows (see corresponding pseudo-code in Fig. 1). It contains 3 stages:

1. The detection stage consists in detecting the time areas where a source is isolated. This stage therefore operates as follows. For each time t (or practical time window), we compute the mean[7] $\overline{|\rho_{x_1x_i}(t)|}$ of $|\rho_{x_1x_i}(t)|$ over all i, with $2 \leq i \leq N$. We then order all times t according to decreasing values of $\overline{|\rho_{x_1x_i}(t)|}$. The first times in this ordered list are then considered as the "best" single-source time areas.

which corresponds to the current section, either the sources are initially assumed to be zero-mean, or their means are assumed to be known and the corresponding centered versions of the signals are considered, and then ii) in practice, centering is performed independently for each considered time window, where the considered signals are assumed to be short-term stationary, by subtracting from the signals the mean estimates computed on these time windows.

[6]Again, in practical implementations of this BSS method, these cross-correlation coefficients are estimated by replacing in (14) the expectations $E\{.\}$ by temporal averages over the considered windows, for the considered signal realization, based on ergodicity assumptions.

[7]Appendix B explains why the mean of $|\rho_{x_1x_i}(t)|$ over i is preferred to its minimum in this approach and in its subsequent TF extensions.

2. The identification stage consists in identifying the columns of B. This is achieved by successively using as follows each of the first and subsequent time areas in the above-defined ordered list. The identification parameters on the left-hand side of (17) yield an estimated column of B. This column is kept only if its distance with respect to all previously identified columns is above a user-defined threshold, showing that the considered time area does not contain the same source as previous areas in the ordered list. The identification procedure ends when the number of columns of B thus kept becomes equal to the specified number N of sources to be separated (this is theoretically guaranteed to occur because all sources are assumed to be accessible in the considered data).

3. The combination stage consists in recombining the mixed signals according to (11), in order to obtain the extracted source signals.

This approach for Linear Instantaneous mixtures is thus a TEMPoral CORRelation-based BSS method, using the Centered version of the signals[8]. We therefore call it "LI-TEMPCORR-C". This method is the basic version of the kind of temporal approaches that we propose in this paper. Various modified versions may then be derived from it. Especially, a possibly more robust version of its identification stage consists in first keeping the identified columns corresponding to *all* available single-source areas, and then clustering them, so as to eventually derive a better estimate of each column of B (e.g. as an average of all its identified occurences). Such clustering-based extensions of our approach may also be used to remove heuristic aspects of its basic version, especially by avoiding the user-defined distance threshold involved in the above identification stage.

4 Proposed standard time-frequency approaches

4.1 Motivations and basic principles

The approach that we introduced in the previous section is attractive because of its simplicity. It may be considered to be of limited practical applicability however, because it assumes all sources to be isolated in associated time areas, which is a restrictive condition. But it opens the way to much more powerful methods if we now take into account the TF distributions of the signals, instead of their plain time distributions considered up to this point. Indeed, the TF extension of the above approach may be briefly defined as follows. Assume that each source is isolated in a TF area. Based on the above presentation, one may then expect the columns of B to be identifiable in such areas, thus allowing one to eventually perform BSS.

The remainder of this section presents this approach in a more formal way. We would like to stress that the resulting methods only request each source to be isolated in a very small bounded TF area, e.g. corresponding to a very limited set of adjacent time windows and one associated frequency. In other words, the proposed time-frequency BSS methods then only request that, for each source, there exist (at least) one very limited set of adjacent time windows and one associated frequency where all other sources are inactive. On the contrary, our temporal approach is based on Assumption 1, so that it requires that, for each source, there exist a time window where all other sources are inactive at *all frequencies*, which is a much more restrictive requirement. This is the reason why

[8]The influence of observation noise on the method obtained at this stage is presented in Appendix C.

the TF versions of the proposed approach have a much broader scope of application than its above temporal version. For instance, mixtures of continuous speech signals meet the assumption considered in this section, because most of the energy of these signals over successive time windows is concentrated in a few bounded time-varying frequency regions, corresponding to formants.

It should also be noted that the considered TF tool, which will now be presented, is originally defined for deterministic signals. This paper uses it in such a framework and concerns two cases i.e: i) either the sources to be separated are deterministic or ii) they are random processes, but in the latter case only a single realization of these processes is considered (this is what is actually available in practice). The following description then concerns this single, deterministic, realization and the tools and properties that we use only require such a realization.

4.2 Time-frequency tool

Many TF representations have been proposed over the last fifty years and are e.g. presented in [12]-[13]. We here use the short-time Fourier transform (STFT), especially because it yields low computional load thanks to FFT algorithms and it does not introduce interference terms (unlike some other TF transforms, such as the Wigner-Ville distribution), thus keeping the linear instantaneous mixing structure when applied to the considered observed signals.

The STFT of a time-domain signal $v(t')$ is obtained by first multiplying that signal by a shifted windowing function $h^*(t'-t)$, centered around time t. This yields the windowed signal $v(t')h^*(t'-t)$. This signal depends on two time variables, i.e. the selected time t where the local spectrum of $v(t')$ is analyzed and the varying time t'. The STFT of $v(t')$ is then defined as the Fourier transform of the above windowed signal, i.e

$$V(t,\omega) = \int_{-\infty}^{+\infty} v(t')h^*(t'-t)e^{-j\omega t'}dt'. \tag{18}$$

$V(t,\omega)$ is then the contribution of the considered signal $v(t')$ in the part of the TF plane corresponding to: i) the short time window centered around t and ii) the angular frequency ω.

4.3 Assumptions and definitions

The first TF method introduced below in Section 4.4 uses the same assumptions and definitions as in Section 3.1, except that: i) we here eventually use a deterministic framework, as explained above, and ii) the temporal concepts considered in Section 3.1 are here replaced by their TF version, i.e:

Definition 1-TF: a source is said to be "isolated" in a TF area if only this source (among all considered mixed sources) has a nonzero variance in this TF area.

Definition 2-TF: a source is said to be "accessible" in the TF domain if there exist at least one TF area where it is isolated.

Assumption 1-TF: each source is accessible in the TF domain.

The TF areas considered in the following practical time-frequency BSS methods are "analysis zones" defined as follows. Each value of a STFT corresponds to a single "TF point", associated to a single angular frequency ω and to a complete time window defined by the selected analysis time t and by the considered finite-length windowing function

$h^*(.)$, as explained in Section 4.2. The BSS methods that we propose below then use means associated to these STFTs, computed over "analysis zones" which consist of TF points. An analysis zone may have any shape in the TF domain. In this first part of our paper, we focus on the case when it forms a "temporal line", i.e. when all its points correspond to the same frequency ω and to adjacent (possibly overlapping) time windows. The latter windows are resp. associated to a discrete set of L analysis times t_p, with $p = 1 \dots L$. This set is denoted as T hereafter. Each analysis zone is then specified in terms of the couple (T, ω), which completely defines the part of the TF domain associated to this analysis zone.

TF analysis is of interest for "non-stationary" signals, i.e. signals whose frequency contents vary over successive time windows. The time-frequency BSS methods presented below apply to such signals, but we would like to stress that they are also suited to various stationary signals (unlike our previous temporal approach). Indeed, assume that the frequency contents of the considered source signals are constant over time, but that these signals do not fully overlap in the frequency domain, so that each of them is isolated at one frequency at least. Then our major assumption, i.e. *Assumption 1-TF*, is met so that the time-frequency BSS methods introduced below apply (provided their other assumptions are also met).

The BSS method presented below in Section 4.4 then uses the following parameters, associated to the above-defined analysis zones. For any time-domain signal $v(t')$, the mean of its TF transform $V(t, \omega)$ over the considered analysis zone is

$$\overline{V}(T, \omega) = \frac{1}{L} \sum_{p=1}^{L} V(t_p, \omega). \tag{19}$$

Similarly, for any couple of signals $v_1(t')$ and $v_2(t')$, whose TF transforms are denoted $V_1(t, \omega)$ and $V_2(t, \omega)$, the cross-correlation of the centered versions of the TF transforms of these signals over the considered analysis zone may be measured: i) either by the TF local non-normalized covariance parameter

$$C_{v_1 v_2}(T, \omega) = \frac{1}{L} \sum_{p=1}^{L} [V_1(t_p, \omega) - \overline{V_1}(T, \omega)][V_2(t_p, \omega) - \overline{V_2}(T, \omega)]^* \tag{20}$$

or ii) by the corresponding covariance coefficient

$$c_{v_1 v_2}(T, \omega) = \frac{C_{v_1 v_2}(T, \omega)}{\sqrt{C_{v_1 v_1}(T, \omega) C_{v_2 v_2}(T, \omega)}}. \tag{21}$$

The source uncorrelation assumption of Section 3 is then replaced by:

Assumption 2-TF: over each analysis zone (T, ω), the (centered) TF transforms of the sources are uncorrelated, i.e: $C_{s_i s_j}(T, \omega) = 0, \ \forall \ i \neq j$.

4.4 Time-frequency BSS method: centered version

The BSS method which results from the above principles is a TF adaptation of the overall temporal approach that we presented in Section 3.2. It is therefore composed of the same 3 stages as the latter approach, adapted to the TF context however, and therefore preceded by a pre-processing stage, i.e:

1. The pre-processing stage consists in deriving the STFTs $X_i(t,\omega)$ of the mixed signals, according to (18).

2. The detection stage aims at finding single-source TF analysis zones. It is based on the following property, shown in Appendix A: a necessary and sufficient condition for a source to be isolated in a TF analysis zone (T,ω) is

$$|c_{x_1x_i}(T,\omega)| = 1 \qquad \forall\ i, \quad 2 \leq i \leq N. \tag{22}$$

   This stage therefore operates as follows. For each analysis zone, we compute the mean $\overline{|c_{x_1x_i}(T,\omega)|}$ of $|c_{x_1x_i}(T,\omega)|$ over all i, with $2 \leq i \leq N$. We then order all analysis zones according to decreasing values of $\overline{|c_{x_1x_i}(T,\omega)|}$. The first zones in this ordered list are then considered as the "best" single-source zones[9].

3. The identification stage consists in identifying the columns of B in single-source analysis zones. This is based on the same approach as in the temporal BSS method, except that the identification parameter in (17) is here replaced by

$$I_i(T,\omega) = \frac{C_{x_ix_1}(T,\omega)}{C_{x_1x_1}(T,\omega)} \qquad i = 2 \ldots N, \tag{23}$$

   which is equal to a_{ik}/a_{1k} when source $s_k(t)$ is isolated in the considered TF analysis zone, as shown in Appendix E.

4. In the combination stage, we eventually recombine the mixed signals according to (11), in order to obtain the extracted source signals.

This approach for Linear Instantaneous mixtures is therefore a TIme-Frequency CORRelation-based BSS method, which uses the Centered version of the STFTs of the signals. We therefore call it "LI-TIFCORR-C" in the remainder of this paper. This approach may then be extended in the same way as the temporal version of this method, e.g. using clustering techniques, as explained in Section 3.2.

4.5 Time-frequency BSS method: non-centered version

The extension of our initial temporal BSS method to TF concepts led in a natural way to the above LI-TIFCORR-C method. A slightly simpler version of the latter method may also be obtained by replacing the centered versions of the parameters by plain, i.e. non-centered, parameters in all the assumptions and definitions of Section 4.3 and in the description of this BSS method provided in Section 4.4. Especially, the covariance parameter $C_{v_1v_2}(T,\omega)$ defined in (20) is here replaced by the TF local non-centered non-normalized correlation parameter

$$R_{v_1v_2}(T,\omega) = \frac{1}{L}\sum_{p=1}^{L} V_1(t_p,\omega)V_2^*(t_p,\omega), \tag{24}$$

so that the covariance coefficient $c_{v_1v_2}(T,\omega)$ defined in (21) is replaced by the correlation coefficient

$$r_{v_1v_2}(T,\omega) = \frac{R_{v_1v_2}(T,\omega)}{\sqrt{R_{v_1v_1}(T,\omega)R_{v_2v_2}(T,\omega)}}. \tag{25}$$

[9] Modified versions of this approach are defined in Appendix D.

This type of parameter is used instead of $c_{v_1 v_2}(T,\omega)$ in the detection stage defined in Section 4.4, especially in its condition (22) used for detecting single-source analysis zones (the validity of the resulting condition may be shown by means of the non-centered version of the proof provided in Appendix A). Similarly, in the identification stage defined in Section 4.4, the non-normalized covariance parameters used in (23) are here replaced by their non-centered version, so that the identification parameter here becomes

$$I_i(T,\omega) = \frac{R_{x_i x_1}(T,\omega)}{R_{x_1 x_1}(T,\omega)} \qquad\qquad i = 2 \dots N. \tag{26}$$

This parameter is again equal to a_{ik}/a_{1k} when source $s_k(t)$ is isolated in the considered TF analysis zone (this may be shown by means of the non-centered version of the proof provided in Appendix E.). This method for Linear Instantaneous mixtures, based on the CORRelation of Non-Centered TIme-Frequency transforms, is called "LI-TIFCORR-NC" hereafter.

4.6 Relationship between the LI-TIFCORR and LI-TIFROM methods

4.6.1 Principles of LI-TIFROM

The overall structure used in both versions of the LI-TIFCORR approach has some similarities with the LI-TIFROM time-frequency BSS approach that we proposed in [5]-[6], i.e. it is composed of the same stages. However, these two types of approaches use completely different parameters in the two stages which are the core of such methods, i.e. the detection and identification stages. More precisely, the LI-TIFROM approach operates as follows:

1. Its pre-processing stage consists in deriving the STFTs $X_i(t,\omega)$ of the mixed signals, according to (18).

2. The detection stage then derives single-source analysis zones as the zones where the variances of ratios of STFTs of observations take the lowest values. More precisely, for each TF point (t,ω), we first compute the ratio of observations[10]

$$\beta(t,\omega) = \frac{X_2(t,\omega)}{X_1(t,\omega)}. \tag{27}$$

   We then derive the mean and variance of $\beta(t,\omega)$ over each analysis zone, i.e.

$$\overline{\beta}(T,\omega) = \frac{1}{L}\sum_{p=1}^{L} \beta(t_p,\omega) \tag{28}$$

$$var[\beta](T,\omega) = \frac{1}{L}\sum_{p=1}^{L} |\beta(t_p,\omega) - \overline{\beta}(T,\omega)|^2. \tag{29}$$

   We then order all analysis zones according to increasing values of $var[\beta](T,\omega)$. The first zones in this ordered list are then considered as the "best" single-source zones.

[10] We may use either the ratio defined in (27) or its inverse. We here consider the ratio in (27) because this yields the version of LI-TIFROM which is the most similar to the above description of LI-TIFCORR. The inverse ratio is used in [5]-[6]. In [6], we also explain why a simple version of the detection stage, which only uses a single couple of observations, is most often acceptable (although it may not be optimal).

3. The identification stage consists in identifying the columns of B in the first single-source analysis zones of the above list. The entries of B are set to the means on these analysis zones of ratios $X_i(t,\omega)/X_1(t,\omega)$ of STFTs of observations, where these means are defined in the same way as in (28). The same method as in Subsection 3.2 is used for deciding which of these identified columns of B are kept.

