15

A Study of Methods for Initialization and Permutation Alignment for Time-Frequency Domain Blind Source Separation

Auxiliadora Sarmiento, Iván Durán, Pablo Aguilera and Sergio Cruces
Department of Signal Theory and Communications, University of Seville, Seville Spain

1. Introduction

The problem of the blind signal separation (BSS) consists of estimating the latent component signals in a linear mixture, referred to as the sources, starting from several observed signals, without relying on any specific knowledge of the sources. In particular, when the sources are audible, this problem is known as to the cocktail-party problem, making reference to the ability of the human ear to isolate the conversation of our interest among several conversations immersed in a noisy environment with many people talking at the same time.

The complexity of the blind separation problem greatly depends on the mixture model, the number of sources and sensors that better adjust to reality and the presence of noise in the mixture. The simplest case regards the linear and instantaneous mixture, that is, when the sources are mixed affected only by some scaling. However, in a real room recording the situation becomes more difficult, since the source signals do not only follow the direct path from the source to the sensor, but there are also other paths coming from the reflections in the walls. Hence, the problem becomes convolutive rather than instantaneous, and the mixture process is then modelled by means of a convolution of the sources by some acoustic mixing filters. In the present chapter we assume that the channel between sources and microphones is time-invariant, the number of sources equals the number of sensors, that is, the $N \times N$ case, and there is no additive noise. In such recording environments the separation is very complex, especially in highly reverberant conditions where the mixing filters can be very long (greater than 250 ms) and can contains strong peaks corresponding to the echoes.

Several component analysis techniques solve the instantaneous and determined case in the time domain. One of the most popular is the Independent Component Analysis (ICA), which is a method to recover statistically independent sources by using implicitly or explicitly high-order statistics Comon (1994). Some of those techniques have been extended to solve the convolutive case in the time domain. However, its use for the separation of real speech recordings is limited because of the high length of the acoustical mixing filters (of the order of hundreds of milliseconds). Since it is required the adjustment of too many parameters, those method present convergence problems and a high computational cost. An extended strategy, referred to as fd-ICA in the literature, consists of formulating the problem in the time-frequency domain instead of the time domain (Smaragdis, 1998). The main reason is that the convolutive mixture can be approximated by a set of instantaneous mixtures, one

for each frequency bin, that can be solved independently by applying several separation algorithms. However, this simplification introduces some additional problems referred to as scaling and permutation problems, since the obtained solutions in each frequency exhibit an arbitrary complex scaling and order. The scaling ambiguity introduces a filtering effect in the estimated sources, that can be removed by introducing some constraint on the separation filters. Nevertheless, the permutation problem leads to non-consistent time-frequency representations of the estimated signals, and needs to be addressed to successfully recover the original sources. During the last years, several algorithms have been proposed in order to solve this problem, although nowadays there is no satisfactory solution, specially for highly reverberant environments. Furthermore, the problem increases in complexity with the number of sources in the mixture, and the developed algorithms usually deal with the case of two source signals and two observations only.

Here, we will focus our attention on the initialization of the separation algorithms and the permutation problem. The initialization procedure is very important since some separation algorithms are very sensitive to initial conditions, and often there are some frequency bins in which the separation algorithm fails to converge. Furthermore, a suitable initialization can achieve a spectacular reduction of permutation misalignments, since it favours the preservation of the order of the separated components in wide frequency blocks, which facilitate the permutation correction. A new permutation algorithm is also proposed based on the spectral coherence property of the speech signal. For that, we will derive a contrast function which maximization achieves the solution to the permutation. For the assessment of the developed algorithms, an exhaustive study has been performed in both synthetic measurements and real recordings, for various mixture environments.

2. Model of BSS of convolutive mixtures in the time-frequency domain

It is well known that any acoustic signal acquired from microphones in a real recording environment suffers from reflections on the walls and surfaces inside the room. In this sense, the recorded signals can be accurately modelled as a convolutive mixture, where the mixing filter is usually considered a high-order FIR filter. The standard convolutive mixing model of N sources, $s_j(n), j = 1, \cdots, N$, in a noiseless situation can be written as

$$x_i(n) = \sum_{j=1}^{N} \sum_{k=-\infty}^{\infty} h_{ij}(k) s_j(n-k), \quad i = 1, \cdots, N \ , \tag{1}$$

where $x_i(n), i = 1, \ldots, N$ are the N sensor signals, and $h_{ij}(n)$ is the impulse response from source j^{th} to microphone i^{th}. In order to blindly recover the original speech signals (sources) one can apply a matrix of demixing filters to the observations $x_i(n)$ that yields an estimate of each of the sources

$$y_i(n) = \sum_{j=1}^{N} \sum_{k=0}^{Q-1} b_{ij}(k) x_j(n-k), \quad i = 1, \cdots, N, \tag{2}$$

where the coefficients $b_{ij}(k)$ denote the impulse response of demixing system filter of Q taps.

The transformation of time-domain signals to the time-frequency domain is usually performed by the short-time Fourier transform (STFT). The main advantage of using the time-frequency domain is that convolutive mixture in Equation (1), can be approximated in

the time-frequency domain by a set of of instantaneous mixtures of complex values, one for each frequency bin, that is an easier problem for which many algorithms have been developed.

Let $X_i(f,t)$ and $S_i(f,t)$ be, respectively, the STFT of $x_i(n)$ and $s_i(n)$, and $H_{ij}(f)$ be the frequency response of the channel $h_{ij}(n)$. From Equation (1) we obtain

$$X_i(f,t) \simeq \sum_{j=1}^{N} H_{ij}(f) S_j(f,t), \quad i = 1, \cdots, N \ , \tag{3}$$

which can be rewritten, in matrix notation, as $\mathbf{X}(f,t) \simeq \mathbf{H}(f)\mathbf{S}(f,t)$, where the observation and source vectors for each time-frequency point are $\mathbf{X}(f,t) = [X_1(f,t), \ldots, X_N(f,t)]^T$ and $\mathbf{S}(f,t) = [S_1(f,t), \ldots, S_N(f,t)]^T$, respectively, and $\mathbf{H}(f)$ is the frequency response of the mixing filter whose elements are $H_{ij}(f) = [\mathbf{H}(f)]_{ij} \ \forall i,j$. The superscript T represents the matrix transpose operator. From now on, we will assume that the mixing matrices $\mathbf{H}(f)$ are full rank. In practice, the approximation (3) is considered valid when the length of the DFT is significantly greater than the length of the mixing filters Parra & Spence (2000). For instance, in fd-ICA context for speech separation it is common that the DFT is twice as long as the length of the mixing filters Araki et al. (2003).

Each of the separation matrices $\mathbf{B}(f)$ can be estimated independently with a suitable algorithm for instantaneous mixtures of complex values, which is computationally very efficient. The vector of outputs or estimated sources $\mathbf{Y}(f,t) = [Y_1(f,t), \ldots, Y_N(f,t)]^T$ is thus given by applying $\mathbf{B}(f)$ to the observations in each frequency bin,

$$\mathbf{Y}(f,t) = \mathbf{B}(f)\mathbf{X}(f,t). \tag{4}$$

Nevertheless, the simplification (3) has some disadvantages that need to be solved to successfully recover the sources. As each instantaneous separation problem is solved independently, the recovered signals will have an arbitrary permutation and scaling in each frequency bin. Those ambiguities are inherent to the problem of the blind source separation. In consequence, $\mathbf{Y}(f,t)$ is usually modelled as

$$\mathbf{Y}(f,t) \approx \mathbf{\Pi}(f)\mathbf{D}(f)\mathbf{S}(f,t) \ , \tag{5}$$

where $\mathbf{\Pi}(\mathbf{f})$ is a permutation matrix and $\mathbf{D}(f)$ is an arbitrary nonsingular diagonal matrix of complex scalars, representing respectively the permutation and scaling ambiguities.

The scaling ambiguity is not a serious problem. In fact, it causes an overall filtering of the sources. However, the correction of the permutation is essential. Even when perfect separation is achieved in all frequency bins, the transformation of the recovered signals into the time domain will be erroneous if the order of the extracted components are not the same in all frequency bins. Therefore, it is necessary to determine the permutation matrix $\mathbf{P}_*(f)$ in each frequency bin in such way that the order of the outputs remains constant over all the frequencies,

$$\mathbf{Y}(f_k,t) \leftarrow \mathbf{P}_*(f_k)\mathbf{Y}(f_k,t). \tag{6}$$

Once the separated components are well aligned, the sources can be finally recovered by converting the time-frequency representations $Y_j(f,t)$ back to the time domain. It is also possible to estimate the sources by first transforming the separation matrices $\mathbf{B}(f)$ to the time domain, correcting previously the ambiguities, and then by applying the Equation (2).

3. The separation stage

The most widely used methods for solving the instantaneous separation problems in the standard fd-ICA approach relies on the statistical independence among different sources and on the notion of contrast function. The statistical independence of the sources is a plausible assumption in real-room recordings, since each speakers acts independently of the others. On the other hand, the notion of contrast function defines a correspondence between the distribution of the estimated sources and the real line which is only maximized when the sources are mutually independent Comon (1994).

In the fd-ICA context, it is important to note that the separation algorithm must be capable of handling complex data, given that the separation problem is formulated in the time-frequency domain. Nowadays, most of the ICA methods that work with complex data often use a preliminary whitening step that leads to $\mathbf{Z} \equiv \mathbf{Z}(f)$ the spatially whitened observations. This preprocessing simplifies the problem and, in some cases, it is also used because it improves the convergence of the algorithm. The whitening procedure consists of a linearly transform of the observed variables to zero mean and unit variance, that can be accomplished by e.g. Principal Component Analysis (Comon, 1994). One of the most widely used algorithm is FastICA Hyvärinen & Oja (1997), which exploits the property of the non-Gaussianity of the sources. The extension to complex data was formulated in Bingham & Hyvärinen (2000). The solution is obtained by finding the extrema of the following contrast function

$$\Psi_{BH}(\mathbf{u}) = E\left[G\left(\left|\mathbf{u}^{\mathrm{H}}\mathbf{Z}\right|^2\right)\right] \quad \text{s. t. } E\left[\left|\mathbf{u}^{\mathrm{H}}\mathbf{Z}\right|^2\right] = 1\,, \tag{7}$$

where E represents expectation, $\mathbf{u}$ is the extraction vector (a row of the separating matrix $\mathbf{U}^{\mathbf{H}}$), while G is a smooth even function whose expectation measures the departure (in a given sense) from the Gaussian distribution. Some usual choices for function G can be found in Bingham & Hyvärinen (2000).

