Initialization method for speech separation algorithms that work in the time-frequency domain

Auxiliadora Sarmiento, Iván Durán-Díaz and Sergio Cruces

Citation: The Journal of the Acoustical Society of America **127**, EL121 (2010); doi: 10.1121/1.3310248
View online: https://doi.org/10.1121/1.3310248
View Table of Contents: https://asa.scitation.org/toc/jas/127/4
Published by the Acoustical Society of America

Initialization method for speech separation algorithms that work in the time-frequency domain

Auxiliadora Sarmiento, Iván Durán-Díaz, and Sergio Cruces[a)]
Departamento de Teoría de la Señal y Comunicaciones, University of Seville, Camino de los Descubrimientos S/N, 41092 Seville, Spain
sarmiento@us.es, iduran@us.es, sergio@us.es

Abstract: This article addresses the problem of the unsupervised separation of speech signals in realistic scenarios. An initialization procedure is proposed for independent component analysis (ICA) algorithms that work in the time-frequency domain and require the prewhitening of the observations. It is shown that the proposed method drastically reduces the permuted solutions in that domain and helps to reduce the execution time of the algorithms. Simulations confirm these advantages for several ICA instantaneous algorithms and the effectiveness of the proposed technique in emulated reverberant environments.

PACS numbers: 43.60.Pt, 43.60.Gk, 43.60.Np [DOS]
Date Received: December 7, 2009 **Date Accepted:** January 11, 2010

1. Introduction

This article considers the problem of the blind separation of speech signals that are recorded in a real room, assuming the same number of microphones and speakers. It is well known that any acoustic signal acquired from microphones in a real recording environment suffers from reflections on the walls and surfaces inside the room. Therefore, the recorded signals can be accurately modeled as a convolutive mixture, where the mixing filter is usually considered a high-order finite impulse response filter.

We focus on the time-frequency domain approach for blind source separation (BSS). In this approach, the convolutive mixture is approximated by a set of parallel instantaneous mixing problems for each frequency, being each of these problems solved independently with a suitably chosen independent component analysis (ICA) algorithm. Since the separated sources can have an arbitrary ordering, with this technique, a postprocessing to align the solutions before reconstructing them in time domain is necessary. The ordering ambiguity for each frequency, which is known as the permutation problem, is ubiquitous when working in the time-frequency domain and is especially important in real recordings, where the length of the room impulse response can be very long (greater than 250 ms) and can contain strong peaks corresponding to the echoes. Several methods have been proposed to overcome the permutation problem, which can be divided into two groups. Some methods solve independently, for each frequency, the instantaneous mixture and then a known property of the signals or of the mixing filter are used in order to fix the permutation ambiguity. Examples of these properties are the following: the assumption of similarity among the envelopes of the source signal waveforms, the estimation of the direction of arrival, and the continuity on the frequency response of the mixing filter. A second group of methods tries to avoid permuted solutions by choosing a suitable initialization of the ICA algorithms for each of the frequencies. Our proposal belongs to this second group of methods and suggests an initialization procedure for those ICA algorithms that use the whitening of the observations in the time-frequency domain. The experiments show

[a)]Author to whom correspondence should be addressed.

that this initialization, which exploits the local continuity of the demixing filter in such domain, reduces drastically the number of the solutions that are permuted and also may be used to reduce the execution time of the ICA algorithms while keeping intact the quality of the separation.

2. Problem formulation

We model the microphone observations $x_i(n)$, $i=1,\ldots,N$, of a real room recording by a convolutive mixture of the speech sources $s_i(n)$, $i=1,\ldots,N$, in a noiseless situation, i.e.,

$$x_i(n)=\sum_{j=1}^{N}\sum_{k=0}^{P-1}h_{ij}(k)s_j(n-k),\quad i=1,\ldots,N, \tag{1}$$

where $h_{ij}(n)$ is the impulse response (of P taps) from the source j to the microphone i. In order to blindly recover the original speech signals (sources), one can apply a matrix of demixing filters to the observations $x_i(n)$ that yields an estimate of each of the sources