4. In the combination stage, we eventually recombine the mixed signals according to (11), in order to obtain the extracted source signals.

Comparing the steps of this method to those of LI-TIFCORR-C, which were detailed in Subsection 4.4, therefore confirms that the main algorithms used in these two types of methods are based on completely different parameters.

As the detection and identification stages are independent one from the other in any of the LI-TIFCORR and LI-TIFROM approaches, we may also derive mixed approaches by using the detection method of one of these two types of BSS approaches and the identification stage of the other type of approaches.

4.6.2 Limitation of LI-TIFROM

The detailed tests that we performed to compare the performance of the LI-TIFCORR and LI-TIFROM methods revealed a limitation of the latter approach. We here provide a theorerical analysis of this phenomenon, while the corresponding experimental results are presented below in Section 6.5. For the sake of clarity, we first consider a configuration involving 2 mixtures of 2 sources, with a symmetrical mixing matrix

$$A = \begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix}, \tag{30}$$

where p is a mixing parameter, whose influence is analyzed hereafter. The parameter of LI-TIFROM for detecting single-source analysis zones, that we defined in (27), here reads

$$\beta(t,\omega) = \frac{pS_1(t,\omega) + S_2(t,\omega)}{S_1(t,\omega) + pS_2(t,\omega)}. \tag{31}$$

First consider an *ideal* analysis zone, i.e. a TF zone where either $S_1(t,\omega)$ or $S_2(t,\omega)$ is strictly zero everywhere. Eq. (31) then shows that $\beta(t,\omega)$ is constant in this zone. Its variance over this zone is then strictly zero and this is precisely the property which is used in LI-TIFROM to detect single-source analysis zones. For *ideal* analysis zones, LI-TIFROM therefore yields the same behavior for the two sources, $S_1(t,\omega)$ and $S_2(t,\omega)$. But, let us now consider a *real* single-source analysis zone associated to a source $S_i(t,\omega)$, i.e. a zone where $S_i(t,\omega)$ is prominent but which also contains slight pollution from the other source $S_j(t,\omega)$, with $S_j(t,\omega) \ll S_i(t,\omega)$. We may expect the behavior of LI-TIFROM in this zone to be possibly different depending whether the prominent source in this zone is $S_1(t,\omega)$ or $S_2(t,\omega)$, because $S_1(t,\omega)$ and $S_2(t,\omega)$ play different roles in the parameter $\beta(t,\omega)$ as shown by (31). More precisely, let us consider the case when p is significantly lower than 1 and analyze the contributions of each source in $\beta(t,\omega)$:

- From the point of view of $S_1(t,\omega)$, $\beta(t,\omega)$ contains the ratio of a small component and of a large component.

- On the contrary, from the point of view of $S_2(t,\omega)$, $\beta(t,\omega)$ contains the ratio of a large component and of a small component.

The parameter $\beta(t,\omega)$ is therefore not symmetrical with respect to the two sources. Beyond this qualitative approach, this phenomenon and its consequences may be analyzed in a more formal way as follows. First consider a real single-source analysis zone associated to $S_1(t,\omega)$, i.e. a zone where $S_2(t,\omega) \ll S_1(t,\omega)$. From the point of view of these source signals, the "single-source quality" of this analysis zone may be measured at each of its TF points (t,ω) by the parameter

$$\epsilon_1 = \frac{S_2(t,\omega)}{S_1(t,\omega)} \tag{32}$$

with $\epsilon_1 \ll 1$. Simple manipulations of (31) then yield the following first-order expansion of $\beta(t,\omega)$ with respect to ϵ_1 for any value of p:

$$\beta(t,\omega) \simeq p\left(1 + \frac{1-p^2}{p}\epsilon_1\right). \tag{33}$$

The first-order term in (33) is responsible for the variations of $\beta(t,\omega)$ over the considered analysis zone. Due to (33), the variance of $\beta(t,\omega)$ in the analysis zone (T,ω) reads as follows, up to a first-order approximation:

$$var[\beta](T,\omega)_{S_1} = p^2\left(\frac{1-p^2}{p}\right)^2 var[\epsilon_1](T,\omega)_{S_1}, \tag{34}$$

where the subscripts $_{S_1}$ mean that these expressions apply to a single-source analysis zone associated to $S_1(t,\omega)$, and where $var[\epsilon_1](T,\omega)_{S_1}$ is defined as in (29) and measures the overall intrinsic "single-source quality" of this analysis zone.

A symmetrical investigation may then be performed for a real single-source analysis zone associated to $S_2(t,\omega)$, i.e. a zone where $S_1(t,\omega) \ll S_2(t,\omega)$. From the point of view of these source signals, the quality of this analysis zone may first be measured at each TF point (t,ω) by the parameter now defined as

$$\epsilon_2 = \frac{S_1(t,\omega)}{S_2(t,\omega)} \tag{35}$$

with $\epsilon_2 \ll 1$. The same manipulations as above here yield, up to a first-order approximation,

$$var[\beta](T,\omega)_{S_2} = \frac{1}{p^2}\left(\frac{1-p^2}{p}\right)^2 var[\epsilon_2](T,\omega)_{S_2}. \tag{36}$$

These results may be interpreted as follows. Consider two real single-source analysis zones, respectively associated to $S_1(t,\omega)$ and $S_2(t,\omega)$, which have the same intrinsic "single-source quality", i.e. which are such that

$$var[\epsilon_1](T,\omega)_{S_1} = var[\epsilon_2](T,\omega)_{S_2}. \tag{37}$$

Then, despite the symmetry of this configuration with respect to the two sources, Eq. (34) and (36) show that the principles used in the LI-TIFROM approach entail different behaviors for this method in these two analysis zones. More precisely, when $p \neq 1$, the detection parameter $\beta(t,\omega)$ does not have the same variance in these two analysis zones. Moreover,

$$\frac{var[\beta](T,\omega)_{S_2}}{var[\beta](T,\omega)_{S_1}} = \frac{1}{p^4}, \tag{38}$$

so that this discrepancy increases drastically when p becomes e.g. much lower than 1. In that case, $\beta(t,\omega)$ has a much higher variance in the single-source analysis zones associated to $S_2(t,\omega)$ than in those associated to $S_1(t,\omega)$. This then has a major influence on the contents of the ordered list of analysis zones created in the detection stage of LI-TIFROM and on its use then performed in the identification stage: in the latter stage, LI-TIFROM first takes into account the analysis zones which are situated at the beginning of the ordered list and which correspond to low variance of $\beta(t,\omega)$. It is thus very likely to identify accurately various representatives of the column of the (scaled permuted) mixing matrix corresponding to $S_1(t,\omega)$. By checking the distances between these successive tentative columns, it only keeps one representative of the column associated to $S_1(t,\omega)$. LI-TIFROM then proceeds by progressively climbing up the list of analysis zones ordered according to increasing values of the variance of $\beta(t,\omega)$. This list contains good-quality zones associated to $S_2(t,\omega)$, i.e. zones where the pollution from $S_1(t,\omega)$ has low magnitude and where LI-TIFROM would be able to identify accurately the column of the mixing matrix associated to $S_2(t,\omega)$. However, when $p \ll 1$, $\beta(t,\omega)$ has very high variances in these analysis zones as shown above, so that these zones are situated very high in the ordered list. Therefore, while LI-TIFROM is climbing in that list, there is a risk that it first performs a poor detection and identification, because it first encounters a zone: i) which has a lower variance than the above-mentioned good-quality zones associated to $S_2(t,\omega)$ and ii) which yields a mixing matrix column far enough from the one which was previously obtained for $S_1(t,\omega)$, so that this new column is also kept. When this situation occurs, LI-TIFROM yields poor separation quality. This theoretical analysis is validated in Section 6.5 by means of experimental tests.

As shown above, this drawback of the LI-TIFROM method results from its asymmetrical behavior with respect to the different sources, depending on the mixture coefficients. While we focused on the specific mixing configuration defined by (30) up to this point, the qualitative analysis that we provided at the beginning of this Section 4.6.2 may now be extended as follows to any mixing conditions. The asymmetry of LI-TIFROM is inherent in its detection parameter: $\beta(t,\omega)$ defined in (27) takes into account in different ways the observed signals (i.e. $X_2(t,\omega)$ appears in its numerator, while $X_1(t,\omega)$ appears in its denominator) and since the latter signals themselves depend in different ways on the sources, depending on mixture coefficients (e.g. in (31), when $p < 1$, $S_1(t,\omega)$ is more prominent in $X_1(t,\omega)$ while $S_2(t,\omega)$ is more prominent in $X_2(t,\omega)$), the parameter $\beta(t,\omega)$ then depends on the sources in different ways. This interpretation is of major importance because, *on the contrary*, the detection parameters of the LI-TIFCORR methods that we introduced in this paper are based on the moduli of the covariance or correlation coefficients defined in (21) and (25). Unlike $\beta(t,\omega)$, these moduli of covariance and correlation coefficients are *symmetrical* with respect to the observed signals for which they are computed, so that the LI-TIFCORR methods are not expected to lead to the drawback that we exhibited above for LI-TIFROM. Indeed, in Section 6 we present tests with low mixture coefficients (i.e. p much lower than 1) where the LI-TIFCORR methods yield very good results while LI-TIFROM fails. The detection stages introduced in this paper are therefore a major advantage as compared to the approach that we previously proposed in [5]-[6].

5 Extensions of proposed time-frequency and temporal approaches

Up to now we only considered the configuration based on the following assumptions:

1. the number P of observations is equal to the number N of sources,

2. all sources are "accessible" (in the above-defined senses).

The proposed BSS methods may be extended beyond this "standard" configuration as follows. Their first extensions concern the situations when the number P of observations is different from the number N of sources. The overdetermined case, i.e. $P > N$, is known to be handled easily in the framework of BSS and is therefore briefly described in Appendix F. On the contrary, the underdetermined case, i.e. $P < N$ is a tougher problem. Classical methods then fail to achieve BSS, i.e. their output signals are mixtures of all N sources [17]-[18]. The solution to this problem that we introduced in [17]-[18] is based on our partial BSS concept, which may be briefly defined as follows (more details about our motivations for introducing this approach may be found in [17]-[18]). We focus on P of the N mixed sources, considered as the signals of interest, while the other $(N - P)$ sources are considered as "noise". We then aim at building a partial BSS system, such that each of its output signals contains a contribution from only one of the sources of interest, plus contributions from the noise sources. We thus achieve the partial BSS of the P sources of interest. Whereas classical BSS methods do not achieve such partial BSS, we introduced in [17]-[18] several statistical methods, based on our general differential BSS concept, which solve this problem.

An attractive feature of the time-frequency and temporal BSS methods that we propose in the current paper is that they also make it possible to achieve partial BSS in a very natural way, as will now be shown. When applying any of the proposed methods to P mixtures of N supposedly accessible sources, with $P < N$, we first obtain an estimate of the scaled permuted matrix B in the same way as in the case when $P = N$, except that this matrix is here rectangular, i.e. composed of N columns which each contain P elements. Let us then select P sources as the sources of interest and keep the corresponding P columns of the estimated matrix B. We thus obtain a square sub-matrix B' of the mixing matrix B. The observed signals may then be considered as mixtures of the P sources of interest associated to the mixing sub-matrix B', plus "noise" composed of contributions from the other $(N - P)$ sources. Let us then transfer these observed signals through the inverse of this square sub-matrix B', as in (11). We thus obtain output signals which separate each of the P sources of interest from the other sources of interest, i.e. output signals which each contain a contribution from only one of the P sources of interest, plus again "noise" consisting of contributions from the other $(N - P)$ sources. In other words, we thus achieve the above-defined partial separation of the selected P sources. Note that we may thus choose arbitrarily which subset of P sources, among the initial N sources, is to be separated. This is an additional advantage with respect to the statistical differential methods that we proposed in [17]-[18] because, in the latter methods, (non-)stationarity constraints on the sources impose which of these sources may be separated.

The second extension of our LI-TIFCORR methods concerns the case when only part of the sources are accessible (the same extension may also be developed for LI-TEMPCORR-C). The basic version of this type of extension may be defined as follows, with $P = N$ just for the sake of simplicity. Assume that at least one of the sources, say $s_k(t)$, is accessible.

The above LI-TIFCORR methods then make it possible to estimate the corresponding column of the scaled permuted mixing matrix B, say column j. This estimated column consists of elements $\widehat{b_{ij}}$ which are estimates of

$$b_{ij} = \frac{a_{ik}}{a_{1k}} \qquad\qquad i = 1 \dots N, \tag{39}$$

as shown by (8). Therefore, if we now compute a modified version of each mixed signal $x_i(t)$, with $i = 2 \dots N$, according to

$$x'_i(t) = x_i(t) - \widehat{b_{ij}} x_1(t) \qquad\qquad i = 2 \dots N, \tag{40}$$

eq. (4) and (39) show that we obtain $N-1$ signals which do not contain any contributions from source $s_k(t)$ (up to errors due to the estimation of b_{ij}). The key point is then that, even if some sources were not accessible from the initial set of N mixed sources, at least one of them may become accessible from the new set of $N-1$ mixed sources involved in the modified mixed signals $x'_i(t)$. This depends on the TF distributions of all sources and happens if some sources were initially hidden, i.e. they were not isolated in any TF analysis zone when considering the initial set of N sources, but they are isolated in at least one zone when considering the set of $N-1$ sources which remains after cancelling the contributions from source $s_k(t)$ in all mixed signals. If at least one source is accessible from this new set of $N-1$ mixed sources, the same procedure may be applied again. This recursive procedure ends when (no more sources are accessible, or when) the number of recombined signals is thus decreased down to one, and this signal contains a single source. This procedure thus succeeds in extracting this source, although not all sources were initially accessible. This procedure may then be applied again, by selecting other sources $s_k(t)$ at each step of its recursion in order to extract other sources.

More advanced versions of this type of procedure may also be defined in order to cancel, at each intermediate stage of the recursion, the contributions from all the sources which are accessible at that stage. This is illustrated in the experimental tests presented in Section 6.6, where the first stage of the recursion removes the only source which is initially accessible and thus reveals the other two sources, which are then both extracted in the second stage of the recursion. Again, the main feature of these advanced recursive versions is their ability to separate all sources in situations when not all these sources are initially accessible but they progressively become accessible at each stage of the procedure, thanks to the previous suppression of the contributions of other sources from the observed mixed signals.

6 Experimental results

6.1 Performance of proposed TF methods for a fixed mixing matrix

In Sections 6.1 and 6.2, we present a large number of tests performed in the following conditions:

- we start from various real English speech signals sampled at 20 kHz,
- we derive various artificial linear instantaneous mixtures of these sources,
- we process these mixed signals with the main BSS methods proposed in this paper, i.e. the standard version of LI-TIFCORR-C and LI-TIFCORR-NC that we defined in Section 4.

The performance achieved in each test is measured by the overall Signal to Interference Ratio (SIR) associated to the outputs of the considered BSS system (denoted SIR^{out} hereafter) and/or by the SIR Improvement achieved by this system (denoted $SIRI$ below). These parameters are defined in Appendix G, together with the input SIR associated to the processed mixed signals (denoted SIR^{in} hereafter).

