Two time-frequency ratio-based blind source separation methods for time-delayed mixtures

Matthieu Puigt and Yannick Deville

Laboratoire d'Astrophysique de Toulouse-Tarbes
Observatoire Midi-Pyrénées - Université Paul Sabatier Toulouse 3
14 Av. Edouard Belin, 31400 Toulouse, France
mpuigt@ast.obs-mip.fr, ydeville@ast.obs-mip.fr

Abstract. We propose two time-frequency (TF) blind source separation (BSS) methods suited to attenuated and delayed (AD) mixtures. They consist in identifying the columns of the (filtered permuted) mixing matrix in Constant-Time TF zones where they detect that a single source occurs, using TIme-Frequency Ratios Of Mixtures (hence their name AD-TIFROM-CT). We thus identify columns of scale coefficients and time shifts. Unlike various previously reported TF-BSS approaches, these methods set very limited constraints on the source sparsity and overlap. They are especially suited to non-stationary sources.

1 Introduction

Blind source separation (BSS) consists in estimating a set of N unknown sources from a set of P observations resulting from mixtures of these sources through unknown propagation channels. Most of the approaches that have been developed to this end are based on Independent Component Analysis [1]. More recently, several methods based on ratios of time-frequency (TF) transforms of the observed signals have been reported. Some of these methods, i.e. DUET and its modified versions, are based on an anechoic mixing model, involving attenuations and delays (AD) (this is not the general convolutive model). However, they require the sources to have no overlap in the TF domain [2], which is quite restrictive. On the contrary, only slight differences in the TF representations of the sources are requested by our Linear Instantaneous (LI) TIFROM method [3]. We here propose two novel TF-BSS methods, inspired by this LI-TIFROM approach, but suited to more general mixtures involving time shifts. We thus avoid the restriction[1] of the DUET method concerning the sparsity of the sources in the TF domain, while addressing the same class of mixtures.

2 Problem statement

In this paper, we assume that N unknown source signals $s_j(n)$ are transferred through AD channels and added, thus providing a set of N mixed observed

[1] Note however that DUET also applies to underdetermined mixtures, which is not, at this stage, the case of the methods that we propose in this paper.

signals $x_i(n)$. This reads

$$x_i(n) = \sum_{j=1}^{N} a_{ij}\, s_j(n - n_{ij}) \qquad i = 1 \ldots N, \tag{1}$$

where a_{ij} are real-valued strictly positive constant scale coefficients and n_{ij} are integer-valued time shifts. We here handle the scale/filter indeterminacies inherent in the BSS problem by extending to AD mixtures an approach that we introduced in another type of LI-BSS method, i.e. LI-TIFCORR [4]. This approach may be defined as follows. We consider an arbitrary permutation function $\sigma(.)$, applied to the indices j of the source signals, which yields the permuted source signals $s_{\sigma(j)}(n)$. We then introduce scaled and time-shifted versions of the latter signals, equal to their contributions in the first mixed signal, i.e.

$$s'_j(n) = a_{1,\sigma(j)}\, s_{\sigma(j)}\left(n - n_{1,\sigma(j)})\right). \tag{2}$$

The mixing equation (1) may then be rewritten as

$$x_i(n) = \sum_{j=1}^{N} a_{i,\sigma(j)}\, s_{\sigma(j)}\left(n - n_{i,\sigma(j)})\right) = \sum_{j=1}^{N} b_{ij}\, s'_j\left(n - \mu_{ij})\right) \tag{3}$$

with

$$b_{ij} = \frac{a_{i,\sigma(j)}}{a_{1,\sigma(j)}} \qquad \text{and} \qquad \mu_{ij} = n_{i,\sigma(j)} - n_{1,\sigma(j)}. \tag{4}$$

The Fourier transform of Eq. (3) reads

$$X_i(\omega) = \sum_{j=1}^{N} b_{ij}\, e^{-j\omega\mu_{ij}}\, S'_j(\omega) \qquad i = 1 \ldots N. \tag{5}$$

This yields in matrix form

$$\underline{X}(\omega) = B(\omega)\, \underline{S'}(\omega) \tag{6}$$

where $\underline{S'}(\omega) = [S'_1(\omega) \cdots S'_N(\omega)]^T$ and

$$B(\omega) = \left[b_{ij} e^{-j\omega\mu_{ij}}\right] \qquad i,j = 1 \ldots N. \tag{7}$$

In this paper, we aim at introducing methods for estimating $B(\omega)$.

