Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 75206, Pages 1–16
DOI 10.1155/ASP/2006/75206

Permutation Correction in the Frequency Domain in Blind Separation of Speech Mixtures

Ch. Servière[1] and D. T. Pham[2]

[1] *Laboratoire des Images et des Signaux, BP 46, 38402 St Martin d'Hère Cedex, France*
[2] *Laboratoire de Modélisation et Calcul, BP 53, 38041 Grenoble Cedex, France*

Received 31 January 2005; Revised 26 August 2005; Accepted 1 September 2005

This paper presents a method for blind separation of convolutive mixtures of speech signals, based on the joint diagonalization of the time varying spectral matrices of the observation records. The main and still largely open problem in a frequency domain approach is permutation ambiguity. In an earlier paper of the authors, the continuity of the frequency response of the unmixing filters is exploited, but it leaves some frequency permutation jumps. This paper therefore proposes a new method based on two assumptions. The frequency continuity of the unmixing filters is still used in the initialization of the diagonalization algorithm. Then, the paper introduces a new method based on the time-frequency representations of the sources. They are assumed to vary smoothly with frequency. This hypothesis of the continuity of the time variation of the source energy is exploited on a sliding frequency bandwidth. It allows us to detect the remaining frequency permutation jumps. The method is compared with other approaches and results on real world recordings demonstrate superior performances of the proposed algorithm.

1. INTRODUCTION

Blind source separation consists in extracting independent sources from their mixtures, without relying on any specific knowledge of the sources. Earlier works have been focused on linear instantaneous mixtures and several efficient algorithms have been developed.

The problem is much more difficult in the case of convolutive mixtures, especially audio mixtures. Although there have been many works on this subject [1–3], the successful application of the proposed algorithms in realistic settings is still elusive [4], due mainly to the long impulse responses of the mixing filters. To blindly separate the sources, one would have to find an "inverse filter" (which would also have long response) such that the recovered sources are as mutually independent as is possible. A direct (time domain) approach would be too computationally heavy, not to mention the difficulty of convergence, since it requires the adjustment of too many parameters. However, by using the Fourier transform, the separation problem of convolutive mixtures can be recast as a set of separation problems of instantaneous mixtures associated with each frequency bin, which can be solved independently. But the discrete Fourier transform tends to produce nearly Gaussian variables, and it is well known that blind separation of instantaneous mixtures requires non-Gaussianity. Fortunately, speech signals are highly non stationary so a promising approach is to exploit this nonstationarity to separate their mixtures using only their second-order statistics [5], which leads to a joint diagonalization problem. This approach has been developed in two earlier papers of the authors [6, 7]. Actually, the idea of exploiting nonstationarity was introduced even earlier by Parra and Spence [1], but these authors used an ad-hoc criterion, while in our papers, a criterion based on the Gaussian mutual information and related to the maximum likelihood is used. Such a criterion has in fact been considered in [3], but without using the nonstationarity idea.

The main advantage of the frequency domain approach is that the calculations can be done in each frequency bin separately and independently, but it comes with a price. As the independence criterion is optimized independently, the separating matrices can be obtained only up to a scale change and a permutation. The scale ambiguity is inherent to the blind separation of convolutive mixtures, since it amounts to applying some filter to each signal and it is clear that such operations do not affect their independence. This ambiguity can be removed by using some *a priori* knowledge of the source signals or by setting constraints to the unmixing filters. So, the original sources cannot be generally recovered and one solution consists in estimating the contribution of the sources recorded on the sensors without the presence of the other sources. The scale ambiguity is fixed such that one output is as close as possible to one sensor by minimizing a mean square error (minimal distortion principle) [8]. This can be realized in the frequency domain by multiplying the outputs by the inverse of the unmixing matrix [9, 10].

The permutation ambiguity must be eliminated or reduced to a global ambiguity not dependent on the frequency. This is the main problem in a frequency domain approach. In the context of blind separation of audio signals, it is the biggest challenge and is still not satisfactorily solved. There have been many proposals to resolve the permutation ambiguity. The earlier works added a constraint to the separation filters by imposing a finite (short) time support [3] as permutations induce filters with infinite or very long tail responses. This idea may be impractical in this audio context, as for long responses the inverse is usually longer [3, 11, 12].

Two other approaches can also be envisaged. They exploit either the continuity of the unmixing filters or the time structure of speech signals. The first idea consists of ensuring the continuity of the separation filter frequency response [2, 3, 6, 13]. This is rather similar to imposing the constraint of short-time support, since such a constraint would entail some smoothness on the filter frequency response. The second idea is to exploit the time envelope structure and to add frequency coupling [2, 7, 9, 14]. These methods rely on the assumption of the comodulation of speech signals. Therefore, the source components belonging to the same source signal, but at different frequencies, should have similar shape in amplitude. Testing all the correlations on amplitude spectrograms [14] could greatly increase the complexity of the algorithm and simpler methods proposed to test only the correlation (or a distance) at one frequency bin with the sum of the aligned frequencies as reference [7, 9, 15] or to process first the channels that have the maximum signal energy [14]. In [16], the permutation is solved in increasing order of similarity and algorithm is implemented in a random frequency sequence. However, calculating the correlations over the whole frequency band is not always efficient as the time-frequency representation coming from the same source can vary considerably across frequency (especially for the higher frequencies) [15, 17]. The work [18] considers the correlation between the envelopes at neighbouring frequency bins, however, it is sensitive to any misaligned frequency bins. Further, the coherency at neighbouring frequencies only exists in a simple environment and does not hold in most cases [15, 19].

Another approach of addressing the problem is to apply beamforming techniques to the permutation alignment [20–27] in a sensor array context. Several methods also combined the previous approaches [10, 15, 20–22]. The work [15] proposed also to add a psychoacoustic filtering process to solve the problem.

This paper focuses on this challenging problem of permutation correction in the frequency domain and introduces a new method based both on the spectral continuity of the mixing filters and on the time variation of the signal energy in each frequency bin as well as its continuity across frequency. It extends earlier papers of the authors [6, 7]. First, the spectral continuity of the mixing (and therefore of the unmixing) filters is used in the initialization of the joint diagonalization algorithm. The exploitation of the continuity of the unmixing filters can perform quite well if the mixing filter does not contain strong echoes [6]. If not, the mixing filter frequency response matrix can be ill-conditioned for isolated frequency bins [6]. For those bins, the above method fails to identify correctly the permutations, as the estimated sources are still mixtures (with similar proportions) so it would be hard to determine to which source they correspond. Nevertheless, this method is efficient for most frequency bins and it tends to fail only on isolated frequency bins, which then produces permutation error on the whole frequency band delimited by those bins as the method forces the spectral continuity of the outputs. So, if there remain some frequency permutations to be corrected after this step, they appear as permutation jumps and not errors occurring on isolated bins.

The originality of this paper is then to introduce a new method based on the consideration of the smoothly time variation of the signal energy across frequency. The proposed algorithm is especially devoted to the detection of permutation jumps. The standard hypothesis of similar time-frequency representations coming from the same source [7, 9, 14, 18] is abandoned in this paper as observations show that they can vary strongly across frequency [15, 17] and that even correlation between the envelopes at neighbouring frequency bin is not always verified on experimental data [15, 19]. So, we only assume that they vary *smoothly* with frequency and that they are continuous across the frequency axis. Thus we work with time variation of the signal energy averaged on a sliding bandwidth around the processed bin, instead of the whole frequency band as in [9]. As only permutation jumps can occur, at each frequency bin, the method tests the continuity of all the averaged time variations of the signal energy across frequency. A short description of the method can also be found in an earlier conference paper [17]. The idea of the continuity of the time variation of the energy arises at the same time in [19] but is exploited in a different way, using reference frequencies.

The paper proposes an original frequency dependent distance in order to compare this continuity. For each bin and output, the time variations of the signal energy are averaged on a bandwidth around the processed bin. We compute first the difference between the averaged time variations of the signal energy as a continuity measure. In short, the method is looking at the bins where a sign change of all these measures appears across the time index. More precisely, the distance compares the continuity measure for the output itself and for the outputs associated with an imposed permutation. The two distances allow to distinguish the two situations and to solve efficiently the permutation ambiguity. The work [19] proposes a frequency-dependent distance between the processed bin f and the most reliable reference frequencies close to f. On the contrary, the proposed method does not need any reference as in [9, 19]. The additional information on the spectral diversity and continuity is powerful for quite short observations where conventional methods based on correlations on amplitude spectrograms [9, 14, 18] fail.

The paper is organized as follows. Section 2 describes the observation model for convolutive mixtures and the separation method based on the joint diagonalization of time varying spectra. Section 3 focuses on the permutation ambiguity problem and the methods to solve it. Finally, performance of the global separation method is investigated with simulation and experimental speech data in Section 4.

2. MODEL AND METHODS

The problem considered corresponds theoretically to the blind separation of convolutive mixtures: the observed sequences $\{x_1(t)\},\dots,\{x_K(t)\}$ are related to the source sequences $\{s_1(t)\},\dots,\{s_K(t)\}$ through a mixing filter with impulse response matrix $\{\mathbf{H}(n)\}$, of general element $\{H_{kj}(n)\}$, as

$$x_k(t) = \sum_{n=-\infty}^{\infty}\sum_{j=1}^{K} H_{kj}(n)s_j(t-n), \quad 1 \le k \le K. \tag{1}$$

The goal is to recover the sources through another filtering operation:

$$\mathbf{y}(t) = \sum_{n=-\infty}^{\infty} \mathbf{G}(n)\mathbf{x}(t-n), \tag{2}$$

where $\mathbf{x}(t) = [x_1(t)\cdots x_K(t)]^{\mathrm{T}}$ (T denoting the transpose), $\{\mathbf{G}(l)\}$ is the impulse response matrix of the separation filter and $\mathbf{y}(t) = [y_1(t)\cdots y_K(t)]^{\mathrm{T}}$ is the recovered source vector.