In our first series of tests, the mixing matrix was set to

$$A_1 = \begin{bmatrix} 1 & 0.9 \\ 0.8 & 1 \end{bmatrix}. \tag{41}$$

SIR^{in} was equal to 1.4 dB in all these tests (which is in agreement with Eq. (85) in Appendix G). Moreover, the two performance parameters, i.e. SIR^{out} and $SIRI$, are linked by the following relationship

$$(SIRI)_{dB} = (SIR^{out})_{dB} - (SIR^{in})_{dB} \tag{42}$$

as demonstrated in Eq. (78) of Appendix G. They here only differ by the constant value of SIR^{in}, i.e. 1.4 dB. Only one of them is therefore considered hereafter, i.e. SIR^{out}.

Each test was performed with two sources, corresponding to one of the following sets:

- Set 1: same male speaker.
- Set 2: different male speakers.
- Set 3: same female speaker.
- Set 4: different female speakers.
- Set 5: one male speaker and one female speaker.
- Set 6: different male speakers.

All these sources consist of 2.5 seconds of continuous speech, except those in Set 6, which last 5 seconds and contain silences. The signals in Set 2 correspond to a 2.5-second window extracted from the signals in Set 6. All these sources were first centered and scaled so that their highest absolute value is equal to 1.

The test results reported here were obtained by independently varying several parameters of the LI-TIFCORR-C and LI-TIFCORR-NC methods as follows:

- The number of samples in each time window was geometrically varied from 128 to 1024 samples, with a step size equal to 2.
- The number of such windows per analysis zone was successively set to 8, 10 or 12.
- The overlap between time windows was successively set to 50%, 75% and 90%.

As a result of these parameter values, the size of the analysis zones was varied from 219 to 6656 samples, i.e. about 11 to 330 ms. The other parameters of these BSS methods were constant in these tests, i.e:

- The windowing function $h^*(.)$ used in STFT computations was a Hanning window.
- The temporal overlap between successive analysis zones was set to 50% of their time windows.

- The distance between two identified columns of the matrix B was measured by the highest absolute difference between elements of these vectors which have the same index, i.e. using l_∞ norm. The distance threshold for accepting a new column of B was set to 0.15 (at this stage, this value was selected because it resulted in good performance in a preliminary set of tests; an automated method for assigning this threshold is preferable, as discussed in the conclusion of this paper).

For each considered BSS method, 216 tests were first performed in the above-defined conditions. Both methods succeeded in identifying all columns of the matrix B, and therefore in separating the considered sources, in all tests. The corresponding results are provided in Table 1. This table shows that both BSS methods yield high performance, i.e. depending on the considered set of sources their mean SIR^{out} over all BSS method parameters range from 57.9 to 72.1 dB. The standard deviation of SIR^{out} is acceptable, i.e. between 4.9 and 8.1 dB (except for the first set of sources, for which is it equal to 11.7 or 11.8 dB depending on the considered BSS method, but this set is more difficult than in practical applications, since it corresponds to two signals from the same speaker which may therefore have a stronger frequency overlap). The minimum SIR^{out} is equal to 38.3 dB over all these tests (except for the first set of sources, for which it is equal to 31.0 dB). The maximum SIR^{out} over all tests is 96.4 dB.

The LI-TIFCORR-NC method yields a slightly better mean SIR^{out} than the LI-TIFCORR-C version for all sets of sources, except for Set 6. It also provides a slightly better average SIR^{out} over all sets of sources, i.e. 65.0 vs. 64.0 dB.

A more detailed analysis of the results of the above tests is provided in Appendix H. It especially shows that the performance of the proposed BSS methods has a low sensitivity with respect to the values of their parameters. The preferred parameter values which result from these tests are: analysis zones consisting of 10 STFT windows, with 256 (or 128) samples per window and 75% overlap. The resulting SIR^{out} are then around 75 to 80 dB, as shown in Appendix H. This proves that both methods estimate the scaled possibly permuted mixing matrix B very accurately.

While all above results were derived by using the performance criteria that we defined in Appendix G, a variety of other criteria have also been proposed in the literature and may be used instead. This e.g. includes the SIR and SDR criteria defined in [19]. We provide in Table 2 the values of the latter criteria corresponding to part of the tests that we reported above, so that a reader more familiar with the latter criteria has an example of the correspondence between them and the criteria that we consider in this paper.

Still considering highly mixed sources, we also checked the applicability of the above methods to a higher number of sources. This investigation is presented in Appendix I. It confirms the main results that we reported above for two sources.

6.2 Performance of proposed TF methods vs. mixing matrix

Our second series of tests in the above-defined conditions aimed at investigating the influence of the mixing matrix on the performance of the proposed TF methods. In addition to the above matrix A_1, we therefore used symmetrical matrices defined as

$$\begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix}, \tag{43}$$

where we successively considered differents real values for the cross-coupling term p, in order to vary the mixture ratio and therefore the SIR^{in} associated to the observed signals

$x_i(t)$. More precisely, the matrix form defined in (43) results in

$$SIR^{in} = \frac{1}{p^2} \tag{44}$$

as shown by Eq. (85) in Appendix G. The values that we considered for p are: $p = 0.1$, 0.5, 0.9. The corresponding mixing matrices are resp. denoted A_2, A_3, A_4, and the associated SIR^{in} are provided in Table 3.

For the sake of simplicity, we only performed these tests for a single set of sources. We selected the above-defined "Set 2" of sources, which corresponds to a difficult but realistic situation, as explained in Section 6.1 and Appendix H. The parameters of the BSS methods were varied in the same way as in Section 6.1.

Each proposed time-frequency BSS method was thus tested in 144 configurations. Each of these methods succeeded in identifying B in all tests. Unlike in Section 6.1, SIR^{in} is varied in the tests considered here. Eq. (42) then shows that the correspondence between SIR^{out} and $SIRI$ is not fixed. The values of both parameters are therefore provided in Table 3. Among these two parameters, the performance of the BSS methods should preferably be assessed in terms of their output behavior, defined by SIR^{out}. Table 3 then shows that the performance of both considered BSS systems has a low sensitivity to the mixing matrix A in the considered range of p, i.e. SIR^{out} roughly ranges from 60 to 65 dB in all cases, which is quite good[11]. Eq. (42) then entails that, on the contrary, $SIRI$ significantly varies with A, which is confirmed by Table 3. Here again, LI-TIFCORR-NC slightly outperforms LI-TIFCORR-C in all cases.

6.3 Performance of proposed temporal method

We also checked the performance of the LI-TEMPCORR-C method for the same source signals and mixing matrix as in the above first series of tests. The parameters of this BSS method were selected as follows. The number of samples in each times window was geometrically varied from 256 to 4096. The overlap between these windows was successively set to 50%, 75% and 90%. The resulting global performance over both parameters is shown in Table 4 and confirms our expectations. First of all, the mean SIR^{out} achieved by our temporal approach is significantly lower than with our two TF extended methods (51.3 dB vs. 64.0 and 65.0 dB). More importantly, this temporal approach identifies the mixing matrix with much lower accuracy in a significant number of configurations. This is reflected both in the much higher standard deviation of its SIR^{out} (20.8 dB vs. 8.9 and 8.6 dB) and in its very low minimum SIR^{out}, i.e. 3.5 dB (vs. 31.0 and 31.1 dB). More precisely, if we first only consider the continuous speech sources, the minimum SIR^{out} for a given set of sources ranges from 3.5 to 14.7 dB, depending on the considered set.

It should also be noted that the only set of sources for which our temporal approach does not yield lower performance than its TF extensions is Set 6: the mean of SIR^{out}

[11] Table 3 shows that SIR^{out} decreases when p is reduced. This was confirmed by additional tests, which also showed that the performance of the proposed methods significantly degrades for very low values of p. For example, when $p = 0.01$, the mean SIR^{out} of LI-TIFCORR-C and LI-TIFCORR-NC over all tests are respectively equal to 14.0 and 17.4 dB. This performance degradation may be due to the fact that: i) the quality of single-source analysis zones and of the identification of the parameters $a_{i,\sigma(j)}/a_{1,\sigma(j)}$ (see (8)) then decreases, and ii) these identification parameters then cover a wide range, which may reduce the performance of our current method for selecting which tentative columns of the mixing matrix are kept. This phenomenon could therefore be investigated in more detail, in connection with the extensions of our methods that we outline in this paper.

is then 72.0 dB (vs. 72.1 and 71.6 dB for the TF methods), its standard deviation is 6.1 dB (vs. 6.8 and 7.2 dB) and its minimum is 60.4 dB (vs. 59.0 and 61.1 dB). This is again in agreement with our above theoretical considerations: Set 6 is the only set of sources consisting of discontinuous speech, where long silence phases therefore exist and are exploited by LI-TEMPCORR-C to achieve high performance.

These tests therefore confirm for speech signals that, except in restrictive situations involving discontinuous speech, our TF methods should be preferred to our basic temporal approach in order to optimize performance.

6.4 Comparison to classical methods

We also compared the performance achieved by our TF methods and by various BSS approaches available from the ICAlab software [14], still in the same conditions as above. The results thus obtained are shown in Table 5. While the SIR^{out} provided by all classical methods range about from 0 to 40 dB, they are higher than 60 dB for our TF methods (and even around 70 dB for their above-defined optimum parameters, when averageing SIR^{out} over all sets of sources). For speech signals, our approaches based on single-source analysis zones therefore highly outperform all methods available in the ICAlab software (including various correlation-based approaches, and including the SONS method, which is also explicitly intended for non-stationary signals, since its name stands for "Second Order Nonstationary Source Separation"). Note that, for discontinuous speech (i.e. Set 6 of sources), our LI-TEMPCORR-C method also highly outperforms all methods in the ICAlab software.

6.5 Comparison to LI-TIFROM

We then compared the performance achieved by our LI-TIFCORR methods and by the LI-TIFROM approach that we defined in Subsection 4.6. To this end, we first applied the latter approach to a single mixing matrix, in the same conditions as in Subsection 6.1. The corresponding results are provided in Table 6. This table shows that the LI-TIFROM approach yields lower performance than the LI-TIFCORR methods in these conditions:

1. Its mean SIR^{out} is a bit lower i) for almost all sets of sources (this may be compared in detail in Tables 1 and 6) and ii) when averaged over all sets of sources (i.e. 61.0 dB vs. 64.0 dB for LI-TIFCORR-C and 65.0 dB for LI-TIFCORR-NC).

2. Its performance has a larger spread around the above mean values than LI-TIFCORR for all sets of sources. This is reflected: i) in the significantly larger standard deviations of its SIR^{out} (i.e. 12.5 dB vs. 8.9 and 8.6 dB over all sets of sources, with even larger discrepancies for some sets of sources, such as 14.4 dB vs. 7.7 and 7.6 dB for Set 2) and ii) in its much lower minimum values (17.3 dB vs. 31.0 and 31.1 dB over all sets of sources).

We then investigated the influence of the mixing matrix on the performance of LI-TIFROM, using the same conditions as in Subsection 6.2. The corresponding results are provided in Table 7. This table first confirms that the mean performance of LI-TIFROM is somewhat lower than that of LI-TIFCORR when the mixing parameter is p is not much lower than 1. Moreover, this table mainly shows that LI-TIFROM yields very bad performance when p is low: for $p = 0.1$, the output SIR provided by this BSS method is lower than the input SIR in the observed signals, i.e. LI-TIFROM degrades the signals,

while the LI-TIFCORR methods still yield high SIRIs in this case (38.7 dB). This clearly illustrates the drawback of LI-TIFROM that we analyzed above in Section 4.6.2.

6.6 Performance of extended TF methods

We here first illustrate the performance of our extended approach for underdetermined mixtures, that we introduced in Section 5. To this end, we performed tests with three of the speech sources defined in Section 6.1, that we mixed by means of the matrix

$$A = \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.8 & 1 & 1 \end{bmatrix}. \tag{45}$$

We varied in the same way as in Section 6.1 the number of samples in each time window, the number of such windows per analysis zone, and the overlap between time windows. Table 8 contains the overall results thus obtained, and their variations with respect to the number of samples in each time window. Since the considered mixture is underdetermined, the sources cannot be extracted exactly but the proposed BSS method still aims at estimating the scaled permuted matrix B, as explained above. The parameter used here to measure performance therefore refers to the quality of the estimation of B and consists of the Frobenius norm of the difference between the actual matrix B and its estimate $\hat{B}$. Table 8 shows that the mean values of this Frobenius norm, as well as its standard deviations and even its maximum values, are much lower than the entries of B. This demonstrates the ability of the LI-TIFCORR-C method to identify accurately the mixing matrix in this underdetermined configuration (the best accuracy is again especially obtained when the STFT window size is set to 128 or 256 samples).

In addition, Table 9 shows the performance of the LI-TIFROM method in the same conditions as above. Here again, the method proposed in this paper most often performs better than our previous approach.

We eventually checked the performance of the recursive version of our approach intended for inaccessible sources, that we defined in Section 5. To this end, we performed tests with six sets of sources. Each of these sets contained three sources, which were mixed according to the matrix

$$A = \begin{bmatrix} 1 & 0.9 & 0.9^2 \\ 0.9^3 & 1 & 0.9^3 \\ 0.9^2 & 0.9 & 1 \end{bmatrix}. \tag{46}$$

Each set of sources consisted of two speech sources corresponding to one of the 6 sets defined in Section 6.1, and of an artificial Gaussian independent identically distributed (i.i.d) source signal. The latter signal was selected because it is present in all the TF plane and thus makes the two speech signals inaccessible. On the contrary, there exist zones of the TF plane where the contributions from both speech sources are negligible with respect to that of the i.i.d source signal. The latter signal is therefore accessible thanks to these zones. The proposed recursive LI-TIFCORR-C method then consists of two stages:

1. We first identify the only identifiable column of the scaled version of the mixing matrix A. This column corresponds to the i.i.d source. We then derive 2 modified mixed signals as explained in (40). These signals only contain the 2 speech sources.

2. We then apply the standard LI-TIFCORR-C method to these 2 mixtures of 2 sources, where these 2 speech sources are now accessible, as in Section 6.1. We thus identify

the corresponding scaled 2×2 mixing sub-matrix and we separate these 2 speech sources[12].

In these tests, we varied in the same way as in Section 6.1 the number of samples in each time window, the number of such windows per analysis zone, and the overlap between time windows. The results thus obtained for each set of sources and the global performance for all sets are shown in Tables 10 and 11, which respectively correspond to the above-defined two stages of the proposed recursive BSS method. We here measure the performance achieved in each stage by means of the same type of parameters as in the above tests for underdetermined mixtures: we compute the norm of the difference between the actual and estimated values of the part of the scaled mixing matrix which is identified in the considered stage of the recursion. Tables 10 and 11 show that these norms are almost always[13] much lower than the entries of the corresponding parts of the scaled mixing matrix. This demonstrates the ability of the recursive LI-TIFCORR-C method to accurately identify the mixing matrix, and therefore to extract the sources, in this configuration where two sources were inaccessible before they were revealed by the first stage of the recursion.