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

where $b_{ij}(k)$ denotes the (i,j) demixing filter of M taps. Let $X_i(f,t)$ and $S_i(f,t)$ be, respectively, the short-time Fourier transform (STFT) of $x_i(n)$ and $s_i(n)$. The time-domain convolutive mixture in Eq. (1) can be approximated in the time-frequency domain by a set of parallel instantaneous mixing problems:

$$\mathbf{X}(f,t)=\mathbf{H}(f)\mathbf{S}(f,t), \tag{3}$$

where $\mathbf{X}(f,t)=[X_1(f,t),\ldots,X_N(f,t)]^T$ and $\mathbf{S}(f,t)=[S_1(f,t),\ldots,S_N(f,t)]^T$ are the observation and source vectors for each time-frequency point, respectively, and $\mathbf{H}(f)$ is the frequency response of the mixing filter whose elements are $H_{ij}(f)=[\mathbf{H}(f)]_{ij}\,\forall\, i,j$. The separation model is given by

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

where $\mathbf{Y}(f,t)=[Y_1(f,t),\ldots,Y_N(f,t)]^T$ is the vector of outputs or estimated sources, and $\mathbf{B}(f)$ are the separating matrices to be estimated for each frequency f.

Due to the decoupled nature of the solutions across different frequencies, the correspondence between the true sources and their estimates suffers from ambiguities in the scaling, phase, and order. Thus, the vector of source estimates can be modeled approximately as

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

where $\mathbf{P}(f)$ is a permutation matrix and $\mathbf{D}(f)$ is a diagonal matrix of complex scalars. $\mathbf{P}(f)$ and $\mathbf{D}(f)$ constitute ambiguities for each frequency that need to be determined before being able to recover the estimated sources in time domain.

3. Initialization procedure for ICA algorithms

The ICA algorithms used for estimating the optimal separation system $\mathbf{B}(f)$ in each frequency are often started at any arbitrary point. However, a suitable initialization of the algorithm has several advantages. For instance, when the algorithm is initialized near the optimal solution, a much faster convergence of the algorithm will be obtained. Furthermore, the initialization can exploit prior information on the mixture in order to avoid permutation ambiguity. One interesting initialization approach considers the continuity of the frequency response of the mixing filter $\mathbf{H}(f)$ and its inverse. Under this assumption, the initialization of the separation system $\mathbf{B}(f)$ from the optimal value of the separation system at the previous frequency $\mathbf{B}_o(f-1)$ seems reasonable. However, we cannot directly apply $\mathbf{B}(f)=\mathbf{B}_o(f-1)$ in those ICA algorithms that whiten

the observations. The whitening is performed by premultiplying the observations vectors with an $N \times N$ matrix $\mathbf{W}(f)$. After that, the new observations $\mathbf{Z}(f,t)$ can be expressed as another mixture of the sources:

$$\mathbf{Z}(f,t) = \mathbf{W}(f)\mathbf{X}(f,t) = \mathbf{U}_*(f)\mathbf{S}(f,t), \tag{6}$$

where the new mixing matrix $\mathbf{U}_*(f) = \mathbf{W}(f)\mathbf{H}(f)$ is unitary. Then, the separation matrix $\mathbf{B}(f)$ can be decomposed as the product of a unitary matrix and the whitening matrix,

$$\mathbf{B}(f) = (\mathbf{U}(f))^H \mathbf{W}(f). \tag{7}$$

Due to the variability of the sources spectra, in general, the whitening matrices $\mathbf{W}(f)$ and $\mathbf{W}(f-1)$ in contiguous frequencies are different. Consequently, when solving $\mathbf{B}(f) = \mathbf{B}_o(f-1)$ directly to obtain $(\mathbf{U}(f))^H = \mathbf{B}_o(f-1)\mathbf{W}^{-1}(f)$, there is no longer guarantee that this matrix is still unitary.

A *classical initialization* technique avoids the previously described problem in the following way. First, it multiplies the observations $\mathbf{X}(f,t)$ by the optimal separation matrix at the previous frequency. Then, it determines the matrix $\mathbf{W}(f)$ which whitens these new observations. Therefore, the overall separation matrix is calculated as

$$\mathbf{B}(f) = (\mathbf{U}(f))^H \mathbf{W}(f)\mathbf{B}_o(f-1). \tag{8}$$

In this work, we propose a new initialization procedure that consists in initializing the separation matrix $\mathbf{B}(f)$ trying to minimize the weighted distance with several of the optimal separation systems previously calculated for nearby frequencies, while the matrix $\mathbf{U}(f)$ is constrained to be unitary. This leads to the constrained minimization problem,