The optimization of the contrast function (7) is performed by the Newton's method, resulting the following update rule for the fixed-point algorithm for one unit

$$\begin{aligned} \mathbf{u}^{(i)} &= E\left[\mathbf{Z}\left(u^{(i-1)\mathbf{H}}\mathbf{Z}\right)^* g\left(|\mathbf{u}^{\mathrm{H}}\mathbf{Z}|^2\right)\right] - E\left[g\left(|\mathbf{u}^{\mathrm{H}}\mathbf{Z}|^2\right) + |\mathbf{u}^{\mathrm{H}}\mathbf{Z}|^2 g'\left(|\mathbf{u}^{\mathrm{H}}\mathbf{Z}|^2\right)\right]\mathbf{u}^{(i-1)} \\ \mathbf{u}^{(i)} &\leftarrow \frac{\mathbf{u}^{(i)}}{\left\|\mathbf{u}^{(i)}\right\|}\,, \end{aligned} \tag{8}$$

where i is the iteration index, and $g(\cdot)$ and $g'(\cdot)$ denote the first and the second derivatives of $G(\cdot)$, respectively. This method can be combined with a deflation procedure to retrieve all the original components. An optimized variant of FastICA consists of introducing an adaptive choice of function G. For this purpose, the distributions of the independent components can be modelled by a generalized Gaussian distribution. The resulting algorithm is called efficient FastICA or simply EFICA Koldovský et al. (2006).

The ICA algorithms previously commented ignore the time structure of source signals. However, for the speech signals, nearby samples are highly correlated and when comparing the statistics for distant samples the nonstationary behaviour is revealed. It is possible to exploit any of these features to achieve the separation using only second order statistics (SOS). One important advantage of the SOS based systems is that they are less sensitive to noise

and outliers. One popular method of this family of algorithms is the second order blind identification (SOBI) algorithm, proposed in Belouchrani et al. (1997).

Under the assumption of spatial decorrelation of the sources, the correlation matrices of the sources $\mathbf{R}_s(\tau) = E[s(t+\tau)s^*(t)]$ for any nonzero time lag τ are diagonal, where superscript * denotes conjugate operation. If we consider now time-delayed correlation matrices of whitened observations, the next relation for prewhitened sensor signals is satisfied

$$\mathbf{R}_z(\tau) = \mathbf{W}\mathbf{R}_x(\tau)\mathbf{W}^H = \mathbf{U}\mathbf{R}_s(\tau)\mathbf{U}^H \tag{9}$$

where $\mathbf{W}$ is the whitening matrix and $\mathbf{U}$ is the unitary mixing matrix. Since $\mathbf{R}_s(\tau)$ is diagonal, the separation matrix $\mathbf{U}^H$ may be estimated by enforcing an unitary diagonalization of a covariance matrix $\mathbf{R}_z(\tau)$ for some non zero lag. Instead of use only one time lag, SOBI approximated jointly diagonalizes a set of covariance matrices computed for a fixed set of time lags.

An extension of this algorithm that jointly exploits the non-stationary and the temporal structure of the source signals is second-order non-stationary source separation (SEONS) algorithm proposed in Choi & Cichocki (2000). This method estimates a set of covariance matrices at different time-frames. For that, the whitened observations are divided into non-overlapping blocks, where different time-delayed covariance matrices are computed. Then, a joint approximate diagonalization method is applied to this set of matrices to estimate the separation matrix. The application of the SEONS algorithm in the simulations of this chapter considers covariance matrices for $\tau = 0$ and one sample in each block.

3.1 The ThinICA algorithm

The higher order cumulants of the outputs have been one of the first class of contrast functions proposed in the context of blind deconvolution Donoho (1981) and later extensively used in the context of blind source separation Comon (1994); Cruces et al. (2004a). In its simpler form, the contrast function takes the form of a sum of the fourth-order cumulants of the outputs

$$\Psi(\mathbf{U}) = \sum_{i=1}^{N} \left|Cum\left(Y_i(t), \cdots, Y_i(t)\right)\right|^2, \tag{10}$$

subject to a unitary constraint on the separation matrix ($\mathbf{U}^H\mathbf{U} = \mathbf{I}$). Indeed, the first implementation of the Fast-ICA algorithm Hyvärinen & Oja (1997) considered the maximization of (10). Nearly at the same time, other authors developed in DeLathauwer et al. (2000) a higher-order power method that consider the separation of the sources with a contrast function based on a least squares fitting of a higher-order cumulant tensor.

The ThinICA algorithm was proposed in Cruces & Cichocki (2003) as a flexible tool to address the simultaneous optimization of several correlation matrices and/or cumulant tensors. These matrices and tensors can be arbitrarily defined, so the algorithm is able to exploit not only the independence of the sources, but also their individual temporal correlations and also their possible large-term non-stationarity Cruces et al. (2004b).

For simplicity, it is assumed that sources are locally stationary and standardized to zero mean and unit variance. In order to determine the statistics that take part in the ThinICA contrast function one should specify the order of the cumulants q and the considered time tuples $\theta = (t_1, \cdots, t_q)$ which are grouped in the set $\Theta = \{\theta_m \in \mathbb{R}^q, m = 1, \cdots, r\}$. The algorithm works

with positive weighting scalars w_θ and with q unitary matrix estimates $\mathbf{U}^{[k]}, k = 1, \cdots, q$, of the mixing system **WA** and their respective linear estimates $\mathbf{Y}^{[k]}(t) = \mathbf{U}^{[k]H}\mathbf{Z}(t),\ k = 1, \ldots, q$, of the vector of desired sources. It was shown in Cruces et al. (2004b) that the function

$$\Phi_\Theta(\mathbf{U}^{[1]}, \ldots, \mathbf{U}^{[q]}) = \sum_{i=1}^{N} \sum_{\theta \in \Theta} w_\theta \left| Cum\left(Y_i^{[1]}(t_1), \cdots, Y_i^{[q]}(t_q)\right)\right|^2 , \tag{11}$$

is a contrast function whose global maxima are only obtained when all the estimates agree ($\mathbf{U}^{[1]} = \cdots = \mathbf{U}^{[q]}$) and the sources of the mixture are recovered. Moreover, the constrained maximization of the previous contrast function is equivalent to the constrained minimization of the weighted least squares error between a set of q-order cumulant tensors of the observations $\{\mathcal{C}_q^{\mathbf{Z}}(\theta), \forall\theta \in \Theta\}$ and their best approximations that take into account the mutual independence statistical structure of the sources $\{\hat{\mathcal{C}}_q^{\mathbf{Z}}(\mathbf{D}_\theta, \mathbf{U}^{[1]}, \ldots, \mathbf{U}^{[q]}), \forall\theta \in \Theta\}$. If $\mathbf{D}_\theta$ denote diagonal matrices, the maximization of (11) is equivalent to the minimization of

$$\epsilon_\Theta(\mathbf{U}^{[1]}, \ldots, \mathbf{U}^{[q]}) = \sum_{\theta \in \Theta} w_\theta \min_{\mathbf{D}_\theta} \|\mathcal{C}_q^{\mathbf{Z}}(\theta) - \hat{\mathcal{C}}_q^{\mathbf{Z}}(\mathbf{D}_\theta, \mathbf{U}^{[1]}, \ldots, \mathbf{U}^{[q]})\|_F^2 \tag{12}$$

with respect to the unitary matrices $\mathbf{U}^{[k]}, k = 1, \cdots, q$. See Cruces et al. (2004b) for more details on the equivalence between the contrast functions (11) and (12).

The optimization of the ThinICA contrast function can be implemented either hierarchically or simultaneously, with respective implementations are based on the thin-QR and thin-SVD factorizations. A MatLab implementation of this algorithm can be found at the ICAlab toolbox icalab (2012), or obtained from the authors upon request.

The ThinICA contrast function and the algorithm has been also extended in Durán & Cruces (2007) to allow the simultaneous combination of correlation matrices and cumulant tensors of arbitrary orders. In this way, the algorithm is able to simultaneously exploit the information of different statistics of the observations, what makes it suitable for obtaining accurate estimates from a reduced set of observations.

The application of the ThinICA algorithm in the simulations of this chapter tries to exploit the non-stationarity behavior of the speech signals by considering $q = 2$ and the set $\Theta = \{(t_m, t_m) \in \mathbb{R}^q, m = 1, \cdots, r\}$, i.e., it uses the information of several local autocorrelations of the observations in frequency domain in order to estimate the latent sources.

4. Initialization procedure for ICA algorithms

The ICA algorithm used for estimating the optimal separation system in each frequency bin, is often randomly initialized. However, there are several advantages to make a suitable initialization of the algorithm. For instance, if the algorithm is initialized near the optimal solution, one can guarantee a high convergence speed. Also, the permutation ambiguity can be avoided if the mixture have some properties.

One interesting approach to develop an appropriate initialization method is to consider the continuity of the frequency response of the mixing filter $\mathbf{H}(f)$ and its inverse. Under this assumption, it seems reasonable to initialize the separation system $\mathbf{B}_{ini}(f)$ from the value of the optimal separation system at the previous frequency $\mathbf{B}_o(f-1)$. However, we can not directly apply $\mathbf{B}(f) = \mathbf{B}_o(f-1)$ in those separation algorithms that whiten the observations as a preprocessing step.

The whitening is performed by premultiplying the observations with an $N \times N$ matrix $\mathbf{W}(f)$ as $\mathbf{Z}(f,t) = \mathbf{W}(f)\mathbf{X}(f,t)$, where $\mathbf{W}(f)$ is chosen so as to enforce the covariance of $\mathbf{Z}(f,t)$ to be the identity matrix $\mathbf{C}_Z(f,t) = \mathbf{I}_N$. The computation of the whitening matrix can be accomplished by e.g. Principal Component Analysis (Comon, 1994). After that, the new observations $\mathbf{Z}(f,t)$ can be expressed as a new mixture of the sources through a new unitary mixing matrix $\mathbf{U}_o(f) = \mathbf{W}(f)\mathbf{H}(f)$,

$$\mathbf{Z}(f,t) = \mathbf{U}(f)\mathbf{S}(f,t). \tag{13}$$

Given an estimate of the unitary mixing matrix $\mathbf{U}_o(f)$, then it is immediate to see that the separation matrix $\mathbf{U}(f)^{-1} = \mathbf{U}(f)^H$ is also unitary. Therefore, the estimated components or outputs

$$\mathbf{Y}(f,t) = \mathbf{U}(f)^H\mathbf{Z}(f,t) = \mathbf{B}(f)\mathbf{X}(f,t) \tag{14}$$

yields the decomposition of the separation matrix $\mathbf{B}(f)$ as the product of an unitary matrix and the whitening system $\mathbf{B}(f) = \mathbf{U}(f)^H\mathbf{W}(f)$. Due to the variability of the sources spectra, even at contiguous frequencies, the whitening matrices $\mathbf{W}(f)$ and $\mathbf{W}(f-1)$ are different. Consequently, in general we violate the unitary assumption of $\mathbf{U}(f)$ by solving directly for $\mathbf{U}_{ini}(f)^H = \mathbf{B}_o(f-1)\mathbf{W}^{-1}(f)$.