7 Discussion and conclusions

In this paper, we proposed two types of CORRelation-based BSS approaches for Linear Instantaneous mixtures. The first approach operates in the TEMPoral domain, on the Centered version of the signals, and is therefore called LI-TEMPCORR-C. It was introduced in a statistical framework. It therefore compares as follows to the taxonomy of statistical methods for BSS and ICA that may be defined in connection with [15]:

1. The most classical class in this taxonomy consists of methods intended for stationary, statistically independent, non-Gaussian, mainly i.i.d[14], sources. These methods are based on the fact that i.i.d signals cannot be separated by only resorting to their second-order statistics. These approaches therefore take into account the assumed independence of the sources beyond second order, either partly by means of some of their higher-order cumulants or moments, or more completely e.g. by using information theoretic criteria. As a consequence, these methods require the sources to be non-Gaussian (expect possibly one of these sources).

[12] We may then also extract the i.i.d source signal by recombining the extracted 2 speech sources with an initial mixture of all 3 sources, using an approach similar to (40).

[13] It may be noted that lower performance is achieved for Set 3 of sources but, again, this corresponds to a more difficult situation than in real applications since: i) the considered two speech sources then correspond to the same speaker and may therefore have a strong overlap in the TF domain and ii) in order to clearly illustrate the performance of our recursive BSS method, we here intentionnally mixed the two speech signals in Set 3 with an artificial source which covers *all* the TF domain, while in many speech applications each source is only present in part of the TF domain, so that the first stage of the recursion used to remove the accessible source then yields smaller residues of this source than here in the resulting modified mixed signals subsequently processed in the second stage of the recursion.

Note on the contrary that Set 6 yields higher overall performance. This confirms the ability of this BSS method to take advantage of realistic situations where some sources contain silences.

[14] These methods are applicable if the sources have a temporal structure (i.e. time correlation) and also if they have no such structure, but anyway they do not exploit that structure. They are therefore especially intended for i.i.d sources. This is to be constrasted with the second class of methods presented below, which requires the source spectra to meet some constraints and which therefore does not apply to i.i.d signals.

Our temporal approach does not belong to that class, since it only uses the second-order statistics of the source signals. It therefore does not require the sources to be independent, but only uncorrelated. The additional constraint that it sets to be able to separate these signals is *Assumption 1*, which requires the sources to be non-stationary as explained above. Second-order BSS methods, which are also involved in the other classes of classical methods to be described hereafter, have been claimed in the literature to avoid the estimation accuracy problems of high-order methods. This claim therefore also concerns our approach. Moreover, by only using second-order statistics, our method is also applicable to Gaussian signals, unlike the first class of classical methods that we defined above.

2. The second class in this taxonomy consists of methods intended for stationary sources, which have time correlation, and which are mutually uncorrelated. These methods only use second-order statistical properties of these stationary sources. These statistics may be defined in terms of the single-variable autocorrelation functions of the sources, or of the Fourier transforms of these functions, i.e. the power spectral densities (PSD) of these sources. These methods then require the sources to have different properties, which may be seen either as conditions on differences between their auto-correlation functions or as conditions on their spectral differences [16].

   The approaches in this class are similar to ours in the sense that, instead of requiring the sources to be independent, they only request them to be uncorrelated and they set an additional specific condition on them. However, these two types of approaches then differ in the type of condition that they consider. Unlike these classical approaches, we do not assume stationary sources with specific PSDs. Instead, we use *Assumption 1*, which requires the sources to be non-stationary. PSDs are not defined for non-stationary sources. They cannot be introduced as monodimensional Fourier transforms of single-lag autocorrelation functions, since the source autocorrelation functions then have two variables. Indeed, PSDs are not used in our method. The constraint in *Assumption 1* concerns the temporal variations of source variances, which is more directly related to the last class of classical approaches to be now described.

3. As in the second class of this taxonomy, the approaches in the third class mainly consist in avoiding the need for source independence, by assuming uncorrelated sources and requesting another condition: the sources are here supposed to be non-stationary. The emphasis is then often put on the case of Gaussian signals with time-varying variances [15].

   The temporal approach that we proposed in this paper may be seen as a new method belonging to this third class. It also takes advantage of assumed time variations of source variances, but sets different conditions on them as compared to various reported methods. It does not require us to focus on Gaussian signals, although it applies to them as well. Moreover, we have to stress again that our main motivation for introducing this temporal approach is that it opens the way to our second type of methods, based on TF analysis, which apply to much more general conditions and are therefore the main result of this paper.

More precisely, we introduced two Linear Instantaneous TIme-Frequency CORRelation-based BSS methods, which resp. use the Centered and Non-Centered versions of the

TF transforms of the signals, and which are therefore resp. called LI-TIFCORR-C and LI-TIFCORR-NC. By only using correlation-based parameters, as our above temporal approach, these TF methods also have the resulting features of that temporal approach that we defined above. Especially, they do not require the sources to be independent but only uncorrelated, and they apply to (realizations of) Gaussian sources, unlike standard ICA methods.

Our TF methods have a major advantage as compared to our temporal approach and to various previously reported time-frequency BSS methods, i.e. they set much more limited constraints on the sparsity of the sources and on the overlap between them. More precisely, they are based on Assumption 1-TF, i.e. they only require each source to be isolated in a *tiny* area of the TF plane. In other words, they only request that, for each source, there e.g. exist (at least) one very limited set of adjacent time windows and one associated frequency[15] where all other sources are inactive. On the contrary, our temporal approach is based on Assumption 1, so that it requires that, for each source, there exist a time window where all other sources are inactive at *all frequencies*, which is a much more restrictive requirement. For instance, when applied to speech sources, our temporal approach is mainly suited to discontinuous speech, so that in some time windows only one speaker is talking while the others are silent. On the contrary, our TF approaches also yield high performance for continuous speech: even if several speakers are talking in any time window, each on them only appears in a few frequency bands in each time window, due to the formant structure of speech; moreover, these bands are not the same for all speakers, at least in some time windows, so that each source is isolated at some frequencies (unless a very high number of sources are mixed).

Our time-frequency BSS methods consist in identifying the columns of the (scaled permuted) mixing matrix in TF areas where these methods detect that a source is isolated. Thanks to this principle, both versions of our LI-TIFCORR method are especially well-suited to non-stationary sources, such as speech signals, but they also apply to stationary sources, provided there exist at least one small frequency band per source where this source is isolated. This is to be contrasted with our purely temporal LI-TEMPCORR-C approach, which requires the sources to be non-stationary, as explained above.

In addition to the standard version of these TF and temporal methods, which aims at completely separating determined mixtures, we introduced extended versions of these approaches, which especially achieve partial BSS when processing underdetermined mixtures.

We presented various aspects of the experimental performance of all versions of the LI-TIFCORR method, derived from a large number of tests performed with continuous and discontinuous speech sources. This showed that these methods yield very good performance for linear instantaneous mixtures of real speech sources. Especially, for 2 mixtures of 2 source signals, their mean output SIRs over BSS parameters are above about 60 dB and their output SIRs in optimum conditions are close to 80 dB. Moreover, these SIRs have a low sensitivity to the values of the parameters of these methods, i.e. to the size and overlap of their STFT windows and to the number of such windows in analysis zones (short STFT windows should preferably be used when a large number of sources are mixed). These methods therefore provide an attractive new way to tackle the BSS problem and associated applications, such as speech separation. The LI-TIFCORR-NC version of our method yields slightly higher performance and a somewhat lower computational cost than

[15] As already noted in Section 4.3, analysis zones may have other shapes in the TF plane. Anyway, our time-frequency BSS methods then still only yield constraints on tiny areas of the TF plane.

LI-TIFCORR-C. It is therefore the preferred version of our approach. It should also be kept in mind that we compared the performance of both proposed TF methods to that of: i) our temporal version, thus confirming the above theoretical considerations, ii) various BSS methods from the literature, thus showing that our TF methods highly outperform all of them, iii) a somewhat related time-frequency BSS method that we previously developed, which was here theoretically and experimentally shown to yield lower performance than our new LI-TIFCORR methods.

The methods that we obtained at this stage still contain some heuristics, especially concerning the selection of the distance threshold for accepting a new column of the scaled permuted mixing matrix. Our future investigations will aim at avoiding such heuristics, e.g. thanks to clustering methods as suggested above, or by using prior information, when available, in a Bayesian framework. Moreover, up to now we only considered the case when the sources are mixed in a linear instantaneous way. The BSS methods that we proposed for that case may be extended so as to handle more general classes of mixtures, especially mixtures which involve time delays, e.g. associated to propagation phenomena. The resulting approaches require a detailed description and are therefore presented in the second part of this paper.

References

[1] A. Hyvarinen, J. Karhunen, E. Oja, "Independent Component Analysis", Wiley, New York, 2001.

[2] A. Cichocki, S.-I. Amari, "Adaptive blind signal and image processing. Learning algorithms and applications", Wiley, Chichester, England, 2002.

[3] J.-F. Cardoso, "Blind signal separation: statistical principles", Proceedings of the IEEE, vol. 86, no. 10, pp. 2009-2025, Oct. 1998.

[4] A. Jourjine, S. Rickard, O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures", Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000), vol. 5, pp. 2985-2988, Istanbul, Turkey, June 5-9, 2000.

[5] F. Abrard, Y. Deville, P. White, "From blind source separation to blind source cancellation in the underdetermined case: a new approach based on time-frequency analysis", Proceedings of the 3rd International Conference on Independent Component Analysis and Signal Separation (ICA'2001), pp. 734-739, San Diego, California, Dec. 9-13, 2001.

[6] F. Abrard, Y. Deville, "A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources", Signal Processing, Vol. 85, Issue 7, pp. 1389-1403, July 2005.

[7] P. Bofill, M. Zibulevsky, "Underdetermined blind source separation using sparse representations", Signal Processing, vol. 81, pp. 2353-2362, 2001.

[8] C. Févotte, C. Doncarli, "Two contributions to blind source separation using time-frequency distributions", IEEE Signal Processing Letters, vol. 11, no. 3, pp. 386-389, March 2004.

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

[10] A. Belouchrani, K. Abed-Meraim, M.G. Amin, A.M. Zoubir, "Joint anti-diagonalization for blind source separation", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 5, pp. 2789-2792, Salt Lake City, USA, May. 2001.

[11] L. Giulieri, N. Thirion-Moreau, P.-Y. Arquès, "Blind sources separation using bilinear and quadratic time-frequency representations", Proceedings of the 3rd International Conference on Independent Component Analysis and Signal Separation (ICA'2001), pp. 486-491, San Diego, California, Dec. 9-13, 2001.

[12] F. Hlawatsch, G.F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations", IEEE Signal Processing Magazine, pp. 21- 67, April 1992.

[13] L. Cohen, "Time-frequency distributions - a review", Proceedings of the IEEE, vol. 77, no. 7, pp. 941-981, July 1989.

[14] A. Cichocki, S. Amari, K. Siwek, T. Tanaka et al., ICALAB Toolboxes, http://www.bsp.brain.riken.jp/ICALAB

[15] J.F. Cardoso, "The three easy routes to Independent Component Analysis: contrasts and geometry", Proceedings of the 3rd International Conference on Independent Component Analysis and Signal Separation (ICA'2001), San Diego, California, Dec. 9-13, 2001.

[16] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, E. Moulines, "A blind source separation technique using second-order statistics", IEEE Transactions on Signal Processing, vol. 45, no. 2, pp. 434-444, Feb. 1997.

[17] Y. Deville, M. Benali, F. Abrard, "Differential source separation for underdetermined instantaneous or convolutive mixtures: concept and algorithms", Signal Processing, Vol. 84, Issue 10, pp. 1759-1776, Oct. 2004.

[18] Y. Deville, M. Benali, "Differential source separation: concept and application to a criterion based on differential normalized kurtosis", Proceedings of the 10th European Signal Processing Conference (EUSIPCO 2000), session TueAmOR1, Tampere, Finland, Sept. 4-8, 2000.

[19] C. Févotte, R. Gribonval, E. Vincent, BSS_EVAL Toolbox User Guide, IRISA Technical Report 1706, Rennes, France, April 2005. http://www.irisa.fr/metiss/bss_eval/.

Acknowledgment

The authors would like to thank the five anonymous reviewers for their very detailed and helpful comments.

A Proofs for detection criteria

We here show the validity of the detection criteria (15) and (22) resp. used in the temporal and TF versions of the proposed BSS approach. In the frame of the temporal version of our approach described in Section 3, we have the following theorem:

Theorem 1 *A source is isolated at time t if an only if*

$$|\rho_{x_1x_i}(t)| = 1 \qquad \forall\ i, \quad 2 \leq i \leq N. \tag{47}$$

Proof For the mixed signals expressed in (4), the cross-correlation coefficient defined in (14) reads

$$\rho_{x_1x_i}(t) = \frac{E\{\left(\sum_{j=1}^N a_{1j}s_j(t)\right)\left(\sum_{j=1}^N a_{ij}s_j(t)\right)^*\}}{\sqrt{E\{\left(\sum_{j=1}^N a_{1j}s_j(t)\right)\left(\sum_{j=1}^N a_{1j}s_j(t)\right)^*\}E\{\left(\sum_{j=1}^N a_{ij}s_j(t)\right)\left(\sum_{j=1}^N a_{ij}s_j(t)\right)^*\}}}. \tag{48}$$

Since the sources are assumed to be centered and uncorrelated, this yields

$$\rho_{x_1x_i}(t) = \frac{\sum_{j=1}^N a_{1j}a_{ij}^* E\{s_j(t)s_j^*(t)\}}{\sqrt{\left(\sum_{j=1}^N a_{1j}a_{1j}^* E\{s_j(t)s_j^*(t)\}\right)\left(\sum_{j=1}^N a_{ij}a_{ij}^* E\{s_j(t)s_j^*(t)\}\right)}}. \tag{49}$$

This coefficient may therefore be expressed as

$$\rho_{x_1x_i}(t) = \frac{< V_1(t), V_i(t) >}{||V_1(t)||.||V_i(t)||}, \tag{50}$$

where the notations $< .,. >$ and $||.||$ resp. stand for the usual inner product and vector 2-norm, and where the j-th component of each N-dimensional vector $V_i(t)$ is equal to $a_{ij}\sqrt{\lambda_j(t)}$, with

$$\lambda_j(t) = E\{s_j(t)s_j^*(t)\}. \tag{51}$$

Note that each parameter $\lambda_j(t)$ is the variance at time t of the source with index j, whose centered version is denoted $s_j(t)$. The Cauchy-Schwarz inequality then yields

$$| < V_1(t), V_i(t) > | \quad \leq \quad ||V_1(t)||.||V_i(t)|| \qquad \forall\ i, \quad 1 \leq i \leq N \tag{52}$$

so that (50) results in

$$|\rho_{x_1x_i}(t)| \leq 1 \qquad \forall\ i, \quad 1 \leq i \leq N, \tag{53}$$

with equality if and only if $V_1(t)$ and $V_i(t)$ are linearly dependent.

Let us now analyze this condition at a given time t, depending on the values of the source variances $\lambda_j(t),\ j = 1, \ldots N$. It should first be noted that when all values $\lambda_j(t)$ are equal to zero, all vectors $V_i(t)$ are equal to zero, so that the cross-correlation coefficients in (50) cannot be defined. This case is therefore excluded from the theoretical analysis for noiseless mixtures presented hereafter. It should be clear that this is not a restriction of our method because: i) this case corresponds to the situation when all sources have zero variance, so that the BSS problem then becomes irrelevant and ii) this case does not limit the applicability of the proposed method to practical situations, taking noise into

consideration, as explained in Appendix C. So, we now analyze the cases when at least one of the values $\lambda_j(t)$ is non zero.