$$\sum_i \alpha_i \|\mathbf{B}_o(f-i) - \mathbf{B}(f)\|_F^2 \quad s.t. \quad (\mathbf{U}(f))^H \mathbf{U}(f) = \mathbf{I}_N, \tag{9}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and α_i are non-negative weighting scalars. The solution of this problem is given by $\mathbf{U}(f) = \mathbf{Q}_L \mathbf{Q}_R^H$, where $\mathbf{Q}_L$ and $\mathbf{Q}_R$ are, respectively, the left and right singular vectors of the singular value factorization which follows:

$$[\mathbf{Q}_L, \mathbf{D}, \mathbf{Q}_R] = \mathrm{svd}(\mathbf{W}(f) \sum_i \alpha_i (\mathbf{B}_o(f-i))^H). \tag{10}$$

4. Experimental results

In this section, we present several experiments illustrating that the proposed initialization produces good quality separation with convolutive mixtures of speech signals by using different ICA algorithms. In addition, we will discuss the ability of the initialization procedure to reduce the permuted solutions, as well as its effectiveness to guarantee a high convergence speed of the ICA algorithm in such reverberant conditions. In order to have the possibility to determine the number of permuted solutions and some objective measures of quality of the separation, it is needed to know the exact room impulse response and the sources without errors. For this reason, we emulated real room recordings by means of synthetic mixtures. Therefore we created 25 synthetic mixtures of two speech sources. The sources were chosen from male and female speakers in a database[1] of 12 individual recordings of 5 s duration and sampled at 10 kHz. Those sources were mixed using a simulated room mixing system, shown in Fig. 1, determined using the ROOMSIM toolbox.[2] We computed the STFT with a finite Fourier transform of 2048 points, 90% overlapping, and Hanning windows of length 1024 samples. Then, we estimated the separation system $\mathbf{B}_o(f)$ by initializing the ICA algorithms with both the *classical initialization* and the proposed initialization. After that, we fixed the permutation and scale ambiguities applying the method described in Ref. 3, and finally filtered the observations to obtain the time-domain estimated sources.

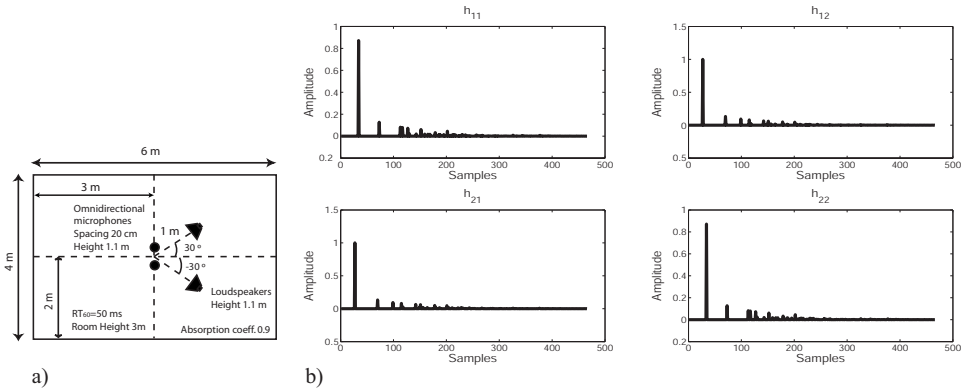

Fig. 1. (a) Microphone and loudspeaker positions for the simulated room recordings and (b) channel impulse responses of the considered filter.

We applied our initialization method to various ICA algorithms which have been proved to be efficient to estimate the separation system $\mathbf{B}_o(f)$. Since speech signals are highly nonstationary, we used two popular ICA algorithms based on nonstationarity of signals, SOBI (Ref. 4) and THINICA (Ref. 5). THINICA was used in two different configurations, by first extracting one source and then reconstructing the other, and by the simultaneous extraction of the two sources, referred hereinafter as THINICA-SIM. To quantify the quality of the estimated sources, each output was decomposed, by the BSS_EVAL toolbox,[6] into three terms $y_i(t)=s_{\text{tar}}+e_{\text{int}}+e_{\text{art}}$, which represent, respectively, the target source, the interference from other sources, and a last component of artifacts. Then, we calculated three performance measures: the source to interference ratio (SIR), the source to artifact ratio (SAR), and the source to distortion ratio (SDR),

$$\text{SIR} = 10\log_{10}\frac{\|s_{\text{tar}}\|^2}{\|e_{\text{int}}\|^2}, \quad \text{SAR} = 10\log_{10}\frac{\|s_{\text{tar}}+e_{\text{int}}\|^2}{\|e_{\text{art}}\|^2}, \quad \text{SDR} = 10\log_{10}\frac{\|s_{\text{tar}}\|^2}{\|e_{\text{int}}+e_{\text{art}}\|^2}. \tag{11}$$