3 Proposed basic TIFROM method for AD mixtures

3.1 Time-frequency tool and assumptions

We recently proposed [3] a LI-BSS method based on TIme-Frequency Ratios Of Mixtures, that we therefore called "LI-TIFROM". Starting from this method, we here develop extensions intended for AD mixtures. These approaches are called AD-TIFROM-CT, since they are shown below to only use "Constant-Time analysis zones". The TF transform of the signals considered in these approaches is the Short-Time Fourier Transform (STFT) defined as:

$$U(n,\omega) = \sum_{n'=-\infty}^{+\infty} u(n')h(n'-n)e^{-j\omega n'} \tag{8}$$

where $h(n'-n)$ is a shifted windowing function, centered on time n. $U(n,\omega)$ is the contribution of the signal $u(n)$ in the TF window corresponding to the short time window centered on n and to the angular frequency ω.

The AD-TIFROM-CT approach uses the following definitions and assumptions.

Definition 1 A source is said to "occur alone" in a TF area (which is composed of several adjacent above-defined TF windows) if only this source has a TF transform which is not equal to zero everywhere in this TF area.

Definition 2 A source is said to be "visible" in the TF domain if there exist at least one TF area where it occurs alone.

Assumption 1 Each source is visible in the TF domain.

Note that this is a very limited sparsity constraint !

Assumption 2 There exist no TF areas where the TF transforms of all sources are equal to zero everywhere[2].

Assumption 3 When several sources occur in a given set of adjacent TF windows, they should vary so that at least one of the moduli of ratios of STFTs of observations, $|X_i(n,\omega)/X_1(n,\omega)|$, with $i = 2 \ldots N$, does not take the same value in all these windows . Especially, i) at least one of the sources must take significantly different TF values in these windows and ii) the sources should not vary proportionally.

3.2 Overall structure of the basic AD-TIFROM-CT method

The AD-TIFROM-CT method aims at estimating the mixing matrix $B(\omega)$ defined in Eq. (7), i.e. the parameters b_{im} and μ_{im}, with $i = 2 \ldots N$ and $m = 1 \ldots N$ ($i = 1$ yields $b_{ij} = 1$ and $\mu_{ij} = 0$: see Eq. (4)). The basic version of this method is composed of a pre-processing stage and 3 main stages:

1. The pre-processing stage consists in deriving the STFTs $X_i(n,\omega)$ of the mixed signals, according to Eq. (8).
2. The detection stage aims at finding "constant-time TF analysis zones" where a source occurs alone, using the method introduced in Section 3.3.
3. The identification stage aims at estimating the columns of $B(\omega)$ in the above single-source analysis zones, using the method proposed in Section 3.4.
4. In the combination stage, we eventually derive the output signals. They may be obtained in the frequency domain by computing $\underline{Y}(\omega) = B^{-1}(\omega)\,\underline{X}(\omega)$ where $\underline{Y}(\omega) = [Y_1(\omega) \cdots Y_N(\omega)]^T$ is the vector of Fourier transforms of the output signals. The time-domain versions of these signals are then obtained by applying an inverse Fourier transform to $\underline{Y}(\omega)$.

[2] This assumption is only made for the sake of simplicity: it may be removed in practice, thanks to the noise contained by real recordings, as explained in [3].

3.3 Detection of single-source constant-time TF analysis zones

As stated above, the BSS method that we here introduce first includes a detection stage for finding single-source TF zones. The frequency-domain mixture equations corresponding to Eq. (1) read

$$X_i(\omega) = \sum_{j=1}^{N} a_{ij}\, e^{-j\omega n_{ij}}\, S_j(\omega) \qquad i = 1 \dots N. \tag{9}$$

This relationship between the observations and sources remains almost exact when expressed in the TF domain if the time shifts n_{ij} are small enough as compared to the temporal width of the windowing function $h(.)$ used in the STFT transform. We here assume that this condition is met and thus that the STFTs of the observations can be expressed wrt. the STFTs of the sources as

$$X_i(n,\omega) = \sum_{j=1}^{N} a_{ij}\, e^{-j\,\omega n_{ij}}\, S_j(n,\omega) \qquad i = 1 \dots N. \tag{10}$$

Let us consider the ratio of STFTs of mixtures

$$\alpha_i(n,\omega) = \frac{X_i(n,\omega)}{X_1(n,\omega)} = \frac{\sum_{j=1}^{N} a_{ij}\, e^{-j\,\omega n_{ij}}\, S_j(n,\omega)}{\sum_{j=1}^{N} a_{1j}\, e^{-j\,\omega n_{1j}}\, S_j(n,\omega)}. \tag{11}$$