If only one of these values is non zero, then all vectors $V_i(t)$ are linearly dependent, because all vectors $V_i(t)$ have exactly one non-zero component, which has the same index j for all these vectors. Equality then holds for all of them in (53) and therefore condition (47) is fulfilled.

The only case that remains to be considered is then the situation when at least two values $\lambda_j(t)$ and $\lambda_k(t)$ are not equal to zero. It may then be shown that if $V_1(t)$ and $V_i(t)$ were linearly dependent for all $i, \quad 2 \leq i \leq N$, then the columns with indices j and k of the mixing matrix A would be linearly dependent, which is not true since A is assumed to be invertible. Therefore, in the considered case, at least one pair of vectors $(V_1(t), V_i(t))$ does not consist of linearly dependent vectors, so that $|\rho_{x_1 x_i}(t)| < 1$ and condition (47) is not fulfilled.

As an overall result, this condition (47) is fulfilled if and only if exactly one of the values $\lambda_j(t)$ is not equal to zero at the considered time t, i.e. if and only if a source is isolated at that time. This yields Theorem 1.

Now consider the centered TF version of our approach described in Section 4.4. We have the following theorem:

Theorem 2 *A source is isolated in a TF analysis zone (T, ω) if an only if*

$$|c_{x_1 x_i}(T, \omega)| = 1 \qquad \forall\ i, \quad 2 \leq i \leq N. \tag{54}$$

Proof For any couple of mixed signals $x_i(t)$ and $x_k(t)$, the corresponding non-normalized covariance parameter (20) over an analysis zone reads[16]

$$C_{x_i x_k}(T, \omega) = \frac{1}{L} \sum_{p=1}^{L} [X_i(t_p, \omega) - \overline{X_i}(T, \omega)][X_k(t_p, \omega) - \overline{X_k}(T, \omega)]^* \qquad i, k = 1 \ldots N. \tag{55}$$

Moreover, the mixed signals are defined by (4) in the time domain. Taking the STFT of this mixture equation (4) yields

$$X_i(t, \omega) = \sum_{j=1}^{N} a_{ij} S_j(t, \omega) \qquad i = 1 \ldots N. \tag{56}$$

Eq. (19) then yields

$$\overline{X_i}(T, \omega) = \sum_{j=1}^{N} a_{ij} \overline{S_j}(T, \omega) \qquad i = 1 \ldots N, \tag{57}$$

so that

$$X_i(t, \omega) - \overline{X_i}(T, \omega) = \sum_{j=1}^{N} a_{ij} [S_j(t, \omega) - \overline{S_j}(T, \omega)] \qquad i = 1 \ldots N. \tag{58}$$

[16]The proof below is presented for the type of analysis zone considered in Section 4, i.e. temporal lines. It could be extended to other types of analysis zones.

Eq. (55) then becomes

$$C_{x_i x_k}(T,\omega) = \frac{1}{L}\sum_{p=1}^{L}\left\{\sum_{j=1}^{N} a_{ij}[S_j(t_p,\omega)-\overline{S_j}(T,\omega)]\right\}\left\{\sum_{l=1}^{N} a_{kl}[S_l(t_p,\omega)-\overline{S_l}(T,\omega)]\right\}^* \quad (59)$$

$$= \sum_{j=1}^{N}\sum_{l=1}^{N} a_{ij}a_{kl}^*\frac{1}{L}\sum_{p=1}^{L}[S_j(t_p,\omega)-\overline{S_j}(T,\omega)][S_l(t_p,\omega)-\overline{S_l}(T,\omega)]^* \quad (60)$$

$$= \sum_{j=1}^{N}\sum_{l=1}^{N} a_{ij}a_{kl}^* C_{s_j s_l}(T,\omega) \qquad i,k=1\ldots N. \quad (61)$$

The centered TF transforms of the sources are assumed to be uncorrelated over each analysis zone, as defined by *Assumption 2-TF*. Eq. (61) therefore reduces to

$$C_{x_i x_k}(T,\omega) = \sum_{j=1}^{N} a_{ij}a_{kj}^* C_{s_j s_j}(T,\omega) \qquad i,k=1\ldots N. \quad (62)$$

Now consider the covariance coefficient, over an analysis zone, of the mixed signals $x_1(t)$ and $x_i(t)$, with $i=1\ldots N$. This coefficient, defined in (21), then reads

$$c_{x_1 x_i}(T,\omega) = \frac{\sum_{j=1}^{N} a_{1j}a_{ij}^* C_{s_j s_j}(T,\omega)}{\sqrt{\left[\sum_{j=1}^{N} a_{1j}a_{1j}^* C_{s_j s_j}(T,\omega)\right]\left[\sum_{j=1}^{N} a_{ij}a_{ij}^* C_{s_j s_j}(T,\omega)\right]}} \qquad i=1\ldots N. \quad (63)$$

This parameter $c_{x_1 x_i}(T,\omega)$ may therefore also be expressed according to the right-hand side of (50), with the same notations except that: i) the vectors in (50) here depend on the considered analysis zone (T,ω), and ii) $\lambda_j(t)$ is here defined as

$$\lambda_j(T,\omega) = C_{s_j s_j}(T,\omega) \quad (64)$$

and is therefore the variance of the STFT of source $s_j(t)$ over the analysis zone (T,ω). This leads to the same discussion as above, except that the considered time t is replaced by the considered analysis zone (T,ω). This eventually shows that condition (54) is fulfilled if and only if a source is isolated in the considered analysis zone.

B Mean vs. minimum in detection criterion

The temporal approach that we proposed in Section 3 detects single-source time areas by combining all values of $|\rho_{x_1 x_i}(t)|$, with $2 \le i \le N$, in an overall detection criterion. We may consider using the mean or minimum of these values as this criterion (note that these two criteria differ only if there exist at least two values $|\rho_{x_1 x_i}(t)|$ to be combined, i.e. only if $N \ge 3$). The preferable criterion among these two alternatives first depends whether we want to select the times t which optimize the mean or worst-case value (vs. i, i.e. over all channels of BSS systems) of the parameter $|\rho_{x_1 x_i}(t)|$ that we use for detecting single-source time areas in each channel i of BSS systems. But the selection between these to alternative detection criteria should also take into account the parameter used to measure the performance of BSS systems. Consider the usual case, when the output performance of a BSS system is first defined by a criterion associated to each source (such as the output Signal to Interference Ratio, or SIR) and global performance is then measured by an overall

criterion defined as the *mean* over all sources of the above single-source criterion. This e.g. includes the criterion $(SIR^{out})_{dB}$ defined in (74) and used throughout this paper. Then, it is more coherent to use the mean of $|\rho_{x_1x_i}(t)|$: using the minimum of $|\rho_{x_1x_i}(t)|$ instead may result in selecting times t when this minimum of $|\rho_{x_1x_i}(t)|$ is higher than it would be if the mean of $|\rho_{x_1x_i}(t)|$ was used for selecting time areas, but when $|\rho_{x_1x_i}(t)|$ is close to this minimum for various channels i. All these channels then yield moderate accuracy when identifying the corresponding column of the scaled permuted matrix, so that they all tend to decrease $(SIR^{out})_{dB}$. This mean parameter then takes a lower value than what would have been obtained by selecting single-source time areas based on the *mean* of $|\rho_{x_1x_i}(t)|$. The same principle applies to our subsequent TF extensions of the proposed method. This analysis was confirmed by our experimental tests: $(SIR^{out})_{dB}$ tends to be significantly lower when using the minimum of $|\rho_{x_1x_i}(t)|$ for detecting single-source areas. We therefore use the mean of $|\rho_{x_1x_i}(t)|$ in this paper. The time or TF areas thus selected are then the "best" ones in the sense that they optimize that criterion.

C Influence of noise

In Section 3, we considered the theoretical noiseless BSS configuration, and we proposed a temporal BSS method. The denominators of the detection and identification parameters (14) and (17) of this method depend on the variances of the observations. If all source variances are null, these denominators are also null and the parameters in (14) and (17) are undefined. So, if we were only to provide a formal presentation for the noiseless configuration, in order to avoid the above singularity we would require that in all considered time areas at least one source has nonzero variance.

Now consider practical signals, which are typically provided by a set of sensors. One may then reasonably assume that the above requirement on sources is met, because practical source signals are not likely to all have strictly zero variance. But anyway, there is then no need to require practical sources to fulfill the above condition, because the sensor noise contained by real recordings avoids the above singularity and makes our BSS method work, as will now be shown. First consider time areas where at least one source has significantly higher variance than sensor noise. In such areas, the influence of noise is negligible, so that the proposed method just operates as explained in Section 3. Now, in time areas where all sources have significantly lower variance than noise, each observed signal $x_i(t)$ becomes restricted to the noise contribution measured by the considered sensor. In classical situations, these sensor noise signals are mutually uncorrelated. The corresponding correlation coefficients $\rho_{x_1x_i}(t)$ are therefore low. Consequently, the corresponding time areas are not inserted at the beginning of the ordered list created in the detection stage of the proposed BSS method, and are therefore not used for identifying a column of the mixing matrix. So, our BSS method handles correctly these time areas where all sources have negligible variance as compared to noise, and even zero variance.

We experimentally validated as follows the above theoretical analysis. Starting from the two 50000-sample sources of Set 2 defined in Section 6.1, we added at the end of each of these sources 50000 samples which were all equal to zero. We mixed these extended source signals with the matrix defined in (41). We then added noise components to these source mixtures. These noise components are mutually independent i.i.d Gaussian centered signals. Their variances were selected so that they are 40 dB lower than the smallest source variance. With this low noise level, we may hope the LI-TEMPCORR-C method to still be able to detect the single-source time areas that it detected in the noiseless case that we

experimentally studied in Section 6.3. We here aim at checking that

1. it succeeds in detecting these time areas, i.e. it still puts them in the ordered list of single-source areas (possibly in a slightly different order than in the noiseless case, because the small noise components in these areas may slightly modify the values of the detection parameter in these areas)

2. and it handles correctly the numerous time areas where both sources have much lower variance than the noise components, i.e. all the areas corresponding to the second part of the source signals, where these signals here have strictly zero variance. Handling these areas correctly means that the LI-TEMPCORR-C method should not put them before the single-source areas in the ordered list.

The results to be derived from these tests therefore concern the detection of single-source time areas. In order to more easily compare these results with those in the noiseless case, we then used as follows the time areas that we detected here. We again considered the 100000-sample source signals and we again mixed them according to (41), but without adding noise. Starting from these observations, we then identified the columns of the scaled permuted mixing matrix in the time areas that we detected above. We then derived the output signals of our BSS system and the associated SIR^{out}. This procedure was repeatedly applied for the same parameter values of our BSS system as in Section 6.3. The resulting mean of SIR^{out} is 51.8 dB, its standard deviation is 17.5 dB and its minimum and maximum values are respectively 16.1 and 70.3 dB. These results should be compared to those for Set 2 in Table 4, which concerns the noiseless case studied in Section 6.3. They are quite similar (considering the limited number of configurations processed in these tests), which confirms that the proposed BSS method handles correctly time areas where the sources have much lower variances than the noise components.

While we considered the LI-TEMPCORR-C method above, the same type of comments also applies to the TF extension of this BSS method introduced in Section 4.

D Alternative detection stages for the centered time-frequency BSS method

In Section 4.4, we defined a version of the detection stage of the considered centered time-frequency BSS method. Modified versions of this approach (and of the corresponding temporal approach described in Section 3) may also be defined. Especially:

- Another method consists in considering the values of $|c_{x_i x_j}(T,\omega)|$ associated to a single, arbitrary, couple of observations with indices i and j. This is acceptable if the considered signals yield perfect single-source zones because, as all mixing coefficients a_{ij} are assumed to be nonzero in this paper, these zones result in $|c_{x_i x_j}(T,\omega)| = 1$ for any couple of observations and may therefore be detected from a single couple. Moreover, this is attractive because it reduces the computational cost as compared to the detection methods which use a large number of couples of observations. However, practical signals are likely to only yield non-ideal "single-source" zones, where residues from other sources result in deviations of $|c_{x_i x_j}(T,\omega)|$ from their ideal value 1. The magnitude of these deviations may depend on the considered mixing coefficients and therefore on the considered couple of observations. Using a single, arbitrary, couple may therefore yield lower performance than the multi-couple method

based on the mean value $\overline{|c_{x_1x_i}(T,\omega)|}$. The single-couple method should therefore only be used when computational cost must be optimized at the expense of performance.

- More elaborate methods for combining the information provided by all couples of observations may of course also be derived, which will result in extended versions of the detection stage of the proposed time-frequency BSS method.

E Proof for identification parameter of time-frequency BSS method

We here derive the value of the centered identification parameter $I_i(T,\omega)$, defined in (23), in a single-source analysis zone. For any couple of mixed signals $x_i(t)$ and $x_k(t)$, the corresponding non-normalized covariance parameter over an analysis zone may be expressed according to (62). If a source $s_j(t)$ is isolated (i.e. only this source has a nonzero variance) in the considered analysis zone, (62) reduces to

$$C_{x_ix_k}(T,\omega) = a_{ij}a_{kj}^* C_{s_js_j}(T,\omega) \qquad\qquad i,k = 1\ldots N. \tag{65}$$

The identification parameter $I_i(T,\omega)$ defined in (23) may then be expressed as follows

$$I_i(T,\omega) = \frac{a_{ij}a_{1j}^* C_{s_js_j}(T,\omega)}{a_{1j}a_{1j}^* C_{s_js_j}(T,\omega)} \tag{66}$$

$$= \frac{a_{ij}}{a_{1j}} \qquad\qquad i = 1\ldots N, \tag{67}$$

where j is the index of the source which is isolated in the considered analysis zone. If we now denote $s_k(t)$ the source which is isolated, as in Section 4.4 which uses the current appendix, (67) becomes

$$I_i(T,\omega) = \frac{a_{ik}}{a_{1k}} \qquad\qquad i = 1\ldots N. \tag{68}$$

F Extension to overdetermined mixtures

This appendix shows how the BSS methods proposed in this paper may be used in the overdetermined case which was defined in Section 5, i.e. in situations when the number of observations may be higher than the number of sources. We may then use the standard technique which consists in first applying a Principal Component Analysis to the available observations, so as the estimate the number N of sources and to project the observations into a N-dimensional subspace. The mixed signals thus obtained are then used as the inputs of the BSS methods that we proposed in this paper. As an alternative, the above description of our BSS methods shows that they may also be used directly to estimate the number N of sources from the original observations. Separation may then be achieved by selecting an arbitrary subset of N signals among the available P observations and applying our BSS methods to them.

G SIR of mixed signals and performance criteria of BSS methods

We here define the Signal to Interference Ratio (SIR) associated to the mixed signals which are processed by our BSS methods and the parameters used to measure the performance of these methods in the tests reported in this first part of our paper.

