Indeterminacy free frequency-domain blind separation of reverberant audio sources

Leandro Di Persia, *Student Member, IEEE,* Diego Milone, *Member, IEEE,* and Masuzo Yanagida

Abstract

Blind separation of convolutive mixtures is a very complicated task that has applications in many fields of speech and audio processing, such as hearing aids and man-machine interfaces. One of the proposed solutions is the frequency-domain independent component analysis. The main disadvantage of this method is the presence of permutation ambiguities among consecutive frequency bins. Moreover, this problem is worst when reverberation time increases. Presented in this paper is a new frequency-domain method, that uses a simplified mixing model, where the impulse responses from one source to each microphone are expressed as scaled and delayed versions of one of these impulse responses. This assumption, based on the similitude among waveforms of the impulse responses, is valid for a small spacing of the microphones. Under this model, separation is performed without any permutation or amplitude ambiguity among consecutive frequency bins. This new method is aimed mainly to obtain separation, with a small reduction of reverberation. Nevertheless, as the reverberation is included in the model, the new method is capable of performing separation for a wide range of reverberant conditions, with very high speed. The separation quality is evaluated using a perceptually designed objective measure. Also, an automatic speech recognition system is used to test the advantages of the algorithm in a real application. Very good results are obtained for both, artificial and real mixtures. The results are significantly better than those by other standard blind source separation algorithms.

Index Terms

Blind Source Separation, Reverberation, Independent Component Analysis, Speech Enhancement.

I. Introduction

The objective of source separation applied to sound sources is to obtain a set of signals approximating the original sound sources, given a set of sound field measurements [1]. When the mixture is produced inside an

Manuscript received __________ __, 2007; revised __________ __, 2008. This work was supported by the National Council for Science and Technology (CONICET), the National Agency for the Promotion of Science and Technology (ANPCyT) under projects ANPCYT-UNER PICT 11-12700 and ANPCYT-UNL PICT 11-25984, and National University of Litoral (UNL) under project CAID 012-72. The associate editor coordinating the review of this manuscript and approving it for publication was __________ __________.

L. Di Persia and D. Milone are with UNL and CONICET. Facultad de Ingeniería y Ciencias Hídricas (UNL): Ciudad Universitaria (CC 217), Ruta Nacional N 168 - Km. 472.4, Santa Fe (CP3000), Argentina. Tel.:+54-342-4575245 ext. 145. Fax:+54-342-4575224. (e-mail: ldipersia@fich.unl.edu.ar)

M. Yanagida is with Intelligent Information Engineering and Science department, Engineering Faculty, Doshisha University,1-3, Tatara-Miyakodani, Kyo-Tanabe, 610-0321, Japan. Tel.:+81-774-65-6981. Fax: +81-774-65-6798 (e-mail: myanagid@mail.doshisha.ac.jp).

enclosed environment, the sound waves are reflected by every solid surface in the room and so each microphone receives not only the direct sound wave but also the reflections of all orders until the energy of the source vanishes. This phenomenon, called reverberation, can be modeled as the output of a linear and time invariant (LTI) system [2]. This is the well known problem of cocktail party, where one is interested in the isolation of some of the sources from several ones, after their mixture in a real room. This kind of algorithms can be applied to all application areas where the sources are remotely located respect to the sensors, including hands-free communications, hearing aid processing, dictation systems, man-machine interfaces, etc.

In the case of blind source separation (BSS), the separation must be obtained without using (almost) any knowledge regarding the source characteristics nor the transmission media. There are many approaches to try to solve this problem. Among them, two have received an important interest in last years, one based on the sparsity of the signals [3], [4] and the other using the statistical independence among the sources [5], [6], [7], [8]. The first approach uses the property of sparsity in time-frequency domain to segregate the sources using some properly designed masks. It has the advantage that it can be used when more sources than sensors are present (i.e., in the under-determined case), but the sparsity assumption collapse quickly when reverberation increases [9], so its applicability to real environments is limited.

The second approach needs a number of sensors equal or greater than the number of sources, and can be applied to environments with some (small) reverberation. In this case the main hypothesis is the statistical independence among sources. This assumption is used to achieve the separation by an iterative optimization of a properly chosen cost function that expresses the independence of the resulting signals. The process can be done in the time domain [10], [7], in the frequency domain [11], [8], or in both the domains [6], [12]. The time domain formulation has an advantage that there are no ambiguities in the solution, but the algorithmic complexity is high due to the convolution operations involved. This limitation usually restricts its usage to simple toy examples. On the contrary, the frequency domain formulation has a lower complexity but it has the so-called permutation problem that makes the solution difficult. There are some formulations in both the domains, in which all processing is done in the frequency domain, but the updating and evaluation of costs functions is done in the time domain. Although this can avoid the permutation indeterminacy, it requires constant switching between domains during the iterations, which makes the complexity and computational cost higher.

In this paper the frequency domain approach based on statistical independence of the sources is adopted. This formulation, called frequency-domain blind source separation (fd-BSS) or frequency domain independent component analysis (fd-ICA) [13], solves a standard instantaneous ICA problem for each frequency bin, after applying a short-time Fourier transform (STFT) to switch to the time-frequency domain. To avoid confusion, this approach will be called fd-ICA in all the manuscript.

The main disadvantage of this approach is the permutation indeterminacy among source identification, as the separated sources can have arbitrary permutations and scalings for each frequency bin. It must be noted that this problem is a direct result of the indeterminacy that arise in ICA from the lack of information about the sources. Thus, to obtain a consistent time-frequency representation for each source, this approach requires to correct the permutations and the arbitrary scaling among frequency bins. Although there are some approaches to solve the permutation problem, based on the correlation between frequency bands [11] or on the estimation of directivity patterns of the separation matrix [14], these approaches tend to fail when reverberation time of the environment increases.

A detailed analysis of the working principles and limitations of fd-ICA for reverberant environments is presented in [15]. In this approach, in order to capture the impulse response effect, a large frame size is required for the STFT analysis. This reduces the amount of data available in each frequency bin, and produces a deficient estimation of the separation matrix. As a consequence, there is a compromise between the need of long frames to deal with reverberation, and the need of short frames to properly estimate the separation matrix. Furthermore, in the same work the BSS processing is compared to a set of null beamformers, and it is shown that for longer reverberation times, the directivity pattern produced by fd-ICA is increasingly deteriorated, mainly in low frequencies, due to wrong estimation of the mixing matrices. This increases the rate of permutation misalignment, producing poorer results.

Considering all this, some means to avoid the permutation problem are required. In previous works, a separation algorithm that is permutation-free was proposed [16], [17]. This algorithm uses only one separation matrix common to all frequency bins, estimated over a stack of all the time-frequency plane in one long vector. This approach has the disadvantage that by using the same separation matrix for all frequency bins, the directivity pattern generated for each frequency is different, and so it does not produce the right separation for all frequencies.

In the present work we propose a new method of fd-ICA based on a simplified mixture model, which assumes a high similarity between impulse responses from a given source to all the microphones. This method can be used to generate a separation matrix for each frequency bin, having no indeterminacy among bins, and with high processing speed. As a consequence, constant directivity patterns are obtained, which improves the separation quality. Also a time-frequency Wiener postfilter is applied to enhance the output by reducing the residual noise. The simplified algorithm includes the reverberation effect in the obtained source waves, and so the performance is less sensitive than other fd-ICA approaches to the effects of reverberation. In this way, the proposed algorithm can solve two of the problems mentioned above.

In the next section a detailed explanation of the mixture model used in this work will be presented. Based on this model, a new separation algorithm will be outlined in section III. Next, in section IV, several experiments to assess the capabilities of the algorithm will be presented, using both synthetic and real mixtures. The performance will be evaluated with objective quality measures. Also the application to a specific task of robust speech recognition will be evaluated and the robustness of the method will be assessed. Finally, conclusions and future works will be presented in section V.

II. Pseudoanechoic Mixture Model

To obtain a robust method of separation, a simplified mixture model is proposed. The mixing model and the separation algorithm derived from it, will be explained for the case of two sources and two microphones. The

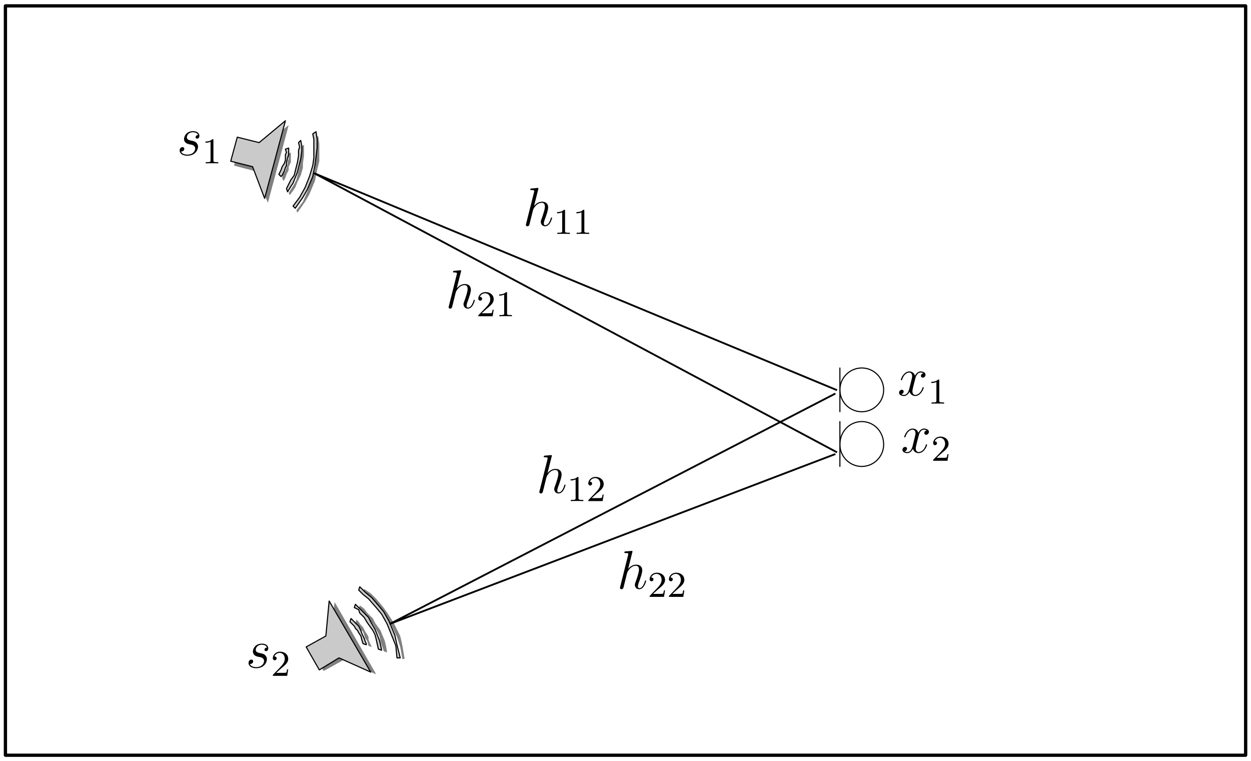

Fig. 1. Environment description and notation for a two sources and two microphones case

generalization to more sources and microphones is straightforward, and will be sketched after presenting the algorithm. In a 2-by-2 configuration, there are four impulse responses (IR), that characterize the transmission path from each source to each microphone. We assume the usage of omnidirectional microphones, both pointing in the same direction to avoid phase inversions. The IR from source i to microphone j will be denoted as $h_{ji}(t)$. The source vector is $\mathbf{s}\left(t\right) = [s_1(t)\, s_2(t)]^T$ and the mixture vector is denoted $\mathbf{x}\left(t\right) = [x_1(t)\, x_2(t)]^T$. Figure 1 shows these variables.