An alternative method to initialize from previous solutions while avoiding the previously described problem, consists of initially preprocessing the observations at frequency f by the separation matrix determined for the previous frequency. This technique, referred on now as classical initialization, first computes the new observations as

$$\mathbf{X}_{new}(f,t) = \mathbf{B}(f-1)\mathbf{X}(f,t), \tag{15}$$

and then, determines the matrix $\mathbf{W}(f)$ which whitens these new observations. Finally, the separation matrices are obtained by any preferred ICA method on those new observations. In brief, this classical initialization method decomposes the overall separation matrix in the following three factors

$$\mathbf{B}(f) = \mathbf{U}(f)^H\mathbf{W}(f)\mathbf{B}(f-1). \tag{16}$$

Instead of this classical initialization, here we aim to exploit the continuity of the frequency response of the separation filter in a different way. We propose to initialize the separation system $\mathbf{B}_{ini}(f)$ from its joint closest value to a set of optimal separation systems already computed at nearby frequencies (Sarmiento et al., 2009; 2010) This leads to the following constrained minimization problem

$$\underset{\mathbf{U}(f)^H}{\arg\min} \sum_i \alpha_i \|\mathbf{B}(f-i) - \mathbf{B}(f)\|_F^2 \qquad \text{s.t. } \mathbf{U}(f)^H\mathbf{U}(f) = \mathbf{I}_N \ , \tag{17}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and α_i are weights assigned to the separation matrices of nearby frequencies. This problem can be solved by applying Lagrange multipliers, where the corresponding Lagrangian function $\mathcal{L}$ is given by

$$\mathcal{L} = \mathrm{Tr}\Big\{\sum_i \alpha_i \left[\Big(\mathbf{B}(f-i) - \mathbf{U}(f)^H\mathbf{W}(f)\Big)^H \Big(\mathbf{B}(f-i) - \mathbf{U}(f)^H\mathbf{W}(f)\Big)\right] \\ -\mathbf{\Lambda}\Big(\mathbf{U}(f)^H\mathbf{U}(f) - \mathbf{I}_N\Big)\Big\}, \tag{18}$$

where $\mathbf{\Lambda}$ is the Hermitian matrix of multipliers and $\mathrm{Tr}\{\cdot\}$ denotes the trace of the argument. The minimization of the Lagrangian is obtained solving for $\mathbf{U}(f)$ from the equation

$$\nabla_{\mathbf{U}(f)^*}\mathcal{L} = -\left[\sum_i \alpha_i \mathbf{W}(f)\left(\mathbf{B}(f-i)^H - \mathbf{W}(f)^H\mathbf{U}(f)\right) + \mathbf{U}(f)\mathbf{\Lambda}\right] = \mathbf{0}_N, \tag{19}$$

where $\mathbf{0}_N$ denotes null matrix of dimension $N \times N$. After some manipulations, one obtains the desired solution

$$\mathbf{U}_{ini}(f)^H = \mathbf{Q}_R\mathbf{Q}_L^H. \tag{20}$$

where $\mathbf{Q}_L$ and $\mathbf{Q}_R$ are, respectively, the left and right singular vectors of the following factorization

$$[\mathbf{Q}_L, \mathbf{D}, \mathbf{Q}_R] = \mathrm{svd}\left(\mathbf{W}(f)\sum_i \alpha_i \mathbf{B}(f-i)^H\right). \tag{21}$$

As we will se below, this initialization procedure helps to preserve the ordering of the separated components across the frequencies. However, we can not guarantee that all the frequencies will be correctly aligned. In fact, in the audio context, the mixing filters, and therefore the demixing filters, can contain strong echoes. Thus, in general, the assumption of continuity of the filter frequency response is not valid in all the frequency bins.

Furthermore, it can exist some isolated frequency bins in which the separation problem is ill conditioned, and in consequence, the estimated separation matrices should not correspond to the optimal solution. Despite those aspects, in practice, the initialization procedure can achieve a spectacular reduction of permutation misalignments when it is applied to various ICA separation algorithms.

In order to corroborate this point, we now present various 2×2 separation experiments. In Figure 1, we show the number of transitions in the ordering of the estimated components when we apply both, the classical and the initialization procedures aforementioned to various standard ICA algorithm that whitens the observations to the estimation of the separation matrices. For comparison, we have selected three representative ICA algorithms: ThinICA, SEONS and EFICA. As it can be seen, the initialization overperforms the classical procedure, achieving a drastic reduction in the number of permutations in all the cases. Although it is possible to take into account the separation matrices from several frequencies to estimate the initial separation matrix, in our experience the best performing initialization is achieved when we use only one preceding frequency.

The initialization procedure also preserves the ordering of the separated components in wide frequency blocks. This last property is illustrated in Figure 2, where it is shown the spectrograms of the original and estimated components from a simulation for separating two speech sources from a synthetic convolutive mixture. The estimated components have been obtained by using the ThinICA algorithm initialized with the procedure described above, but without correcting the permutation ambiguity. In this simulation there are only four transitions in the order of the estimated components, where it is easy to see that the components are well aligned in wide frequency blocks. This property is particularly very interesting for our purposes, because it could be used to alleviate the computational burden of the algorithms that solve the permutation problem, although this issue will not be discussed in this chapter.

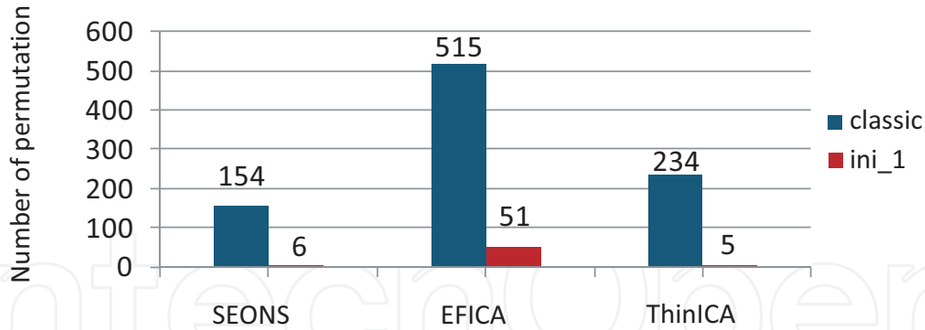


Fig. 1. Number of transitions in the ordering of the estimated components by applying the classical and the initialization procedure to several ICA separation algorithms, explored on synthetic 2×2 convolutive mixtures. Results are the average number over 10 different mixtures.

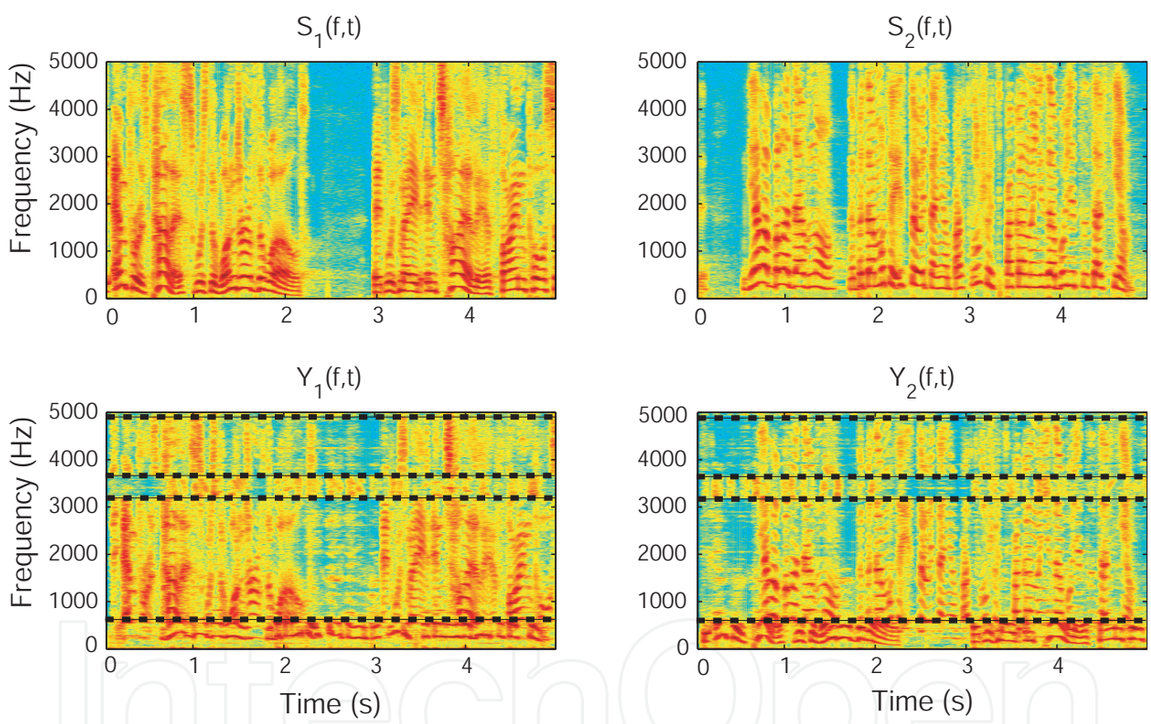


Fig. 2. Spectrograms estimated by the ThinICA algorithm using the initialization procedure. In the first row it is shown the spectrogram of the two original speech sources, whereas in the second row it is shown the estimated spectrograms. There are only four frequencies in which the order is not preserved, indicated by a dotted line.

5. Avoiding the indeterminacies

As we described above, due to the decoupled nature of the solutions across different frequencies, the correspondence between the true sources and their estimates, in general suffers from scaling and ordering ambiguities. Hereinafter, we describe some existing method to try to avoid these ambiguities.