As one does not have any specific knowledge either of the source distributions or of the mixing filter, the idea is to adjust the separating filter such that the recovered sources are as independent as is possible. A direct time domain approach would mean minimizing some independence criterion (for the sequences $\{y_1(t)\},\dots,\{y_K(t)\}$), with respect to the matrix sequence $\{\mathbf{G}(n)\}$, assuming that one has truncated it to some finite sequence. The difficulty is that in audio applications the mixing filter often has a quite long impulse response which contains strong peaks corresponding to echoes, so the separating filter should also have long impulse response, hence there would be too many parameters to adjust. This would be computationally too heavy, not to mention the difficulty of ensuring the convergence of the optimization algorithm. In this context, the frequency domain approach seems to be more interesting (and is often adopted), since it reduces the problem to a set of independent separation problems of instantaneous mixtures associated with each frequency bin. Indeed, let $\mathbf{X}(t,f)$ (resp., $\mathbf{S}(t,f)$) be the vector composed of the N-points sliding discrete Fourier transforms (DFT) of the data block $[\mathbf{x}(t)\cdots\mathbf{x}(t+N-1)]$ (resp., $[\mathbf{s}(t)\cdots\mathbf{s}(t+N-1)]$) along the time axis t. With these notations, the mixing model (1) can be written approximately as

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

where $\mathbf{H}(f)$ denotes the frequency response of the mixing filter. The approximation comes from the fact that the DFT is based on finite stretches of data; it becomes exact as the data length N goes to infinity. The above model is an instantaneous mixing model for each frequency bin. Further, since the DFT at different frequencies tends to be independent, it is justified to treat the separation of instantaneous mixture problems independently. But the DFT also tends to produce nearly Gaussian variables while blind separation of instantaneous mixtures requires non-Gaussianity.[1] Fortunately, speech signals are highly nonstationary and one can exploit this feature to achieve separation using only second-order statistics. By adopting a second-order approach, we are in fact focused on the interspectra between the reconstructed sources at every frequency. But since we are dealing with nonstationary signals, we will consider the time varying spectra, that is the localized spectra around each given time point. It is precisely the time evolution of these spectra which helps us to separate the sources.

2.1. Joint diagonalization criterion

From (3), the time varying spectrum of the vector observation sequence $\{\mathbf{x}(t)\}$ is

$$S_{\mathbf{x}}(t,f) = \mathbf{H}(f)S_{\mathbf{s}}(t,f)\mathbf{H}^*(f), \tag{4}$$

where $S_{\mathbf{s}}(t,f)$ is the diagonal matrix with diagonal elements being the time varying spectra of the sources and $*$ denotes the transpose conjugated. The spectrum of the reconstructed source vector, which equals $\mathbf{G}(f)S_{\mathbf{x}}(t,f)\mathbf{G}^*(f)$, should be diagonal. Thus to perform the separation, a natural idea is to find matrices $\mathbf{G}(f)$ such that for each frequency f the matrices $\mathbf{G}(f)\widehat{S}_x(t,f)\mathbf{G}^*(f)$, at different time points t, are as close to diagonal as is possible, where $\widehat{S}_x(t,f)$ are estimates of $S_x(t,f)$. This idea has been exploited by Parra and Spence [1, 13], but they use a different diagonality criterion from ours. The one we use is the same as in [5] in the instantaneous case and comes from the maximum likelihood and/or the mutual information approach. A similar criterion also in the instantaneous case has been proposed in [28] but without link to the maximum likelihood. This criterion has also been considered in [3] in the convolutive case but without using the nonstationarity idea. Experiments realized in the case of instantaneous mixtures show that it is a powerful criterion [5]. Besides, we have developed a simple and very fast algorithm to perform joint approximate diagonalization based on minimizing this criterion [29]. For a single matrix $\mathbf{G}(f)\widehat{S}_x(t,f)\mathbf{G}^*(f)$, the diagonality measure is given by

$$\begin{aligned}\frac{1}{2}\Big\{&\log\det\operatorname{diag}\big[\mathbf{G}(f)\widehat{S}_{\mathbf{x}}(t,f)\mathbf{G}^*(f)\big]\\ &-\log\det\big[\mathbf{G}(f)\widehat{S}_{\mathbf{x}}(t,f)\mathbf{G}^*(f)\big]\Big\},\end{aligned} \tag{5}$$

[1] This does not mean that one cannot separate the sources but only that higher (than second) order moments of the DFT are of little use and one has to consider also cross higher order moments between the DFT at different frequencies. But this would require treating all the separation of instantaneous mixture problems simultaneously and not independently.

where $\operatorname{diag}(\cdot)$ denotes the operator which builds a diagonal matrix from its argument. But the last term equals $2\log|\det\mathbf{G}(f)| + \log\det\hat{S}_x(t,f)$ and the term $\log\det\hat{S}_x(t,f)$ being constant, can be dropped. Therefore a global diagonality criterion can be written as

$$\sum_t \left\{ \frac{1}{2}\log\det\operatorname{diag}\left[\mathbf{G}(f)\hat{S}_x(t,f)\mathbf{G}^*(f)\right] - \log|\det\mathbf{G}(f)|\right\}, \tag{6}$$

where the summation is over the time points of interest. This criterion is to be minimized with respect to $\mathbf{G}(f)$ to obtain the frequency response of the separation filter. Note that such minimization can be done in each frequency bin separately and independently, using the fast joint diagonalization algorithm [29].

2.2. Spectral estimation

The first step in the separation procedure is to estimate the (time varying) spectral matrix of the observation sequences appearing in the criterion (6). It is important to have good estimators since the quality of the separation depends on their accuracy, as all subsequent calculations are based on these estimators. Specifically, we will need a very high frequency resolution, as the mixing filter frequency responses present rapid variations (due to their long impulse responses) and this forces us to work with very narrow frequency bins. We also need a good time resolution in order to fully exploit the nonstationarity of the source signals (and also for the "profile" method in Section 3 to work well). Of course both high frequency and time resolutions would result in a larger variance of the estimator, so some compromise must be reached. But in the present situation, high resolutions should be given more importance than low variance.

There are several ways to estimate the spectrum of a (multivariate) signal [30]. We focus on frequency domain methods as time domain methods are too costly since a large number of lags would be needed. Since we are dealing with time varying spectra, the simplest way is to subdivide the data sequence into consecutive blocks and estimate the spectrum as if the data inside each block came from a stationary process. A common (frequency domain) estimation method is to compute the DFT of the data block, forming the periodogram and then averaging it over consecutive frequencies. In practice, we find that this method lacks flexibility since we have few choices for the number of frequencies to average: due to the required high resolution, the choices reduce to 3 and 5. Also, the block length should be a power of 2 in order to benefit from the fast Fourier transform, so its choice is also very limited. Therefore, we will adopt another method which is also common in the case of nonstationary signals. We will work with shorter block lengths and further introduce a taper before applying the DFT. The tapered periodogram is now averaged not over frequency but over time using *sliding data blocks*. The number of data blocks to be averaged is related to the time resolution and can be easily fine tuned. The block length is related to the frequency resolution and can also be adjusted to a large degree, since this length is not so large and the use of a taper makes it possible to have an effective block length of any size. We first form the short term *sliding* periodogram using a *Hanning taper window*

$$\begin{aligned} P_{\mathbf{x}}(\tau,f) = {} & \frac{1}{\|H_N\|^2}\left[\sum_t H_N(t-\tau)\mathbf{x}(t)e^{2\pi ift}\right] \\ & \times \left[\sum_t H_N(t-\tau)\mathbf{x}(t)e^{2\pi ift}\right]^*, \end{aligned} \tag{7}$$

where H_N is the Hanning taper window of length, N: $H_N(t) = 1 - \cos(2\pi t/N + \pi/N)$ for $0 \le t < N, 0$ otherwise, and $\|H_N\|^2 = \sum_{t=0}^{N-1} H_N^2(t)$ (which equals $3N/2$). This periodogram will be averaged over m consecutive equispaced points $\tau_1, \dots, \tau_m$ yielding the estimated spectrum at time $(\tau_1 + \tau_m + N - 1)/2$:

$$\hat{S}_{\mathbf{x}}\left(\frac{\tau_1 + \tau_m + N - 1}{2}, f\right) = \frac{1}{m}\sum_{k=1}^{m} P_x(\tau_k, f). \tag{8}$$

The frequencies are taken to be of the form $f = n/N, n = 0, \dots, N/2$, with N being chosen to be a power of 2, to take advantage of the fast Fourier transform. Thus the spectrum is estimated at a frequency spacing of $1/N$, but the real frequency resolution is lower due to tapering. The use of tapering also helps to reduce the bias of the estimator. It is also possible to choose N, not to be a power of 2, by padding zeros to the tapered data block to increase its length to the next power of 2. This doesn't change the real frequency resolution but only increases the number of frequency points at which the spectrum is estimated. The time resolution is determined by $m\delta$, where $\delta = \tau_i - \tau_{i-1}$ is the spacing between the τ_i. Using $\delta \gg 1$ helps to reduce the computational cost but slightly degrades the estimator: actually δ can be a small fraction of N without a significant degradation. Of course a compromise between time and frequency resolution has to be made to get a reasonably low variance of the estimator. The interest of the chosen spectral estimation is that this compromise is easier to obtain than with other spectral estimations [6, 7].

2.3. The scale and permutation ambiguity problems

The frequency domain approach has the great advantage that the calculations can be done in each frequency bin separately and independently. This is very important since in the present application the number of these bins must be very large as the response of the separation filter could be very long. A time domain approach would require the minimization of some criteria with respect to a very large number of parameters, which is too costly. By contrast, in our approach, for each frequency bin, one only has a small minimization problem, which can be solved very quickly. There is however a price to be paid for this. The joint diagonalization of the time varying spectra $S_{\mathbf{s}}(t,f)$ only provides the matrices $\mathbf{G}(f)$ up to a scale change and a permutation: if $\mathbf{G}(f)$ is a solution, then so is $\mathbf{\Pi}(f)\mathbf{D}(f)\mathbf{G}(f)$ for any diagonal matrix $\mathbf{D}(f)$ and any permutation matrix $\mathbf{\Pi}(f)$. Thus, one only gets a separation filter of frequency response matrix of the form

$$\mathbf{G}(\mathbf{f}) = \mathbf{\Pi}(\mathbf{f})\mathbf{D}(\mathbf{f})\hat{\mathbf{H}}^{-1}(\mathbf{f}), \tag{9}$$

where $\widehat{H}(f)$ is a consistent estimator of $H(f)$, but $\mathbf{\Pi}(f)$ and $\mathbf{D}(f)$ are *arbitrary* permutation and diagonal matrices.