First consider a single source, with a given index[17] k. We define the input SIR of our BSS system associated to source k by using the following two-stage approach. As a first stage, we consider a single input with index i of the BSS system, which receives the mixed signal $x_i(t)$. This signal consists of:

1. A contribution from the source with index k. This contribution is considered as the signal of interest contained by input i of the BSS system and is equal to $a_{ik}s_k(t)$.

2. Contributions from all others sources with indices $j \neq k$ (we here study the situation when no noise is added to the source signals). These sources are considered as interfering signals contained by input i of the BSS system. Their overall contribution in $x_i(t)$ is equal to $x_i(t) - a_{ik}s_k(t)$.

The elementary input SIR of the BSS system, associated to its input i and to source k, is then defined as the ratio of the powers of the above signal and interference contributions[18], i.e

$$SIR_k^{in}(i) = \frac{E\{|a_{ik}s_k(t)|^2\}}{E\{|x_i(t) - a_{ik}s_k(t)|^2\}}. \tag{69}$$

As a second stage, we define the overall input SIR of the BSS system associated to source k, i.e. when taking into account all observed signals $x_i(t)$ used as the inputs of this system. This overall SIR is defined as

$$SIR_k^{in} = \max_{i=1\ldots N}(SIR_k^{in}(i)), \tag{70}$$

i.e, for the considered source k, we take into account the observed signal where this source has the highest SIR.

We then use the same approach for defining the output SIR of the BSS system. Therefore, we first consider source k and output i of the BSS system, which provides the signal $y_i(t)$. This signal consists of:

1. The useful contribution associated to output i of the BSS system. This contribution is defined as the ideal value of output $y_i(t)$ when the source extracted on that output is source k. Due to the principle of the considered BSS methods which was presented in Section 2, this ideal output is equal to the contribution of source k in the first mixed signal, i.e. $a_{1k}s_k(t)$.

2. In the same way as in input signals, the interference contribution in output i is then defined as the remainder of $y_i(t)$, i.e. it is equal to $y_i(t) - a_{1k}s_k(t)$.

The elementary output SIR of the BSS system, associated to its output i and to source k, is then defined as the ratio of the powers of the above signal and interference contributions, i.e

$$SIR_k^{out}(i) = \frac{E\{|a_{1k}s_k(t)|^2\}}{E\{|y_i(t) - a_{1k}s_k(t)|^2\}}. \tag{71}$$

[17] The index of each source signal is known when testing our BSS methods with given source signals.

[18] In our tests, we first centered each overall time series defining one source signal.

We then define the overall output SIR of the BSS system associated to source k as

$$SIR_k^{out} = \max_{i=1...N}(SIR_k^{out}(i)). \tag{72}$$

We then define the SIR Improvement (SIRI) achieved by the BSS system with respect to source k as

$$SIRI_k = \frac{SIR_k^{out}}{SIR_k^{in}}. \tag{73}$$

All above parameters only refer to a single source. The corresponding overall features of the BSS system are eventually defined as the geometrical means of each considered parameter over all sources k, i.e. as the arithmetic means of this parameter expressed in dB. This yields explicitly

$$SIR^{in} = \left(\prod_{k=1}^{N} SIR_k^{in}\right)^{\frac{1}{N}} \quad \text{and} \quad (SIR^{in})_{dB} = \frac{1}{N}\sum_{k=1}^{N}(SIR_k^{in})_{dB}, \tag{74}$$

$$SIR^{out} = \left(\prod_{k=1}^{N} SIR_k^{out}\right)^{\frac{1}{N}} \quad \text{and} \quad (SIR^{out})_{dB} = \frac{1}{N}\sum_{k=1}^{N}(SIR_k^{out})_{dB}, \tag{75}$$

$$SIRI = \left(\prod_{k=1}^{N} SIRI_k\right)^{\frac{1}{N}} \quad \text{and} \quad (SIRI)_{dB} = \frac{1}{N}\sum_{k=1}^{N}(SIRI_k)_{dB}. \tag{76}$$

Note that this also entails

$$\frac{SIR^{out}}{SIR^{in}} = \left(\prod_{k=1}^{N}\left[\frac{SIR_k^{out}}{SIR_k^{in}}\right]\right)^{\frac{1}{N}} = \left(\prod_{k=1}^{N} SIRI_k\right)^{\frac{1}{N}} = SIRI \tag{77}$$

and therefore

$$(SIRI)_{dB} = (SIR^{out})_{dB} - (SIR^{in})_{dB}. \tag{78}$$

The parameters SIR^{out} and/or $SIRI$ are used to measure the performance of the considered BSS system, whereas SIR^{in} indicates to which extent the signals processed by this system are mixed. It should be noted that SIR^{in} has a simple expression in the configuration involving $N = 2$ mixtures of $N = 2$ sources, as will now be shown. First consider the source with index $k = 1$. Eq. (69) yields

$$SIR_1^{in}(i) = \frac{E\{|a_{i1}s_1(t)|^2\}}{E\{|a_{i2}s_2(t)|^2\}}. \tag{79}$$

Denoting m the value of the input index i which corresponds to the highest $SIR_1^{in}(i)$, eq. (70) yields

$$SIR_1^{in} = \frac{E\{|a_{m1}s_1(t)|^2\}}{E\{|a_{m2}s_2(t)|^2\}}. \tag{80}$$

Similarly, for the source with index $k = 2$, eq. (69) yields

$$SIR_2^{in}(i) = \frac{E\{|a_{i2}s_2(t)|^2\}}{E\{|a_{i1}s_1(t)|^2\}}. \tag{81}$$

Denoting n the value of the input index i which corresponds to the highest $SIR_2^{in}(i)$, eq. (70) yields

$$SIR_2^{in} = \frac{E\{|a_{n2}s_2(t)|^2\}}{E\{|a_{n1}s_1(t)|^2\}}. \tag{82}$$

Therefore

$$SIR^{in} = \sqrt{SIR_1^{in}\ SIR_2^{in}} \tag{83}$$

$$= \left|\frac{a_{m1}a_{n2}}{a_{m2}a_{n1}}\right|. \tag{84}$$

Especially, in the standard situation when source 1 is prominent in input 1 and source 2 is prominent in input 2, we have: $m = 1$ and $n = 2$. Eq. (83) then reads

$$SIR^{in} = \left|\frac{a_{11}a_{22}}{a_{12}a_{21}}\right|. \tag{85}$$

H Additional test results for two mixtures of two sources

We here provide a more detailed analysis of the results of the tests considered in Section 6.1.

Let us first analyze the influence of the considered set of sources on performance. First consider continuous speech sources. One may expect the performance of time-frequency BSS methods to be higher for sources which have lower spectral overlap, and therefore: i) higher for different male (resp. female) speakers than for the same speaker and ii) higher when mixing a male and a female speakers than when mixing the same or different male (resp. female) speakers. Similarly, one may expect Set 6 to yield better performance than Set 2, because it contains silence phases in addition, thus making the sources more accessible (at any frequency). The mean values of SIR^{out} in Table 1 confirm all these expectations, except one of them i.e: the performance achieved when mixing male and female speech is slightly lower than when mixing two male speakers or two female speakers (but still higher than when mixing two signals from the same speaker). However, a few permutations with respect to the expected respective merits of the considered sources could be foreseen because all considered sets of sources yield very high and often relatively similar mean SIR^{out}, and such permutations are not guaranteed to be statistically significant due to the limited number of signals considered at this stage.

Tables 12 to 14 provide a more detailed analysis of some aspects of the above tests: they only contain the overall values of the considered performance criteria over all sets of sources, but each of these tables details the variations of these criteria vs one of the parameters of the BSS methods (while averaging over the others). They first show that the mean of SIR^{out} has a relatively low sensitivity to the size and overlap of STFT windows and to the number of such windows in analysis zones. In addition, Table 12 shows that the number of samples in STFT windows should preferably be set to 256 (or 128) in order to optimize the mean, standard deviation and minimum value of SIR^{out}. A trade-off between these parameters may be obtained by also using 10 STFT windows per analysis zone[19] and a 75% overlap between these windows. These tables also show that the LI-TIFCORR-NC version yields a slightly better mean SIR^{out} than LI-TIFCORR-C in

[19]Performance may be lower when using a higher number of STFT windows per analysis zone for the following reason. Consider the best single-source TF areas, i.e. the areas where the interfering sources have the smallest STFT values as compared to the source of interest. These are the areas where the mixing

almost all considered cases, which is in agreement with the overall respective performance of these methods derived in Section 6.1 from the results contained in Table 1.

As an example, we present in more detail the results achieved by these methods when they are operated with the preferred parameter values that we selected above, i.e. analysis zones consisting of 10 STFT windows, with 256 samples per window and 75% overlap. These methods are here applied to the "Set 2" of sources, since this corresponds to a difficult but realistic situation, as explained above. The LI-TIFCORR-C and LI-TIFFCORR-NC methods then resp. yield $SIR^{out} = 74.8$ dB and 77.7 dB [20]. As expected, both values are quite high and SIR^{out} is somewhat better for LI-TIFCORR-NC. These high values prove that both methods estimate the scaled possibly permuted mixing matrix B very accurately. This may also be checked directly by analyzing the estimates $\hat{B}$ of this matrix provided by these BSS methods. When rounding their elements with 10^{-4} accuracy, these matrices are equal to

$$\hat{B} = \begin{bmatrix} 1.0000 & 1.0000 \\ 1.1112 & 0.8000 \end{bmatrix} \qquad \text{and} \qquad \hat{B} = \begin{bmatrix} 1.0000 & 1.0000 \\ 1.1111 & 0.8000 \end{bmatrix} \tag{86}$$

resp. for the LI-TIFCORR-C and LI-TIFCORR-NC methods. Both values should be compared to the actual matrix B defined by (8), which is here equal to

$$B = \begin{bmatrix} 1.0000 & 1.0000 \\ 0.8000 & 1.1111 \end{bmatrix} \qquad \text{or} \qquad B = \begin{bmatrix} 1.0000 & 1.0000 \\ 1.1111 & 0.8000 \end{bmatrix}, \tag{87}$$

depending whether it corresponds to a non-permuted or a permuted version of the source signals. Both estimated values in (86) are therefore extremely close to the permuted version of B, which clearly demonstrates the high separation capability of the proposed approaches. This capability may also be checked as follows from the temporal and TF representations of the considered signals. The source signals used in the test detailed here are shown in Fig. 2. Although the TF transforms of these source signals have significant differences (see Fig. 3 and 4), the TF transforms of the resulting mixed signals are almost identical (see Fig. 5 and 6), due to the considered hard mixing conditions. Nevertheless, consider e.g. the estimated output signals provided by the LI-TIFCORR-NC method, which are shown in Fig. 7. These signals are identical to the (scaled permuted) sources, which confirms that the proposed time-frequency BSS methods succeed in separating these signals with a high accuracy.

I Test results for four mixtures of four sources

In addition to the tests with two mixtures of two sources reported in Section 6.1, we checked the applicability of the above methods to a higher number of sources. To this end, we used a single set of 4 sources, corresponding to 2 male and 2 female speakers,

matrix may be identified with the highest accuracy. Consider a situation where these areas are "small". They are exploited in our BSS methods when the size of the analysis zones used in these methods is smaller than these single-source TF areas. High performance is then achieved. On the contrary, if our methods are operated with analysis zones larger than these single-source areas, they cannot identify the mixing matrix in these areas. The mixing matrix is then identified in other areas (which may include these areas), thus resulting in lower performance. Increasing the number of STFT windows per analysis zone also increases the size of these analysis zones and may therefore lead to such performance degradation.

[20] We also checked the performance of both methods for a lower number of STFT windows per analysis zone: their SIR^{out} remain higher than 60 dB for 8, 6 or 4 windows, but degrade significantly for 2 windows.

and selected from the above continuous speech signals. We mixed them with a 4x4 matrix defined as an extension of the matrix form introduced in (43), i.e

$$A = \begin{bmatrix} 1 & p & p^2 & p^3 \\ p & 1 & p & p^2 \\ p^2 & p & 1 & p \\ p^3 & p^2 & p & 1 \end{bmatrix}. \tag{88}$$

We here focus on the results obtained for a value of p corresponding to highly mixed signals, i.e. $p = 0.9$. The parameters of the two considered time-frequency BSS methods were varied in the same way as above.

Both methods again succeeded in identifying all columns of the matrix B in all these tests. Since we consider a single set of sources and a single mixing matrix, all tests are performed with the same SIR^{in}, which is equal to - 3.65 dB. As in Section 6.1, the difference between SIR^{out} and $SIRI$ is thus constant and we only consider SIR^{out} hereafter. Its mean value over all considered tests is equal to 41.5 dB for LI-TIFCORR-C and 43.2 dB for LI-TIFCORR-NC. The latter method therefore again yields slightly better performance.

Tables 15 to 17 resp. detail the dependence of SIR^{out} with respect to each BSS method parameter, while averaging over the other parameters. These tables show that, here again, the proposed methods should be operated with 10 short STFT windows per analysis zone, with 75% overlap (or 90% for LI-TIFCORR-NC). As for the size of STFT windows:

- If we take into account all test configurations reported in Tables 15 to 17, decreasing the STFT window size down to 128 samples yields a somewhat better value for the mean of SIR^{out} (and for its standard deviation, which is not detailed here for the sake of brevity), whereas overall performance was somewhat better for 256-sample windows for the tests with two sources reported in Section 6.1.

- But performance is here again better with 256 samples than with 128 samples if we focus on the case when the above-defined optimum values are used for the other parameters of the BSS methods, i.e. 10 STFT windows with 75% overlap in analysis zones. More precisely, the SIR^{out} resp. achieved by LI-TIFCORR-C and LI-TIFCORR-NC are 47.7 and 46.9 dB for 128-sample windows, 49.8 and 49.3 dB for 256-sample windows (LI-TIFCORR-C performs slightly better than LI-TIFCORR-NC in these specific configurations).

These results are therefore coherent with those obtained in Section 6.1 for mixtures of two sources. The possibility to get slightly better performance with shorter (i.e. 128-sample) STFT windows when mixing four sources may be explained as follows. When the observed signals are mixtures of a larger number of arbitrary sources, the TF areas where a source is isolated may tend to get smaller, because more sources overlap in the considered observations. Shorter STFT windows should then be used in order not to miss these small single-source areas in the detection stage of our BSS approaches.

BSS method	perf. criterion	set of sources						
		1	2	3	4	5	6	1-6
LI-TIFCORR-C	SIR^{out}: mean	61.1	64.1	57.9	65.8	62.8	72.1	64.0
	dev.	11.8	7.7	7.3	7.2	4.9	6.8	8.9
	min.	31.0	47.6	38.3	50.8	54.5	59.0	31.0
	max.	76.7	75.9	70.8	78.4	74.3	88.6	88.6
LI-TIFCORR-NC	SIR^{out}: mean	61.2	65.7	61.5	66.8	63.3	71.6	65.0
	dev.	11.7	7.6	8.1	6.5	5.0	7.2	8.6
	min.	31.1	47.0	40.6	52.7	56.3	61.1	31.1
	max.	76.6	78.3	80.2	76.5	77.5	96.4	96.4

Table 1: Performance of both time-frequency BSS methods for each set of 2 speech sources and global performance for all 6 sets. Performance criteria: mean value, standard deviation, minimum and maximum of SIR^{out} (in dB) over all parameter values of BSS methods.