5.1 The scaling ambiguity

The scale ambiguity can be fixed by setting some constraints to the separation filters or by using some *a priori* knowledge of the source signals. One option is to constraint the separating matrices to have unit determinant Smaragdis (1998), whereas another one is to constraint the diagonal elements of the separating matrices to unity Parra & Spence (2000). However, the most extended option is based on the minimal distortion principle introduced in Matsuoka & Nakashima (2001). The goal of this procedure is to obtain the signals as received by the microphones, that is, including the distortion of the mixing system while not adding other distortion effects. The solution consists of multiplying the separation matrices in each frequency bin by the diagonal of the matrix $\mathbf{B}(f)^{-1}$

$$\mathbf{B}(f) \leftarrow \mathrm{diag}\{\mathbf{B}(f)^{-1}\}\mathbf{B}(f). \tag{22}$$

5.2 The permutation ambiguity

Nowadays the permutation ambiguity constitutes the main problem in fd-ICA of acoustic signals, and it is still not satisfactory solved in high reverberating environments or for a large number of sources and observations. In order to tackle the problem, it is necessary to take some aspects into consideration. First, it is important to note that, when there are N sources in the mixtures, there are $N!$ possible permutations in each frequency bins, so the problem becomes difficult as the number of sources increase. In fact, a great number of the existing methods work only on the 2×2 case, and unfortunately, such methods cannot be directly extended to the general $N \times N$ case. On the other hand, in general, we can not guarantee the optimal solution of the instantaneous separation problem in all frequency bins, since the source signals are not homogeneous in their statistical properties along different frequencies. Therefore, there will be some frequencies in which the estimated sources do not correspond to the original sources. This can affect hardly the robustness of the permutation correction algorithms, and often it causes fatal errors in some of the existing methods.

The general structure of the permutation correction algorithms is presented below. The main goal of the permutation correction algorithms consist of estimating a set of permutation correction matrices, one for each frequency bin, $\mathbf{P}_{:} = \left\{\mathbf{P}_{f_1},\ \mathbf{P}_{f_2}, \cdots,\ \mathbf{P}_{f_{n_F}}\right\}, \mathbf{P}_{f_k} \in \mathcal{P}$, where n_F is the number of frequency bins and $\mathcal{P}$ represents all the possible permutation matrices of dimension $N \times N$. Those permutation matrices are applied either to the outputs $\mathbf{Y}(f,t)$ or to the separation filters $\mathbf{B}(f)$ to fix the permutation problem.

If we denote $\mathbf{\Pi}_{:} = \left\{\mathbf{\Pi}_{f_1}, \mathbf{\Pi}_{f_2}, \cdots,\ \mathbf{\Pi}_{f_{n_F}}\right\}, \mathbf{\Pi}_{f_k} \in \mathcal{P}$ a set of permutation matrices, one for each frequency bin, that describes mathematically the permutation ambiguity, then it is possible to define the set of global permutation matrices $\mathbf{Q}_{:} = \left\{\mathbf{Q}_{f_1},\ \mathbf{Q}_{f_2}, \cdots,\ \mathbf{Q}_{f_{n_F}}\right\}$, whose elements are $\mathbf{Q}_{f_k} = \mathbf{P}_{f_k}\mathbf{\Pi}_{f_k}$.

Then, it is immediate to deduce that the set $\mathbf{P}_{:}$ will be an optimal solution to the permutation problem if the corresponding set of global permutation matrices $\mathbf{Q}_{:}$ satisfy the following condition,

$$\mathbf{Q}_{f_1} = \mathbf{Q}_{f_2} = \cdots = \mathbf{Q}_{f_{n_F}} = \mathbf{Q}, \qquad ,\forall \mathbf{Q} \in \mathcal{P}, \tag{23}$$

which implies that the permutation problem has $N!$ possible optimal solutions.

5.2.1 Brief review of existing methods

Here, we present the main ideas emerged in last years to solve the permutation problem, putting special emphasis on the drawbacks and limitations of the techniques. Many different approaches have been proposed during last years. Basically, those methods are based on one of the following two assumptions, or in a combination of both (Pedersen et al., 2008): consistency of the spectrum of the recovered signals and consistency of the filter coefficients.

The first set of methods use the consistency of the spectrum of the recovered signals, which relies on the property of amplitude modulation correlation Anemüller & Kollmeier (2000) or simply co-modulation, of speech signals. This property refers to the spectrogram of a speech signal reveals that there is a pattern in the changes in amplitude at different frequency bins. This can be explained since the energy seems to vary in time in a similar way over different frequency bins, up to a gain factor. In fact, when a speaker starts talking, the power of the signal increases in a similar way at all the frequencies, and the same happens when the speaker stops talking, that is, the power decreases in a similar way in all the frequencies. This pattern is in general different for different speakers, at least at some parts of the recording. Therefore, it is possible to propose a permutation correction algorithm based on some evaluation procedure of the similarity between the envelopes of separated signals.

This idea has been used extensively to propose different methods. One option consists of adjusting the permutation between either adjacent or sorted frequency bins in a sequential order. The method was first proposed in Ikeda & Murata (1998); Murata et al. (2001), where the output components are ordered according to the highest correlation between the frequency bin to be order and a global envelope calculated with the already ordered frequency bins. However, the sequential approach has a major drawback, since an error while estimating the correct permutation at a frequency bin can be propagated to the rest of frequencies to be ordered. In Rahbar & Reilly (2005) it is proposed a dyadic hierarchical sorting scheme to prevent this situation.

Rather than solving the permutation problem *a posteriori*, some methods try to avoid it. One option consists of introducing some constraint that penalizes the permuted solutions in the separation problem in each frequency bin. For instance, in Anemüller & Kollmeier (2000) the separation and permutation problems are solved simultaneously, so the computational cost is limited. However, it does not work well in high reverberation environments. Another option proposed in Kim et al. (2006) is based on the concept of Independent Vector Analysis (IVA), which is an extension of ICA from univariate components to multivariate components. This method models the time-frequency representations of speech signals with a multivariate probability density function, and separates the fellow source components together. The contrast proposed is a multidimensional extension of the maximum likelihood (ML) approach. The method performs successfully in most conditions, recovering high quality speech signals. However, the convergence to local minima limits the robustness of the method, since in those cases it does not successfully separate all the components.

The second set of methods, which are based on the spectral coherence of the separation filters, includes methods based on the continuity and smoothness of the frequency response of the separation filters and methods based on the sparsity of the separation filters. The property of continuity and smoothness refers to the fact that the frequency response of the separation filters has not got abrupt transitions. Under this assumption, in Pham et al. (2003) the permutation is solved by checking if the ratio $\mathbf{R}(f, f-1) = \mathbf{B}(f)\mathbf{B}^{-1}(f-1)$ is close to a diagonal matrix, in which case the frequencies f and $f-1$ are well aligned. In a similar way,

in Asano et al. (2003) the permutation is corrected by minimizing a distance measure between the filters evaluated at contiguous frequencies. The main weakness of those methods is the propagation of error, since an error in one frequency bin can lead to wrong permutations over the rest of frequency bins to be solved.

The continuity of the separation filters is equivalent to constraint the separation filters to have short support in the time domain. This idea, proposed in Parra & Spence (2000) is based on the observation that the existence of permutations will produce time domain filters with greater lengths. Therefore, if we impose a short length on the separation filters in the separation stage, one can assume that the estimated filters will preserve the same order in all the frequencies. Unfortunately, this method tends to fail in reverberant acoustic environments since the acoustic filters are already quite long. A recent method introduced in Sudhakar & Gribonval (2009) uses the temporal sparsity of the filters to solve the permutation problem, where the sparsity means that the filters have few non-zero coefficients. The main idea is that the permutation errors decreases the sparsity of the reconstructed filters in the time domain, so it is possible to solve the permutation problem by maximizing the sparsity of the time domain demixing filters. This method has a high computational cost, and also only works in absence of the scaling ambiguity, which is not a realistic assumption.

Another family of methods, closed to the beamforming techniques, are based on the different direction of arrival (DOA) of source signals Parra & Alvino (2002). For that, it is necessary the knowledge of the geometry of the sensor array, and the distance between the microphones to be small enough to prevent the problem of spatial aliasing. Those methods assume that the direct path dominates the mixing filter response, and therefore the frequency response of the mixing filter from the i source to the j sensor can be approximately modelled as an anechoic model,

$$H_{ji}(f) = e^{j2\pi f \tau_{ij}}, \quad \tau_{ij} = \frac{d_{ij} \sin \theta_i}{c}, \tag{24}$$

where θ_i is the direction of arrival of the i source, d_{ij} is the distance between the microphones i and j, and c is the propagation speed of sound. Due to the coherence of the separation filter, some authors, as in Kurita et al. (2000); Saruwatari et al. (2003), assume that the quotient of the frequency response of the mixing filters between a given source and whatever two sensors will present a continuous variation with frequency, so this property is exploited to match the order of the components. However, a correct estimation of the DOAs is not always possible and the method tends to fail in high reverberation conditions or when the sources are near.

Finally, some methods combine the direction of arrival estimation with signal inter-frequency dependence to provide robust solutions, as in Sawada et al. (2004). The method, first fix the permutations by using the DOA approach in those frequencies where the confidence of the method is high enough. Then the remaining frequencies are solved by a correlation approach on nearby frequencies, without changing the permutation fixed by the DOA approach. This method has been extended when the geometry of the sensor array is unknown, in Sawada et al. (2005), and when spatial aliasing happens, in Sawada et al. (2006).

6. A coherence contrast for solving the permutation

In this section we present a method for solving the permutation problem in the general $N \times N$ case, based on the amplitude modulation correlation property of speech signals. For that, we will define a global coherence measure of the separated components that constitutes a contrast for solving the permutation problem Sarmiento et al. (2011). Then, the set of permutation

matrices to align the separated components are estimated in an iterative way by using a block-coordinate gradient ascent method that maximize the contrast.