If a source $S_k(n,\omega)$ occurs alone in the considered TF window (n_p,ω_l) then

$$\alpha_i(n_p,\omega_l) = \frac{a_{ik}}{a_{1k}} e^{-j\omega(n_{ik}-n_{1k})} = b_{im} e^{-j\omega\mu_{im}} \tag{12}$$

with b_{im} and μ_{im} defined by Eq. (4) and $k = \sigma(m)$. Since we assumed all mixing coefficients a_{ik} to be real and positive, all resulting scale coefficients b_{im} are also real and positive. The modulus of the parameter value $\alpha_i(n_p,\omega_l)$ provided in Eq. (12) is therefore equal to b_{im}. If only source $S_k(n,\omega)$ occurs in several frequency-adjacent windows (n_p,ω_l), then $|\alpha_i(n_p,\omega_l)|$ is constant over these adjacent windows. On the contrary, it takes different values over these windows for at least one index i if several sources are present, due to Assumption 3. To exploit this phenomenon, we compute the sample variance of $|\alpha_i(n,\omega)|$ on "constant-time analysis zones" that we define as series of M frequency windows corresponding to adjacent ω_l, applying this approach independently to each time index n_p. This set of frequency points ω_l is denoted Ω hereafter and the corresponding TF zone is therefore denoted (n_p,Ω). We respectively define the sample mean and variance of $|\alpha_i(n_p,\omega_l)|$ on (n_p,Ω) as

$$\overline{|\alpha_i|}(n_p,\Omega) = \frac{1}{M}\sum_{l=1}^{M} |\alpha_i(n_p,\omega_l)|\,, \tag{13}$$

$$var\left[|\alpha_i|\right](n_p,\Omega) = \frac{1}{M}\sum_{l=1}^{M} \left| |\alpha_i(n_p,\omega_l)| - \overline{|\alpha_i|}(n_p,\Omega)\right|^2. \tag{14}$$

We first compute these parameters independently for each i, with $i = 2 \dots N$. We then derive the mean over i of these variances $var\left[|\alpha_i|\right](n_p, \Omega)$, i.e.

$$MVAR\left[|\alpha|\right](n_p, \Omega) = \frac{1}{N-1} \sum_{i=2}^{N} var\left[|\alpha_i|\right](n_p, \Omega). \tag{15}$$

Similarly, we compute the inverse ratios and their means and variances on each considered analysis zone, i.e.

$$\beta_i(n, \omega) = \frac{1}{\alpha_i(n, \omega)} = \frac{X_1(n, \omega)}{X_i(n, \omega)} \tag{16}$$

$$\overline{|\beta_i|}(n_p, \Omega) = \frac{1}{M} \sum_{l=1}^{M} |\beta_i(n_p, \omega_l)| \tag{17}$$

$$var\left[|\beta_i|\right](n_p, \Omega) = \frac{1}{M} \sum_{l=1}^{M} \left| |\beta_i(n_p, \omega_l)| - \overline{|\beta_i|}(n_p, \Omega) \right|^2. \tag{18}$$

The mean over i of these variances $var\left[|\beta_i|\right](n_p, \Omega)$ then reads

$$MVAR\left[|\beta|\right](n_p, \Omega) = \frac{1}{N-1} \sum_{i=2}^{N} var\left[|\beta_i|\right](n_p, \Omega). \tag{19}$$

This mean $MVAR\left[|\beta|\right](n_p, \Omega)$ has lower or higher values than the above mean $MVAR\left[|\alpha|\right](n_p, \Omega)$, depending on mixing scale coefficients. The best single-source TF zones are those where $min\left\{MVAR\left[|\alpha|\right](n_p, \Omega), MVAR\left[|\beta|\right](n_p, \Omega)\right\}$ takes the lowest values.