Using this setting, the mixtures can be expressed as

$$\begin{aligned} x_1\left(t\right) &= s_1\left(t\right) * h_{11}\left(t\right) + s_2\left(t\right) * h_{12}\left(t\right) \\ x_2\left(t\right) &= s_1\left(t\right) * h_{21}\left(t\right) + s_2\left(t\right) * h_{22}\left(t\right) \ . \end{aligned} \tag{1}$$

where $*$ stands for convolution. As can be seen, each signal x_i is produced by the addition of two terms, one generated by each source s_i, after convolution with different IRs. In the general case of arbitrary microphone locations, the IRs from a source, say s_1, to both microphones, h_{11} and h_{21}, will be quite different, and thus after convolution with them the results will have very different waveform. This means that the contributions of the same source on each microphone would behave as completely different signals. Thus the problem behaves as under-complete, as it is like a 4 sources and 2 microphones mixture. This is why instantaneous independent component analysis fails to solve the problem.

Now assume that the microphones, instead of having arbitrary positions, are restricted to remain "near" to each other. The sound generated by some source corresponds to local changes in pressure from a steady, stable value. Thus, what we are interested in, is the evolution along time and space of the pressure, relative to the steady value. This pressure variation, denoted by $p(\mathbf{v})$, where $\mathbf{v} = [x, y, z, t]$ represents the concatenation of space coordinates and the time evolution, can be modeled using the classic fluid mechanics theory. For a flow at small velocity (which is the usual case for sound at normal power levels), the sound field is characterized by the classical wave equation [18]. As this equation includes second order partial derivatives in time and space, the pressure function that solves the equation must be a C^2 class function, that is, a twice continuously-differentiable function of space and time coordinates. Therefore, both the pressure and its first derivative must be continuous. This continuity imply that the limit of the pressure must exist in all space-time coordinates of the domain. In other words, at two "near enough" points of the space-time coordinates, the pressure cannot be too different. In this phrase, the terms "near enough" refer to the Euclidean norm $||\mathbf{v}_1 - \mathbf{v}_2||$ being small. The meaning of this, is that if the microphones are enough near, the IR measured from the same source at both microphones will have a similar waveform, possibly affected by some delay and scaling. This observation motivates the following assumption that we will use to simplify the mixture model: Given enough near located microphones, the impulse responses from one source to all the microphones are similar in shape, and are only modified by a delay and a scaling factor. That is,

$$\begin{aligned} h_{21}(t) &\simeq \alpha h_{11}\left(t - d_1\right) \\ h_{12}(t) &\simeq \beta h_{22}\left(t - d_2\right) \ . \end{aligned} \tag{2}$$

To simplify the notation let $h_1(t)$ and $h_2(t)$ denote $h_{11}(t)$ and $h_{22}(t)$, respectively. Denoting $z_1 = s_1 * h_1$ and $z_2 = s_2 * h_2$, we can rewrite (1) as

$$\begin{aligned} x_1\left(t\right) &= z_1\left(t\right) + \beta z_2\left(t - d_2\right) \\ x_2\left(t\right) &= \alpha z_1\left(t - d_1\right) + z_2\left(t\right) \ . \end{aligned} \tag{3}$$

After a STFT, and assuming the time invariance of the impulse responses (as usual for static or short duration sources), this can be written as

$$\mathbf{X}\left(\omega, \tau\right) = A\left(\omega\right) \mathbf{Z}\left(\omega, \tau\right) \ , \tag{4}$$

where the mixing matrix A has the form

$$A\left(\omega\right) = \left[\begin{array}{cc} 1 & \beta e^{-jd_2\omega} \\ \alpha e^{-jd_1\omega} & 1 \end{array}\right] . \tag{5}$$

In this model, the parameters α, β, d_1 and d_2 are related to the relative attenuations and delays of the impulse responses arriving at different microphones, and the effect of the room is included in $\mathbf{Z}\left(\omega, \tau\right)$. The separated sources (convolved by the impulse responses h_1 and h_2) can be obtained using the inverse $W\left(\omega\right)$ of the mixing matrix $A\left(\omega\right)$ for each frequency bin. In this way, we have a specific mixing matrix for each frequency bin, and thus a specific separation matrix, which will produce the specific directivity patterns.

In the standard fd-ICA formulation, the problem consists of estimating a 2-by-2 complex separation matrix for each frequency bin (that can be hundreds). Using this new model, the problem has been reduced to the estimation of four parameters, named α, β, d_1 and d_2. If one can obtain a reliable estimate of these parameters for some frequency bin, then they can be used to build the mixing matrix $A(\omega)$ for other frequency bins. Given $A(\omega)$, the separation matrix $W(\omega)$ is obtained as its inverse, and the separation is realized by applying it to each mixed frequency bin. This matrix works in a similar way than that of standard fd-ICA methods, and can be interpreted as a pair of null beamformers.

The main assumption of this pseudoanechoic model, that is, the similar waveforms of the impulse responses from the same source in all the microphones, has also been observed in other works. In a recent work [19], in the context of underdetermined BSS methods for echoic environments, the authors analyze the IR for closely spaced microphones (2.5 cm). They present some graphics that show very similar impulse responses for 4 consecutive microphones, and then state that these suggest that the impulse responses are merely delayed and scaled versions of each other. Moreover, in [20] a pair of microphones are located with their tips almost coincident, and so the authors simplify the mixture model because they consider the IR to be identical. In this case the authors use directional microphones, and this directionality is what allows the separation. These works, although propose the usage of closely spaced microphones, does not explore the theoretical bases for their use.

Two aspects must be noted: by using this model, there are no amplitude nor permutation ambiguities among bins, thus there is no need for permutation correction stages after the separation. Also, an algorithm based on this approach is expected to have a low computational cost, as only one optimization (to estimate the parameters) would be needed. This is the opposite for standard fd-ICA approaches, that need one ICA optimization for each frequency bin, and after that needs to solve the permutation and amplitude ambiguities.

This approach is quite different from the anechoic model used by example in [21]. In the anechoic model, the effect of reflections is disregarded, and no restrictions on the microphones location are imposed. In this way the anechoic model only works for rooms with very small reverberation time. On the contrary, our model takes reflections into account, and consider that their effect can be grouper into some latent variables z_i, which are obtained as outputs. This is clearly explained in Fig. 2. In part a) of this figure, the anechoic model is shown in the left, and the block diagram in the right side shows the usual transformation to relative parameters. Both models are completely equivalent, and they can be applied only if the mixture was anechoic. If also a far field simplification is used, $\alpha = \beta = 1$, as in [21]. On the other side, in part b) a fully echoic model is shown in the left, which is valid no matter how long the filters are. The transformation in the right, which yields relative parameters, is possible for near enough microphones. In that case, both models are completely equivalent, and so the one in the right models a fully echoic mixture. Although the right sides of both models are similar in structure, they clearly differ in they principles and conditions of applicability. Given the similitude to the anechoic model we called this the "pseudoanechoic" model.

In the pseudoanechoic model the reverberation time is not a limiting aspect, because as can be seen in Fig. 2, the transformation to relative parameters only depends on the validity of the assumption of similar waveforms of

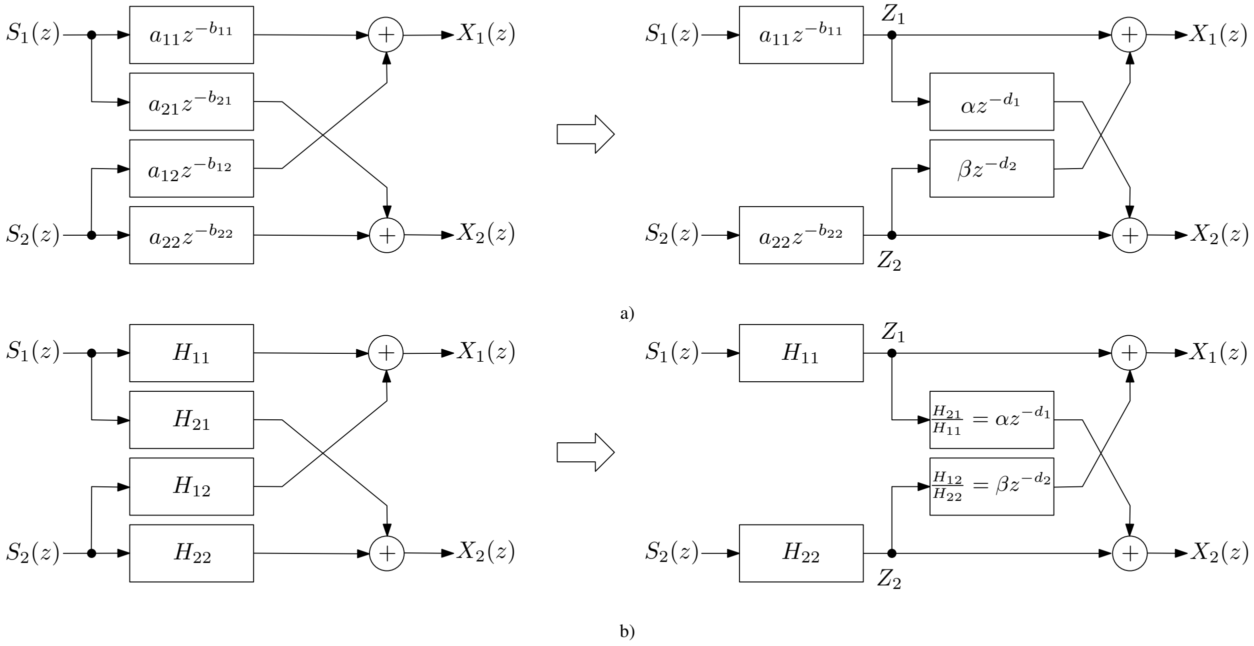

Fig. 2. Block diagrams comparing a) the anechoic, and b) the pseudoanechoic models. For both cases, in the left is the general case, and in the right the equivalent model using relative parameters.

the impulse responses. In [21], the anechoic model is used to synthesize null beamformers that does not take into account the amplitude attenuations, using some closed formulation. On the contrary, our method takes reflections into account, considers both delays and attenuation factors, synthesize the mixing matrix for each frequency bin using the estimated parameters, and calculates the separation matrices by direct inversion of the estimated mixing matrix. This yields very different equations for the separation matrix coefficients with respect to those obtained by synthesizing null beamformers with constant attenuations.

III. Separation Algorithm

According to the previous section, the key point to achieve separation using the pseudoanechoic model is to be able to find a good estimate for matrix $A(\omega)$ for a given frequency bin. This allows the estimation of the mixing parameters, and thus we can build a separation matrix for each frequency bin. To realize this idea, an algorithm is designed as shown in Fig. 3. In this algorithm, there are several subjects that need to be clarified. We will detail all the steps in the following.

Step 1) Transformation to the time-frequency domain: This transformation is done by means of a standard STFT using a Hanning window [22]. Let $x(n)$ be a digital signal and $x(m,\tau) = \phi(m)x(m+\tau R)$ the windowed and time-shifted version of $x(n)$. The STFT $\mathcal{X}(\omega_k,\tau)$ is given by

$$\mathcal{X}(\omega_k,\tau) = \sum_{m=0}^{N-1} x(m,\tau)e^{-j\omega_k m}\,, \tag{6}$$

1) *Apply a STFT to switch to the time-frequency domain.*
2) *Choose some frequency bin ω_ℓ.*
3) *Estimate the separation and the mixing matrix for ω_ℓ by ICA.*
4) *Convert the mixing matrix to the normalized form of (5).*
5) *Use the obtained matrix to calculate the four parameters: α, β, d_1 and d_2.*
6) *Separate each frequency bin. For each ω:*
    a) *Calculate $A(\omega)$ according to (5).*
    b) *Calculate separation matrix $W(\omega)$ by inversion of $A(\omega)$.*
    c) *Obtain the estimated source contributions $\tilde{\mathbf{Z}}(\omega,\tau) = W(\omega)\mathbf{X}(\omega,\tau)$.*
7) *Apply the time-frequency Wiener filters to enhance the output signals.*
8) *Reconstruct the temporal signals by inverting the STFT.*

Fig. 3. Separation algorithm based on the proposed mixture model

where $\omega_k = \frac{2\pi k}{N}$ is the discrete normalized frequency, with bin index $k = 0, \ldots, N-1$, frame index $\tau = 0, \ldots, L-1$, $\phi(n)$ is a Hanning window of length N and R is the frame shifting interval for the analysis. This formula is used to obtain the time-frequency representations for all the mixtures.