BSS method	perf. criterion	set of sources: 2
LI-TIFCORR-C	SIR : mean	71.4
	dev.	9.3
	min.	51.5
	max.	93.6
LI-TIFCORR-C	SDR : mean	71.4
	dev.	9.3
	min.	51.5
	max.	93.6

Table 2: Performance of the LI-TIFCORR-C method for Set 2 of speech sources. Performance criteria: mean value, standard deviation, minimum and maximum (in dB), over all parameter values of BSS method, of the version of SIR and SDR defined in [19].

BSS method	criterion	mixing matrix			
		A_2 ($p = 0.1$)	A_3 ($p = 0.5$)	A_1	A_4 ($p = 0.9$)
	SIR^{in}	20.0	6.0	1.4	0.9
LI-TIFCORR-C	SIR^{out}	58.7	63.1	64.1	64.1
	$SIRI$	38.7	57.1	62.7	63.2
LI-TIFCORR-NC	SIR^{out}	58.7	64.2	65.7	65.7
	$SIRI$	38.7	58.1	64.3	64.8

Table 3: SIR^{in} (in dB) and performance of both time-frequency BSS methods vs. mixing matrix. Performance criteria: mean values (in dB) of SIR^{out} and $SIRI$ over all parameter values of BSS methods.

BSS method	perf. criterion	set of sources						
		1	2	3	4	5	6	1-6
LI-TEMPCORR-C	SIR^{out}: mean	42.4	49.6	39.3	53.0	51.8	72.0	51.3
	dev.	19.5	17.1	20.3	24.3	18.5	6.1	20.8
	min.	10.2	14.7	6.1	14.4	3.5	60.4	3.5
	max.	66.4	66.9	63.6	79.7	67.7	82.0	82.0

Table 4: Performance of temporal BSS method for each set of 2 speech sources and global performance for all 6 sets. Performance criteria: same as Table 1.

BSS method		SIR^{out}
LI-TIFCORR-C	mean param.	64.0
	opt. param.	69.5
LI-TIFCORR-NC	mean param.	65.0
	opt. param.	71.7
AMUSE		30.5
EVD2		31.5
EVD24		23.6
SOBI		31.5
SOBI-RO		35.7
SOBI-BPF		28.4
SONS		36.2
JADE-op		2.4
JADETD		34.1
FPICA	hyper tangent	39.5
	Gauss.	41.0
	Cubic	41.9
	5th-order Cum.	25.1
	6th-order Cum.	28.4
PEARSON opt.		42.2
SANG		40.5
NG-FICA		35.7
ThinICA		39.0
ERICA		37.3
SIMBEC		38.8
UNICA		37.3
FOBI-E		16.6
SYM-WHITE		20.1

Table 5: 1) SIR^{out} (in dB) of both proposed time-frequency BSS methods: i) mean SIR^{out} over all parameter values, ii) SIR^{out} for optimum parameter values. 2) SIR^{out} (in dB) of classical methods. This table contains the global performance for all 6 sets of sources.

BSS method	perf. criterion	set of sources						
		1	2	3	4	5	6	1-6
LI-TIFROM	SIR^{out}: mean	57.1	60.2	57.5	60.0	64.1	67.4	61.0
	dev.	13.0	14.4	11.6	13.8	10.5	8.5	12.5
	min.	17.3	27.8	21.5	33.0	30.9	42.5	17.3
	max.	72.9	85.9	71.8	80.4	83.0	80.6	85.9

Table 6: Performance of the LI-TIFROM BSS method for each set of 2 speech sources and global performance for all 6 sets. Performance criteria: same as Table 1.

BSS method	criterion	mixing matrix			
		A_2 ($p = 0.1$)	A_3 ($p = 0.5$)	A_1	A_4 ($p = 0.9$)
	SIR^{in}	20.0	6.0	1.4	0.9
LI-TIFROM	SIR^{out}	19.8	58.4	60.2	60.2
	$SIRI$	- 0.2	52.4	58.8	59.3

Table 7: SIR^{in} (in dB) and performance of the TIFROM method vs. mixing matrix. Performance criteria: same as Table 3.

BSS method	perf. criterion	window size				
		128	256	512	1024	all
LI-TIFCORR-C	$\|\|B - \hat{B}\|\|_F$: mean	0.0018	0.0019	0.0045	0.0061	0.0036
	dev.	0.0008	0.0011	0.0024	0.0054	0.0034
	min.	0.0005	0.0008	0.0020	0.0021	0.0005
	max.	0.0027	0.0033	0.0098	0.0194	0.0194

Table 8: Performance of the LI-TIFCORR-C method vs. STFT window size (in samples) and global performance for all STFT window sizes. Performance criteria: mean value, standard deviation, minimum and maximum of Frobenius norm of difference between actual matrix B and its estimate $\hat{B}$, over all other parameter values of BSS method.

BSS method	perf. criterion	window size				
		128	256	512	1024	all
LI-TIFROM	$\|\|B - \hat{B}\|\|_F$: mean	0.0015	0.0024	0.0043	0.0269	0.0088
	dev.	0.0010	0.0022	0.0048	0.0353	0.0201
	min.	0.0002	0.0006	0.0010	0.0029	0.0002
	max.	0.0035	0.0070	0.0168	0.0992	0.0992

Table 9: Performance of the LI-TIFROM method vs. STFT window size (in samples) and global performance for all STFT window sizes. Performance criteria: same as Table 8.

BSS method	perf. criterion	set of sources						
		1	2	3	4	5	6	1-6
LI-TIFCORR-C	norm: mean	5.21 e-5	4.61 e-5	14.15 e-5	9.98 e-5	9.66 e-5	5.20 e-5	8.13 e-5
	dev.	3.12 e-5	3.71 e-5	11.72 e-5	9.13 e-5	7.63 e-5	2.86 e-5	7.91 e-5
	min.	0.65 e-5	0.42 e-5	2.01 e-5	0.76 e-5	1.12 e-5	0.99 e-5	0.42 e-5
	max.	13.03 e-5	12.88 e-5	45.07 e-5	37.37 e-5	33.58 e-5	13.75 e-5	45.07 e-5

Table 10: Performance of the LI-TIFCORR-C method for each set of sources and global performance for all 6 sets. Performance criteria: mean value (over all parameter values of BSS method), standard deviation, minimum and maximum of norm of difference between actual and estimated values of the single column of scaled mixing matrix identified in first stage of recursion.

BSS method	perf. criterion	set of sources						
		1	2	3	4	5	6	1-6
LI-TIFCORR-C	norm: mean	0.0094	0.0032	0.1069	0.0029	0.0041	0.0012	0.0213
	dev.	0.0182	0.0036	0.5984	0.0035	0.0031	0.0008	0.2446
	min.	0.0007	0.0004	0.0006	0.0001	0.0008	0.0002	0.0001
	max.	0.0866	0.0181	3.5973	0.0178	0.0125	0.0033	3.5973

Table 11: Performance of the LI-TIFCORR-C method for each set of sources and global performance for all 6 sets. Performance criteria: mean value (over all parameter values of BSS method), standard deviation, minimum and maximum of Frobenius norm of difference between actual and estimated values of the 2×2 scaled mixing sub-matrix identified in second stage of recursion.

BSS method	perf. criterion	window size			
		128	256	512	1024
LI-TIFCORR-C	SIR^{out}: mean	65.6	65.5	64.1	60.6
	dev.	8.0	6.3	9.9	10.2
	min.	46.4	50.8	31.0	34.0
	max.	78.4	75.6	86.5	88.6
LI-TIFCORR-NC	SIR^{out}: mean	67.2	66.8	63.7	62.3
	dev.	6.4	5.1	8.8	11.8
	min.	50.2	54.7	31.1	34.1
	max.	78.8	78.3	76.6	96.4

Table 12: Performance of both time-frequency BSS methods vs. STFT window size (in samples). Performance criteria: same as Table 1.

BSS method	perf. criterion	nb. windows		
		8	10	12
LI-TIFCORR-C	SIR^{out}: mean	63.7	63.8	64.5
	dev.	7.8	8.9	10.0
	min.	38.2	42.3	31.0
	max.	79.1	86.5	88.6
LI-TIFCORR-NC	SIR^{out}: mean	65.8	64.7	64.6
	dev.	7.6	8.9	9.3
	min.	39.4	39.9	31.1
	max.	81.6	96.4	84.1

Table 13: Performance of both time-frequency BSS methods vs. number of STFT windows in analysis zones. Performance criteria: same as previous table.

BSS method	perf. criterion	overlap		
		50%	75%	90%
LI-TIFCORR-C	SIR^{out}: mean	62.1	65.7	64.1
	dev.	10.9	8.1	7.1
	min.	31.0	34.0	46.4
	max.	86.5	79.1	88.6
LI-TIFCORR-NC	SIR^{out}: mean	62.6	66.4	66.1
	dev.	10.5	7.7	6.9
	min.	31.1	34.1	50.2
	max.	80.2	79.1	96.4

Table 14: Performance of both time-frequency BSS methods vs. overlap between STFT windows. Performance criteria: same as previous table.

BSS method	window size			
	128	256	512	1024
LI-TIFCORR-C	46.9	42.7	40.4	36.0
LI-TIFCORR-NC	48.1	45.1	41.4	38.2

Table 15: Mean values of SIR^{out} (in dB) of both time-frequency BSS methods vs. STFT window size (in samples), for a set of 4 speech sources.

BSS method	nb. windows		
	8	10	12
LI-TIFCORR-C	41.3	42.6	40.6
LI-TIFCORR-NC	43.0	44.3	42.3

Table 16: Mean values of SIR^{out} (in dB) of both time-frequency BSS methods vs. number of STFT windows in analysis zones, for a set of 4 speech sources.

BSS method	overlap		
	50%	75%	90%
LI-TIFCORR-C	38.3	43.3	42.9
LI-TIFCORR-NC	39.0	43.7	46.9

Table 17: Mean values of SIR^{out} (in dB) of both time-frequency BSS methods vs. overlap between STFT windows, for a set of 4 speech sources.

1. For each time t, compute $\overline{|\rho_{x_1 x_i}(t)|}$.
2. Create list: order all times t according to decreasing values of $\overline{|\rho_{x_1 x_i}(t)|}$.
3. Successively for first and subsequent times t in above list:
   (a) Estimate a column of B using: $E\{x_i(t)x_1^*(t)\}/E\{x_1(t)x_1^*(t)\}$.
   (b) Keep column if its distance vs all previously identified columns $>$ threshold.
   (c) End if number of kept columns $= N$.
4. Compute estimated sources: $y'(t) = \hat{B}^{-1}x(t)$.

Figure 1: Pseudo-code of proposed temporal BSS method (each time t corresponds to a time window in practice).

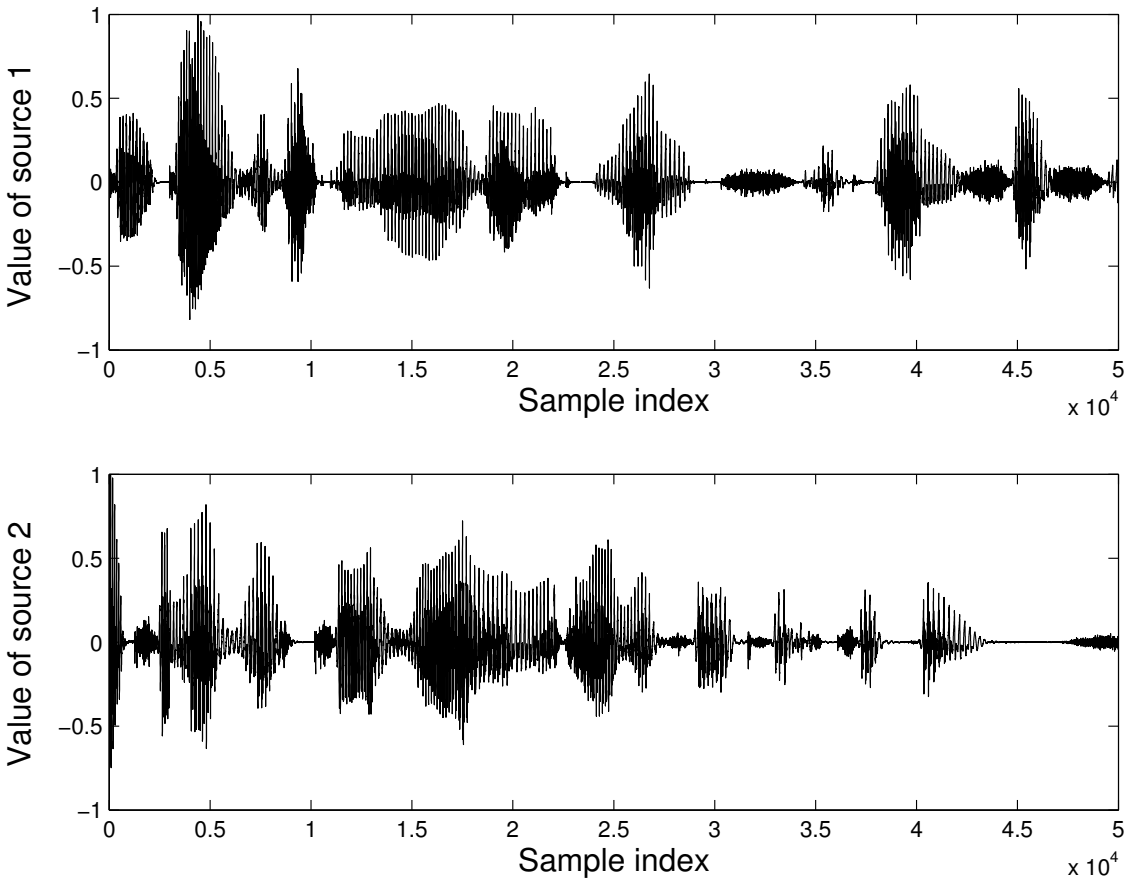

Figure 2: Sample values of both sources.