3.4 Identification stage

Thanks to expression (12) of the parameters $\alpha_i(n, \omega)$ in single-source analysis zones, a natural idea for estimating the time shifts μ_{im} consists in taking advantage of the phase of $\alpha_i(n, \omega)$. We consider independently each time position n_p associated to TF windows and for each such position, we unwrap the phase of $\alpha_i(n_p, \omega)$ over all associated frequency-adjacent TF points. If we assume that $S_k(n, \omega)$ occurs alone in an analysis zone (n_p, Ω) and we consider the unwraped phase $\phi_i(n_p, \omega_l)$ of $\alpha_i(n_p, \omega_l)$ in this zone, due to Eq. (12) we have

$$-\omega_l \mu_{im} = \phi_i(n_p, \omega_l) + 2q_{im}(n_p)\pi, \tag{20}$$

where $q_{im}(n_p)$ is an unknown integer. Eq (20) shows that the curve associated to the variations of the phase $\phi_i(n_p, \omega_l)$ wrt. ω_l in a single-source zone (n_p, Ω) is a line and that its slope, equal to $-\mu_{im}$, does not depend on the value of $q_{im}(n_p)$. This slope therefore allows us to identify μ_{im}, with no phase indeterminacy. Our method for identifying the set of parameters μ_{im} associated to a column of $B(\omega)$ therefore operates as follows. In the selected constant-time single-source analysis

zone, for each observed signal with index i, we consider the M points which have two coordinates, resp. defined as the frequencies ω_l and the corresponding values $\phi_i(n_p, \omega_l)$ of the unwraped phase of the identification parameter. We determine the least-mean square regression line associated to these points. The estimate of the parameter μ_{im} is then set to the integer which is the closest to the opposite of the slope of this regression line.
The overall identification stage consists in successively considering the analysis zones ordered according to increasing values of $min\left\{MVAR\left[|\alpha|\right](n_p, \Omega), MVAR\left[|\beta|\right](n_p, \Omega)\right\}$. For each such zone, the estimates of b_{im} associated to a column of $B(\omega)$ are set to the values of $\overline{|\alpha_i|}(n_p, \Omega)$ or $1/\overline{|\beta_i|}(n_p, \Omega)$, depending whether respectively the parameter $MVAR\left[|\alpha|\right]$ or $MVAR\left[|\beta|\right]$ takes the lowest value in this zone. A new column of b_{im} is kept if its distance wrt. each previously found column of b_{im} is above a user-defined threshold ϵ_1. If a column of b_{im} is identified and kept, the corresponding column of μ_{im} is simultaneously identified, by using regression lines in the same analysis zone as explained above. The identification procedure ends when the number of columns of $B(\omega)$ thus kept becomes equal to the specified number N of sources to be separated.

4 Proposed improved TIFROM method for AD mixtures

For $N > 2$ or when the time shifts μ_{im} are non-negligible wrt. to the length of STFT windows in the case $N = 2$, the above basic method turned out to yield false results in a significant number of experimental tests: on the one hand, we obtained columns of scale coefficients which did not correspond to actual columns of $B(\omega)$. On the other hand, we only achieved a coarse identification of the associated time shifts. Both problems can be solved thanks to clustering techniques. We now detail such an approach. In this approach, we form clusters of "points" where each point consists of a tentative column of parameters b_{im}. To this end, we first compute the parameters $MVAR\left[|\alpha|\right](n_p, \Omega)$ and $MVAR\left[|\beta|\right](n_p, \Omega)$ for all analysis zones and we then only consider the zones which are such that

$$min\left\{MVAR\left[|\alpha|\right](n_p, \Omega), MVAR\left[|\beta|\right](n_p, \Omega)\right\} \leq \epsilon_2, \tag{21}$$

where ϵ_2 is a small positive user-defined threshold. We thus only keep single-source zones, which correspond to the beginning of the ordered list created in the detection stage. We successively consider each of the first and subsequent analysis zones in this beginning of the ordered list and we use them in a slightly different way than in the basic identification procedure that we described above. Here again, for each considered analysis zone, the estimates of the parameters b_{im} are set to the values of $\overline{|\alpha_i|}(n_p, \Omega)$ or $1/\overline{|\beta_i|}(n_p, \Omega)$, depending on which of the parameters $MVAR\left[|\alpha|\right]$ and $MVAR\left[|\beta|\right]$ takes the lowest value in this zone. The estimated column associated to the first zone in the ordered list is kept as the first point in the first cluster. Each subsequently estimated column is then used as follows. We compute its distances wrt. all clusters created up to this stage,

where the distance wrt. a cluster is defined as the distance wrt. the first point which was included in it. If such a distance is below a user-defined threshold ϵ_1, this new column is inserted as a new point in the corresponding cluster. Otherwise, this new column is kept as the first point of a new cluster. This is repeated for all analysis zones which fulfill condition (21). If the threshold ϵ_1 is low enough, the number of clusters thus created is at least equal to the specified number N of sources to be extracted. We then keep the N clusters which contain the highest numbers of points. For each cluster, we eventually derive a representative, by selecting its point which corresponds to the lowest value of $min \left\{MVAR\left[|\alpha|\right](n_p,\Omega), MVAR\left[|\beta|\right](n_p,\Omega)\right\}$ and thus presumably to the best single-source zone. This yields the N columns of estimates of b_{im}.
We estimate the parameters μ_{im} as follows. Independently, for each of the above N clusters of columns of b_{im}, we first compute the parameters μ_{im} in the same TF zones as these scale coefficients b_{im}. We then derive the histograms of these parameters μ_{im}, independently for each index i. We eventually keep the peak value in each histogram as the estimate of μ_{im}.