The two relevant parameters in this transformation are the window length N and the frame shifting interval R. As in this method the impulse response is considered as a part of the signal to obtain, these parameters are not so critical. In usual fd-ICA, a large window length is used to capture the impulse response characteristics. This increases the number of frequency bins to be processed. As the reverberation is included in the model in this new approach, a relatively small window length can be used without significant degradation of separation. This will speed-up the algorithm as less frequency bins need to be processed. Regarding the frame shifting interval, in standard fd-ICA a small value is used mainly to increase the amount of available data. This increase of the data length does not necessarily imply a better separation, because the data is highly redundant and the convergence may be slow. On the contrary, in this new algorithm the amount of data is decided by other aspects (see Step 2), and so the frame shifting interval can be increased (even to half of the window length) to reduce computational costs.

Step 2) Selection of frequency bin: This selection is not trivial, the ideal frequency bin would be one which presents a good signal to noise ratio, and for which the ICA algorithm will produce good directivity patterns. We have selected the frequency bin based on knowledge of source characteristics (see section IV for details), but some better designed automatic decision algorithms can be developed. To give robustness to the method we use not only one frequency bin, but we select a number Δ of frequencies to each side of the chosen one, and pack all in one long vector. We have verified that sometimes ICA fails to converge with isolated bins. To avoid these fails, a number of lateral bins is concatenated to the selected ω_ℓ, thus ensuring the production of an usable separation matrix. This also gives a robust estimation in the case that one of the signals has no contents for a given frequency bin. The use of lateral bins not only makes estimation for the central bin more robust, but also increases the amount of available data (which improves the convergence properties of ICA), so we can fix a desired number K of training samples for the ICA algorithm. If the number of frames used in the STFT is L (this is the number of time elements in a frequency bin), we set $\Delta = \max\left(3, \lceil (K/L-1)/2 \rceil\right)$, where $\lceil \cdot \rceil$ means rounding to the nearest higher integer. That is,we use a number of bins enough to have, combined, K samples for ICA training, and if this number is less than 3, we fix it to 3. The separation matrix obtained from this process will be assigned to the central bin of index ℓ.

Step 3) Estimation of mixing and separation matrices: We use the complex version of FastICA algorithm, as proposed in [23]. This algorithm uses a deflationary approach where each source is extracted sequentially. It searches for the extrema of $E\left\{G\left(\left|\mathbf{w}_i^H\mathbf{x}\right|^2\right)\right\}$ where $\mathbf{w}_i^H$ corresponds to i-th row of separation matrix W. For extraction of each source, $\mathbf{w}_i$ is updated as

$$\begin{aligned}\mathbf{w}_i^{+} &= E\left\{\mathbf{x}\left(\mathbf{w}_i^H\mathbf{x}\right)^* g\left(\left|\mathbf{w}_i^H\mathbf{x}\right|^2\right)\right\} - \\ &\quad E\left\{g\left(\left|\mathbf{w}_i^H\mathbf{x}\right|^2\right) + \left|\mathbf{w}_i^H\mathbf{x}\right|^2 g'\left(\left|\mathbf{w}_i^H\mathbf{x}\right|^2\right)\right\}\mathbf{w}_i \qquad (7)\\ \mathbf{w}_i^{new} &= \frac{\mathbf{w}_i^{+}}{\|\mathbf{w}_i^{+}\|}\,. \qquad (8)\end{aligned}$$

After finding the separating vectors $\mathbf{w}_i$ for p sources, a Gram-Schmidt-like decorrelation is applied for vector $\mathbf{w}_{p+1}$ in each iteration to avoid the convergence to the previous optima. In these equations we have used $G(y) = \log(\gamma + y)$, its derivative $g(y) = \frac{1}{\gamma+y}$ and second derivative $g'(y) = \frac{-1}{(\gamma+y)^2}$, with $\gamma = 0.1$. After finding the separation matrix, the mixing one is calculated as its inverse.

Step 4) Conversion of mixing matrix to normalized form: The normalized form consists of ones in the main diagonal, and in general this will not be the case with the estimated mixing matrix. To obtain the normalized form of (5), all elements in column i must be divided by the i-th element of the diagonal. This step is responsible for the elimination of the amplitude ambiguity because all scaling effects of the mixing matrix are absorbed into $\mathbf{z}$.

Step 5) Estimation of the mixing parameters: Once the mixing matrix in normalized form is obtained, the parameters can be calculated as:

$$\begin{aligned}\alpha &= |a_{21}|\,, \quad d_1 = -\frac{N}{2\pi\ell f_s}\Im m(\ln(a_{21}))\\ \beta &= |a_{12}|\,, \quad d_2 = -\frac{N}{2\pi\ell f_s}\Im m(\ln(a_{12}))\,, \qquad (9)\end{aligned}$$

where ℓ is the index of the central frequency bin used in step 2, f_s denotes the sampling frequency, and $\Im m(\cdot)$ is the imaginary part function. It must be noted that the delay estimations will be valid only if $\frac{2\pi\ell f_s}{N}d_i < \pi$, which follows from the periodicity of the complex exponentials. This requirement will be satisfied if the microphone spacing is small enough to avoid spatial aliasing, and of course, if the mixing matrix is successfully estimated.

Note that this robust estimation of the parameters is quite different from the direction of arrival (DOA) estimation used in the field of fd-ICA [14], [21]. For ICA-based DOA estimation, an ICA problem is solved in each frequency bin, and after solving the permutation problem, each global DOA is estimated by averaging the DOAs estimated on each frequency bin, all under an anechoic model (as in eq. (13) of [21]). The estimations for each frequency bin are affected by many disturbances like different noise powers, bad convergence of the ICA algorithm, and residual permutations. All of these noise sources affect the estimation of each bin and thus the resulting average estimates will lacks robustness. The robustness of our approach is a consequence of the use of several frequency bins in step 2 and the absence of permutations.

Step 6) Separation: In this step, a specific mixing matrix for each frequency bin is calculated using the estimated parameters. A separation matrix is obtained as its inverse, and the separated sources are calculated using it. It must be noted that the structure of the mixing matrix can be exploited to speed-up the calculation of the separation matrix, without using the inverse. After this step we obtain an estimation $\tilde{\mathbf{Z}}(\omega_k, \tau) = [\tilde{\mathcal{Z}}_1(\omega_k, \tau)\ \tilde{\mathcal{Z}}_2(\omega_k, \tau)]^T$ of the sources $\mathbf{Z}(\omega_k, \tau)$.

Step 7) Wiener filtering: Due to the behavior of the separation algorithm as a pair of null beamformers [15], the estimated sources will still have some residual noise. To obtain an estimate of source 1, a beam with a null pointed towards source 2 is formed, and vice-versa. As only 2 microphones are being used, the depth of this null will not be enough to eliminate the jammer signal. Moreover, the null eliminates all signals coming from one specific direction, but due to the reverberation, all the echoes coming from other different directions will remain. This means that some residual noise will be always left undeleted, and the residual will be larger for environments with strong reverberation. To reduce this residual we propose to use a pair of non-causal time-frequency Wiener filters as post-processing [24]. The short-time Wiener filter $F_{\mathcal{W},1}$ to enhance source 1 is

$$F_{\mathcal{W},1}(\omega_k, \tau) = \frac{\left|\tilde{\mathcal{Z}}_1(\omega_k, \tau)\right|^2}{\left|\tilde{\mathcal{Z}}_1(\omega_k, \tau)\right|^2 + \left|\tilde{\mathcal{Z}}_2(\omega_k, \tau)\right|^2} . \tag{10}$$

where $\tilde{\mathcal{Z}}_2(\omega_k, \tau)$ in the denominator is used as an estimation of the residual noise. The short-time Wiener filter to improve source $\tilde{\mathcal{Z}}_2$, $F_{\mathcal{W},2}(\omega_k, \tau)$ is calculated in a similar way to (10), with the roles of $\tilde{\mathcal{Z}}_1$ and $\tilde{\mathcal{Z}}_2$ interchanged.

This Wiener filter in which the filter characteristics depend both on time and frequency, using each separated signal as the noise estimation to enhance the other one, has not been previously used in the context of fd-ICA for convolutive mixtures (although similar strategies have been proposed in the context of single channel BSS). Its behavior will depend of the capability of having a good estimation of source and noise, that is, it depends on a good result for the previous separation stage.

Step 8) Signal reconstructions: We use the overlap-add inverse STFT to reconstruct the source signals [25]. For a given $\mathcal{Z}(\omega_k, \tau)$, the windowed and time-shifted inverse for each frame is obtained as

$$z(m, \tau) = \frac{1}{N} \sum_{k=0}^{N-1} \mathcal{Z}(\omega_k, \tau) e^{j\omega_k m} . \tag{11}$$

Using each reconstructed frame we form the sum of all them, correcting the time-shift

$$z^*(n) = \sum_{\tau=0}^{L-1} z(n - \tau R, \tau)$$

$$
\begin{aligned}
&= \sum_{\tau=0}^{L-1} \phi(n-\tau R) z(n-\tau R+\tau R) \\
&= z(n) \sum_{\tau=0}^{L-1} \phi(n-\tau R) \\
&= z(n)\Phi(n) \qquad (12)
\end{aligned}
$$

where $\Phi(n)$ denotes the shifted sum of the windows. From the first equation to the second, the windowed and delayed signal $z(n-\tau R, \tau)$ was replaced by the product of the delayed window and signal, in a similar way that it was used in equation (6). Using this, the signal can be reconstructed as $z(n) = z^*(n)/\Phi(n)$. For some windows, with certain frame shifting intervals, $\Phi(n) = C$ where C is a constant (except at the start and the end of the signal), but for general windows and shifting intervals, $\Phi(n)$ will have important oscillations. As can be seen, these formulas are valid only for frame shifting intervals smaller than the window length (otherwise $\Phi(n) = 0$ for some ranges of n and the division will produce an indeterminacy). This reconstruction formula is applied to each estimated source $\mathcal{Z}_i(\omega, \tau)$ to obtain its time-domain representation $z_i(n)$.

It must be noted that as we are searching for $\mathbf{z}$ but not for $\mathbf{s}$, the algorithm will achieve separation but not reverberation reduction (with the exception of a small reduction of reverberation introduced by the Wiener filter). As the reverberation is considered as part of the target signals, the algorithm will be less sensitive to it, and thus will achieve better separation for the cases where standard fd-ICA methods fail. If reverberation reduction is also desired, a second processing stage should be employed. In this way we split the problem into two simpler ones. Also note that the words "indeterminacy free" in the title refer to the elimination of the need to solve the usual fd-ICA permutation and amplitude ambiguities among different bins, but the time-domain reconstructed signals will still contain an arbitrary scale and permutation.

