A Geometrically Constrained ICA Algorithm for Blind Separation in Convolution Environments

Michele SCARPINITI, Francesco DI PALMA, Raffaele PARISI and Aurelio UNCINI

Infocom Department, “Sapienza” University of Rome

Abstract. In this paper a blind source separation algorithm in convolutive environment is presented. In order to avoid the classical permutation ambiguity in the frequency domain solution, a geometrical constraint is considered. Moreover a beamformer algorithm is integrated with the proposed solution: in this way the directivity pattern of the proposed architecture can take into account the residual permutation at low frequencies and the scaling inconsistency. Several experimental results are shown to demonstrate the effectiveness of the proposed method.

Keywords. Blind Source Separation, Convolutional environment, Frequency domain algorithms, Geometrical constraints, FastICA.

Introduction

An increasingly interest on Blind Source Separation (BSS) has arisen from signal processing researchers in the last fifteen years and a huge number of works were published [4,6]. A particular and emerging field of applications for signal processing is hands-free communication particularly useful in a lot of practical situations, such as a teleconference or a vehicle environment. In these applications a speech enhancement is necessary, because usually several sources are captured. In this way BSS is a promising technique for speech enhancement in adverse environments, like highly reverberant rooms.

While there are a lot of studies in the linear and instantaneous case, much poor is the range of works in convolutional environment. To achieve BSS of convolutive mixtures, several methods have been proposed [2,6]. In order to solve the BSS problem in a reasonable amount of time the problem is transformed into the frequency domain: the algorithm solves an instantaneous BSS problem for every frequency simultaneously [11,21].

Unfortunately in frequency domain two trivial ambiguities occur that could be particular troublesome [9,10]. The permutation ambiguity is particularly tiresome: when converting signal to time domain, contributions from different sources will appear into a single channel, thus destroying the separation achieved in the frequency domain; in addition the scaling indeterminacy at each frequency bin will result in an overall filtering of the sources. Different solutions to these problems can be found in literature [12,16].

1Corresponding Author: Infocom Department, “Sapienza” University of Rome, via Eudossiana 18, 00184 Roma; Tel. +39 0644585869; e-mail: michele.scarpiniti@uniroma1.it.
In order to solve these indeterminacies a geometrical constraint is introduced in this paper. The proposed solution is based on two steps: first of all a geometrically constrained algorithm, introduced and fully explained in [13], is adopted to lead the solution given by the standard FastICA algorithm towards the true solution; subsequently the scaling ambiguity and the residual permutation indeterminacy are solved by combining the algorithm with a beamforming solution, like proposed in [18]. This paper utilizes the proposed algorithm to enhance the speech signals captured from a microphone array.

The paper is organized as follows: section 1 introduces the BSS problem in the convolutive environment. Section 2 describes the proposed geometrical constraint that can solve the permutation indeterminacy; in addition an integration with a beamformer, that can solve the residual permutation at low frequencies and the scaling ambiguity, is described. Section 3 shows some experimental results while section 4 concludes the work.

1. Blind Source Separation for Convolutional Mixtures

Let us consider a set of $N$ unknown and independent sources $s(n) = [s_1(n), \ldots, s_N(n)]^T$, such that the components $s_i(n)$ are zero-mean and mutually independent. Signals received by an array of $M$ sensors are denoted by $x(n) = [x_1(n), \ldots, x_M(n)]^T$ and are called mixtures. For simplicity we consider the case of $N = M$.

The convolutive model introduces the following relation between the $i$-th mixed signal and the original source signals

$$x_i(n) = \sum_{j=1}^{N} \sum_{k=0}^{K-1} a_{ij}(k)s_j(n-k), \quad i = 1, \ldots, M \quad (1)$$

The mixed signal is a linear mixture of filtered versions of the source signals, $a_{ij}(k)$ represents the $k$-th mixing filter coefficient and $K$ is the number of filter taps. The task is to estimate the independent components from the observations without resort to a priori knowledge about the mixing system and obtaining an estimate $u(n)$ of the original source vector $s(n)$:

$$u_i(n) = \sum_{j=1}^{M} \sum_{l=0}^{L-1} w_{ij}(l)x_j(n-l), \quad i = 1, \ldots, N \quad (2)$$

where $w_{ij}(l)$ denotes the $l$-th mixing filter coefficient and $L$ is the number of filter taps.