5 Experimental results

We now present tests performed with $N = 2$ sources of English speech signals sampled at 20 kHz. These signals consist of 2.5 s of continuous speech from different male speakers. The performance achieved in each test is measured by the overall signal-to-interference-ratio (SIR) Improvement achieved by this system, denoted $SIRI$ below, and defined as the ratio of the output and input SIRs of our BSS system. The mixing matrix is set to

$$A(\omega) = \begin{bmatrix} 1 & 0.9\,e^{-j\omega 75} \\ 0.9\,e^{-j\omega 75} & 1 \end{bmatrix}. \tag{22}$$

The input SIR is thus equal to 0.9 dB. The number d of samples per STFT window is varied geometrically from 1024 to 16384. The number M of windows per analysis zone is set to 8 when $d = 1024$. This value of M is then increased geometrically with d. Thus, the absolute width of the frequency bands associated to the frequency domain Ω of the analysis zones (n_p, Ω) takes the same value whatever d. This value is 156.25 Hz. In each test, the temporal overlap between STFT windows is fixed to 75%. The resulting $SIRI$s are given in Table 1.
The cluster-based method yields better or same results as the basic one. The mean $SIRI$s are resp. equal to 11.2 and 24.3 dB with the basic and cluster-based approaches. When $d = 16384$, one source is not visible in the TF plane. Two results illustrate the usefulness of clustering techniques in our approaches: when $d = 1024$, with the basic method, the Frobenius norm of the difference between the actual and theoretical matrices of scales coefficients b_{im} is equal to 1.3, while this norm is only equal to 5.4e-2 with the cluster-based approach. We explain this phenomenon as follows: with the basic method, the parameters b_{im} were identified in analysis zones which were selected because they were at the beginning of the ordered list created in the detection stage, but these identified

Method	STFT window size d				
	1024	2048	4096	8192	16384
Basic	-2.8	14.8	6.2	26.8	invisible
Improved	20.2	14.8	23.9	26.8	invisible

Table 1. Performance ($SIRI$ in dB) for both methods vs STFT window size d.

columns did not correspond to the actual (filtered permuted) mixing matrix, so that the outputs of our BSS system did not provide well separated sources. As only a few occurrences are obtained for each false column value, clustering techniques solved this problem. The case when $d = 4096$ is interesting too: we obtain the same matrix of scale coefficients with both methods (the above-defined Frobenius norm is equal to 3.6e-2). The estimated values of time shifts are equal to the theoretical ones with the cluster-based method, while we have slight differences with the basic approach: the estimated time shifts are equal to 74 and -76 while theoretical ones are ±75. This clearly demonstrates the usefulness of clustering techniques.

6 Conclusion and extensions

In this paper, we proposed two TF BSS methods for AD mixtures. They avoid the restrictions of the DUET method, which needs the source to be (approximately) W-disjoint orthogonal. Our methods consist in first finding the TF zones where a source occurs alone and then, identifying in these zones the parameters of the (filtered permuted) mixing matrix. Thanks to this principle, these approaches apply to non-stationary sources, but also to stationary and/or dependent sources (we could extend the discussion in [3] to AD mixtures), provided there exists at least a tiny TF zone per source where this source occurs alone. We experimentally showed the usefulness of clustering techniques in our methods. Our future investigations will consist in a more detailed characterization of the experimental performance of the proposed approaches. We will also aim at extending these methods to general convolutive mixtures.

References

1. A. Hyvärinen, J. Karhunen, E. Oja: Independent Component Analysis, Wiley-Interscience, New York, 2001.
2. A. Jourjine, S. Rickard, O. Yilmaz : Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures, Proceedings of ICASSP 2000, IEEE Press, Istanbul, Turkey, June 5-9, 2000, vol. 5, pp. 2985-2988.
3. 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.
4. Y. Deville, Temporal and time-frequency correlation-based blind source separation methods, Proceedings of ICA 2003, pp. 1059-1064, Nara, Japan, April 1-4, 2003.