This algorithm can easily be generalized for the q-by-q mixture case. Up to step 4, no modifications are needed. For step 5, in the presented algorithm the amplitudes and delays are relative to the reference signal (the one in the main diagonal). For the general case, as in the normalized form the main diagonal has always ones, the parameters to estimate are the amplitude of the off-diagonal elements, and the pairwise relative delays in the exponents, that is, two parameters for each off-diagonal element. So, there are $2q(q-1)$ parameters to estimate, and their estimation is straightforward from the normalized form, using the corresponding analogous to (9). Step 6 does not need any change. For step 7, the Wiener filter formulation needs to be modified, as the noise spectrum estimate will be the sum of all the other estimated source spectrums. Step 8 remains unchanged. Although this generalization is straightforward, we will restrict our experiments to the 2-by-2 case, leaving for future works the analysis of the general case behavior.

IV. Results and discussion

The pseudoanechoic model was built using that the impulse responses from one source to all microphones should have approximately the same waveform. We need to investigate how the spacing between microphones will affect its performance. Furthermore, there are two parameters: window length and frame shifting interval, that need to be calibrated. These aspects will be studied in the first and second part of the current section.

We are interested in the application of this algorithm to automatic speech recognition (ASR), and also in applications where the processed speech is presented to a human listener, like in hearing aids. Then we need to evaluate our algorithm regarding both, speech recognition tests and perceptual quality. In the following subsections, all these issues will be explored in detail.

For all the experiments we have used speech sentences extracted from Albayzin Spanish database [26]. Also we have used white noise source from Noisex-92 database [27]. All the signals were resampled to 8000 Hz. These sources were mixed in different conditions, using both speech and white noise as competing sources, to generate appropriate data sets for the experiments.

The mixtures were separated by the algorithm of Fig. 3. In all experiments, for this algorithm we used a different central frequency ω_c in the case of speech noise and white noise. For speech-speech mixtures, a high frequency bin was used, as in general the separation matrix are better estimated in high frequencies, as shown in [15]. For speech signals with telephony quality bandwidth, the maximum frequency of interest present in the signals is 4000 Hz. A frequency band located at $5/8$ of the maximum frequency was selected in this case. On the other hand, for speech-white mixtures a low frequency bin was used. This is due to the fact that white noise presents equal power in all frequencies, whereas speech tends to have more power in low frequencies due to the low-pass characteristics of the glottal response [22]. In this way, the signal to noise ratio is degraded with increasing frequencies. A bin located at $3/8$ of the maximum frequency was selected in this case. The desired number of data samples to use with FastICA was fixed at $K = 5000$.

For the evaluation of the method in robust speech recognition, we performed some test with a state-of-the-art continuous speech recognition system. For this task, a three-state semi continuous hidden Markov model for context-independent phonemes and silences was used [28]. Gaussian mixtures were utilized as observation probability densities. After four reestimations using the Baum-Welch algorithm, tying was applied in the model, to reduce the number of parameters. For this, a pool of 200 Gaussians for each model state was selected. After tying, another 12 reestimations were performed. A bigram language model was used for recognition, estimated from transcriptions of the training sentences [29]. For the front-end, we performed parametrization with standard mel frequency cepstral coefficients (MFCC), including energy and the first derivative of these features [30]. This recognition system was built using the HTK toolkit [31]. The results were evaluated using the word recognition rate (WRR), defined as

$$WRR\% = \frac{T - D - S}{T} 100 \tag{13}$$

where T is the total number of words in the reference transcription, D is the number of deletion errors and S is the number of substitution errors.

To evaluate the separation results we have used the PESQ raw score in narrow band mode as defined in [32]. This measure is known by its high correlation with subjective perceptual quality measured by MOS. Also, in [33] a high correlation between PESQ scores and recognition rates of an automatic speech recognition system is reported, using different speech enhancement techniques for additive noise, with artificial voices. In previous works [34], [35], we have evaluated several objective quality measures as predictors of recognition rate of an ASR system, after application of fd-ICA for convolutive mixtures. The best results were obtained for perceptually designed measures, in particular the PESQ.

Given the large amount of parameters to explore, if we varied all them in an exhaustive search, the number of required experiments would grown exponentially. To avoid this we have sorted the parameters according to its influence on the algorithm, and then explored the variation of each independently, with the other parameters fixed. Although this would produce a sub-optimal set of parameters, it will allows us to explore a larger area of the parameter space, with a reasonable number of experiments and time.

A. *Effects of Microphone Spacing*

The proposed algorithm uses a physically plausible assumption to simplify the mixture model. The key question to be analyzed in this section is how plausible is that hypothesis in real cases.

As already discussed, the motivation for the assumption comes from the physics of sound propagation, taking into account the continuity of the sound field. Intuitively, if the microphones are “near enough”, then they should measure similar variations of the sound field, and thus the IR measured at those point should have approximately the same shape, but affected by some delay and scaling. This produces two main aspects that needed to be determined. One is how much near the microphones must be for the hypothesis to be applicable. And the second one is how sensitive the algorithm is with respect to poor adjustment to the hypothesis.

As an example to illustrate the first point, Fig. 4 shows two impulse responses recorded in the room of Fig. 5. The impulse responses were measured from source 1 to microphones 1 and 5, which are spaced by 4 cm. The distance from the source to the microphones was about 113 cm, with an angle respect to the array center of 26 degrees, in a room with 343 ms of reverberation time.

The top part of Fig. 4 shows the impulse responses. It shows that in a general view the IRs seems to have similar global characteristics, although due to the scale it is difficult to realize how similar they really are. In the central panel, a zoom of the initial 256 samples of the IRs is shown. In this image it is easier to see the similitude between the two impulse responses. Although there are some parts showing small differences, most of them can be attributable to the combined effect of the fractional delay and the sampling. This can be seen in the bottom panel, where a zoom of the first 64 samples is shown. To generate this plot, a resampling using bandlimited interpolation was used, to elevate the sampling frequency to 10 times the original (i.e., from 8 kHz to 80 kHz). Also the original samples are shown with dark dots. In this plot, the fractional delay can be clearly seen. The delay corresponds to about $2/5$ of the original sampling time, which agrees with the spatial setup. The bottom panel also shows that most of the local differences in the waveforms have disappeared, showing that the morphological differences were artifacts produced by the sampling. Considering this example, the assumption about the similarity of the IR waveforms seems to be very plausible. It must be noted that in this example the microphone spacing is of 4 cm, and even with such a “wide” separation, the similitudes are evident. This is supported also by the results in [19],

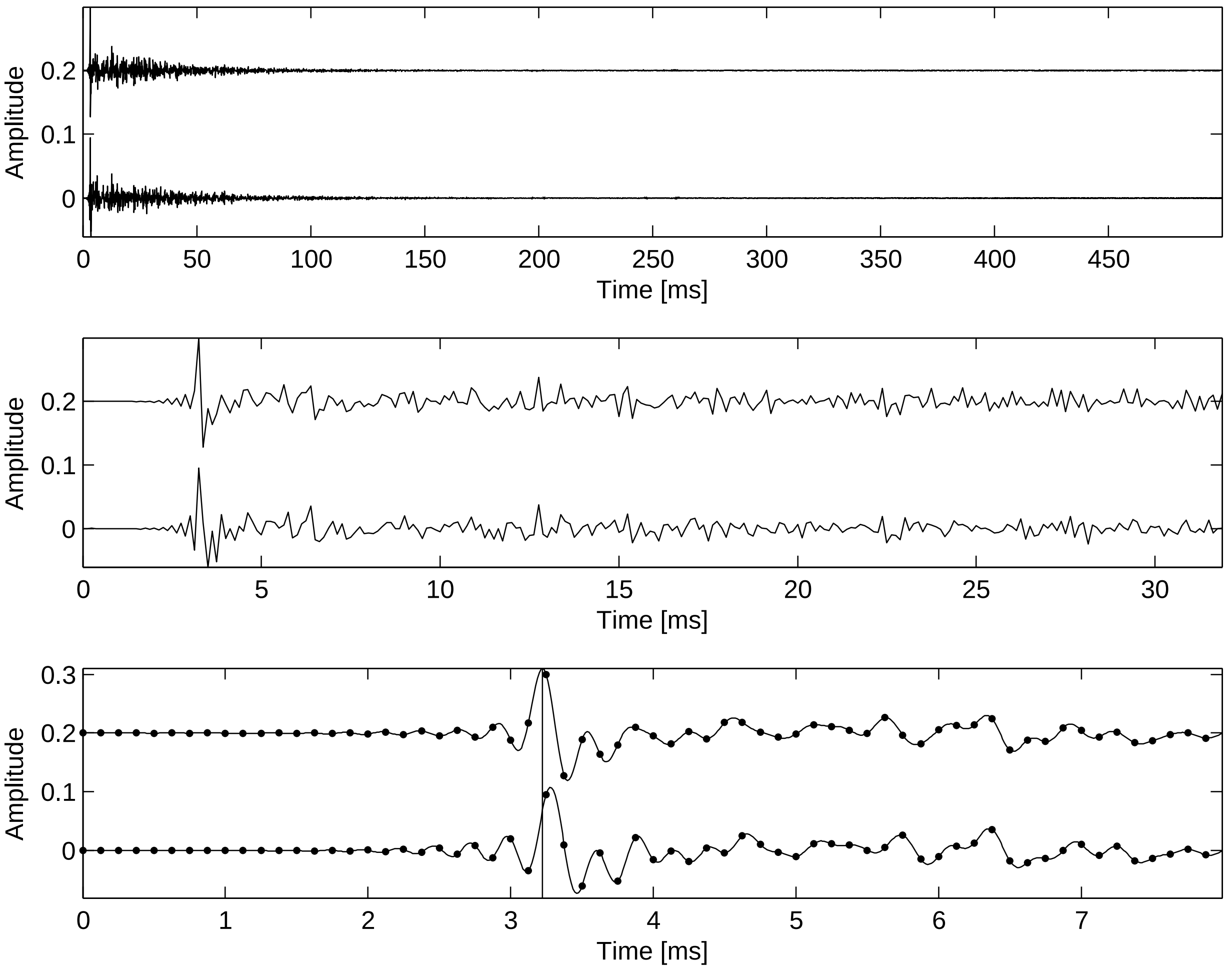

Fig. 4. Impulse response characteristics for 4 cm spacing, recorded as in Fig. 5. Top: first 0.5 seconds. Center: zoom showing the first 256 samples. Bottom: first 64 samples, resampled at 10 times the original sampling frequency. In all panels, one of the signals has a constant of 0.2 added, to separate the two plots.

in which four impulse responses measured with a 2.5 cm uniform spacing are found to be very similar in shape. The authors conclude that the IR can be possibly considered as delayed and scaled copies.

Considering the second aspect (the sensibility of the algorithm with respect to the hypothesis), we need to evaluate how the separation performance is modified by dissimilar impulse responses. Thus we explore the effect of microphone spacing. If the spacing is too large, the impulse responses from one source to the microphones will be too different and the hypothesis will become invalid. Also, spatial aliasing can be produced for large microphone distance. The maximum allowable distance to avoid spatial aliasing is related to the sampling frequency by $d_{max} = \lambda_{min}/2 = c/(2f_{max}) = c/f_s$, where λ_{min} is the minimum wavelength of the signal used and c is the speed of sound. For a 8000 Hz sampled signal, with $c = 340$ m/s, this is $d_{max} = 4.25$ cm [36]. On the other hand, too small spacing will cause the relative amplitudes to be very near to one, and the relative delays to be very small. An accurate estimation of these parameters will thus be difficult to obtain, mainly because of the limited measurement precision and the microphone noise.

We have explored this issue using synthetic mixtures of speech. To produce the mixtures we have selected five sentences from Albayzin database spoken each one by two male and two female speakers, for a total of 20 utterances. Also one sentence spoken by a male and a female speakers were selected to be used as competing noise. This sentence was used because it was longer than any of the other sources, and so the same sentence could be used to interfere with all the target sources. To compete with male speech, female speech was used, and vice-versa. The utterance duration ranged from 2.26 s to 4.65 s, with an average duration of 3.55 s. We also used white noise from Noisex database.