It should be noted that the above ambiguity problem is not really related to the frequency domain approach but to the use of a criterion such as (6) which expresses the mutual dependence of the signals in a decoupling way in the frequency domain. The scale ambiguity can be removed by reconstructing the *i*th output as close as is possible to the contribution of the *i*th source on the *i*th sensor (or minimal distortion principle) [8–10]. The scale ambiguity is solved in the experimental results by applying frequency domain Wiener filtering between outputs and sensors, where outputs act as reference signals. However, the permutation ambiguity is a more difficult problem which is still open. The main novelty of this work is a method to resolve this crucial problem. The algorithm is described in detail in the next section.

3. RESOLVING THE PERMUTATION AMBIGUITY

Several ideas have been introduced to resolve the permutation ambiguity, as detailed in the introduction. The first one consists in constraining the separating filters with short support FIR structures in the time domain [2, 3]. It may be not useful, as the mixing filter response is already quite long and for long responses the inverse is usually longer [3, 11, 12]. Other ideas are to exploit a continuity assumption on the frequency response of the unmixing filters [2, 3, 13] or to add frequency coupling [2, 7, 9, 14, 15, 17–19, 31], for example, in the adaptation parameters to preserve the same permutation [2, 14].

Several methods also used geometric information such as beam patterns [20–22, 25] direction of arrival and source location [24, 27]. It seems to be an effective approach without too much multi-path propagation and with distinct localization of sources. Unfortunately, classification based on the estimated location tends to be inconsistent especially in a reverberant environment [24] and needs additional methods such as inter-frequency correlation for neighbouring bins [18] to solve the permutation problem for all bins [24].

In [6] we have proposed a method to solve the permutation ambiguity problem based on the continuity of the frequency response of the separation filter, which is more or less equivalent to constraining this filter to have short support in the time domain [2, 3, 13]. It has the advantage that it relies only on the weak assumption that the frequency response $\mathbf{H}(\mathbf{f})$ of the mixing filter is continuous and requires a very little computational cost. However, it has a main weakness that it can leave wrong permutations over a block of contiguous frequency bins. In this paper, a method is proposed to address this weakness.

3.1. Overview of our earlier works

The method in [6] assumes that $\mathbf{H}(f)$ is continuous and hence the frequency response $\mathbf{G}(f)$ of the separating filter should also be continuous. But a permutation function cannot be continuous unless it is a constant function, this constraint reduces the ambiguity with respect to a permutation *varying with the frequency* to that with respect to a global fixed permutation. This global permutation ambiguity is unavoidable, since it corresponds to simply permuting the recovered sources. In practice, $\mathbf{G}(\mathbf{f})$ will be available only over a finite regular grid of frequencies $f_0 < \cdots < f_L$, say. To detect permutation change, one may look at the "ratio" $\mathbf{G}(\mathbf{f_l})\mathbf{G^{-1}}(\mathbf{f_{l-1}})$ and test for its closeness to a diagonal matrix. Indeed, by using the representation (9), this ratio can be written as:

$$\mathbf{\Pi}(\mathbf{f_l})[\mathbf{D}(\mathbf{f_l})\widehat{\mathbf{H}}^{-1}(\mathbf{f_l})\widehat{\mathbf{H}}(\mathbf{f_{l-1}})\mathbf{D}^{-1}(\mathbf{f_{l-1}})]\mathbf{\Pi}^{-1}(\mathbf{f_{l-1}}). \tag{10}$$

Since the function $\mathbf{H}(\cdot)$ is continuous, $\widehat{H}^{-1}(f_l)\widehat{H}(f_{l-1})$ is nearly the identity matrix, hence the matrix product in the above square bracket [] is nearly a diagonal. Left and right multiplying this matrix by $\mathbf{\Pi}(f_{l-1})$ and $\mathbf{\Pi}^{-1}(f_{l-1})$ results in the same matrix with its rows and columns permuted by the same permutation, which is thus also nearly diagonal. Therefore $\mathbf{G}(f_l)\mathbf{G}^{-1}(f_{l-1})$ appears as the product of $\mathbf{\Pi}(f_l)\mathbf{\Pi}^{-1}(f_{l-1})$ with a nearly diagonal matrix. Thus a permutation change can be detected by examining all permutations of the rows of $\mathbf{G}(f_l)\mathbf{G}^{-1}(f_{l-1})$ and picking the one for which the resulting matrix is closest to diagonal in some sense. If the obtained permutation is not an identity then there is a permutation change, which can then be corrected using this obtained permutation.

The above method is quite simple and cheap (except when the number of sources is large). In practice however we find that one can achieve comparable performance by another simpler and cheaper method, relying on the particular behaviour of the joint (approximate) diagonalization algorithm. This algorithm operates iteratively by transforming successively the matrices to be diagonalized by left and right multiplying them by an appropriate matrix and its transpose conjugated, and each time between two candidates for such a matrix, differing only by a permutation, the one which is closer to the identity matrix (in some sense) is chosen [29]. Thus, instead of jointly diagonalizing the matrices $\widehat{S}_{\mathbf{x}}(t, f_l)$ we jointly diagonalize the matrices $\mathbf{G}(f_{l-1})\widehat{S}_{\mathbf{x}}(t, f_l)\mathbf{G}^*(f_{l-1})$, where $\mathbf{G}(f_{l-1})$ is the solution to the previous problem of joint diagonalization of the $\widehat{S}_{\mathbf{x}}(t, f_{l-1})$. By continuity, we expect that the matrices $\mathbf{G}(f_{l-1})\widehat{S}_{\mathbf{x}}(t, f_l)\mathbf{G}^*(f_{l-1})$ are already rather close to diagonal so that a solution to their joint diagonalization problem is nearly the identity matrix and the algorithm would pick this solution (up to possibly a row scale change). Thus, the algorithm would produce a matrix ratio $\mathbf{G}(f_l)\mathbf{G}^{-1}(f_{l-1})$ close to a diagonal matrix and hence no subsequent permutation correction is needed. A side advantage of this method is that the joint diagonalization algorithm converges faster since it is better initialized, thus reducing the computational cost.

Although the above method can correct most frequency permutation errors, its weakness is that even a single wrong correction (e.g., in non invertible bins) can cause wrong permutations over a large block of frequency, that is, permutation jumps. If, at one frequency f_l, a source has been wrongly permuted versus frequency bin f_{l-1}, then the solution will remain on that permuted source in frequency bins $f_{l+1}, f_{l+2}, \ldots$ by forcing the continuity assumption.

To avoid this problem and eliminate these frequency permutation jumps, a complementary method based on an idea similar to that in [2, 9, 14, 18], which introduces some frequency coupling, is proposed in [7]. The glottis is the main source of energy for speech production and emits a broadband sound with spectral peaks at the harmonics of the speaker's pitch frequency. Then the vocal tract filters this broadband sound and the resulting speech signal can be seen as an amplitude modulation due to the succession of phonemes which constitutes speech. Based on this observation, the main idea is that, for a speech signal, the energy over different frequency bins appears to vary in time in a similar way, up to a gain factor. For example, one would expect that its energy would be nearly zero in all frequency bins in a period of pause and be maximum in all frequency bins for speech periods. Several papers evaluate the similarity (or correlations) between the envelopes of separated signals. To check this similarity, [14] proposes to recover the permutation ambiguity by considering correlations on amplitude spectrograms, that is, the modulus of the time varying spectra. But this is awkward and very time consuming as there are $K^2L(L-1)/2$ correlations to be computed, L denoting the number of frequency bins. The method can be also implemented in an iterative way by first processing the channels that have the maximum signal energy [14]. The sequence of frequency bins used to solve the permutation ambiguity is determined in [16] by sorting the similarity in an increasing order. In [9], the correlation is tested at each frequency bin and the sum of the aligned frequencies is taken as a reference.

In the same way, the method proposed in [7] simplifies the problem by associating each frequency bin with a profile (of relative variation of the spectral energy) and compares it with a reference profile. More specifically, after joint diagonalization, the spectra of the reconstructed sources $\widehat{S}_y(t,f)$ can be computed as the kth diagonal element of $\mathbf{G}(f)\widehat{S}_x(t,f)\mathbf{G}^*(f)$. As each spectrum is recovered up to a gain factor, we consider the "profiles" $E(f,k,\cdot)$, defined as the logarithm of the kth diagonal element of $\mathbf{G}(f)\widehat{S}_x(\cdot,f)\mathbf{G}^*(f)$. Thus, they are defined up to an additive constant. Hence by centering all profiles by subtracting their time averages, the additive constant is eliminated and the notation E' will be used for centered profiles. In [7], these profiles are compared with reference profiles associated with each source (but not dependent on the frequency) to determine which sources they come from. The reference profiles are not fixed as in [9], but, in turn, are constructed iteratively by averaging profiles associated with different frequencies and previously identified as coming from the same sources. The basic assumption is that profiles from the same sources, but at different frequencies, are still more similar than those from other sources. Therefore, the iterative algorithm determines the permutation corrections such that the sum of squared distances between profiles coming from a source (after permutation correction) to its reference profiles is minimum. The algorithm however needs a good initialization for the reference profiles, and for this end the method based on the continuity assumption of the frequency response of the mixing filter is used.

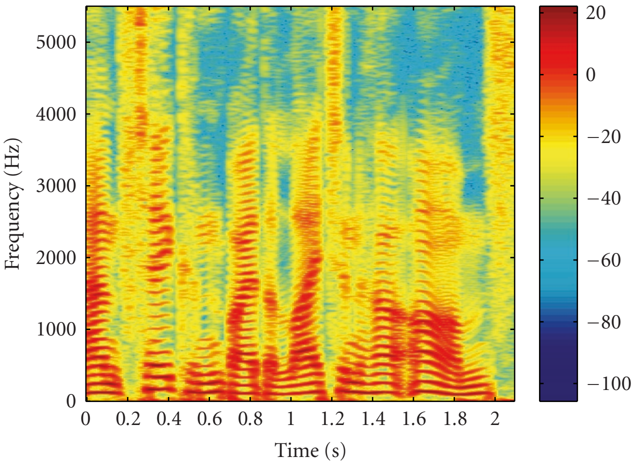


Figure 1: Time-frequency representation of a speech signal in dB.

3.2. *The proposed method*

The method in [7] assumes that profiles coming from the same sources, but at different frequencies, are still more similar than those from other sources. It is the implicit idea of methods relying on the correlations on amplitude spectrograms or on neighbouring frequency bins [2, 9, 14, 18]. It implies that the time-frequency representation (or profiles) of distinct sources must be different enough. For example, speakers should have different speech periods and pause periods (and not synchronous ones), at least at some part of the processed observations. This may not be completely true for short signals. A second problem is that, in fact, profiles coming from the same source can vary considerably with frequency (see Figure 1) [15, 17]. Further, the coherency at neighbouring frequencies can exist only in a simple environment and this hypothesis does not hold in most cases [15, 19]. For these reasons, considering the correlations between the envelopes over the whole frequency band or even at neighbouring frequency bins is not always efficient.