First, we transform the profiles of the separated components in a logarithmic scale, since it will exhibit clearly the coherence property of speech signals. Given a source signal $s_i(k)$ and its STFT $S_i(f,t)$, the spectrogram in dB $|S_i|_{dB}(f,t)$ is defined as

$$|S_i|_{dB}(f,t) = 10\log_{10}|S_i(f,t)|^2 . \tag{25}$$

Consider now two source signals $s_i(k)$ and $s_j(k)$. The correlation coefficient between i component at f_k frequency, $|S_i|_{dB}(f_k,t)$, and j component at f_p frequency, $|S_j|_{dB}(f_p,t)$ is given by

$$\rho_{ij}(f_k,f_p) = \rho(|S_i|_{dB}(f_k,t),|S_j|_{dB}(f_p,t)) = \frac{r_{ij}(f_k,f_p)-\mu_i(f_k)\mu_j(f_p)}{\sigma_i(f_k)\sigma_j(f_p)}, \quad \in [-1,1], \tag{26}$$

where the cross correlation, mean and variance of the spectrograms are respectively

$$r_{ij}(f_k,f_p) = E\left[|S_i|_{dB}(f_k,t),|S_j|_{dB}(f_p,t)\right] \tag{27}$$

$$\mu_i(f_k) = E\left[|S_i|_{dB}(f_k,t)\right] \tag{28}$$

$$\sigma_i(f_k)^2 = E\left[|S_i|^2_{dB}(f_k,t)\right] - \mu_i^2(f_k). \tag{29}$$

Although, in general, the speech signals fulfil the co-modulation property , several authors have stated that the direct comparison between the separated components at different frequencies is not always efficient to solve the permutation problem, mainly owing to the fact that the inter-frequency correlation is degraded in certain conditions. In fact, one speech signal will have high correlation coefficients in nearby frequency bins, but this assumption is not always correct if the frequencies are far apart or the correlation is evaluated in certain frequency range, mainly at very low frequencies or very high frequencies (approximately over 5 kHz).To overcome this, we define the mean correlation coefficient $\rho_{ij}(f_k)$, as an averaged measure of the correlation coefficients, in other words, a measure of similarity between the i component at frequency f_k and the j component in all the frequencies

$$\rho_{ij}(f_k) = \frac{1}{n_F}\sum_{p=1}^{n_F}\rho_{ij}(f_k,f_p), \quad \in [-1,1], \tag{30}$$

where n_F is the number of frequency bins.

Due to the spectral properties of speech signals, it is reasonable to expect that the mean correlation coefficient between one source at any f_k and itself will be greater than if we compare with another different source. Therefore, given the set of sources $\mathbf{s}(k)$, one can deduce that the following property will be satisfied $\forall f_k$

$$\rho_{ii}(f_k) > \rho_{ij}(f_k), \qquad \forall i,j = 1,\cdots,N, j \neq i. \tag{31}$$

This property is illustrated in Figure 3, where we it is shown the mean correlation coefficients of a set of 3 sources, evaluated in all frequency bins. From Figure 3, we can see that the assumption of Equation (31) is clearly valid in most frequency bins, except at lower frequencies, as it was expected.

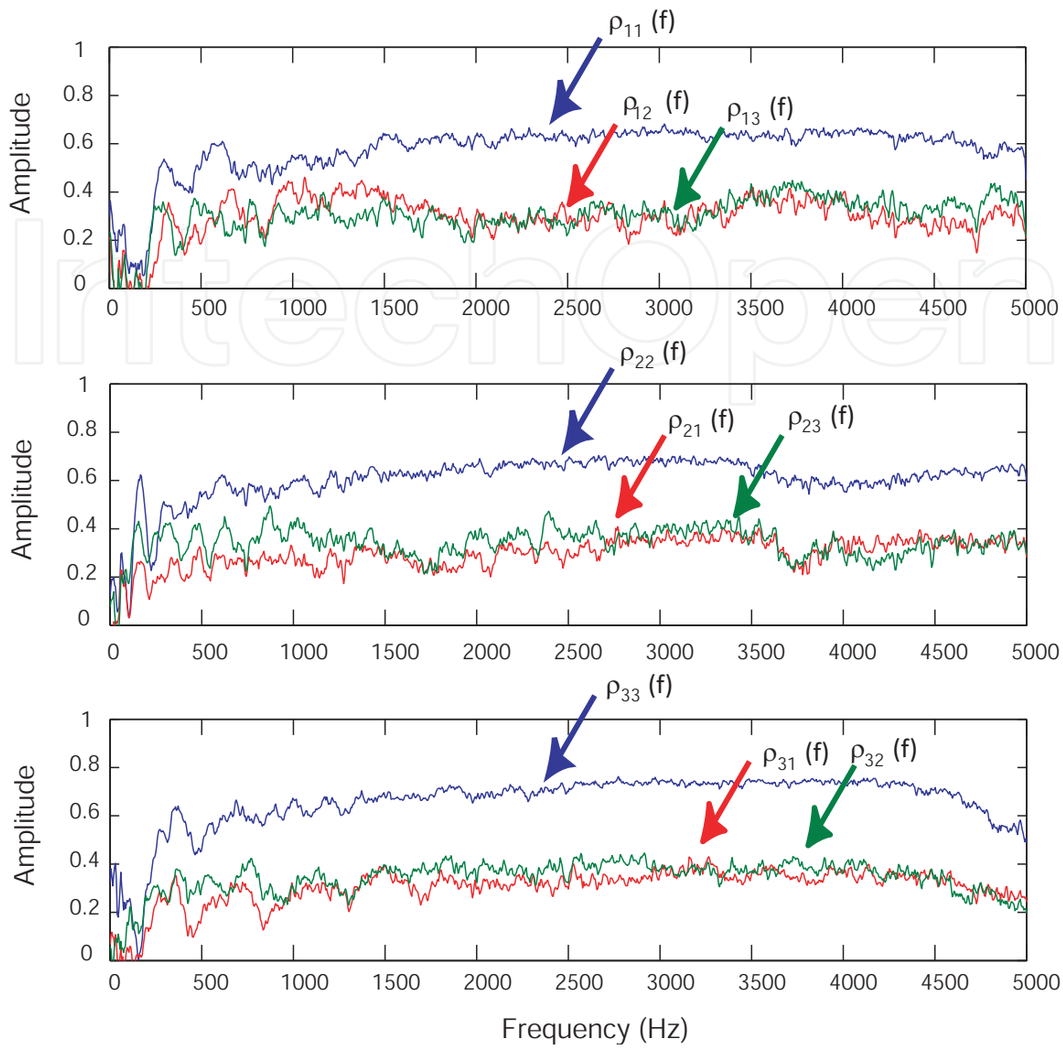


Fig. 3. *Mean correlation coefficients between one speech and itself, denoted as* $\bar{\rho}_{11}$*, and* 2 *other speech signals, denotes as* $\bar{\rho}_{12}$ *and* $\bar{\rho}_{13}$*.*

Considering the mean correlation coefficient, we define the global coherence of the source vector $\bar{\rho}$ as the average as in frequency as in components of the mean correlation coefficients, that is

$$\bar{\rho} = \frac{1}{N}\frac{1}{n_F}\sum_{i,f_k}\rho_{ii}(f_k) \qquad \in [-1,1]. \tag{32}$$

6.1 Description of the permutation correction algorithm

Consider the ordering vector $\pi_{f_k} = [\pi_{f_k}(1), \cdots, \pi_{f_k}(N)]$ associated to the existing permutation matrix $\mathbf{\Pi}_{f_k}$, defined in such way that its i element represents the non-nule row index of the i column of $\mathbf{\Pi}_{f_k}$. Therefore, the i component of the estimated source vector $\hat{S}_i(f_k,t)$ at frequency f_k corresponds to the component $\pi_{f_k}(i)$ of the output vector, that is

$$Y_{\pi_{f_k}(i)}(f_k,t) = \hat{S}_i(f_k,t). \tag{33}$$

In order to clarify this point, an example for $N = 3$ sources it is presented,

$$\mathbf{\Pi}_{f_k} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \rightarrow \pi_{f_k} = [3\ 1\ 2], \tag{34}$$

which means that $\hat{S}_1(f_k) = Y_3(f_k)$, $\hat{S}_2(f_k) = Y_1(f_k)$, y $\hat{S}_3(f_k) = Y_2(f_k)$. The global coherence of the outputs with alignment errors is given then by

$$\bar{\rho}(\mathbf{\Pi}_:) = \frac{1}{N}\frac{1}{n_F}\sum_{i,f_k}\left(\frac{1}{n_F}\sum_{f_p}\rho_{\pi_{f_k}(i),\pi_{f_p}(i)}(f_k,f_p)\right) \qquad \in [-1,1]. \tag{35}$$

From Equation (31) we can deduce that the global coherence of the outputs with alignment errors will be lower than the global coherence of the sources. Hence, it is possible to derive a contrast based on the global coherence which maximization achieves to solve the permutation problem. For that, we define, analogously to Equation (35), a global coherence of the corrected outputs

$$\bar{\rho}(\mathbf{Q}_:) = \frac{1}{N}\frac{1}{n_F}\sum_{i,f_k}\left\{\frac{1}{n_F}\sum_{f_p}\rho_{q_{f_k}(i)q_{f_p}(i)}(f_k,f_p)\right\} \qquad \in [-1,1], \tag{36}$$

where $q_{f_k}(i) = p_{f_k}(\pi_{f_k}(i))$, $i = 1,\ldots,N$ are the elements of the ordering global vector at frequency f_k. This global coherence will be maximum when the global permutation matrices satisfy the condition of Equation (23). Hence, the Equation (36) constitutes a coherence contrast for solving the permutation problem.

In order to calculate the permutation matrices that correct the alignment, it is necessary to solve the next constrained optimization problem

$$\mathbf{P}_: = \left\{\mathbf{P}_{f_1},\ \mathbf{P}_{f_2},\cdots,\ \mathbf{P}_{f_{n_F}}\right\} = \arg\max_{\mathbf{P}_:}\ \{\bar{\rho}(\mathbf{Q}_:)\}\,. \tag{37}$$

However, since it is not possible to find an analytical solution to the optimization problem, it is necessary to estimate the permutation correction matrices in an iterative way. Here we adopt a block coordinated ascent method, since it provides good permutation corrections with an efficient computational cost. In this method, at each iteration, the correction permutation matrices are calculated in a independent manner in all the frequencies as follows,

- Step 1: Calculate the mean correlation coefficients $\rho_{ij}^{(l)}(f_k)$ for all N separated components and for all frequency bins. The superindex (l) denotes the iteration index of the algorithm.

$$\begin{aligned}\rho_{ij}^{(l)}(f_k) &= \frac{1}{n_F}\sum_{p=1}^{n_F}\rho_{ij}^{(l)}(f_k,f_p) \\ &= \frac{1}{n_F}\sum_{p=1}^{n_F}\rho(|Y_i|_{dB}^{(l)}(f_k,t),|Y_j|_{dB}^{(l)}(f_p,t)).\end{aligned} \tag{38}$$

- Step 2: Find at each f_k the permutation matrix $P^{(l)}(f_k)$ that maximizes the sum of the mean correlation coefficients

$$\mathbf{P}_{f_k}^{(l)} = \underset{\mathbf{P}_{f_k} \in \mathcal{P}}{\arg\max} \sum_{i=1}^{N} \frac{1}{N} \rho_{p_{f_k}^{(l)}(i),i}(f_k). \tag{39}$$

- Step 3: If $\mathbf{P}_{f_k}^{(l)} \neq \mathbf{I}_N$ for any f_k, reorder the estimated components as

$$\mathbf{Y}^{(l+1)}(f_k, t) = \mathbf{P}^{(l)}(f_k)\mathbf{Y}^{(l)}(f_k, t), \tag{40}$$

  set the iteration index $l = l + 1$ and go to step 1. Otherwise, it is considered that the estimated components are well aligned and end the algorithm.