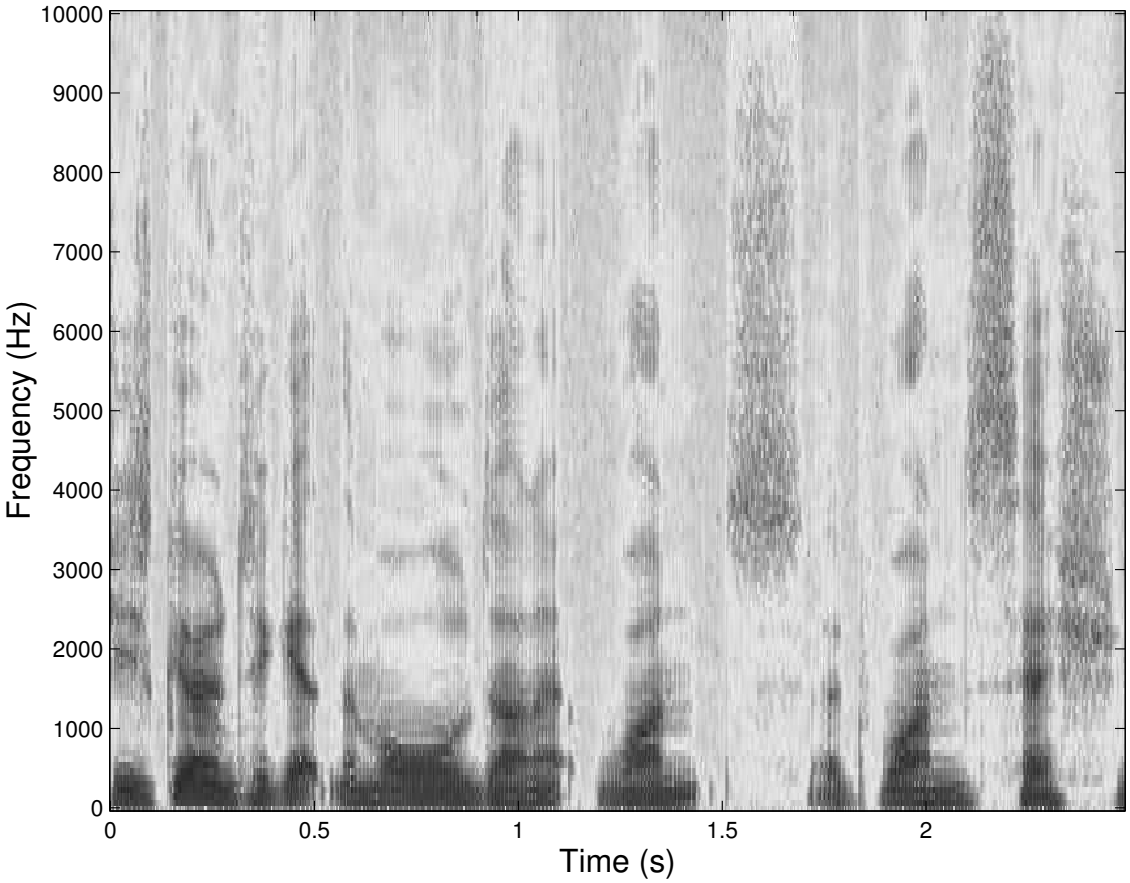

Figure 3: Time-frequency transform of first source.

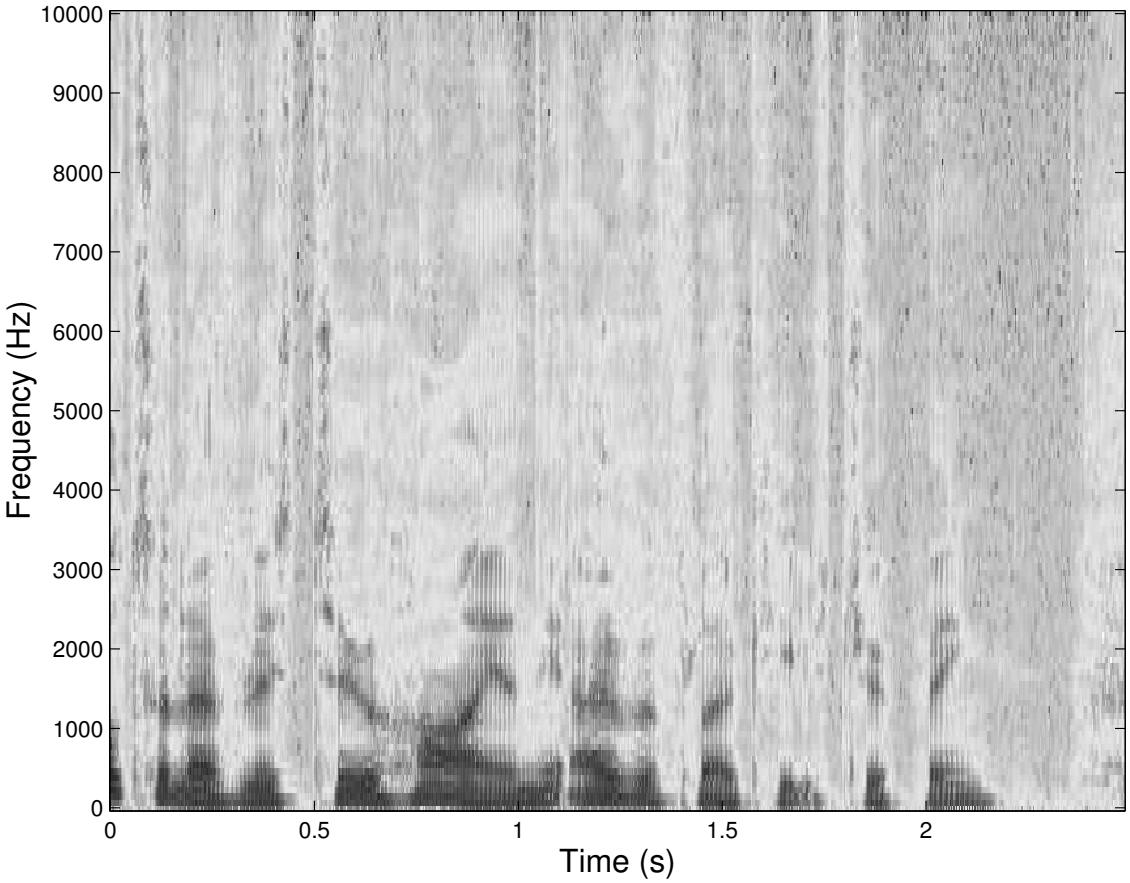

Figure 4: Time-frequency transform of second source.

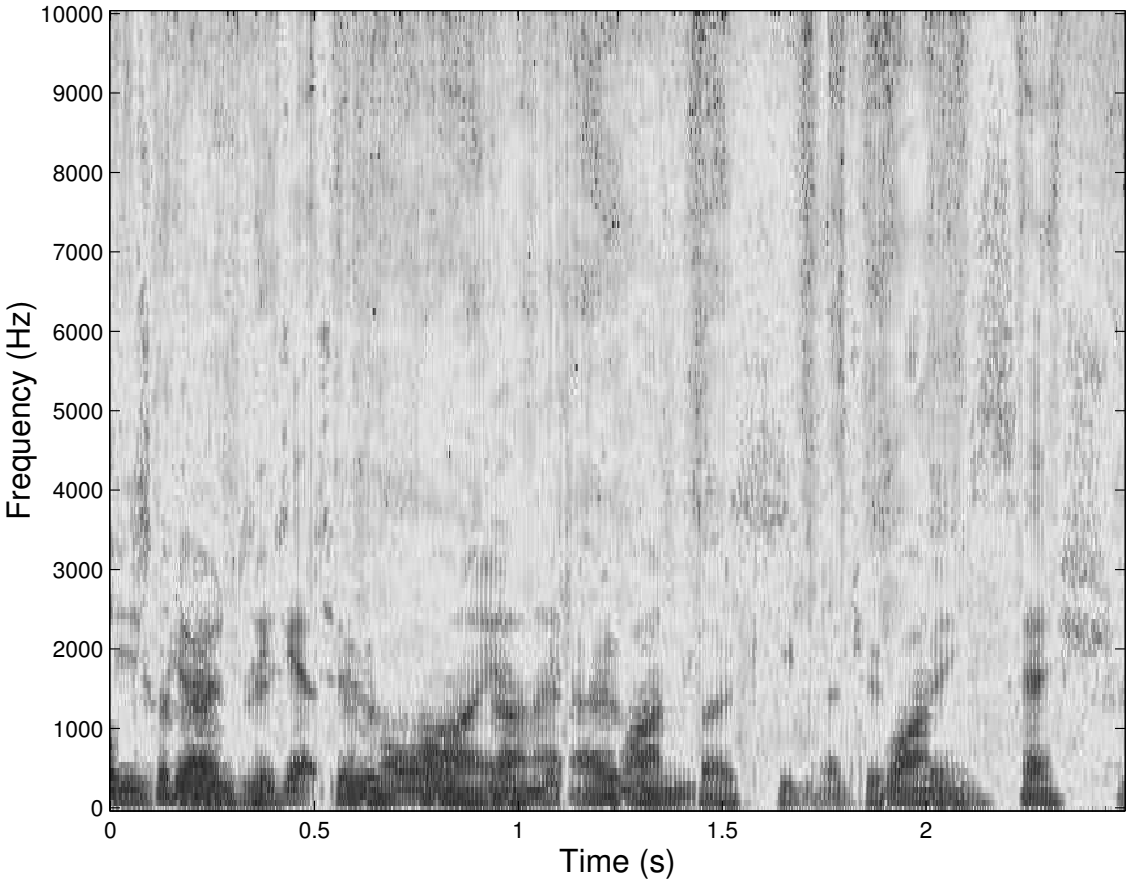

Figure 5: Time-frequency transform of first mixed signal.

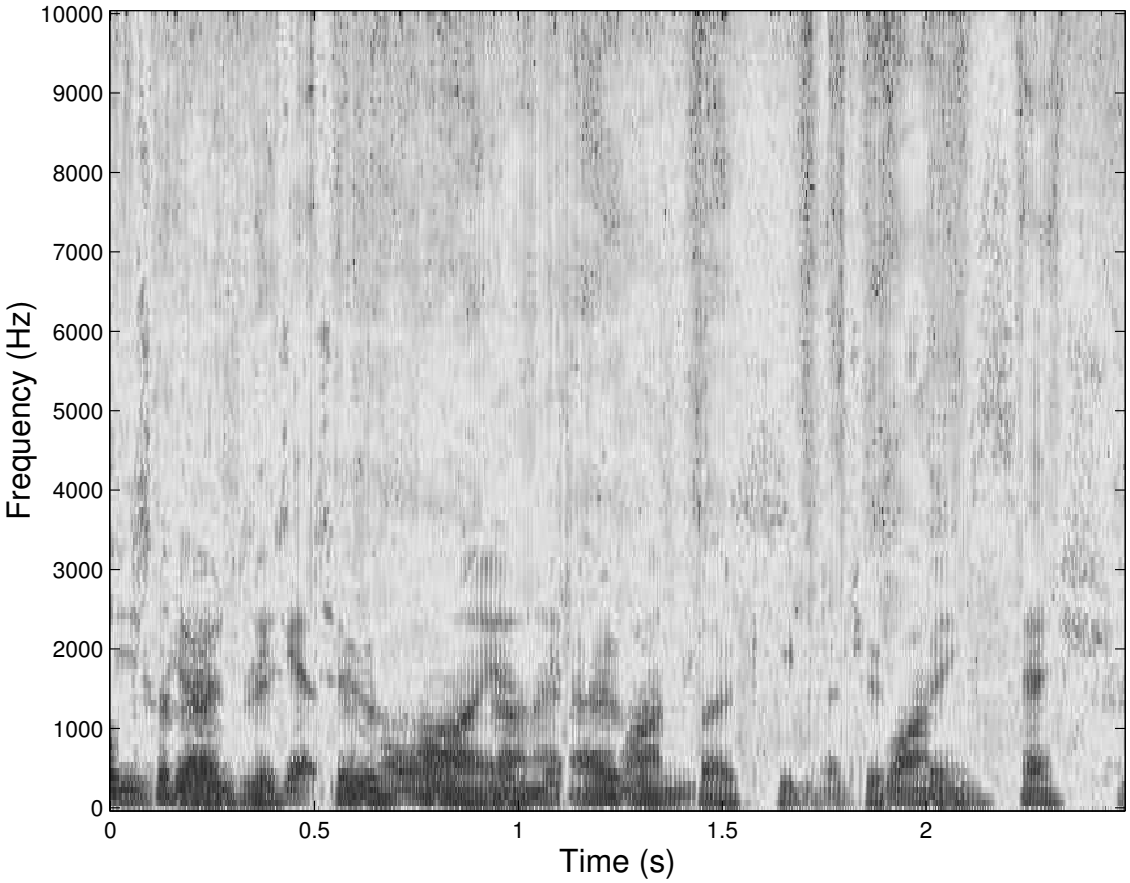

Figure 6: Time-frequency transform of second mixed signal.

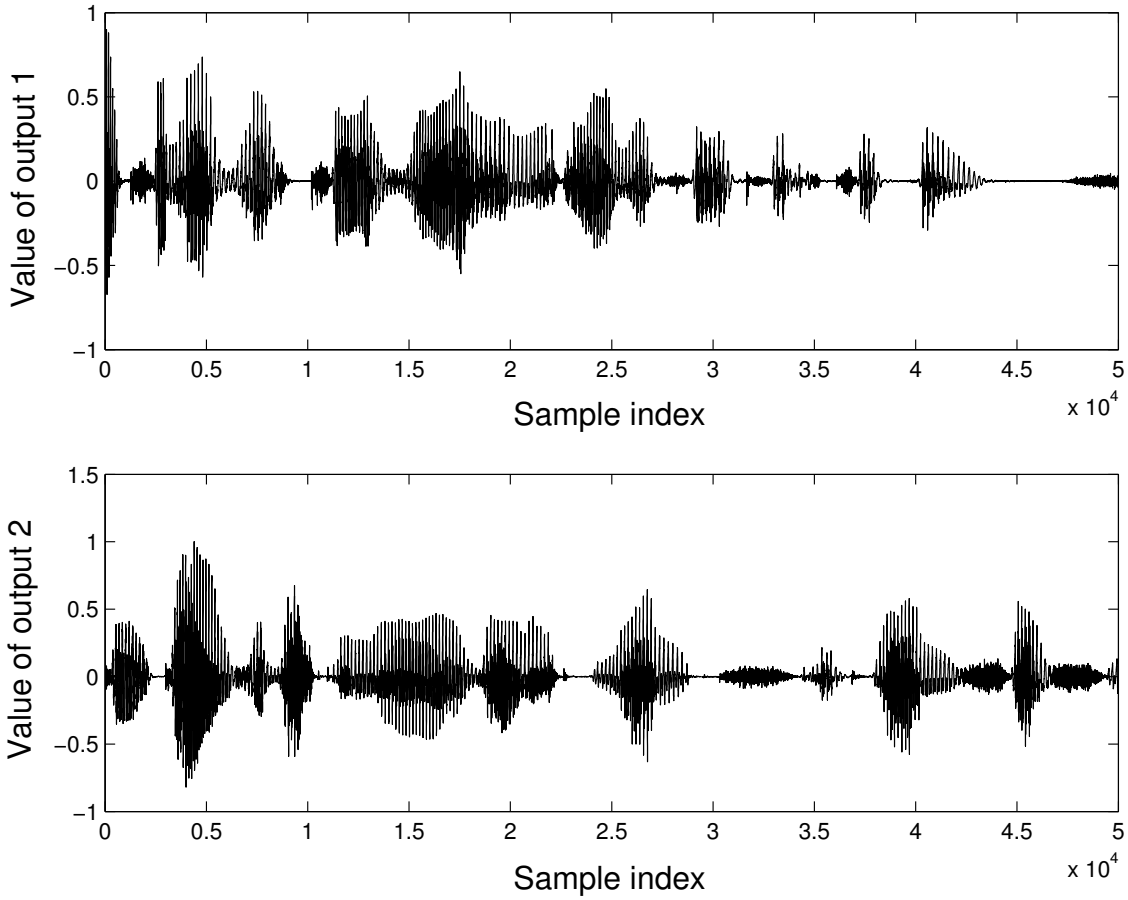

Figure 7: Sample values of both estimated source signals.