In this paper we abandon this assumption and only assume that profiles vary *smoothly* with frequency. The hypothesis of the continuity of the time variation of the source energy also arises in [19], but is exploited in a different way, using reference frequencies. The great interest of the proposed method is that no frequency reference or profile reference is needed to introduce a distance. This additional information on the spectral diversity and the spectral continuity will allow us to use shorter observations. Thus we work with profiles averaged on a bandwidth $[f_{l-M}, f_{l+M}]$ instead of profiles averaged on the whole frequency band:

$$F_{\mathbf{y}}(f_l,k;\cdot)=\frac{1}{2M+1}\sum_{n=l-M}^{l+M}E'(f_n,k;\cdot). \qquad (11)$$

These averaged profiles are used to detect the block permutation errors arising after the stage of joint diagonalization of time varying spectra [6] with adaptation to ensure continuity of the frequency response of the separating filter, as explained in the previous subsection. Thus, after this stage,

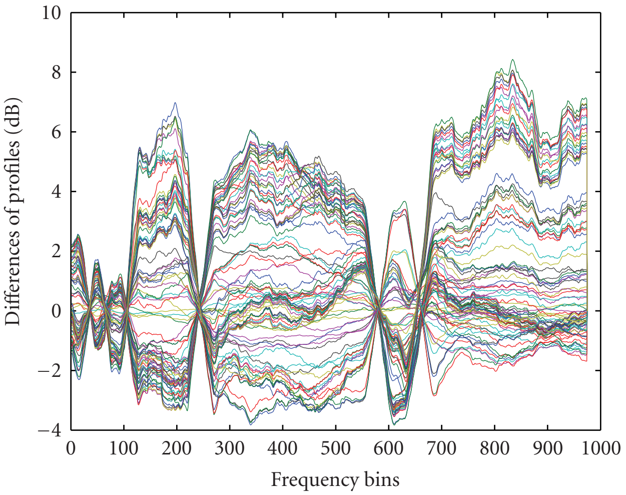


Figure 2: Differences between averaged profiles in function of frequency bin for each time index.

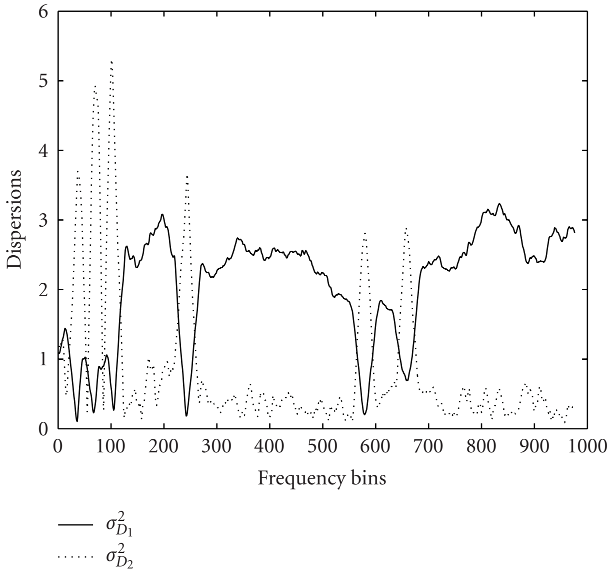


Figure 3: Dispersions $\sigma^2_{D_1}$ (solid) and $\sigma^2_{D_2}$ (dotted) before permutation correction in function of frequency index k.

there can remain only some frequency permutation jumps to detect. Such jumps may happen at the frequency bins where the mixing filter frequency response matrix is ill-conditioned [6].

Consider for simplicity the case of two sources and two sensors, we look at the difference between the profiles of the two reconstructed sources after the above stage of separation:

$$D_1(f,k) = F_{\mathbf{y}}(f,k;1) - F_{\mathbf{y}}(f,k;2). \tag{12}$$

Suppose there is a permutation of the separation filter $\mathbf{G}(f)$ at frequency bin f_l. Between f_{l-M} and f_{l+M}, the two outputs correspond to two different sources and the profiles are also permuted,

$$\begin{aligned} D_1(f_{l-M},k) &= F_{\mathbf{S}}(f_{l-M},k;1) - F_{\mathbf{S}}(f_{l-M},k;2), \\ D_1(f_{l+M},k) &= F_{\mathbf{S}}(f_{l+M},k;2) - F_{\mathbf{S}}(f_{l+M},k;1). \end{aligned} \tag{13}$$

If we assume that the averaged profiles are changing slowly enough, the difference $D_1(f_{l-M},k)$ and $D_1(f_{l+M},k)$ will be of opposite sign, whatever the time index k. To illustrate the assumption, two speech signals have been convolved with premeasured room responses (detailed in Section 4). After the step of joint diagonalization, the averaged profiles have been computed for these outputs as well as functions $D_1(f,k)$. We know that six frequency jumps remain since the mixing system is accessible. The curves $D_1(f,k)$ are plotted in Figure 2 as a function of f, for each time index k. These curves change sign correctly at the six frequencies where the sources must be permuted. If we examine the same curves after elimination of the permutations (not shown here), we notice that all the sign changes have disappeared. It can be deduced from this, that at each frequency bin f_l where the sources are permuted, the dispersion of the values $D_1(f_l,k)$ will be minimum. The minima can then detect the beginning and the end of a frequency block to permute. Suppose that the time-frequency representation is computed on L time blocks. As the profiles are centered by construction, the mean value of $D_1(f_l,k), k = 1,\dots,L$ is zero and its dispersion is

$$\sigma^2_{D_1(f_l)} = \sum_{k=1}^{L} D_1^2(f_l,k). \tag{14}$$

The dispersion $\sigma^2_{D_1(f)}$ of the data $D_1(f,\cdot)$, shown in Figure 2, is plotted by the solid line in Figures 3 and 4, before and after performing permutation correction. In Figure 3, the six minima are actually permutation (jump) frequencies. They occur correctly at the six sign changes (see Figure 2). After permutation correction, these minima disappear, as can be seen in Figure 4.

In order to detect a possible permutation at any frequency bin f_l, we introduce a second function difference $D_2(f,k)$ based on new profiles $H_{\mathbf{y}}(f,k;\cdot)$ of outputs $\mathbf{y}(t)$. Similar to $F_{\mathbf{y}}(f,k;\cdot)$, they are constructed by averaging on the bandwidth $[f_{l-M}, f_{l+M}]$, but we impose a permutation on the second part of the band $[f_{l+1}, f_{l+M}]$. The outputs are permuted on the band $[f_{l+1}, f_{l+M}]$ versus the outputs on the band $[f_{l-M}, f_l]$:

$$\begin{aligned} H_{\mathbf{y}}(f_l,k;\cdot) =& \frac{1}{2M+1} \\ & \times \left(\sum_{n=l-M}^{l} E'(f_n,k;\cdot) + \sum_{n=l+1}^{l+M} E'(f_n,k;\pi) \right), \end{aligned} \tag{15}$$

where π denotes the permutation between the two outputs. A second difference $D_2(f,k)$ and its dispersion $\sigma^2_{D_2(f_l)}$ can be

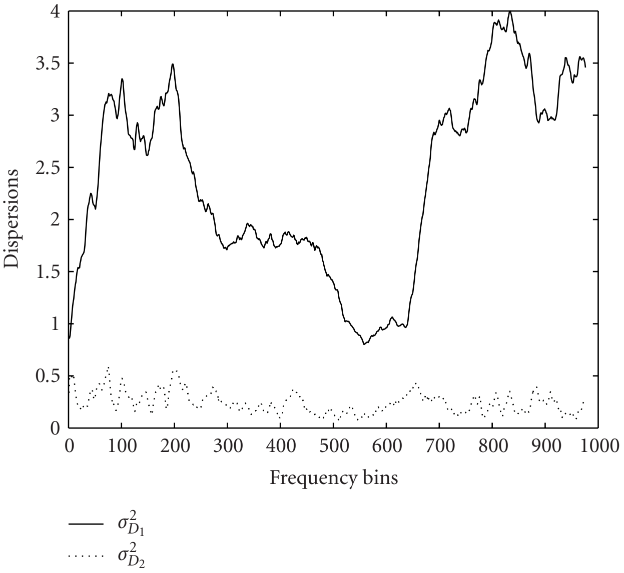


Figure 4: Dispersions $\sigma^2_{D_1}$ (solid) and $\sigma^2_{D_2}$ (dotted) after permutation correction in function of frequency index k.

calculated with the new averaged profiles:

$$\begin{aligned} D_2(f,k) &= H_{\mathbf{Y}}(f,k;1) - H_{\mathbf{Y}}(f,k;2), \\ \sigma^2_{D_2(f_l)} &= \sum_{k=1}^{L} D_2^2(f_l,k). \end{aligned} \tag{16}$$

The dispersion $\sigma^2_{D_2(f_l)}$ is plotted by the dotted line before (Figure 3) and after (Figure 4) elimination of the permutation. If f_l is a permutation frequency, $H_{\mathbf{Y}}(f_l,k;\cdot)$ will be the profiles of the corrected sources and the dispersion $\sigma^2_{D_2(f_l)}$ will be bigger than $\sigma^2_{D_1(f_l)}$ as there will be no sign change in the difference of profiles $H_{\mathbf{Y}}(f_l,k;\cdot)$. The two curves $\sigma^2_{D_1(f_l)}$ and $\sigma^2_{D_2(f_l)}$ cross when permutation must be detected. On the contrary, when a frequency band is correctly permuted, the profiles $F_{\mathbf{Y}}(f,k;\cdot)$ are good and the dispersion $\sigma^2_{D_1(f)}$ is maximum in this band and bigger than $\sigma^2_{D_2(f)}$. The curves do not cross in this band. When all permutations are corrected, the profiles $H_{\mathbf{Y}}(f,k;\cdot)$ only add false permutations and impose sign changes in the function $D_2(f,k)$. The dispersion $\sigma^2_{D_2(f)}$ is then always smaller than $\sigma^2_{D_1(f)}$.

The permutation detection can be done in an iterative way as follows.