It is important to note that convergence to the optimal solution is not guaranteed. However, in practice, the convergence to local optima that provide highly erroneous solutions is highly improbable.

6.2 Performance in a perfect separation situation

Here we present some experiments that were conducted to illustrate the robustness of the proposed method in perfect separation situation. For that, we artificially applied randomly selected permutation matrices to a set of spectrograms of speech sources $\mathbf{S}(f, t)$ at each frequency bin. The result corresponds with the outputs of a frequency domain blind source separation scheme, when the separation is achieved perfectly in all the frequencies. We used speech sources of 5-seconds long sampled at 10 kHz, randomly chosen from the database of 12 individual male and female sources in sources (2012). The parameter for the calculus of the STFT were FFT length of 2048 points, Hanning windows of 1024 samples and 90 % overlap. Then, we used the correction permutation algorithm in order to recover the original spectrograms. In Table 1, it is presented the average number of unsolved permutations for a set of 30 different simulation for each configuration from $N = 2, \cdots, 8$ speech sources.

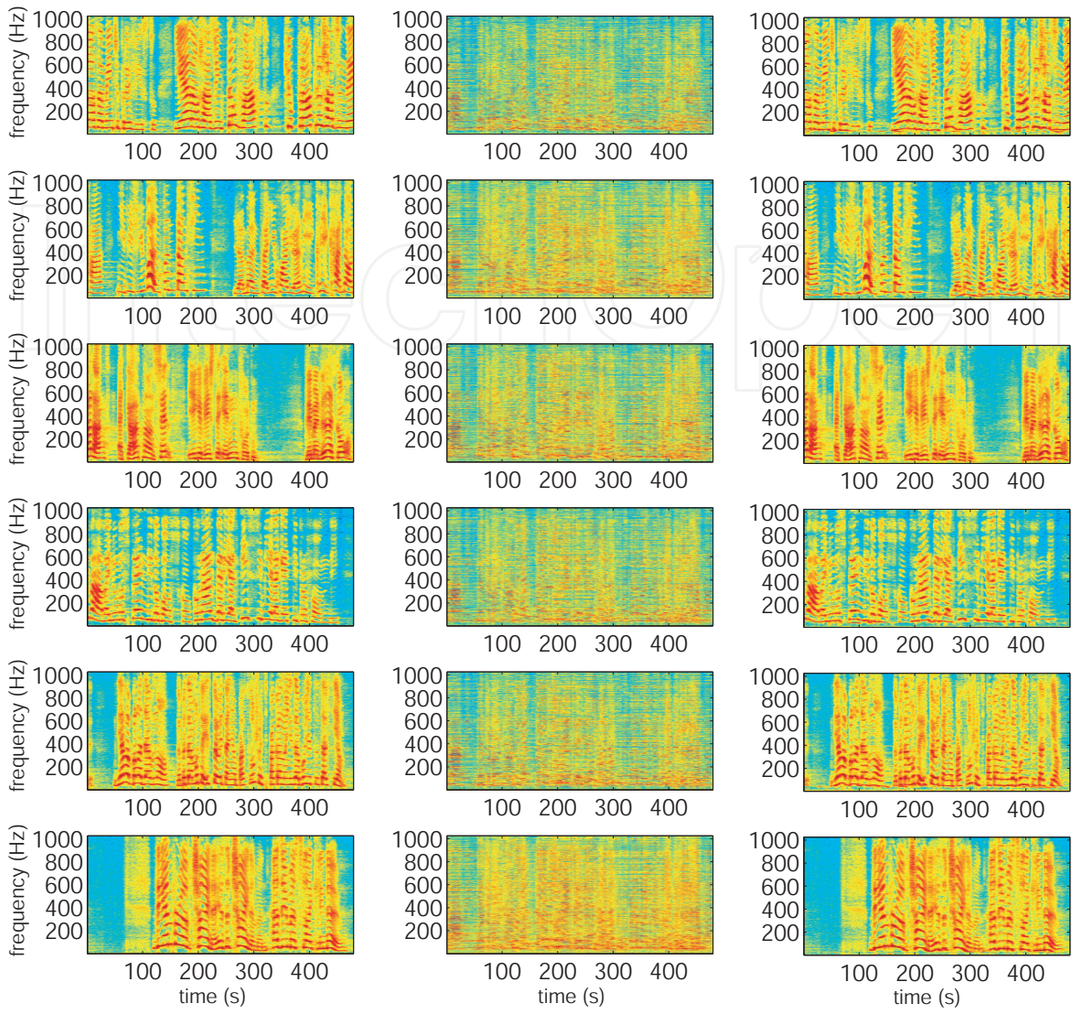
	2x2	3x3	4x4	5x5	6x6	7x7
Errors	1.87	2.67	7.87	12.73	14.07	24.13

Table 1. Performance of the permutation correction algorithm in perfect separation situation. Results are the averaged remaining unsolved permutations (errors) when the number of sources are $N = 2, \cdots, 8$ over 30 simulations.

In all the simulations the algorithm correctly order the frequency components, remaining some permuted solutions at lower frequencies as we expected, since the speech sources do not always satisfy the property of spectral coherence. However, those errors do not affect the quality of the recovered speech sources, since they are always located at very low frequencies. In Figure 4, we show the spectrograms of the original, permuted and recovered signals for one simulation of 6 speech sources, where it can be corroborated the robustness of the permutation correction algorithm. Another important feature of the algorithm is its capacity to order the source components by using a reduced number of iterations. For instance, in the previous experiment, the convergence was achieved in only four iterations.

7. Simulations

In this section we are going to test the performance of the initialization procedure and the permutation correction algorithm with both simulated and live recording by means of the



(a) Original sources (b) Permuted sources (c) Recovered sources

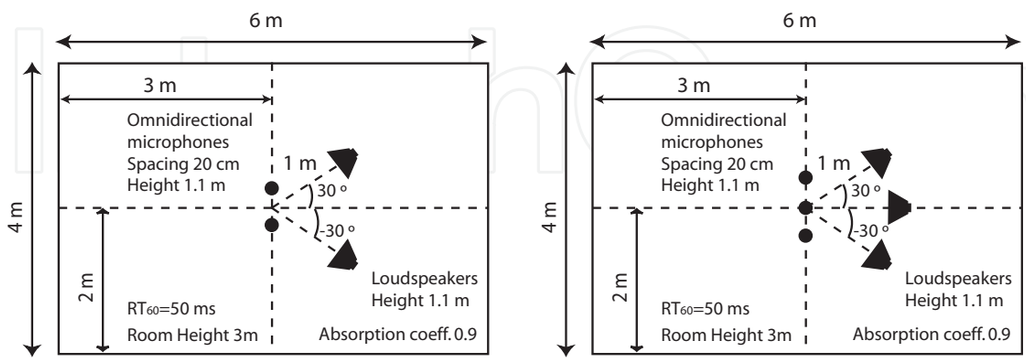
Fig. 4. Performance of the proposed correction permutation algorithm in perfect separation situation for $N = 6$ speech sources. For clarity, we have arranged the outputs according to the original sources.

quality of the recovered sources. This quality was measured in terms of both objective and perceptually measures. The objective measures are the Source to Distortion Ratio (SDR), the Source to Interferences Ratio (SIR) and the Source to Artifacts Ratio (SAR) computed by the BSS_EVAL toolbox, Fèvotte et al. (2006), whereas the perceptually measure is the Perceptual Evaluation of Speech Quality (PESQ) with a maximum value of 4.5. The Matlab code for calculating the PESQ index can be found in Loizou (2007).

7.1 Performance for simulated recording

For the synthetic mixtures, we considered the 2×2 and 3×3 mixing system for the configuration of microphones and loudspeakers showed in Figure 5. The corresponding channel impulse responses were determined by the Roomsim toolbox roomsim (2012). The

sources were randomly chosen from male and female speakers in a database of 12 individual recordings of 5 s duration and sampled at 10 KHz, available in sources (2012), for the 2×2 case, and in a database of 8 individual recordings from the Stereo Audio Source Separation Evaluation Campaign 2007 Vincent et al. (2007) of 10 second long sampled at 8 KHz for the 3×3 case.



(a) 2×2 simulated room recording (b) 3×3 simulated room recording

Fig. 5. Microphones and loudspeakers positions for the simulated recording rooms.

The separation experiments have been carried out as following. For the computation of the STFT, the parameters chosen were: Hanning windows of length 1024 samples, FFT of 2048 points and 90% overlap. Then, we estimated the separation matrices by initializing the ThinICA algorithm with the two procedures presented in Section 4: the classical initialization and the initialization with $k = 1$ preceding frequency, which will be referred from now on as $\text{ThinICA}_{\text{classic}}$ and $\text{ThinICA}_{\text{ini1}}$, respectively. After that, we fixed the permutation problem applying the method described in Section 6, and the scaling ambiguity by the Minimal Distortion Principle. Finally, we transformed the separation matrices back to the time domain and filtered the observations to obtain the time domain estimated sources and the quality of those signals was computed with the aforementioned methods. For comparison, we also carried out the same separation experiments by using IVA algorithm. The obtained results are presented In Table 2.

In Figure 6 is depicted an example of original sources, mixtures and demixed sources of one 3×3 separation experiment by using $\text{ThinICA}_{\text{ini1}}$ simulation configuration.

Note that, in the simplest case, the 2×2 case, the initialization procedure does not seem to introduce any significant improvement in the quality of the recovered sources respect to the classical procedure. This can be explained when the separation algorithm and the permutation correction algorithm adequately converge in all the experiments. Nevertheless, in the most complex 3×3 case, where the convergence of the separation algorithms can be more difficult to achieve, the initialization procedure over perform the classical procedure. Thus, one can conclude that the initialization procedure can obtain better performances in hard situations, mainly as the increment of the number of sources, or when it is available a reduced number of data. It is important to note that IVA method failed in three of the fifteen simulations. In those experiments IVA recovered only one sources, remaining the other two mixed.