When a mixing environment is quite complex, filters of the ICA network may require thousands of taps to appropriately invert the mixing. In such cases, the time domain methods have a large computational load to compute convolution of long filters and are expensive for updating filter coefficients. The methods can be implemented in the frequency domain using the Fast Fourier Transform (FFT) [15] in order to decrease the computational load because the convolution operation in the time domain can be performed by element-wise multiplication in the frequency domain. Note that the convolutive mixtures can be expressed as

$$x(f,k) = A(f)s(f,k), \quad \forall f \quad (3)$$
where $x(f, k)$ and $s(f, k)$ are the frequency components of mixtures and the independent sources at frequency $f$, respectively. $A(f)$ denotes a matrix containing elements of the frequency transforms of mixing filters at frequency $f$. From (3), it is clear that convolutive mixtures can be represented by a set of instantaneous mixtures in the frequency domain. Thus, the independent components can be recovered by applying ICA for instantaneous mixtures at each frequency bin and then transforming the results in the time domain:

$$u(f, k) = W(f)x(f, k), \quad \forall f \tag{4}$$

where $W(f)$ denotes the demixing matrix in the frequency domain. Note that $s(f, k)$, $x(f, k)$ and $u(f, k)$ are vectors of complex elements.

In order to solve the BSS in the convolutive environment Bingham & Hyvarinen in [3] have proposed a complex-valued version of the well-known and best performing FastICA algorithm introduced in [7,8]. After a canonical centering and whitening preprocessing, the column $w$ of the $W(f)$ matrix are obtained by maximizing the negentropy function [10], using the following equations

$$w^+ = E\left\{x(w^Hx)g(|w^Hx|^2)\right\} - E\left\{g(|w^Hx|^2) + |w^Hx|^2g'(|w^Hx|^2)\right\}w, \tag{5}$$

$$w = \frac{w^+}{\|w^+\|}, \tag{6}$$

where $g(\cdot)$ is a suitable nonlinear function and $g'(\cdot)$ its first derivative. We adopt the following function $g(y) = 1/(y + a)$, with $a$ a constant value, usually set to $a = 0.1$. Note that it is necessary to choose no learning rates in this algorithm.

2. The Geometrically Constrained Algorithm

The fundamental limitations of the convolutive BSS problem solved by a frequency domain approach are the frequency and scaling ambiguities: if the permutation is not consistent across frequency then converting the signal back to the time-domain will combine contributions from different sources into a single channel, and thus destroying the separation achieved in the frequency-domain (FD). In this paragraph we attempt to propose a solution exploiting some a priori information given by the knowledge of the geometrical position of the sources.

2.1. Geometrical Constraints

The aim of recovering permutations is a severe problem, especially when the number $N$ of sources is large. Several solution exist to the previous problem. Some authors proposed to consider the consistency of the filter coefficients, that can be achieved by requiring continuity of filter values in frequency domain [12]. An alternative way is to consider some geometrical informations on source positions.

Recently several works, like [1], show that frequency-domain blind source separation is equivalent to a set of frequency-domain adaptive beamformers (ABF) under certain conditions. This equivalence suggests a geometrical constrained approach: the so-
Figure 1. Assumptions on microphones geometry: $M_j$ is the $j$-th microphone while $S_k$ is the $k$-th source.

The solution of the ICA algorithm is projected on the ABF solution. In order to compute the beamformer solution, it is necessary to estimate the Direction of Arrivals (DOA). DOA estimation can be performed by Generalized Cross-Correlation (GCC) [14], MUSIC [19] or other techniques described in literature [5].

The introduction of the geometrical constraint gives a robustness to the algorithm that can not be achieved by the only ABF, that is very sensitive to small variations of parameters. On the other side ICA algorithm can suffer of the local minima problem that gives a solution far from the true one. The synergy between ICA and ABF can overcome these problems.