(1) Computation of $\sigma^2_{D_1(f)}$ and $\sigma^2_{D_2(f)}$, and detection of the global minimum of $\sigma^2_{D_1(f)}$, which occurs at f_l, say.
(2) Permutation of the two outputs for all frequencies higher than f_l.
(3) Computation of the new profiles $F_{\mathbf{Y}}(f,k;\cdot)$ and $H_{\mathbf{Y}}(f,k;\cdot)$, the new functions $\sigma^2_{D_1(f)}$ and $\sigma^2_{D_2(f)}$, redetection of the new global minimum of $\sigma^2_{D_1(f)}$, and so on until $\sigma^2_{D_1(f)} > \sigma^2_{D_2(f)}$ for all f.

This method is easy to implement and shows quite good results even for short signals. The number of iterations is exactly the number of permutation corrections to adjust, which is usually small, as in the diagonalization stage we have made use of the continuity of the mixing filter frequency response.

4. DESIGN AND RESULTS

The first subsection is devoted to the illustration of the improvement of the method with simulation results. It shows the behaviour of the permutation correction when the source profiles vary strongly with frequency (see Figure 1). Such sources were artificially mixed with premeasured room impulse responses. The resulting mixtures have been already used in Section 3 to illustrate how the proposed method for solving the permutation ambiguity operates. In the second subsection, real-room recordings are exploited to compare the proposed method to some of the state-of-the-art methods for convolutive BSS.

4.1. Simulation results

We considered mixtures of real sound sources from premeasured room impulse responses of a conference room. The last are provided by the Matlab routine roommix.m of Alex Westner (found at http://sound.media.mit.edu/ica-bench), which uses a library of impulse responses measured in a real 3.5 m×7 m×3 m conference room. Two and a half walls of the room are covered with whiteboards, one wall is covered with a projection screen and a large table sits in the middle of the room. There are eight microphones hanging from the lighting grid of the room, spaced about half-meter apart from one another (the experiment is detailed in [12]). The user specifies the positions of the sensors and the sources (using 8 preset positions). We chose distances between sources and sensors around 50 cm and 1 m. Two speech signals of 2 s sampled at 11 kHz (24000 samples) are convolved with the premeasured room impulse responses to build up two observations. These responses are quite long, up to 8192 lags, but become quite small at high lags so that we can truncate them to 256 lags and still retain all echoes. The four impulse responses are shown in Figure 5.

We also used these two mixtures in Section 3 to illustrate how the proposed method for solving the permutation ambiguity operates. The time-frequency representation of the first source is represented in Figure 1. Figures 2, 3, and 4 show the profiles and their dispersions of the separated sources after the stage of joint diagonalization. The spectral matrices are estimated as detailed in Section 2, using a block length of $N = 2048$ with an overlap of $1 - (\delta - 1)/N = 75\%$ (yielding 41 time blocks). The averaged profiles $F_{\mathbf{Y}}(f,k;\cdot)$ are constructed by averaging on 50 frequency bins ($M = 25$). After the above stage of separation by joint diagonalization, certain permutation errors have been eliminated by way of forcing the continuity of the frequency responses. Yet, there can still remain permutation jumps. As we know the mixing systems, we can consider the separation index, defined as

$$r(f) = \left|(\mathbf{GH})_{12}(f)(\mathbf{GH})_{21}(f)/\left[(\mathbf{GH})_{11}(f)(\mathbf{GH})_{22}(f)\right]\right|^{1/2}, \tag{17}$$

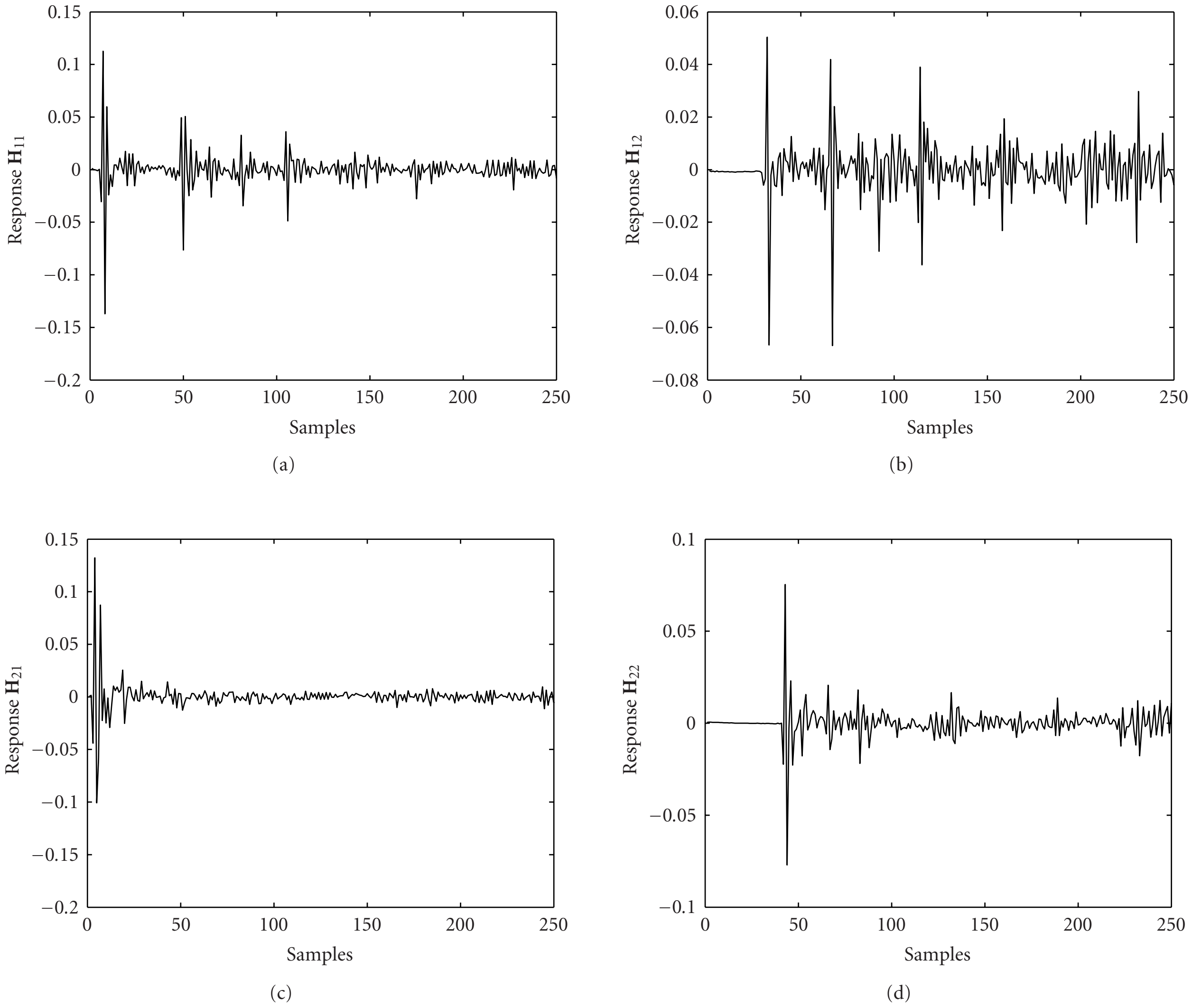


FIGURE 5: The four impulse responses of the mixing filter.

where $(\mathbf{GH})_{ij}(f)$ is the ij element of the matrix $\mathbf{G}(f)\mathbf{H}(f)$. For a good separation, this index should be close to 0 or infinity (in this case the estimated sources are permuted). When r crosses the value 1, this means that a permutation has occurred. Therefore we plot both $\min(r, 1)$ and $\min(1/r, 1)$ versus frequency (in Hz), using different line styles (dots and solid) to distinguish them. Figure 6 shows these curves, before and after applying the new method of frequency permutation correction. It is clear from the first curve that six frequency jumps are present after the separation step. It can also be mentioned that the two curves $\min(r, 1)$ and $\min(1/r, 1)$ are quite distinct. One is close to zero whereas the second one is close to 1. This means that the separation has been well achieved up to a permutation, except at some isolated frequency bins. Moreover, the second plot (corresponding to the separation index after the permutation correction) shows that the new method eliminates all permutation errors (relative to a global permutation) since the two curves do not cross.

To validate the whole BSS method (e.g., separation and permutation correction), we reconstructed the four impulse responses of the global filter $(\mathbf{G} * \mathbf{H})(n)$ between the two sources and the two sensors. They are plotted in Figure 7. One can see that $(\mathbf{G} * \mathbf{H})_{11}(n)$ is much higher than $(\mathbf{G} * \mathbf{H})_{12}(n)$ and $(\mathbf{G} * \mathbf{H})_{22}(n)$ is also bigger than $(\mathbf{G} * \mathbf{H})_{21}(n)$, meaning that the sources are well separated (and permuted). This will be also revealed afterwards by calculating the noise-reduction rate.

The efficiency of the whole separation procedure can be confirmed by looking at the original sources, the mixtures, and the separated sources, displayed in Figure 8. To quantify the performance, signal-to-noise ratio (SNR) is computed before and after separation. For one observation, one source is considered as “signal” and the second one as “noise”. In that sense, the SNR values of the two mixtures were equal to 3.3 dB and −3.7 dB. The SNR values of the outputs have been improved until 20.4 dB and 17.7 dB with the proposed method. Usually, BSS is compared with the noise-reduction

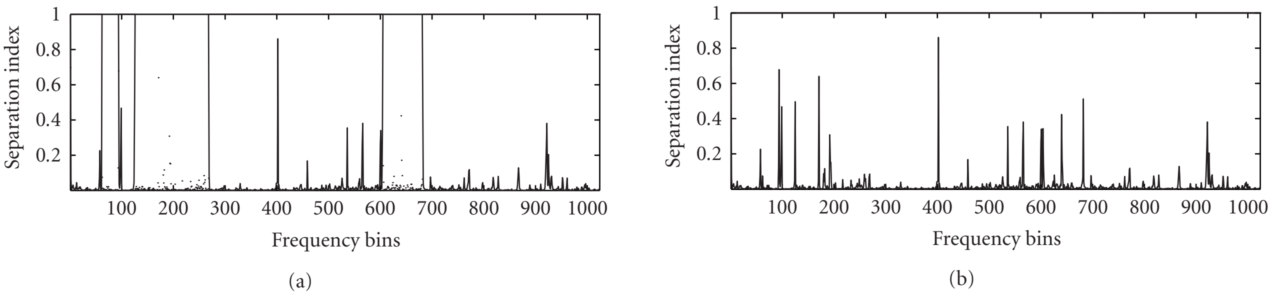


Figure 6: Separation index (dots) and its inverse (solid) truncated at 1 (a) before and (b) after applying the proposed permutation correction algorithm.

rate, defined as the output SNR in dB minus the input SNR. In that experiment, the noise-reduction rates were equal to 16.7 dB and 21.4 dB, which are really efficient on such short observations (here 2 s).