The mixtures were generated by convolving each source with an impulse response measured in a real room and adding the results to generate each microphone signal. The impulse responses were measured using the method of time stretched pulses (TSP) [37]. A condenser desktop microphone with omnidirectional flat frequency response from 20 Hz to 20 kHz was used. The measurements were made in a room depicted in Fig. 5. We used 5 microphone locations with a spacing of 1 cm, with a careful synchronization to preserve relative amplitude and delays between impulse responses. The average reverberation time[1], measured by the Schroeder backintegration method [38], was of $\tau_{60} = 343$ ms. The impulse responses measured from pairs of microphones with spacings of 1, 2, 3 and 4 cm were used (longer spacings would introduce spatial aliasing). According to the naming convention of Fig. 5, the pairs of microphones were 2-3 and 3-4, 1-3 and 3-5, 1-4 and 2-5, and 1-5 for 1, 2, 3 and 4 cm spacing, respectively.

The effect of different noise powers was also explored. For each noise kind (speech or white) we used two different power ratios (0 dB and 6 dB) by properly scaling the source signals. Thus, we performed mixtures for a total of 16 mixing conditions.

First, we wanted to investigate the optimal spacing, so we performed separation for the different spacings. We have repeated the separation using 3 window lengths (128, 256 and 512 samples) and two different frame shifting intervals for each window length (one quarter and one half of the window length). Figure 6 shows the average PESQ scores over the 20 sentences. In this figure, we have averaged the results for the four noise conditions to produce a single value for each spacing, window length and shift interval. It can be seen that the optimal spacing is 4 cm in all cases. A too short spacing makes difficult to accurately estimate the parameters and thus the algorithm also fails. Longer spacings will cause spatial aliasing. This behavior is repeated if the analysis is discriminated for each noise condition, showing always the best separation at 4 cm. According to this result, we have fixed the spacing in 4 cm for the following experiments.

[1] The reverberation time τ_{60} is the time interval in which the sound pressure level of a decaying sound field drops by 60 dB, that is to one millionth of its initial value [18].

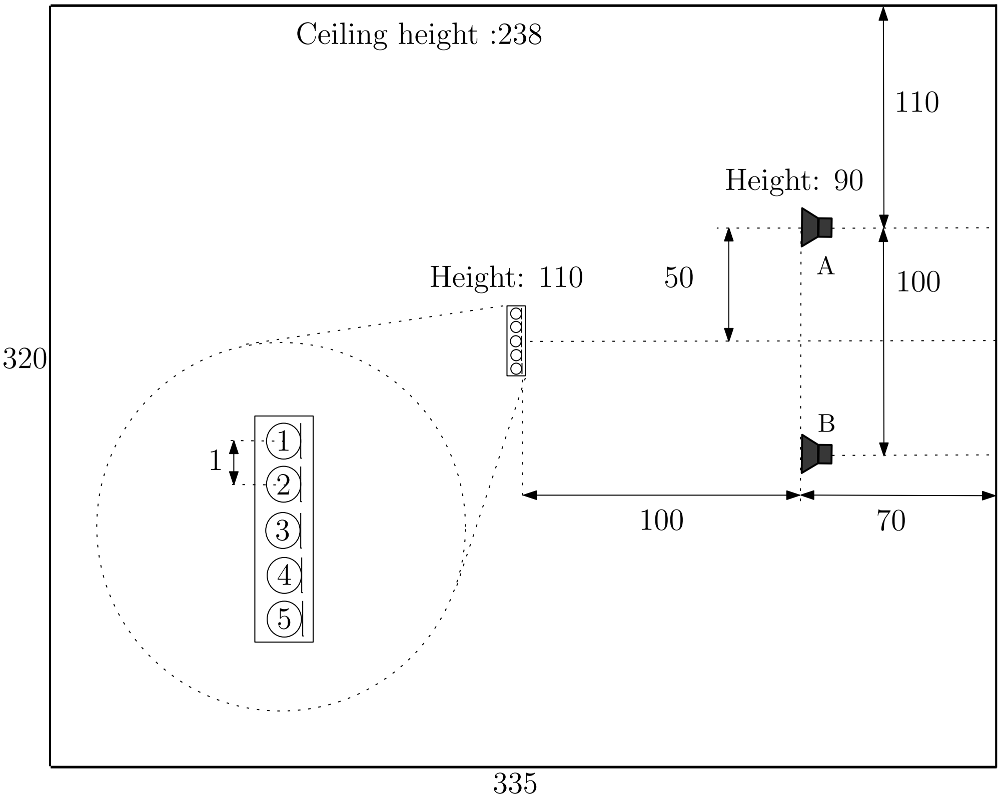

Fig. 5. Experimental setup for two sources and five locations of the microphone. All dimensions are in cm.

B. *Effect of Window Parameters*

Once fixed the optimum spacing, we have explored the effect of window length and frame shifting interval. For this evaluation we proceeded in a similar way to the previous experiment, but only for the case of 4 cm spacing. We used five window lengths (128, 256, 512, 1024 and 2048 samples) with a frame shifting size fixed on half of the window length. For the evaluation we used PESQ and also the average processing time. As a fast separation algorithm with a good quality performance is required, we used the ratio of time to quality as a cost-to-benefit function to determine the optimum window length.

It could be argued that the processing time is not a good index of complexity, because different implementations of the same algorithm would yield different times. Nevertheless, the time requirements of our algorithm are not caused by high complexity tasks that could be performed with different implementations of algorithms, but by simpler task that are repeated many times. The FastICA algorithm used is the same, and is performed only once, with the same amount of data samples, so its influence in the calculation time for different frame lengths is equivalent. Thus, the two processes that have a strong effect in processing time are the calculation of the separation matrices (which involves a matrix inversion for each frequency bin) and the separation of the data itself, which imply a matrix-matrix multiplication for each time-frequency tile.

With increasing frame length, the number of bins to process is increased, which means that more matrix inversions

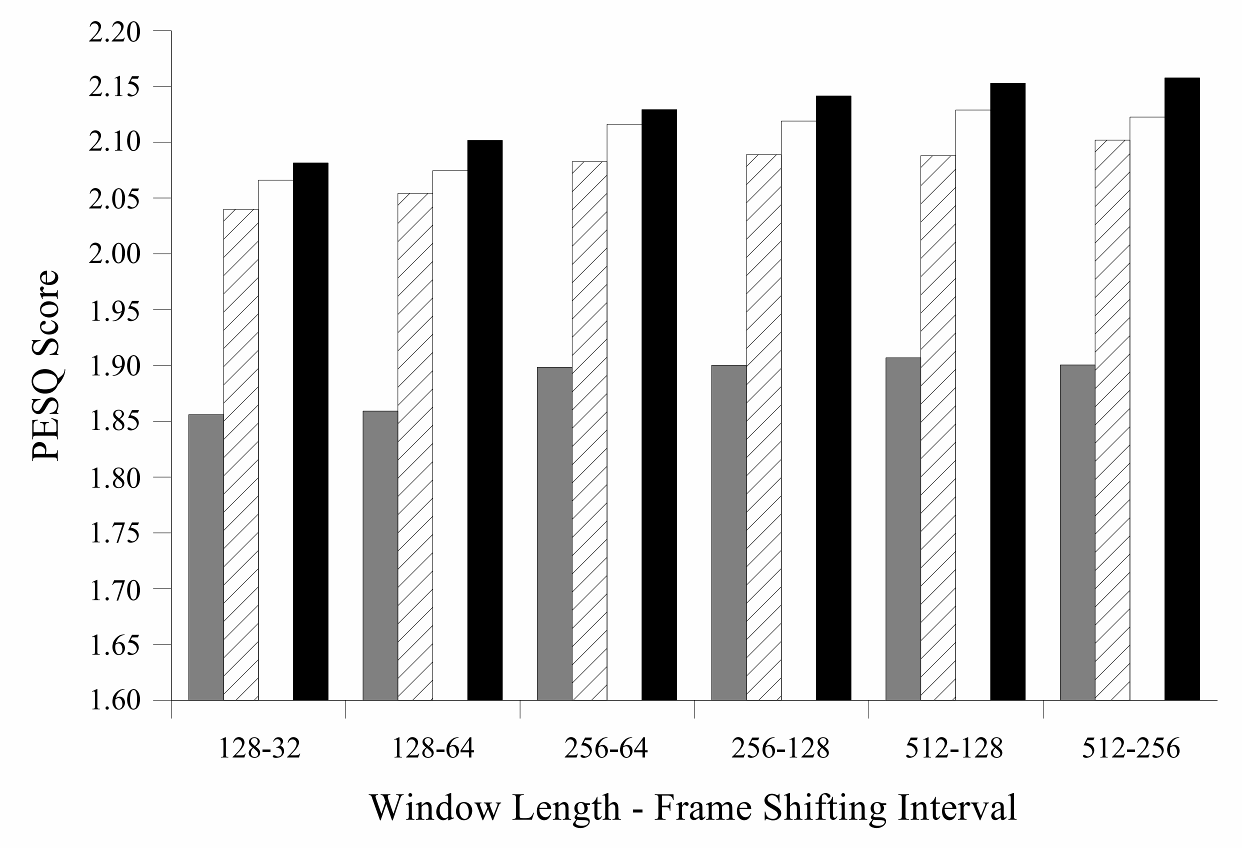

Fig. 6. Effects of microphone spacing. Gray bar: 1 cm spacing, dashed bar: 2 cm spacing, white bar: 3 cm spacing, black bar: 4 cm spacing

TABLE I

EFFECTS OF THE WINDOW LENGTH. COST IS THE RATIO OF TIME TO PESQ SCORE

Window	PESQ	Time [s]	Cost
128	2.166	0.550	0.254
256	2.215	0.558	0.251
512	2.234	0.647	0.290
1024	2.177	0.616	0.283
2048	2.094	0.716	0.342

need to be calculated, but the amount of data to separate is the same. So the change of computation time is mainly due to the increased number of matrix inversions needed. Thus, a larger window will have to perform more matrix inversions, so its processing time would be directly increased. Table I shows the results. It can be seen that the optimum (the minimum of the Time/PESQ ratio) is obtained for a window of 256 samples.