From Table 2 we find another interesting result. The quality of the separated source obtained with fd-ICA methods by means of SIR, SAR and SDR ratios are better than those obtained

		SIR(dB)	SAR(dB)	SDR(dB)	PESQ
2x2	$\text{ThinICA}_{\text{classic}}$	19.87	17.57	15.35	3.05
	$\text{ThinICA}_{\text{ini1}}$	20.34	18.01	15.83	3.05
	IVA	14.93	13.46	10.75	2.94
3x3	$\text{ThinICA}_{\text{classic}}$	19.44	11.33	10.45	2.42
	$\text{ThinICA}_{\text{ini1}}$	22.32	11.88	11.31	2.54
	IVA	12.86	8.87	6.54	2.53
	IVA*	15.54	9.48	8.23	2.68

Table 2. Quality evaluation for 2×2 and 3×3 cases using various separation methods. Results were averaged over 23 mixtures in the 2×2 case, and 15 mixtures in the 3×3 case, except in IVA* case, where results from 3 simulations, in which IVA method failed, have been retracted.

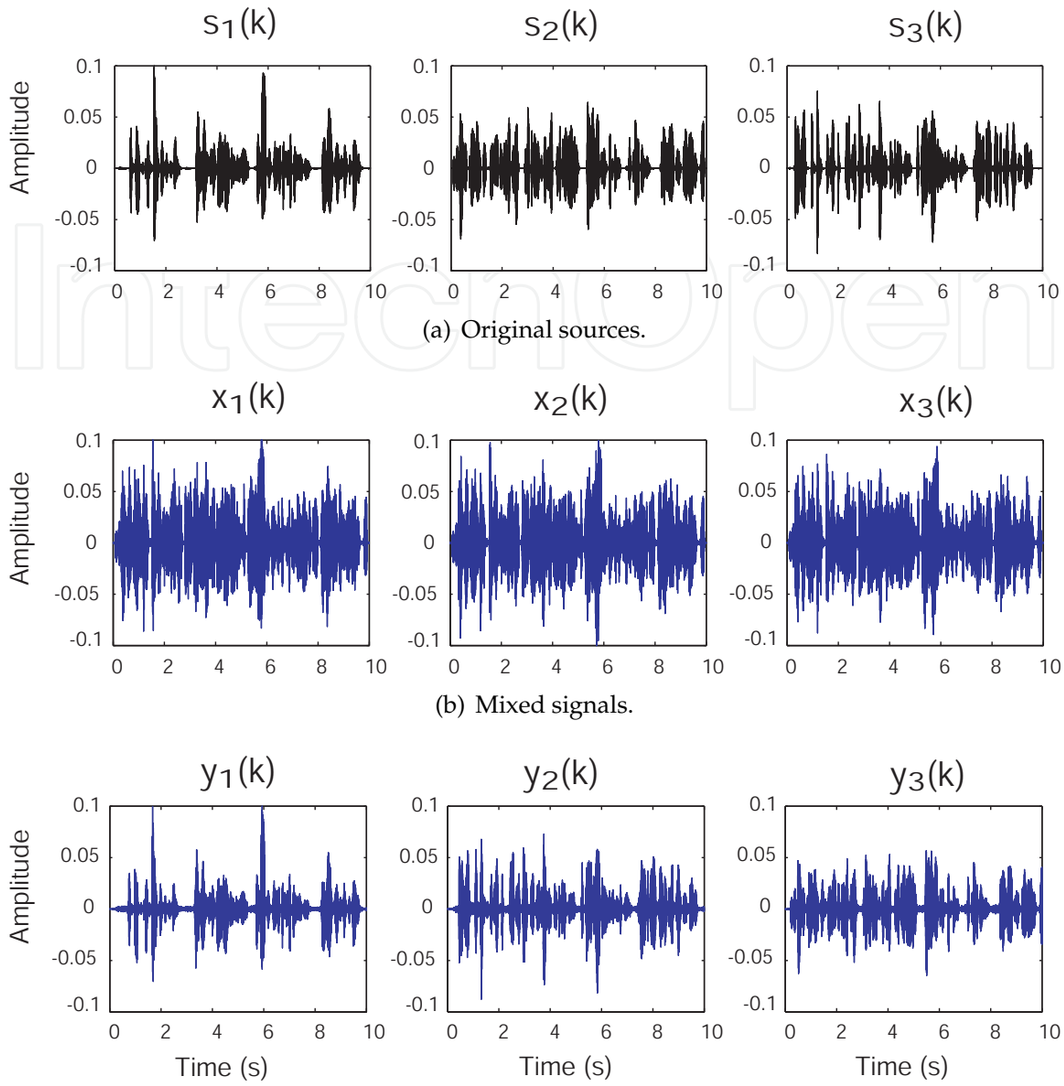
with IVA method. However, the PESQ quality measure obtained in all the cases is similar. This discrepancy can be explained by the conceptual differences between the different quality measures. In general, simulations show that fd-ICA methods obtain better separation ratios than the IVA method, despite hearing the recovered speech reveals that fd-ICA introduce more reverberation in the estimated sources than IVA method. This reverberation degrades the perceived quality of resulting sound, which explains the similar PESQ score of fd-ICA and IVA methods.

7.2 Performance for live recording

In this study, we have reproduce two clean speech sources in a typical office room to obtain a real recording in a noisy environment. A sampling frequency of 10 kHz has been used. The recording setup includes Logitech 5.1 Z-5450 loudspeakers, Shure Lavalier MX_ 180 Serie Microflex microphones and a Digi003Rack recording interface. The source signals were estimated by using both fd-ICA by means of ThinICA algorithm and IVA method. In this case, the STFT were computed using Hanning windows of length 2048 samples, FFT of 4096 points and 90% overlap.

For correctly interpret the results, it is important to note that the mixing conditions on live recordings present significant differences respect to the synthetic mixtures. One of the most important feature is the presence of additive noise coming from different noise sources in the recording environment, such us computers, air conditioning, etc. As a consequence, the estimated components will not correspond to the original sources, since they will have also a component of additive noise. Thus, we have included a component of additive noise in the objective quality measure. This noise component have been estimated in the silence periods of the recording. In Table 3 we show the obtained results. Due to computational limitations in the BSS EVAL toolbox, we present only the SIR and SNR ratios.

As it can be seen in Table 3, the three methods perform well in this situation, although the best performance is achieved by the ThinICA method including the initialization procedure. Moreover, fd-ICA methods present better SIR ratios than IVA method, as in the synthetic mixtures experiments. Also, in Figure 7 is depicted the original sources, real recordings and demixed sources by using $\text{ThinICA}_{\text{ini1}}$ simulation configuration.



(a) Original sources.

(b) Mixed signals.

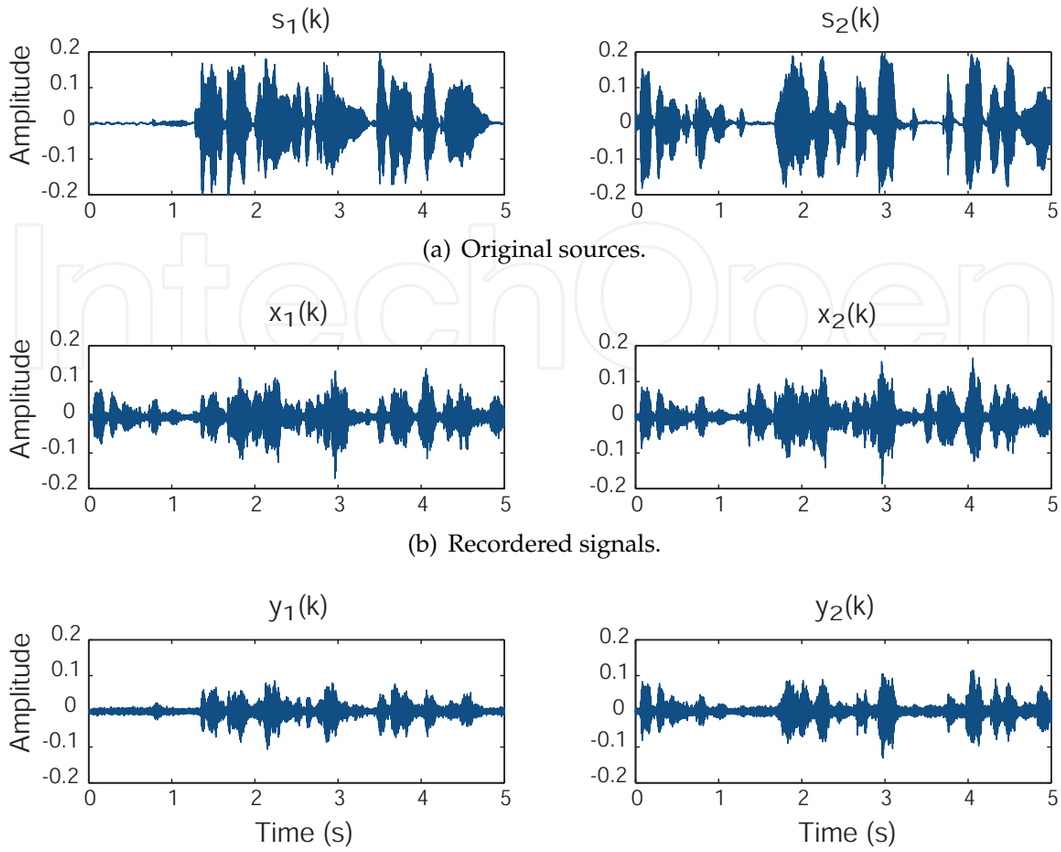
(c) Recovered signals.

Fig. 6. Example of 3×3 separation experiment by using $\text{ThinICA}_{\text{ini1}}$ simulation configuration.

	SIR(dB)	SNR(dB)	PESQ
$\text{ThinICA}_{\text{classic}}$	13.37	14.73	1.96
$\text{ThinICA}_{\text{ini1}}$	16.28	18.88	2.07
IVA	8.84	14.27	1.77

Table 3. SIR (dB), SAR (dB), SDR (dB) and PESQ index for 2×2 real recording separation.

To conclude, we have also applied the complete method to a live recording of 3 sources and 3 mixtures provided in SISEC (2012). The quality of the estimated sources was measured in terms of Source to Interferences Ratio (SIR) by E. Vincent, since the original sources are not public. An average SIR of 10.1 dB was obtained.



(a) Original sources.

(b) Recordered signals.

(c) Recovered signals.

Fig. 7. Example of a separation experiment with real recordered signals by using $\text{ThinICA}_{\text{ini1}}$ simulation configuration.