The obtained results, presented in Table 1, show that for the THINICA case the initialization improves up to 8 dB the SAR and SDR in comparison with the classical method. In the other cases, the initialization does not achieve a significant improvement of the estimated sources quality.

As an example, two sources can be listened to in Mm. 1 and Mm. 2. Mixtures from these sources by means of a mixing system like that described in Fig. 1 are in Mm. 3 and Mm. 4. Finally, in Mm. 5 and Mm. 6, the sources recovered by the algorithm THINICA-SIM (Ref. 4) with the proposed initialization can be listened to.

Mm. 1. [First source: female voice sampled at 10 kHz. This is a file of type "wav" (99 kbytes).]

Mm. 2. [Second Source: male voice sampled at 10 kHz. This is a file of type "wav" (99 kbytes).]

Mm. 3. [First observation of the mixture of the two sources. This is a file of type "wav" (99 kbytes).]

Mm. 4. [Second observation of the mixture of the two sources. File of type "wav" (99 kbytes).]

Mm. 5. [First recovered source. This is a file of type "wav" (99 kbytes).]

Mm. 6. [Second recovered source. This is a file of type "wav" (99 kbytes).]

Table 1. Comparison of the average SIR, SAR, and SDR for different ICA algorithms by initializing with both the classical and the proposed initialization method.

	SIR (dB)		SAR (dB)		SDR (dB)	
	Classic	Ini-1	Classic	Ini-1	Classic	Ini-1
THINICA	20.16	22.03	0.86	9.32	0.70	8.92
THINICA-SIM	21.97	22.15	12.70	12.79	12.02	12.13
SOBI	22.48	22.15	13.02	12.83	12.43	12.16

In order to prove the effectiveness of the initialization to guarantee a high convergence speed of the algorithm, we calculated the average CPU time that each algorithm uses to solve the separation in each frequency. We also analyzed the performance of our initialization procedure in terms of the number of permutations. The results, summarized in Fig. 2, corroborate that the proposed initialization reduces both the computational effort of the ICA algorithms and the number of permutations. The reduced number of permutations is particularly very interesting because it allows us to design new algorithms to solve the permutations based on this reduction. Also, it could be used to alleviate the computational burden of the algorithms that solve the permutation problem.

We investigated those frequencies in which our initialization is not able to preserve the permutation order. Without loss of generality, we considered the simulation results provided by the THINICA-SIM algorithm using the proposed initialization. In Fig. 3, we represent the normalized modulo of the frequency response of the optimal demixing filter from source 1 to microphone 1 (upper plot), and the normalized histogram of the frequencies in which the solutions remained permuted (lower plot). It could be noted that echoes in the impulse response of the mixing filter introduce rapid oscillations on the frequency response, so our main assumption about the continuity of the mixing filter is not valid in all the frequencies. For this reason, it can be observed that frequencies presenting a high number of permutations correspond to those in which the frequency response of the optimal demixing filter exhibits strong peaks. However, there are also a set of frequencies in which, although the mixing filter does not exhibit those oscillations, the solutions are still permuted. This can be explained when the source separation problem is ill determined at these frequency bands or when the profiles across time of the second order statistics used by the chosen algorithms are similar for both sources, since they can fail to separate the sources in these situations.