Finally we investigated the effect of the frame shifting interval. Fixing the microphone space in 4 cm and the window length in 256 samples, we have explored the shifting parameter from 8 to 128 samples with increments

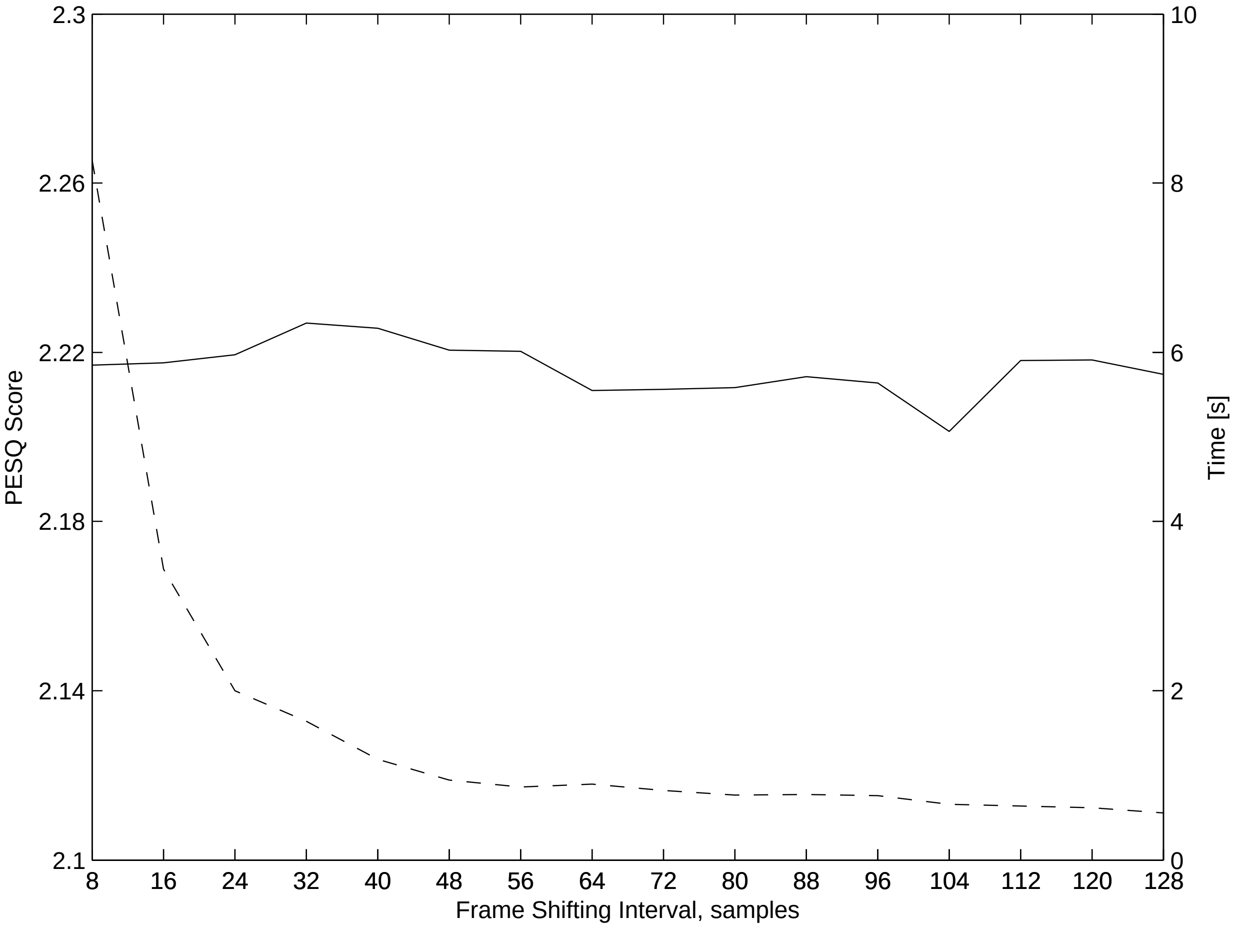

Fig. 7. Effects of the frame shifting interval. Solid line: PESQ score, Dashed line: average processing time

of 8 samples. The maximum shifting interval was fixed to 128 samples because at least a half window redundancy is necessary to obtain a proper reconstruction of time domain signals. Figure 7 shows the PESQ score and the average processing time for this analysis. As can be noted from the figure, there is no significant quality change for different shifting intervals. Nevertheless, with small shifting values the processing times grow very fast. With small shifts, the number of frames in each frequency bin increases. As we fixed a value of at least $\Delta = 3$ for the lateral bins, as the shift interval diminishes, the amount of data used in FastICA increases and the algorithm is slowed down. On the other hand, when the shift interval increases, as the amount of data is fixed at $K = 5000$, the redundancy is reduced and the convergence is faster. According to this, we selected 128 samples as the best frame shifting interval.

C. Evaluation on Artificial Mixtures

Once obtained the optimum values for spacing, windows length and frame shifting size, we wanted to check the performance of the separation algorithm in a larger database. In this test, the separation algorithm was used

TABLE II

RECOGNITION RATE AND PESQ SCORE FOR SYNTHETIC MIXTURE

		Mixtures		Separated	
Power ratio	Noise	WRR%	PESQ	WRR%	PESQ
6 dB	Speech	32.44	1.89	63.94	2.33
	White	16.73	1.74	63.36	2.30
0 dB	Speech	20.96	1.56	51.35	2.15
	White	12.26	1.50	43.09	2.05
Average		20.60	1.67	55.44	2.21
Reverberant				70.22	2.33
Clean				93.98	4.50

as a pre-processing stage for an automatic speech recognizer. Also the objective quality evaluation by means of PESQ score was carried out. For this task we used a larger subset of Albayzin database. This subset consists of 600 sentences spoken by 6 male and 6 female speakers, with a vocabulary of 200 words, from the central area of Spain. The average duration of sentences was 3.95 s, with a minimum of 1.88 s and a maximum of 8.69 s.

The sentences were mixed artificially with impulse responses for microphones 1 and 5 in Fig. 5. As noise sources we used competing speech and white noise. For speech, we selected two Albayzin sentences (different from the 600 used as sources), one from a female speaker to interfere with male speakers and one from a male speaker to contaminate female speech. The noise powers were adjusted to produce a power ratio of 0 dB and 6 dB at the simulated speakers. In this way, 4 sets of mixtures (2 noise kinds with 2 noise powers) were generated. After mixture, the separation algorithm with the optimum window length and optimum frame shifting size obtained in previous experiments were used.

For the speech recognition task we used an ASR system like the one described in section IV. The leave-k-out cross validation method with 10 partitions of the 600 sentences was used to test the acoustic model of the ASR system. For each partition 20% of the sentences were selected randomly to form a test set and the other 80% used as train sets. The results of the 10 partitions were averaged.

It is known that reverberation reduces the automatic recognition rates [39], even if the recognition system is trained with speech recorded in the same reverberant room [40]. According to this, as our algorithm is not aimed to reduce reverberation, we cannot expect the recognition rate to be equivalent to that of clean speech. The maximum obtainable performance would be near to that of reverberant speech, without interference. We have used the artificially reverberated signals to evaluate this maximum performance (Table II).

Table II shows the resulting PESQ scores and WRR, for the different noise kind and powers. Besides the results of the separation algorithm, we present the results for the mixtures, the sources, and the reverberant (but clean) sources.

As can be appreciated, an average improvement of more that 35% in WRR and more than 0.5 in PESQ score is obtained by pre-processing the speech with the proposed algorithm. Also, it can be seen that for the case of 6 dB mixtures, the separated signal achieves a WRR that is near to the maximum attainable for our kind of algorithms, given by the WRR of the reverberant signals. The average PESQ score for processed signals is also similar to that of reverberant ones.

Something interesting can be seen in the case of speech noise with 6 dB of power ratio. For this case, PESQ is the same as for reverberant speech. Nevertheless, speech recognition is lower than that of reverberant speech. We know that the separated signals still have some residual interference, and so the quality should be degraded. However the PESQ is higher than expected because some reverberation has also been removed (this can be noted by inspection of spectrograms and by listening carefully). This reverberation reduction is an effect of the Wiener filter, as the signal used as estimation of the noise has some echoes of the desired source arriving from different directions. The Wiener filter will thus reduce in some amount the reverberant echoes. In this way, a small quality improvement in reverberation compensates the reduction due to residual noise and PESQ is the same, although the remaining noise degrades speech recognition.

D. Evaluation on Real Mixtures

For this experiment we recorded the same 20 sentences used in sections IV-A and IV-B, contaminated with the same noises, but in a real room as shown in Fig. 8. The environment is an acoustically isolated room that naturally has a reverberation time of about $\tau_{60} = 120$ ms. To increase the reverberation time, plywood reflection boards were added in two of the walls. The average reverberation time for this case was about $\tau_{60} = 200$ ms.

After mixture, separation was performed using the proposed separation algorithm, the one proposed in [8] (we will call this Parra), and the one proposed in [11] (we will call this Murata). Both algorithms are fd-ICA methods, that obtain independence exploiting the nonstationary structure of the speech signals, using second-order statistics. Murata's algorithm uses the correlation among envelopes of the frequency bins to solve the permutation problem, and Parra's algorithms avoid permutation by imposing constrains in the structure of the time-domain separation filters. For Murata algorithm, we used a window length of 256 samples (32 ms), a frame shifting interval of 10 samples (1.25 ms) and we selected 40 correlation matrices to diagonalize, as suggested by the authors. For Parra algorithm, as the signals used here were of very different durations than the ones reported by the author, we tried several combinations of filter and window lengths (128/1024, 256/512, 256/2048 and 512/3072). The best results were obtained for a filter length of 256 samples (32 ms) with a window length of 512 samples (64 ms).

An implementation of the Parra algorithm was obtained from the author web page[2], and the implementation of Murata algorithm was obtained from Shiro Ikeda web page[3]. All the algorithms were programmed in Matlab language, and the separation tests were ran in a Pentium 4 EMT64 of 3 GHz, with 1GB of RAM. For the proposed algorithm we used two variants, without including the time-frequency Wiener postfilter of step 7 (in the following

[2] `http://newton.bme.columbia.edu/~lparra/publish/`

[3] `http://www.ism.ac.jp/~shiro/research/index.html`

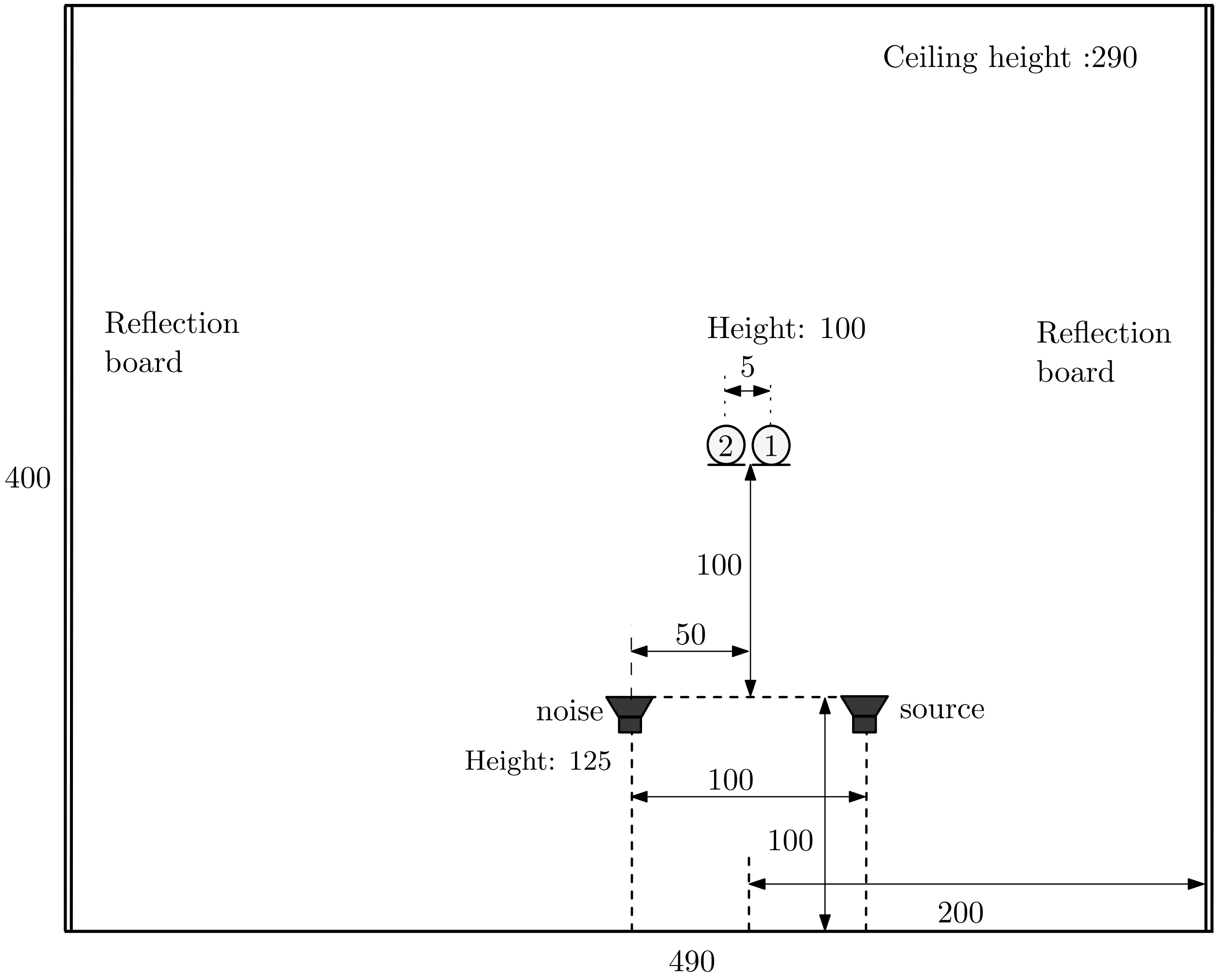

Fig. 8. Experimental setup for a two sources-two microphones case

Pr. I) and with the Wiener filter (Pr. II). Another variant is the related to the central bin selection. As we used some heuristics based on knowledge of the source characteristics to decide which could be a good candidate for central bin, it can be argued that the method is not completely blind. Thus we also run Pr. II, but using as central bin, that located in the middle of the frequency range, for all kind of noises (Pr. III).