8. Conclusions

In this chapter we have considered the problem of the blind separation of speech signals recorded in a real room, when the number of speakers equals the number of simultaneous recordings. We have adopted the time-frequency approach, focusing our attention in the initialization of the separation algorithms and the permutation problem which is ubiquitous to fd-ICA methods. In order to improve the performance of the existing methods we have incorporated an initialization procedure for those ICA algorithms that work in the time-frequency domain and require the whitening of the observations as a preprocessing step. This initialization exploit the local continuity of the demixing filter in the frequency domain, which is a valid property for reverberant filters in a wide range of frequencies. For that, the separation matrix in one frequency bin is initialized from its joint closest value to a set of separation systems already computed at nearby frequencies. Computer simulations show that this initialization, when it is incorporated to the existing ICA algorithms, reduces drastically the number of permutations, preserving the separated components well aligned in wide frequency blocks. Simulations with more than two sources reveal that the proposed initialization also helps to the convergence of the ICA algorithms that solve the separation in each frequency.

The permutation problem becomes a severe problem when the number of sources is large or in high reverberant environments. Nowadays, it is still considered an open problem. For solving the permutation problem, we have present a method, based on the amplitude correlation modulation property of speech signals, that arises the general case of N sources and N observations. We have defined for each frequency bin a measure of coherence based on the amplitude modulation correlation property of speech signals. This measure has been used to formulate a coherence contrast function which maximization allows to successfully arrange the estimated components. An iterative method has been provide for searching the maxima of the contrast. The robustness of the algorithm has been illustrated for artificially permuted sources, which corresponds with a situation of perfect separation. Results show that the algorithm is able to reorder completely the frequency components, except for some very low frequencies that in some cases remained permuted. However, this does not affect to the quality of the recovered sources. Finally, experiments with simulated and live recording in a room with reverberation, for the case where two or three sources are mixed, show that the complete method improves considerably the performance of classical fd-ICA method, as well as IVA method, by means of both objective and perceptually measures.

9. Acknowledgements

This work has been supported by the Ministry of Science and Innovation project of the Government of Spain (Grant TEC2011-23559) and the Andalusian Government (Grant TIC-7869). We thank Emmanuel Vincent collaboration for the evaluation of the results.

10. References

Comon, P., Independent Component Analysis a new concept?, *Signal Processing*, Vol. 36, pp 287-314, 1994.

Smaragdis P. (1998),Blind separation of convolved mixtures in the frequency domain, *Neurocomputing*, Vol. 2, Nov. 1998, pp. 21-34.

Parra, L. & Spence, C., Convolutive blind source separation of non-stationary sources, *IEEE Trans. on Speech and Audio Processing*, May. 2000, pp. 320-327.

Araki, S., Mukai, R., Makino, S., Nishikawa, T. & Saruwatari, H. , The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech, *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 2, Mar. 2003, pp. 109-116.

Hyvärinen, A. & Oja, E., A fast fixed point algorithm for independent component analysis, *Neural Computation*,Vol. 9, 1997, pp. 1483-1492.

Bingham, E. & Hyvärinen, A., A fast fixed point algorithm for independent component analysis of complex valued signals, *International Journal of Neural Systems*, Vol. 10, No. 1, 2000, pp. 1-8.

Koldovský, Z., Tichavský, P. & Oja, E., Efficient variant of algorithm FastICA for independent component analysis attaining the Cramé Rao lower bound, *IEEE Trans. on Neural Networks*, Vol. 17, No. 5, Sep. 2006, pp. 1265-1277.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.F., Moulines, E., A blind source separation technique using second-order statistics, *IEEE Trans. on Signal Processing*, Vol. 45, No. 2, Feb. 1997, pp. 434-444.

Choi, S. & Cichocki, A., Blind separation of nonstationary sources in noisy mixtures, *IEEE Workshop on Neural Networks for Signal Processing (NNSP'2000)*, Sydney, Australia, Dec. 2000, pp. 11-13.

D. Donoho, On Minimun Entropy Deconvolution, *Applied Time Series Analysis II*, D. F. Findley Editor, Academic Press, New York, 1981,pp. 565-608.

Cruces, S., Cichocki, A. & Amari, S., From Blind Signal Extraction to Blind Instantaneous Signal Separation: Criteria, Algortihms and Stability, *IEEE Trans. on Neural Networks*, vol 15(4), July 2004, pp. 859-873.

De Lathauwer, L., De-Moor, B. & Vandewalle, J., On the Best Rank-1 and Rank-(R1;R2;...;RN) Approximation of Higher-order Tensors, *SIAM J. Matrix Anal. Appl.*, vol. 21(4), 2000, pp. 1324-1342.

S. Cruces & A. Cichocki, Combining Blind Source Extraction with Joint Approximate Diagonalization: Thin Algorithms for ICA, *Proceedings of the 4rd International Symposium on Independent Component Analysis and Blind Signal Separation*, Japan, 2003, pp. 463-468.

Cruces, S., Cichocki, A. & De Lathauwer, L., Thin QR and SVD factorizations for simultaneous Blind Signal Extraction, *Proceeding of the European Signal Processing Conference (EUSIPCO'04)*, Viena Austria, 2004, pp. 217-220

Available: http://www.bsp.brain.riken.jp/ICALAB/, Accessed 2012 Feb. 1.

Durán Díaz I. & Cruces, S., A joint optimization criterion for blind DS-CDMA detection, *EURASIP Journal of Applied Signal Processing, Special Issue: Advances in Blind Source Separation*, 2007, pp. 1-11.

Sarmiento, A.; Cruces, S. & Durán, I., Improvement of the initialization of time-frequency algorithms for speech separation,*Proceedings of Int Conf. on Independent Component Analysis and Blind Source Separation (ICA'09)*, 2009, pp. 629-636.

Sarmiento, A., Durán-Díaz & I., Cruces S., Initialization method for speech separation algorithms that work in the time frequency domain, *The Journal of the Acoustical Society of America*, Vol. 127, No. 4, 2010, pp. 121-126.

Matsuoka, K. & Nakashima, S., Minimal distorsion principle for blind source separation,*Proceedings of Int. Conf. on Independent Component Analysis and Blind Source Separation*, 2001, pp. 722-727.

Pedersen, M.S., Larsen, J., Kjems, U. & Parra, L.C., A survey of convolutive blind source separation methods, *Multichannel Speech Processing Handbook*, Eds. Jacob Benesty and Arden Huang, Springer 2007, Chapter 51, pp. 1065-1084.

Anemüller, J. & Kollmeier, B. , Amplitude modulation decorrelation for convolutive blind source separation, *Proceedings of Second International Workshop on Independent Component Analysis and Blind Signal Separation*, Jun. 2000, pp. 215-220.

Ikeda, S. & Murata, N., An approach to blind source separation of speech signals, *Proceedings of International Conference on Artificial Neural Networks (ICANN'98)*, Sep. 1998, Sweden, pp.761-766.

Murata, N., Ikeda,S. & and Ziehe, A., An approach to blind source separation based on temporal structure of speech signals, *Neurocomputing*, Vol. 41, Issue 1-4, Oct. 2001, pp.1-24.

Rahbar, K. & Reilly J.P., A frequency domain method for blind source separation of convolutive audio mixtures, *IEEE Transactions on speech and audio processing*, Vol. 13,No. 5, 2005, pp. 832-844.

Kim, T., Lee, I. & Lee T.W., Independent Vector Analysis: definition and algorithms, *Proceeding of Fortieth Asilomar Conference on Signals, Systems and Computers (ACSSC '06)*, 2006, pp.1393-1396.

Pham, D.T., Serviére, C. & Boumaraf, H. , Blind separation of convolutive audio mixtures using nonstationarity, *Proceedings of Int Conf. on Independent Component Analysis and Blind Source Separation (ICA'03)*, Nara, Japan, Apr. 2003.

Asano, F., Ikeda, S., Ogawa, M., Asoh, H.& Kitawaki, N., Combined approach of array processing and independent component analysis for blind separation of acoustic signals, *IEEE Transactions on Speech and Audio Processing*,Vol.11, No. 3, May. 2003, pp. 204- 215.

Sudhakar, P. & Gribonval, R., A sparsity-based method to solve permutation indeterminacy in frequency-domain convolutive blind source separation, *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation (ICA '09)*, pp. 338-345.

Parra, L.C. & Alvino, C.V., Geometric source separation: merging convolutive source separation with geometric beamforming, *IEEE Transactions on Speech and Audio Processing*, Vol.10, No.6, Sep. 2002, pp. 352- 362.

Kurita, S., Saruwatari, H., Kajita, S., Takeda, K. & Itakura, F., Evaluation of blind signal separation method using directivity pattern under reverberant conditions, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, Vol. 5, 2000, pp. 3140-3143.

Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T. & Shikano, K., Blind source separation combining independent component analysis and beamforming, *EURASIP Journal on Applied Signal Processing*, Jan. 2003, pp. 1135-1146.

Sawada, H., Mukai, R., Araki, S. & Makino, S., A robust and precise method for solving the permutation problem of frequency-domain blind source separation, *IEEE Transactions on Speech and Audio Processing*, Vol. 12, No. 5, Sept. 2004, pp. 530- 538.

Sawada, H., Mukai, R., Araki, S. & Makino, S., Frequency-domain blind source separation without array geometry information,*Proceedings of Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA'05)*, Mar. 2005.

Sawada, H., Araki, S., Mukai, R. & Makino, S., Solving the permutation problem of frequency-domain BSS when spatial aliasing occurs with wide sensor spacing,*Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, Vol. 5, May. 2006, pp. 77-80.

Sarmiento, A., Durán-Díaz, I., Cruces, S. & Aguilera, P., Generalized method for solving the permutation problem in frequency-domain blind source separation of convolved speech signals, *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH'11)*, Aug. 2011, pp. 565-568.

Available: http://www.imm.dtu.dk/pubdb/p.php?4400, Accessed 2012 Feb. 1.

Fèvotte, C., Gribonval, R., Vincent, E. ,BSS_ EVAL toolbox user guide, *Tech. Rep. 1706*, IRISA, Rennes, France, 2005, Available: http://www.irisa.fr/metiss/bss_ eval, Accessed 2012 Feb. 1.

Loizou, P.C., *Speech Enhancement. Theory and Practice*, CRC Press, 2007.

Campbell, D., Roomsim Toolbox, Available. http://media.paisley.ac.uk/ campbell/Roomsim/, Accessed 2012 Feb. 1

Vincent,E., Sawada, H., Bofill, P., Makino, S. & Rosca J. P., First Stereo Audio Source Separation Evaluation Campaign: Data, algorithms and results,*Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation (ICA'07)*, 2007, pp. 552–559.

Available: http://sisec2010.wiki.irisa.fr/tiki-index.php, Accessed 2012 Feb. 1.