After a canonical centering and a whitening pre-processing with an orthogonal matrix $Q$

$$z = Qx,$$  

the demixing matrix can be written as $W = T^H \cdot Q$, where $T = [t_1, \ldots, t_N]^H$ is a rotation matrix that must be estimated by the FastICA algorithm in Eqs. (5) and (6).

Since the equivalence of the FD-ICA and FD-ABF described in [1], the demixing vector must be a scaled and whitened version of the ABF delay vector:

$$t_{opt} = t_{ABF} = \delta V \hat{h}_{target}.$$  

This follows by the constraint

$$w_{target}^H \hat{h}_{target} (f) = t_{target}^H Q \cdot \hat{h}_{target} (f) = c.$$  

The estimated delay vector, for each bin, has the form

$$\hat{h}_{target} (f) = \begin{bmatrix} \frac{1}{c} \\ \vdots \\ e^{j2\pi(N-1)d \delta \cos \hat{\theta}_{target}} \end{bmatrix},$$

where $d$ is the distance between microphones, $c$ the sound speed and $\hat{\theta}_{target}$ is the DOA estimated by a beamformer. Figure 1 describes the geometry of the sources and the microphone array.

Because we are interested only in the direction of the demixing vector $t_{target}$ and not to its norm, we project this solution to the constraint in Eq. (9), thus Eq. (6) becomes:
\[
t = \frac{t^+}{\|t^+ H \tilde{h}\|}. \tag{11}
\]

The performance of the proposed geometrically constrained algorithm is addressed by evaluating the angle between the estimated \( \hat{h}_k \) and true \( h_k \) delay vector, by

\[
\mu_k(f) = \cos^{-1} \left( \frac{h_k^H \hat{h}_k}{|h_k| \hat{h}_k} \right), \tag{12}
\]

and the angle between two delay vectors relative to different sources:

\[
\gamma_{kj}(f) = \cos^{-1} \left( \frac{h_k^H h_j}{|h_k| \hat{h}_j} \right). \tag{13}
\]

The proposed algorithm converges, if

\[
\mu_k(f) < \frac{\gamma_{kj}(f)}{2} = \mu_c(f) \quad \forall k, j. \tag{14}
\]

Unfortunately at low frequencies the \( \mu_c(f) \) parameter is very small and a certain number of permutations occur.

### 2.2. Combining ICA and Beamforming

In order to overcome the previous problem at low frequencies a DOA estimation is performed [17,18]. Such estimation is done analyzing the directional patterns which allow us to associate a single source to a local minimum, as described in [18].

The \( l \)-th directional pattern can be expressed, for the non-restrictive case of \( N = 2 \) sources and \( M = 2 \) microphones, as follows

\[
F_l(f, \theta) = \sum_{k=1}^{2} W_{lk}(f) \exp \left[ \frac{j2\pi f d \sin \theta}{c} \right], \tag{15}
\]

where \( c \) is the sound speed, \( d \) is the distance between microphones and \( W_{lk}(f) \) is the \( k \)-th entry of the \( w_l \) column vector. These directional patterns have a minimum value in correspondence of an estimated disturbing source [17,18]. In particular the two source directions are evaluated as

\[
\theta_1(f) = \min \left[ \arg \min_{\theta} |F_1(f, \theta)| , \arg \min_{\theta} |F_2(f, \theta)| \right],
\]

\[
\theta_2(f) = \max \left[ \arg \min_{\theta} |F_1(f, \theta)| , \arg \min_{\theta} |F_2(f, \theta)| \right]. \tag{16}
\]

Assuming that the minimum value of the pattern \( F_1(f_k, \theta) \) is \( \theta_2 \) for the \( k \)-th frequency bin \( f_k \), if for another frequency bin \( f_j \) the minimum value is \( \theta_1 \), then a permutations occurred (as shown in Figure 2) and the filter coefficients must be swapped. In addition the value of the directional pattern in correspondence of the source, the red points in Figure 2, can be used as scaling factor in order to solve the scaling ambiguity. An optional beamformer can be used for the evaluation of the directions \( \tilde{\theta}_1 \) and \( \tilde{\theta}_2 \).
The use of the directional patterns allows us to solve both the permutation and scaling ambiguity: the patterns decide to choose the ICA solution or a swapped one, if a permutation occurs. Then a generalization for $N > 2$ sources can be easily derived.