To test the performance of the algorithm, we have used a speech recognition system similar to that described in section IV. The acoustic model was trained with 585 sentences from a subset of Albayzin database (the training set does not includes any of the sentences used in the test, nor the used as interfering voices).

Also we performed two additional evaluations, one using the PESQ to have a perceptual objective quality evaluation, and the other the average processing time for each of the algorithms. Table III shows the WRR for this experiment, and Tables IV and V the PESQ scores and the average processing time, respectively.

As can be seen, Murata algorithm produces some degradation of all, WRR and PESQ. This is due to the fact that

TABLE III

WRR% FOR THE EVALUATED ALGORITHMS AND THE MIXTURES. PR IS THE POWER RATIO IN THE LOUDSPEAKERS.

PR	Noise	Mix.	Murata	Parra	Pr. I	Pr. II	Pr. III
6 dB	Speech	44.50	25.00	49.50	83.07	85.50	85.50
	White	19.54	15.00	27.50	61.00	85.50	82.50
0 dB	Speech	30.00	27.00	49.00	62.50	83.00	85.00
	White	7.20	11.00	20.00	24.00	67.50	65.00
Ave.		25.31	19.50	36.50	57.64	80.38	79.50

TABLE IV

PESQ SCORES FOR THE EVALUATED ALGORITHMS AND MIXTURES. PR IS THE POWER RATIO IN THE LOUDSPEAKERS.

PR	Noise	Mix.	Murata	Parra	Pr. I	Pr. II	Pr. III
6 dB	Speech	2.11	1.97	2.22	2.51	2.83	2.83
	White	1.98	1.86	2.37	2.57	2.83	2.82
0 dB	Speech	1.73	1.71	2.19	2.26	2.59	2.61
	White	1.64	1.67	2.16	2.25	2.54	2.53
Ave.		1.86	1.80	2.23	2.40	2.70	2.70

TABLE V

AVERAGE PROCESSING TIME IN SECONDS FOR THE EVALUATED ALGORITHMS AND MIXTURES. PR IS THE POWER RATIO IN THE LOUDSPEAKERS.

PR	Noise	Murata	Parra	Pr. I	Pr. II	Pr. III
6 db	Speech	9.49	6.48	0.36	0.43	0.43
	White	8.79	7.01	0.26	0.27	0.30
0 db	Speech	9.56	6.60	0.31	0.42	0.46
	White	8.98	6.48	0.24	0.28	0.30
Ave.		9.21	6.64	0.29	0.35	0.37

the algorithm cannot handle the reverberation times involved on this tests. The only effect of the processing is to introduce distortions that degrade the performance. For Parra algorithm, some improvement is noted, although it is not enough. The proposed algorithm can handle the separation in this environment and produces a large improvement in WRR and PESQ. Even if we do not use the Wiener postfilter, the quality of our algorithm outperform the other evaluated alternatives, showing that the separation stage indeed works better than the previous approaches. Also, when using a fixed central bin (Pr. III), the performance is slightly changed, suggesting that the method is robust to central bin selection.

For comparison, we have also evaluated the use of a different postfilter. In other works in the area, a binary mask postfilter was used after a first stage of fd-ICA separation, to improve the results. The main assumption for binary masks is that each time-frequency sample has information of only one source. But for reverberant signals this assumption collapses, and so a continuous mask should produce better results. We have implemented the binary mask postfilter presented in [41] and used it instead of our Wiener postfilter (with the same first stage), evaluating the PESQ scores. For speech noise, the PESQ scores obtained were 2.69 and 2.49 for 6 dB and 0 dB of power ratios, respectively. For white noise, the PESQ scores were 2.73 and 2.45 for 6 dB and 0 dB, respectively. These results represent in average only 52.2% of the improvement obtained with our Wiener postfilter.

Regarding the processing times, the proposed algorithm is more than 26 times faster than Murata's and more than 18 times faster than Parra's one. It should be noted that no special optimization was made in the implementation to make a faster code.

E. Robustness of the ICA estimation

There are two other subjects to be explored. One is how the quality and the robustness of the method are affected by the selected number of lateral bins. And the other is how sensitive the algorithm is to the central bin selection.

To answer the first question we have performed an experiment using the same data as the experiment in section IV-D (the real mixtures data). We have selected a bin located in the center of the frequency range, and performed the separation using no lateral bins, one lateral bin to each side, two lateral bins to each side, and so on. Then we averaged the separation results for the 20 test sentences, for the case of white noise and 6 dB of power ratio (similar results were obtained for other noise and powers). In the x axis of Fig. 9 we put the number of lateral bins[4], and we used two y axis, at the left we put the average PESQ obtained (in solid line), and in the right side the average processing time used for separation, in seconds (dashed line).

As can be seen in this figure, the addition of lateral bins is beneficial, as the quality is always improved with respect to the case when using only the central bin. Also, it can be seen that adding more bins increases the quality, up to a limit, where it starts decreasing. This agrees with our assumption that to some extent, by adding lateral bins, the effects of the upper an lower bins are cancelled, and the quality is augmented, but if too much bins are added, the discrepancies have more weight and the quality is lowered.

Furthermore, it can be seen that the processing time initially is decreased, proving that the addition of lateral bins, even if the ICA algorithm has to process more data, produced faster convergence than using only the central bin. When more and more bins are added, the processing time is increased, so it would be desirable to keep the lateral bins as low as possible for reducing the processing time, but large enough to provide a good quality. This shows that the addition of lateral bins provides an improvement both in terms of convergence speed and estimation quality of the ICA algorithm.

Now for the second problem, in all the experiments up to now we used a fixed central bin for the estimation, selected using some a priori knowledge about the source characteristics, or fixed at the center of the frequency range. Clearly, the best case would be to have some method that could determine automatically the optimal central

[4] A frame of length 256 samples was used, so the bin index goes from 0 to 128, we used the index 64 as central bin, and added lateral bins to each side from 0 to 63 bins.

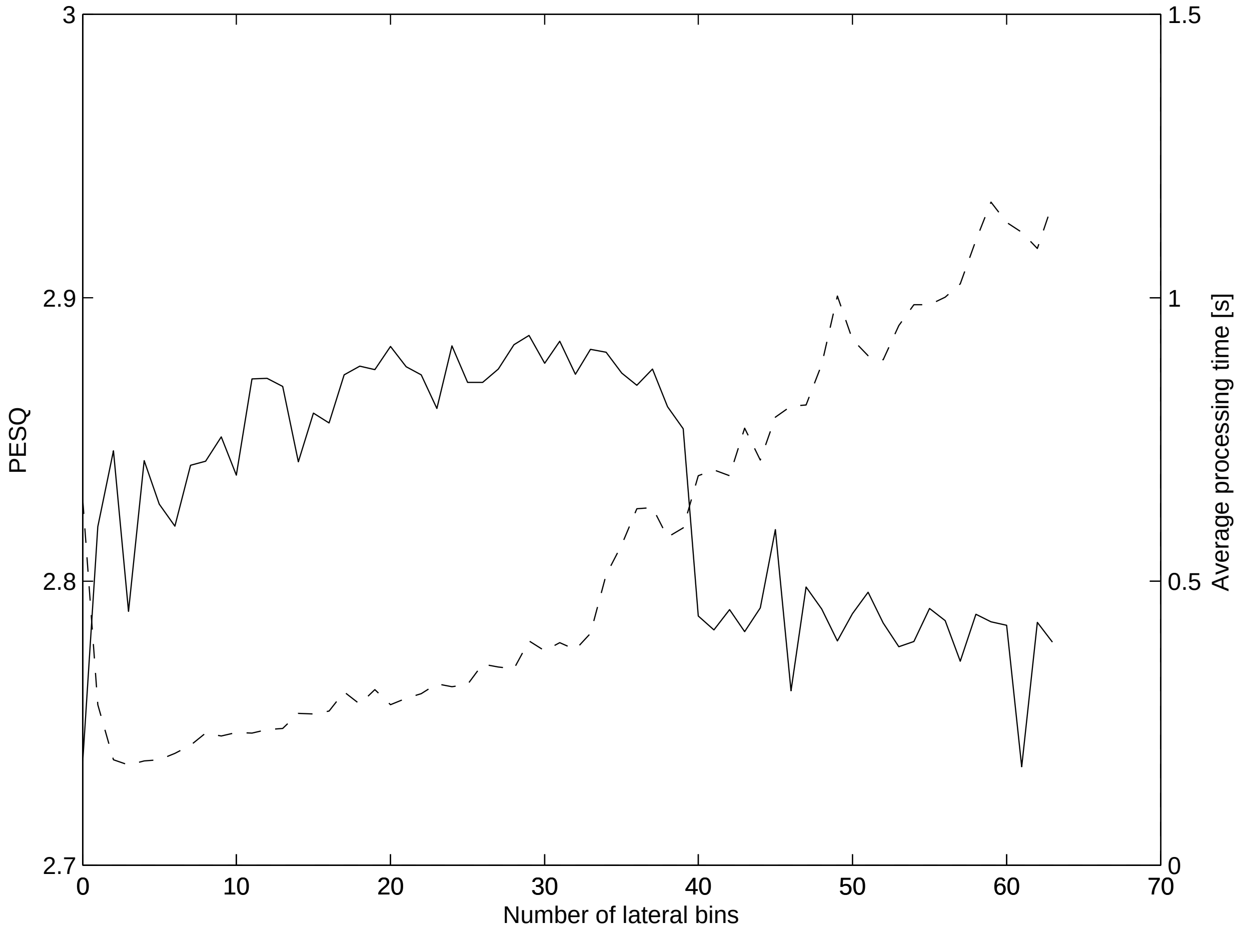

Fig. 9. Effect of the number of lateral bins used in the ICA algorithm. Left axis and solid line: PESQ score; right axis and dashed line: average processing time.

bin. Although more research is needed before producing such a method, it is important to know how sensitive is the method to the central bin selection. To verify this robustness, we performed the following experiment.

We used a sentence from the last experiment in section IV-D (the case of real mixtures), contaminated with white noise at 0 dB of power ratio. We used a frame length of 256 samples with 128 samples of step size for the STFT. This mean that we have the bin index changing from 0 to 128. We repeated this, for three cases: using no lateral bins, using 5 lateral bins, and using 10 lateral bins (we have excluded the extreme cases where the central bin was lower than the number of lateral bins, because we need to use the desired number symmetrically). The results are presented in Fig. 10.