4.2. *Experimental results*

Experiments were conducted at the McMaster University in the context of hearing aid design. McMaster University recorded in the BLISS project a database of real-room recordings: live-capture audio mixtures and a realistic hearing in noise test environment (R-HINT-E) (http://www.lis.inpg.fr/pages_perso/bliss/). A human head and torso model called KEMAR were placed in the centre of three rooms. KEMAR has in each ear a small microphone. A single loudspeaker was moved to different locations around KEMAR with different angles from 0° to 180°. For each of the seven locations, six sentences were played and recorded on the two microphones. In addition, for each location, the room impulse response was measured. The database created by McMaster University is very useful for comparison studies of algorithms as it provides real-room mixtures as well as the true sources.

Several BSS algorithms have been evaluated and compared in a 2-source 2-microphone system, using the real convolved sources captured on the two microphones and coming from two loudspeakers. The loudspeakers were moving from 0° to 180° around the human model at distance of 1.4 m. This corresponds to 21 different mixtures (without repetitions and without equal angles). The chosen room is a reverberant classroom with dimensions 5.3 m by 10.3 m. The reverberant time is around 130 ms.

Several approaches have been developed to solve the permutation ambiguity: in short, exploiting the continuity of the spectra of recovered signals or the separation matrix [2, 13], exploiting the time structure of the source components [9, 14], or applying beamforming techniques if enough sensors are available. In a 2-source 2-microphone system, methods using beamforming alignment cannot be employed. Thus, the proposed method is compared to some of the state-of-the-art methods for convolutive BSS exploiting either the spectral continuity (algorithm of Parra and Spence [13]) or the time envelope structure (algorithm of Murata et al. [9]). The algorithm of Murata et al. [9] is found at http://www.ism.ac.jp/~shiro/. The implementation for the Parra-Spence algorithm has been provided by S. Harmeling.[2]

In the case of synthetic data (artificially convolved with premeasured impulse responses), the BSS performance is commonly evaluated in terms of the signal-to-interference ratio (SIR) and signal-to-distortion ratio (SDR) of each output $\mathbf{y}(t) = [y_1(t) \cdots y_K(t)]^T$, where

$$y_i(t) = \sum_{k=1}^{K} G_{ik} * x_k(t) = \sum_{j=1}^{K} (G * H)_{ij} * s_j(t) = \sum_{j=1}^{K} y_{ij}(t). \tag{18}$$

A solution for solving the scaling problem can be obtained by the minimal distortion principle. The output $y_i(t)$ is calculated to be as close as is possible to the contribution of the ith source on the ith sensor. As the outputs are uncorrelated, $y_i(t)$ can be reconstructed by minimizing a quadratic error between $y_i(t)$ and $x_i(t)$. In the experiment, the quadratic error was defined in the frequency domain. The output $y_i(t)$ is so calculated such that $\sum_t \|X_i(t, f) - Y_i(t, f)\|^2$ is minimized for each frequency bin. It leads to the classical Wiener filter between $y_i(t)$ and $x_i(t)$, expressed in the frequency domain. Therefore, $y_i(t)$ aims at the reconstruction of the contribution of the ith source on the ith sensor.

The SIR for $y_i(t)$ is then defined as the ratio of the power of the portion of $y_i(t)$ coming from source i, $y_{ii}(t)$, to the power from jammer signals, $y_{ij}(t)$:

$$\mathrm{SIR}\, i = 10 \log \frac{\sum_t y_{ii}(t)^2}{\sum_t \sum_{j \neq i} y_{ij}(t)^2}. \tag{19}$$

In the case of real world situations, we have generally no access to the source signals. However, the SIR can still be computed if just one of the sources is active during a certain time interval. In the database, we have also access to the microphone signals $x_{ki}(t)$ $k = 1, \ldots, K$, recorded when only the ith source is present. Therefore, the SIR will be calculated

[2] http://ida.first.gmd.de/~harmeli/.

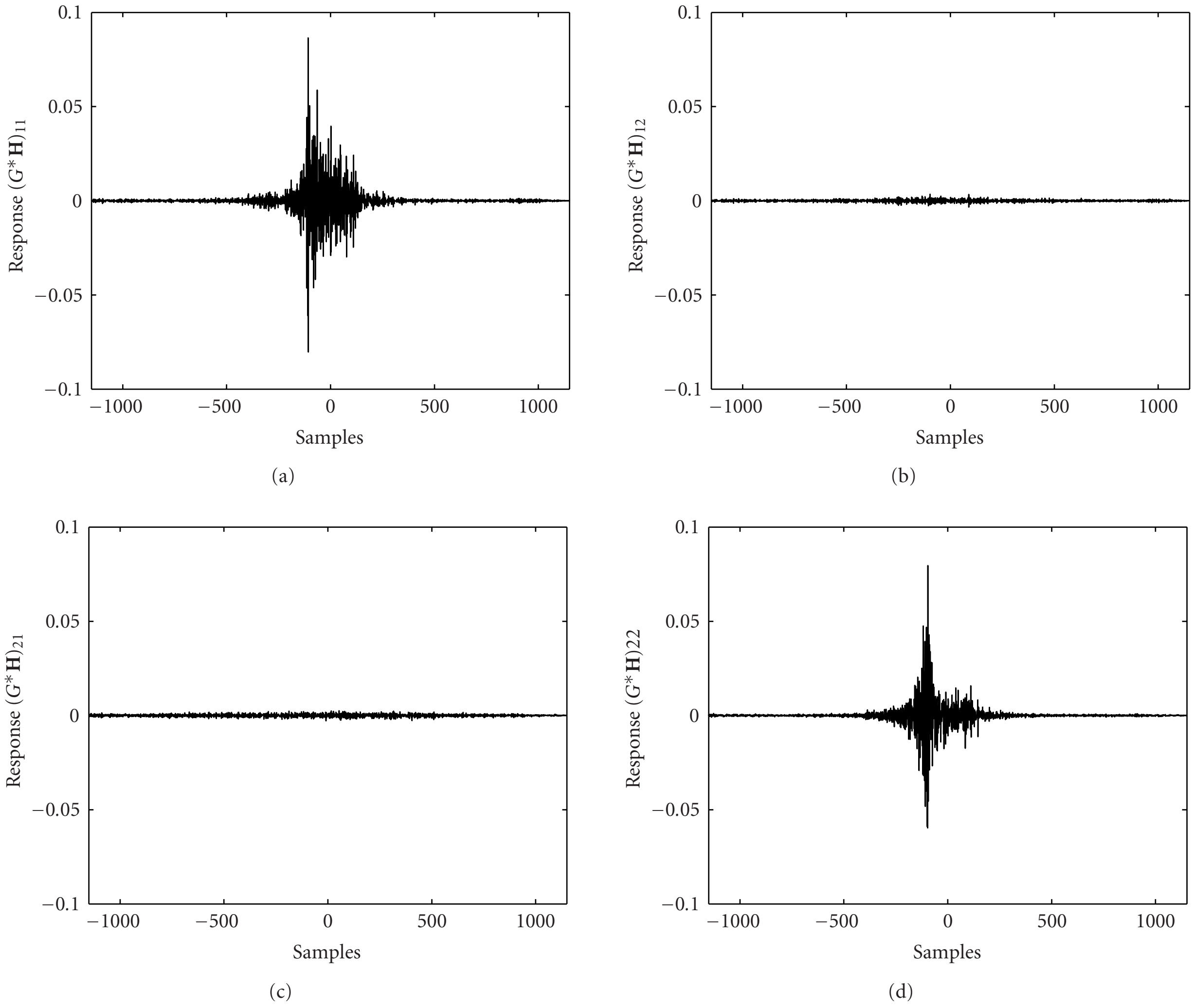


Figure 7: The four impulse responses of the global filter $(\mathbf{G} * \mathbf{H})(n)$.

here by

$$\text{SIR}\, i = 10 \log \frac{\sum_t \left(\sum_{k=1}^{K} G_{ik} * x_{ki}(t) \right)^2}{\sum_t \left(\sum_{k=1}^{K} G_{ik} * \sum_{j \neq i} x_{ki}(t) \right)^2}, \tag{20}$$

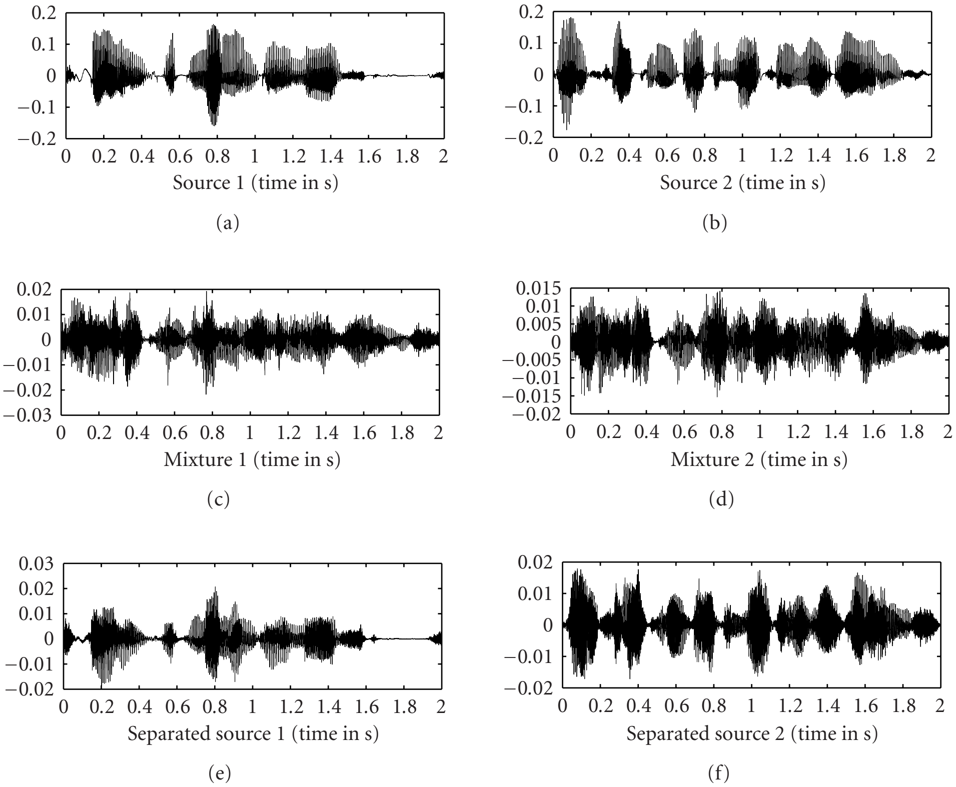
and the SIR is averaged on both channels.