The proposed algorithm can be summarized as proposed in Table 1 and a block diagram is shown in Figure 3.

<table>
<thead>
<tr>
<th>Remove mean value from mixtures;</th>
<th>Whitening;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization: $t = Q \cdot h_{\text{target}}$;</td>
<td></td>
</tr>
<tr>
<td>repeat</td>
<td></td>
</tr>
<tr>
<td>FastICA;</td>
<td></td>
</tr>
<tr>
<td>Projection: $t = \frac{x^k}{|x^k \cdot Q \cdot h_{|}}$</td>
<td></td>
</tr>
<tr>
<td>until convergence</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.** Summary of the proposed algorithm.

3. Results

We tested our architecture in a number of real room environments characterized by a reverberation time $T_{60}$ in a range of $0 - 300$ ms. The impulse responses of the environment were simulated with the Matlab tool RoomSim².

In order to provide a mathematical evaluation of the output separation, different indexes of performance are available in literature. In this paper the signal to interference ratio (SIR) $SIR_j$ of the $j$-th source was adopted [20]:

$$SIR_j = 10 \log \left[ E \left\{ \left( |y_{\sigma(j),j} | \right)^2 \right\} / E \left\{ \sum_{k \neq j} \left( |y_{\sigma(j),k} | \right)^2 \right\} \right]. \quad (17)$$

²Roomsim is a MATLAB simulation of shoe-box room acoustics for use in teaching and research. Roomsim is available from http://media.paisley.ac.uk/~campbell/Roomsim/
In (17) \( y_{i,j} \) is the \( i \)-th output signal when only the \( j \)-th input signal is present, while \( \sigma(j) \) is the output channel corresponding to the \( j \)-th input. In addition we use the Performance Index, defined as:

\[
r = \frac{1}{N} \sum_{f=1}^{N_f} \left[ \left( \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{|H_{ij}(f)|}{\max_{j}|H_{ij}(f)|} \right) + \left( \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{|H_{ji}(f)|}{\max_{i}|H_{ji}(f)|} \right) - 2N \right],
\]

where \( \mathbf{H}(f) = \mathbf{W}(f)\mathbf{A}(f) \) is the matrix describing the overall mixing-demixing system, which ideally should be a permutation matrix, and \( N_f \) is the total number of frequency bins. A perfect separation is achieved if the performance index in (18) approaches zero. In this paper we propose some experimental tests in a standard 5.80 × 3.20 m\(^2\) room. The sources are far 1.15 m from the microphone array and microphones are spaced by 2 cm.

A first test with two male speakers and two microphones is proposed. The two speakers and microphones are located as shown in Figure 4 and their directions are \( \theta_1 = 45 \) and \( \theta_2 = 135 \) degrees. The directivity patterns for this experiment is proposed in the Figure 5. Table 2 summarizes the SIR of the two recovered sources and the Performance Index \( r \). As we can see from Table 2 the proposed architecture is able to recover speech signals very well, but the performance falls down dramatically when the reverberation time increases toward 300 ms.

A second test is conducted with three sources. The third source is an interference white noise located in front of the microphone array and thus at \( \theta_3 = 90 \) degrees, as shown in Figure 6. The algorithm is tested for a reverberation time into the range 0 ÷ 300 ms and results are summarized in Table 3. Also in this second test, as we can see from
Figure 5. Directivity pattern, for $\theta_1 = 45^\circ$ (first row) and $\theta_2 = 135^\circ$ (second row).

Figure 6. Placement of microphones in the test room in the case $N = 3$.