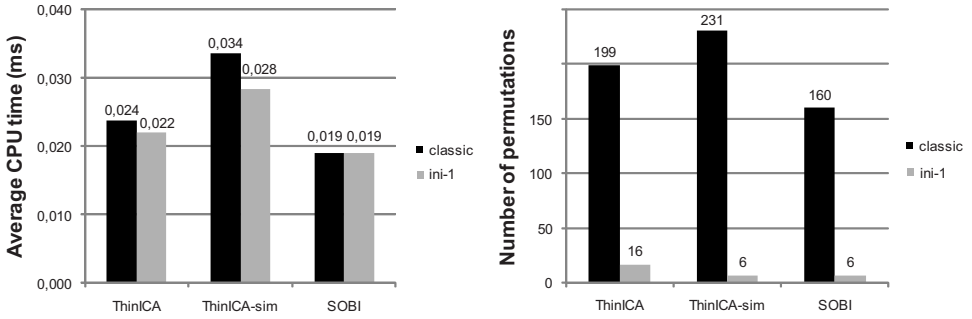

Fig. 2. (a) Average CPU time (ms) to run different ICA algorithms in each frequency by initializing with both the classical and the proposed initialization method. (b) Number of permutations for different ICA algorithms by initializing with both the classical and the proposed initialization method.

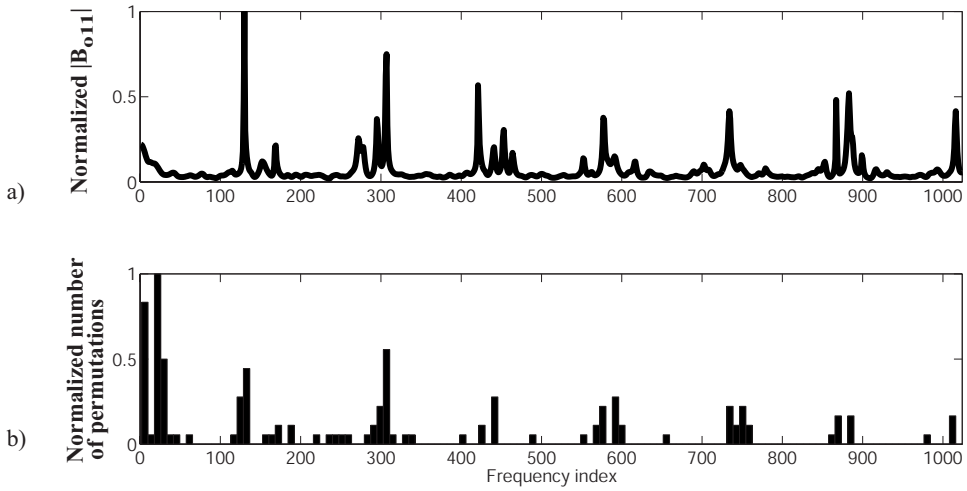

Fig. 3. (a) Normalized modulo of the frequency response of the optimal demixing filter from source 1 to microphone 1 and (b) normalized histogram of the number of permutations for 25 experiments and using THINICA-SIM algorithm.

5. Conclusions

In this article we have considered the problem of the blind separation of speech signals in reverberant scenarios. We have presented an initialization procedure for those ICA algorithms that work in the time-frequency domain and use a whitening of the observations as a preprocessing step. Computer simulations show that this initialization, when incorporated to the existing ICA algorithms, reduces drastically the number of permutations. In addition, the proposed initialization helps to alleviate the computational execution time of the ICA algorithms that solve the separation in each frequency, while preserving the quality of the separated speech sources.

Acknowledgments

This work was supported by MCYT Spanish Project No. TEC2008-06259.

References and links

[1]M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Two-microphone separation of speech mixtures," IEEE Trans. Neural Netw. **19**, 475–492 (2008).
[2]D. Campbell, ROOMSIM toolbox, http://media.paisley.ac.uk/~campbell/Roomsim/ (Last viewed Dec. 7, 2009).
[3]A. Sarmiento, S. Cruces, and I. Durán, "Improvement of the initialization of time-frequency algorithms for speech separation," in Proceedings of the 8th International Conference on ICA and Signal Separation, Paraty, Brazil (2009), pp. 629–636.
[4]A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," IEEE Trans. Signal Process. **45**, 434–444 (1997).
[5]S. Cruces, A. Cichocki, and L. De Lathauwer, "Thin QR and SVD factorizations for simultaneous blind signal extraction," in Proceedings of the European Signal Processing Conference (EUSIPCO), Vienna, Austria (2004), pp. 217–220.
[6]C. Fèvotte, R. Gribonval, and E. Vincent, BSS_EVAL Toolbox User Guide, Technical Report No. 1706, IRISA, Rennes, France (2005).