The sound quality is measured with the distortion between the portion of $y_i(t)$ coming from source i, $y_{ii}(t)$, and the microphone signal $x_{ii}(t)$ recorded when only the ith source is present. $x_{ii}(t)$ can be decomposed as $ay_{ii}(t-l)+e_i(t)$, where a and l are the values that minimize the power of the error $e_i(t) = x_{ki}(t) - ay_{ii}(t-l)$. Then, the SDR is defined by

$$\text{SDR}\, i = 10 \log \frac{\sum_t (x_{ii}(t))^2}{\sum_t \left(x_{ii}(t) - ay_{ii}(t-l) \right)^2}. \tag{21}$$

Figure 9 visualizes the SIRs of the observations, and the SIRs of the unmixed signals. The algorithms of Murata et al. [9], Parra and Spence [13] and the proposed method were tested. The SIRs are shown in grey level for all different angle combinations and are given in dB between 0 dB and 20 dB. The values have been set to 0 dB on the main diagonal since they correspond to the same directions of sources and so the signals are not separable in that case. The parameters of the three algorithms have been optimized to obtain a better SIR for each one (T = 1024, Q = 128, K = 3, N = 5 for Parra's method, NFFT = 512, overlap = 492, N = 40 for Murata's method, and N = 1024, m = 5 for the proposed method). The speech signals (about 18000 samples) were sampled to 11025 kHz (1.6 s), and the SIRs were averaged on the six speakers.

For all angle combinations, the SIRs of input signals are low (dark areas), indicating that the two sources arrive very well mixed at the ears. These plots represent the initial situation. The three other figures show the results after applying one of the BSS algorithms. We improve upon the initial situations when a plot in every box is lighter in the off diagonal. The algorithm of Murata et al. fails on the dataset and we observe that the squares change towards a lighter grey for the



(a)

(b)

(c)

(d)

(e)

(f)

Figure 8: Sources, mixtures, and estimated sources.

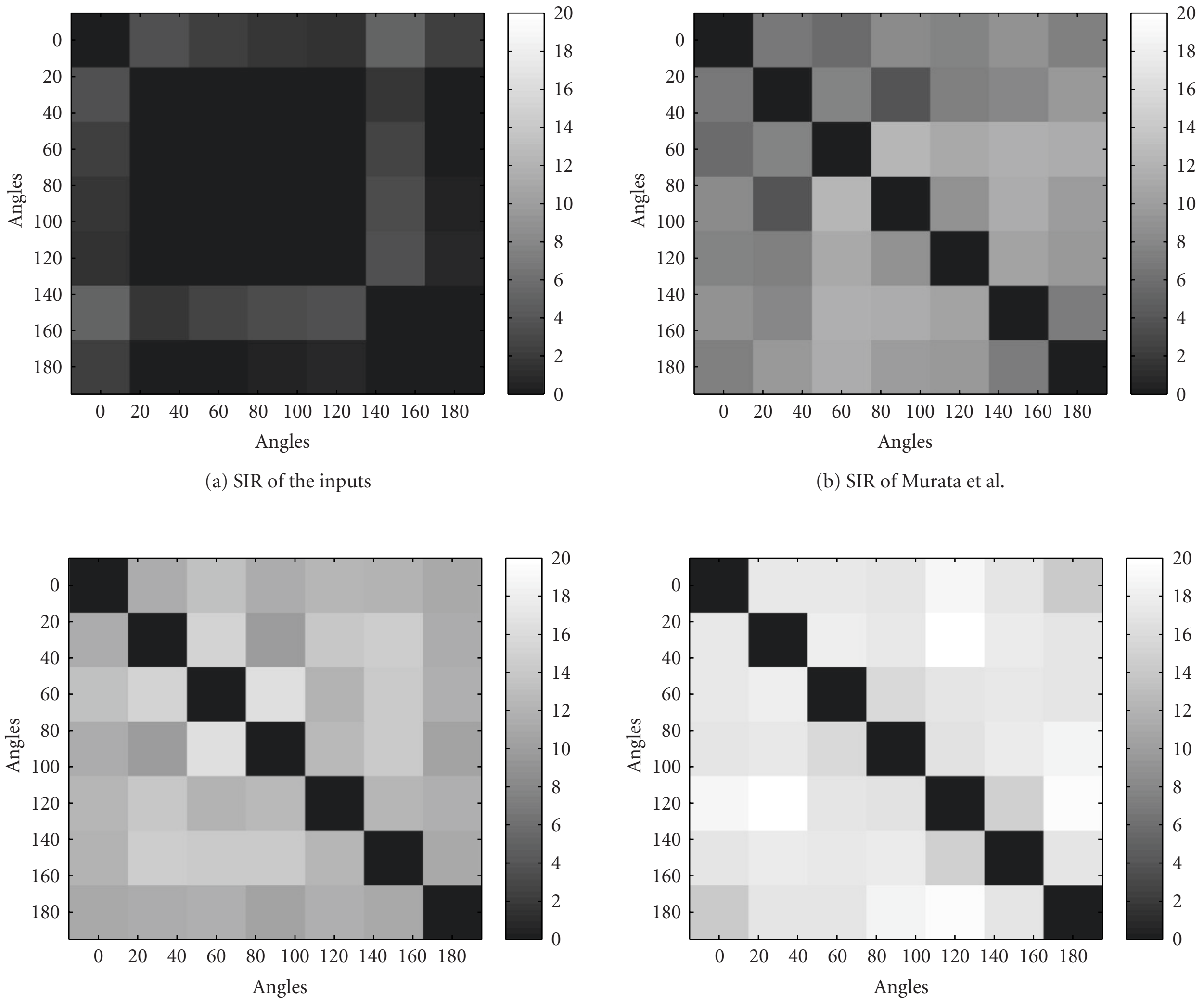
Parra and Spence algorithm. It is able to improve the separation in all cases. The proposed method leads clearly to better results and is able to largely improve the degree of separation.

To confirm the previous comments and evaluate each method, the SIRs have been averaged on all positions (without the diagonal terms) and are reported in the Table 1. The SIR value of the Murata algorithm is low while the Parra algorithm gives more satisfactory results. The proposed method performed best and there was 4.6 dB SIR enhancement on the average versus the Parra and Spence method.

Figure 10 visualizes the SDRs computed for the algorithm of Murata et al. [9], the algorithm of Parra and Spence [13], and the proposed method. As previously, the SDRs are averaged on all positions (without the diagonal terms) in Table 2. Figure 10 shows that the proposed method is able to obtain high SDR. With the algorithms of Murata and Parra, the SDR values are unsatisfactory on the dataset. If the permutations are not correctly aligned, the recovered source components may have different permutations along the frequency axis so that the reconstructed source signals are strongly distorted in the time domain.

Finally, from these experimental results we can say that the proposed algorithm has a superior performance over conventional methods [9, 13] for SIR values as well as SDRs. The algorithm [9] failed in recovering the permutation ambiguity on that dataset while the method [13] gives acceptable results. The reason for such behaviour of [9] might be that the method, which should solve the permutation problem, fails due to the correlations among the envelopes of the sources. Indeed, it seems that calculating the correlations over the whole frequency band or even on neighbouring bins does not give an accurate alignment on that data. It is confirmed by low and strictly similar results obtained for the algorithm [14] (not seen here), which is also based on the same hypothesis. The point has also been reported in [15].

Additional results can be found on the BLISS project website for two less reverberant rooms (http://www.lis.inpg.fr/pages_perso/bliss/). They have been obtained by S. Harmeling, P. Bunau, A. Ziehe (FhG FIRST), and D.T. Pham (LMC) on the McMaster database. The algorithms of Murata et al., Parra and Spence, Anemüller [14], and the proposed method have been compared. The results obtained with the algorithm of Murata et al. [9], Parra and Spence [13], and the proposed method are similar to those obtained in this paper and confirm that [9] failed on that dataset. The reason might be the correlations among the envelopes of the sources. Indeed, the algorithm of Anemüller [14] is based on the observation, that for a speech signal, amplitude variations in frequency channels are correlated but not intercorrelated across different sources. The results are really similar to those obtained with the Murata algorithm [9]. The reason for the failure might be that the used speech signals are quite short so that there might not be enough statistics to estimate the cross-frequency correlations properly. Besides, the hypothesis of correlations on the amplitude spectrogram is not verified on the whole frequency band for the tested data



(a) SIR of the inputs

(b) SIR of Murata et al.

(c) SIR of Parra et al.