<table>
<thead>
<tr>
<th>$T_{00}$ [ms]</th>
<th>$SIR_1$ [dB]</th>
<th>$SIR_2$ [dB]</th>
<th>$SIR_3$ [dB]</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>21.68</td>
<td>25.73</td>
<td>22.14</td>
<td>0.21</td>
</tr>
<tr>
<td>100</td>
<td>18.51</td>
<td>17.23</td>
<td>17.89</td>
<td>0.32</td>
</tr>
<tr>
<td>150</td>
<td>15.64</td>
<td>13.17</td>
<td>14.73</td>
<td>0.41</td>
</tr>
<tr>
<td>300</td>
<td>6.54</td>
<td>4.21</td>
<td>5.03</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Table 3. Summary of the second proposed experimental test with $\theta_1 = 45^\circ$, $\theta_2 = 135^\circ$ and $\theta_3 = 90^\circ$. 
Table 4. Summary of the third proposed experimental test with $\theta_1 = 25^\circ$ and $\theta_2 = 155^\circ$.

<table>
<thead>
<tr>
<th>$T_{60}$ [ms]</th>
<th>$SIR_1$ [dB]</th>
<th>$SIR_2$ [dB]</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>32.46</td>
<td>30.17</td>
<td>0.22</td>
</tr>
<tr>
<td>100</td>
<td>20.13</td>
<td>21.72</td>
<td>0.38</td>
</tr>
<tr>
<td>150</td>
<td>15.66</td>
<td>13.89</td>
<td>0.51</td>
</tr>
<tr>
<td>300</td>
<td>6.18</td>
<td>6.32</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Table 5. Summary of the third proposed experimental test with $\theta_1 = 45^\circ$, $\theta_2 = 135^\circ$ (top rows), and $\theta_1 = 25^\circ$, $\theta_2 = 155^\circ$ (bottom rows) and an inaccuracy of $\pm 5^\circ$.

<table>
<thead>
<tr>
<th>$T_{60}$ [ms]</th>
<th>$SIR_1$ [dB]</th>
<th>$SIR_2$ [dB]</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>36.00</td>
<td>32.37</td>
<td>0.19</td>
</tr>
<tr>
<td>100</td>
<td>23.98</td>
<td>25.59</td>
<td>0.23</td>
</tr>
<tr>
<td>150</td>
<td>19.17</td>
<td>17.44</td>
<td>0.32</td>
</tr>
<tr>
<td>300</td>
<td>9.38</td>
<td>8.26</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Table 3, the separation is achieved. The introduction of a third source results in a small worsening of the overall performances.

A third test is proposed for two male speech located at $\theta_1 = 25^\circ$ and $\theta_2 = 155^\circ$ respectively. Results are summarized in Table 4. The experimental test for a different couple of directions shows similar results with respect the first one, as can be seen from a comparison of Tables 4 and 2. Hence the proposed approach is able to perform separation of sources in this case as well.

Finally we want to observe what happens if the estimation of the DOA is not so accurate. The following fourth experiment proposes a separation test for two male speech when the estimated DOAs $\hat{\theta}_1$ and $\hat{\theta}_2$ have an inaccuracy of $\pm 5^\circ$. Table 5 summarizes the obtained results. Comparing Table 5 with Tables 2 and 4 we can argue that the proposed solution is robust to a not so accurate estimate of the DOA directions.

4. Conclusions

This paper introduced a geometrically constrained algorithm for the frequency-domain blind source separation in convolutive environments. The proposed algorithm is able to reduce the permutation inconsistency, typical of the frequency-domain approaches.

Moreover an integration with a beamformer algorithm is made in order to solve completely the residual permutation ambiguity at the low frequencies, where the geometrical constraint is not enough. In addition this last addition can also solve the scaling ambiguity.

Several tests have been performed to verify the effectiveness of the proposed approach and demonstrate that it can solve the permutation and scaling indeterminacies very successfully. The quality of the separation has been evaluated in terms of the SIR Index and the Performance Index, which are widely diffused in literature.
Acknowledgements

This work has been partially supported by the Italian National Project: Wireless multiplatfOrm mimo active access netwoRks for QoS-demanding muLtimedia Delivery (WORLD), under grant number 2007R989S.

References