As can be seen in this figure, for 0 lateral bins, there are a lot of deep valleys in the PESQ score. These valleys corresponds to bins where the ICA algorithm failed to converge. This behaviour is typical of standard fd-ICA approaches (which estimate the separation matrix in each frequency bin). For these cases the quality of the convergence of ICA is not uniform for all bins, so even if the other problems (permutations, scalings) are solved,

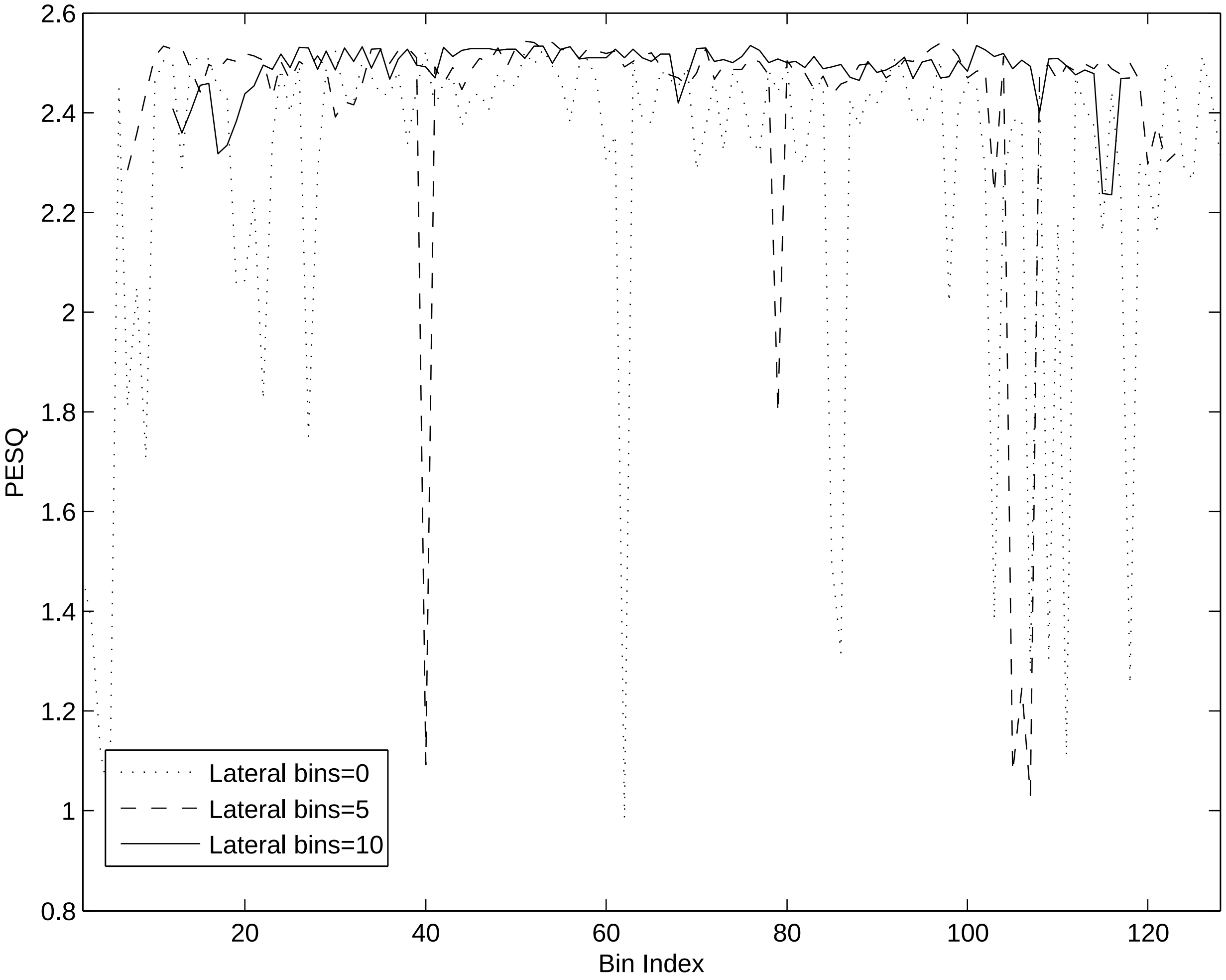

Fig. 10. Effect of the central bin selection on the quality of separation, for different numbers of lateral bins.

the quality will be very variable for different bins. In contrast, using 5 lateral bins, most of the valleys have been eliminated, and using 10 bins (solid line), the quality is quite smooth for any selection of the central bin.

This experiment shows, on one side, that the addition of lateral bins provides for robustness, as it produces a good estimation for cases were (using only the central bin) the ICA algorithm fails. On the other side, it shows that the method is quite robust to a wrong selection of central bin, which confirm our previous findings.

V. Conclusions and Future Works

In this work we have introduced a simplified mixing model for convolutive mixtures of audio sources in reverberant rooms. Based on this model a new separation algorithm has been developed. The new algorithm overcomes most of the problems presented in standard fd-ICA formulations. It has superior separation capabilities, as is shown by the experimental results with both, synthetic and real mixtures. To sum up, the following novelties have been presented:

1) A new pseudoanechoic mixing model for reverberant rooms that includes the effect of delays and amplitude scalings.
2) An indeterminacy-free method for separation in the frequency domain.
3) A robust method to estimate the mixing parameters, using complex ICA.
4) A post-processing method using the separated signals as estimations of clean sources and noise, for a time-frequency Wiener filter.
5) An extremely fast algorithm, between 15 to 20 times faster than standard and well known fd-ICA algorithms.

The capabilities of our algorithm were evaluated for two different frameworks. One is the capability to produce a good subjective quality. This capability was evaluated through a perceptually derived objective quality measure. The other is the capability of enhance the speech for a specific computer automatic task. It was evaluated through an automatic speech recognition system. The results for both cases are far superior to the evaluated alternatives. At the same time, the processing requirements are far lower than that of the alternatives. The field of application is thus very wide, from those aimed at human listening, like hands-free communication or hearing aid processing, to those related to man-machine interfaces and speech-to-text translation.

Though almost all the critical parameters of the algorithm were analyzed, there are some others that should be explored to produce even better results. The center frequency bin ω_ℓ was selected empirically to a fixed value. Instead, some automatic selection based on measures of the separability of the sources in each frequency bin could be developed. This could lead to a better estimation of the mixing matrix, and thus a more robust result would be produced. Also, a reverberation reduction stage could be used as post-processing to improve the quality even more, for applications like speech recognition in which reverberation is also undesirable.

Finally, it must be noted that the algorithm is very fast, even in a non-optimized implementation in an interpreted programming language. This encourages us to explore a real-time semi-online version of the algorithm.

REFERENCES

[1] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing. Learning Algorithms and Applications.* England: John Wiley & Sons, 2002.

[2] M. Kahrs and K. Brandenburg, Eds., *Applications of Digital Signal Processing to Audio and Acoustics*, ser. The Kluwer International Series In Engineering and Computer Science. New York, Boston, Dordrecht, London, Moscow: Kluwer Academic Publishers, 2002.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[4] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, Nov. 2001.

[5] S. Makino, H. Sawada, R. Mukai, and S. Araki, "Blind source separation of convolutive mixtures of speech in frequency domain," *IEICE Trans Fundamentals*, vol. E88-A, no. 7, pp. 1640–1655, 2005.

[6] T. Lee, A. Bell, and R. Orglmeister, "Blind source separation of real world signals," in *Proceedings of IEEE International Conference Neural Networks*, 1997, pp. 2129–2135.

[7] S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in *Proceedings of Signal Processing Advances in Wireless Communication Workshop*, 1997, pp. 101–104.

[8] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[9] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[10] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proceedings of IEEE Workshop on Neural Networks and Signal Processing*, 1996, pp. 423–432.

[11] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, 2001.

[12] M. Joho and P. Schniter, "Frequency-domain realization of a multichannel blind deconvolution algorithm based on the natural gradient'," *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation*, pp. 543–548, 2003.

[13] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, Chichester, Weinheim, Brisbane, Singapore, Toronto: John Wiley & Sons, Inc., 2001.

[14] H. Sawada, R. Mukai, S. Araki, and S. Makino, "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, pp. 530–538, Sep. 2004.

[15] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "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, pp. 109–116, March 2003.

[16] L. Di Persia, K. Ohta, and M. Yanagida, "A method for solving the permutation problem in ICA," in *Technical Report of IEICE*, no. SP2005-193, 2006, pp. 53–58.

[17] L. Di Persia, T. Noguchi, K. Ohta, and M. Yanagida, "Performance of permutation-free ICA," in *Technical Report of IEICE*, no. SP2006-1, 2006, pp. 1–6.

[18] H. Kuttruff, *Room Acoustics*, 4th ed. London: Taylor & Francis, 2000.

[19] T. Melia and S. Rickard, "Underdetermined Blind Source Separation in Echoic Environments Using DESPRIT," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, p. 19 pages, 2007.

[20] Y. Katayama, M. Ito, A. K. Barros, Y. Takeuchi, T. Matsumoto, H. Kudo, N. Ohnishi, and T. Mukai, "Closely arranged directional microphone for source separation," in *Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation*, 2004, pp. 129–135.

[21] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[22] J. Deller, J. Proakis, and J. Hansen, *Discrete Time Processing of Speech Signals*. New York: Macmillan Publishing, 1993.

[23] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *International Journal of Neural Systems*, vol. 10, no. 1, pp. 1–8, 2000.

[24] E. J. Diethorn, "Subband Noise Reduction Methods for Speech Enhancement," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems.*, Y. A. Huang and J. Benesty, Eds. New York, Boston, Dordrecht, London, Moscow: Kluwer Academic Press, 2004, ch. 4, pp. 91–115.

[25] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[26] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Mariño, and C. Nadeu, "Albayzin speech database design of the phonetic corpus," Universitat Politècnica de Catalunya (UPC), Dpto. DTSC, Tech. Rep., 1993.

[27] A. Varga and H. Steeneken, "Assessment for automatic speech recognition II NOISEX- 92: A database and experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[28] X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition*. Edinburgh: Edinburgh University Press, 1990.

[29] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambrige, Masachussets: MIT Press, 1999.

[30] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. New Jersey: Prentice Hall, 1993.

[31] S. Yung, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book (for HTK Version 3.4)*, Cambridge University Engineering Department, Cambridge, 2006.

[32] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862*, 2001.

[33] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective

quality measures and artificial voice," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2006–2013, Nov. 2006.

[34] L. E. Di Persia, D. H. Milone, M. Yanagida, and H. L. Rufiner, "Objective quality evaluation in blind source separation for speech recognition in a real room," *Signal Processing*, vol. 87, no. 8, pp. 1951–1965, 2007.

[35] L. E. Di Persia, D. H. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, 2008, accepted for publication.

[36] M. Brandstein and D. Ward, Eds., *Microphone Arrays. Signal Processing Techniques and Applications*, ser. Digital Signal Processing. Springer, 2001.

[37] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *The Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1119–1123, 1995.

[38] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[39] B. Kinsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 1259–1262.

[40] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, ser. Signals and Communication Technology. Berlin, Heidelberg, New York: Springer, 2005.

[41] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, and T. Morita, "Real-time implementation of two-stage blind source separation combining SIMO-ICA and binary masking," in *Proceedings of 2005 International Workshop on Acoustic Echo and Noise Control*, 2005, pp. 229–232.

PLACE
PHOTO
HERE

Leandro Di Persia The biography text and photograph are submitted in a separate sheet.

PLACE
PHOTO
HERE

Diego Milone The biography text and photograph are submitted in a separate sheet.

PLACE
PHOTO
HERE

Masuzo Yanagida The biography text and photograph are submitted in a separate sheet.