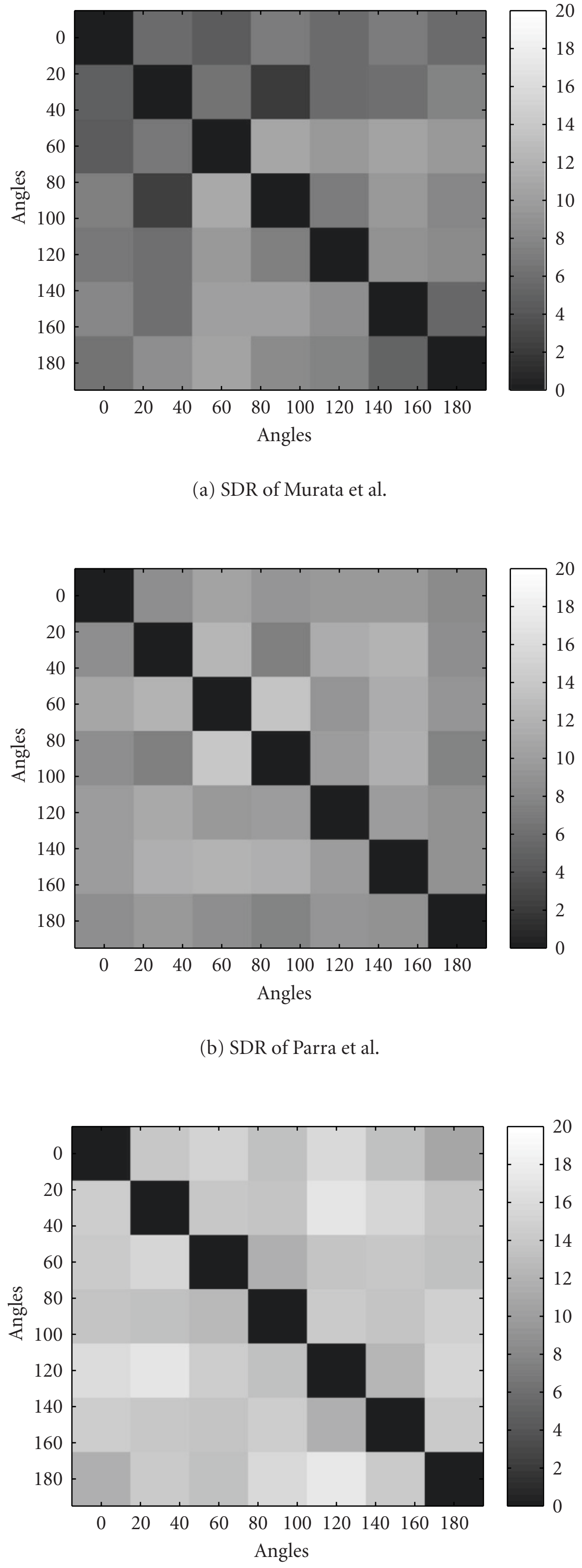
(d) SIR of proposed method

Figure 9: SIRs of the inputs and unmixed signals by BSS algorithms.

Table 1: SIRs averaged of the inputs and unmixed signals by BSS algorithms.

SIR (input signals)	SIR (Murata)	SIR (Parra)	SIR (of the proposed method)
1.3 dB	8.5 dB	12.2 dB	16.8 dB

(see, e.g., the spectrogram of one source in Figure 1). The results obtained with the Parra method [13] could be also explained by its slow convergence method for the joint diagonalization part and not just because of the permutation ambiguity. Parra and Spence's method utilizes a joint diagonalization of time-shifted cross-power spectra which is carried out by gradient-based optimization. The results are improved, if not so much short signals are used (see the other results at http://www.lis.inpg.fr/pages_perso/bliss/). These reasons prove the interest of the proposed method which is able to provide high SIRs and SDRs in real-room conditions even for quite short signals. Another interest is also its low computation complexity, due to a simple and very fast algorithm to perform joint approximate diagonalization [29]. In the case of two sources, the solution for solving the permutation ambiguity is also simple as it is an iterative algorithm where the number of iterations is exactly the number of permutation corrections to adjust. The number of permutation jumps is generally small, as in the diagonalization stage we have made use of the continuity of the mixing filter frequency response. For more than two sources, the permutation should be tested by pairs of outputs which could be difficult. It is clear that for a large number of sensors, methods relying on beamforming are more suitable.



(a) SDR of Murata et al.

(b) SDR of Parra et al.

(c) SDR of proposed method

Figure 10: SDRs of the inputs and the unmixed signals by BSS algorithms.

Table 2: Average of the SDRs of the unmixed signals by BSS algorithms.

SDR (Murata)	SDR (Parra)	SDR (of the proposed method)
7.1 dB	9.7 dB	13.5 dB

5. CONCLUSION

We have developed a method for blind separation of speech signals, which exploits the property of nonstationarity and the presence of pauses. The separation itself is achieved by joint diagonalization of the time varying spectral matrices of the observation records. To solve the permutation ambiguity, which is the main and still largely open problem in a frequency domain approach, we have introduced a new method based on the time variations of the source energy in different frequency bins. Sometimes, the correlation between the time variations of the signal energy in different frequency bins does not hold for real data or short signals even on neighbouring frequency bins. Thus, we assume only that the energy can vary smoothly with frequency and that it is continuous across the frequency axis. A measure of continuity of the speech spectrogram is computed over a limited frequency band, which is sliding across the frequency axis. This new kind of continuity is exploited to correct the block permutation problem.

The method is compared to conventional approaches with real-room recordings and the results show the improvement of the separation in terms of SIR and SDR versus other algorithms. However, there are some limitations on the impulse responses of the mixing filters. The source signals must be sufficient long and nonstationary enough. These conditions ensure a good result in the separation stage, but not sufficient to resolve the frequency permutation ambiguity. The latter needs source signals to have different time variation of energy distributions over frequency bins. For example, it would be difficult to separate synchronous speakers with the same periods of pauses and speech.

REFERENCES

[1] L. C. 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.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," in *Proceedings of the International ICSC Workshop on Independence & Artificial Neural Networks (I&ANN '98)*, pp. 9–10, Tenerife, Spain, February 1998.

[3] H.-C. Wu and J. C. Principe, "Simultaneous diagonalization in the frequency domain (SDIF) for source separation," in *Proceedings of the 1st International Conference on Independent Component Analysis and Signal Separation (ICA '99)*, pp. 245–250, Aussois, France, January 1999.

[4] R. Mukai, S. Araki, and S. Makino, "Separation and dereverberation performance of frequency domain blind source separation," in *Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA '01)*, pp. 230–235, San Diego, Calif, USA, December 2001.

[5] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[6] D.-T. Pham, Ch. Servière, and H. Boumaraf, "Blind separation of convolutive audio mixtures using nonstationarity," in *Proceedings of the 4th International Conference on Independent Component Analysis and Blind Signal Separation (ICA '03)*, pp. 981–986, Nara, Japan, April 2003.

[7] D.-T. Pham, Ch. Servière, and H. Boumaraf, "Blind separation of speech mixtures based on nonstationarity," in *Proceedings of 7th International Symposium on Signal Processing and Its Applications (ISSPA '03)*, vol. 2, pp. 73–76, Paris, France, July 2003.

[8] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA '01)*, pp. 722–727, San Diego, Calif, USA, December 2001.

[9] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[10] H. Sawada, S. Winter, R. Mukai, S. Araki, and S. Makino, "Estimating the number of sources for frequency-domain blind source separation," in *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA '04)*, pp. 610–617, Granada, Spain, September 2004.

[11] K. Torkkola, "Blind separation for audio signals—Are we there yet?" in *Proceedings of the 1st International Workshop on Independent Component Analysis and Signal Separation (ICA '99)*, pp. 239–244, Aussois, France, January 1999.

[12] A. Westner, *Object-based audio capture: separating acoustically mixed sounds*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass, USA, 1998.

[13] L. C. Parra and C. Spence, "On-line convolutive blind source separation of non-stationary signals," *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 26, no. 1-2, pp. 39–46, 2000.

[14] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proceedings of the 2nd International Workshop on Independent Component Analysis and Blind Signal Separation (ICA '00)*, pp. 215–220, Helsinki, Finland, June 2000.

[15] W. Wang, J. A. Chambers, and S. Sanei, "A novel hybrid approach to the permutation problem of frequency domain blind source separation," in *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA '04)*, pp. 532–539, Granada, Spain, September 2004.

[16] S. Ikeda and N. Murata, "A method of blind separation based on temporal structure of signals," in *Proceedings of the 5th International Conference on Neural Information Processing (ICONIP '98)*, pp. 737–742, Kitakyushu, Japan, October 1998.

[17] Ch. Servière and D.-T. Pham, "A novel method for permutation correction in frequency-domain in blind separation of speech mixtures," in *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA '04)*, pp. 807–815, Granada, Spain, September 2004.

[18] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "A combined approach of array processing and independent component analysis for blind separation of acoustic signals," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 5, pp. 2729–2732, Salt Lake City, Utah, USA, May 2001.

[19] K. Kamata, X. Hu, and H. Kobatake, "A new approach to the permutation problem in frequency domain blind source separation," in *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA '04)*, pp. 849–856, Granada, Spain, September 2004.

[20] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 1, pp. 881–884, Orlando, Fla, USA, May 2002.

[21] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained ICA for robust separation of sound mixtures," in *Proceedings of the 4th International Conference on Independent Component Analysis and Blind Signal Separation (ICA '03)*, pp. 951–956, Nara, Japan, April 2003.

[22] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 5, pp. 3140–3143, Istanbul, Turkey, June 2000.

[23] N. Mitianoudis and M. Davies, "Permutation alignment for frequency domain ICA using subspace beamforming methods," in *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA '04)*, pp. 669–676, Granada, Spain, September 2004.

[24] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency domain blind source separation for many speech signals," in *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA '04)*, pp. 461–469, Granada, Spain, September 2004.

[25] L. C. Parra and C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[26] H. Saruwatari, T. Kawamura, and K. Shikano, "Fast-convergence algorithm for ICA-based blind source separation using array signal processing," in *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA '01)*, pp. 91–94, New Platz, NY, USA, October 2001.

[27] V. C. Soon, L. Tong, Y. F. Huang, and R. Liu, "A robust method for wideband signal separation," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '93)*, vol. 1, pp. 703–706, Chicago, Ill, USA, May 1993.

[28] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[29] D.-T. Pham, "Joint approximate diagonalization of positive definite Hermitian matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 22, no. 4, pp. 1136–1152, 2001.

[30] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, USA, 1975.

[31] H. Sawada, R. Muaki, 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, 2004.

Ch. Servière was born in France in 1963. She received the Engineering degree in 1986 and the Ph.D. degree in signal processing in 1989 from the Institut National Polytechnique de Grenoble (France). Since 1991, she has been a researcher with the Centre National de Recherche Scientifique (CNRS) and with the Laboratoire des Images et des Signaux (LIS). Her research interests include statistical signal processing, noise cancellation, and blind source separation. Currently, she works on blind source separation of convolutive mixtures, fault detection, and diagnostic.

D. T. Pham was born in Hanoï, Vietnam, on 10 February 1945. He graduated from the School of Applied Mathematics and Computer Science (ENSIMAG) of the Polytechnic Institute of Grenoble in 1968. He received the Ph.D. degree in statistics in 1975 from the University of Grenoble. He was a Postdoctoral Fellow at Berkeley (Department of Statistics) in 1977–1978 and a Visisting Professor at Indiana University (Department of Mathematics) at Bloomingtion in 1979–1980. He is currently Director of Research at the French Centre National de Recherche Scientifique. His researches include time series analysis, signal modeling, blind source separation, nonlinear (particle) filtering, and biomedical signal processing.