We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

7

Convolutive ICA for Audio Signals

Masoud Geravanchizadeh and Masoumeh Hesam
Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran

1. Introduction

The goal of Blind Source Separation (BSS) is to estimate latent sources from their mixed observations without any knowledge of the mixing process. Under the assumption of statistical independence of hidden sources, the task in BSS is to obtain Independent Components (IC) from the mixed signals. Such algorithms are called ICA-based BSS algorithms [1, 2]. ICA-based BSS has been well studied in the fields of statistics and information theory for different applications, including wireless communication and biomedicine. However, as speech and audio signal mixtures in a real reverberant environment are generally convolutive mixtures, they involve a structurally much more challenging task than instantaneous mixtures, which are prevalent in many other applications [3, 4]. Such a mixing situation is generally modeled with impulse responses from sound sources to microphones. In a practical room situation, such impulse responses can have thousands of taps even with an 8 kHz sampling rate, and this makes the convolutive problem difficult to solve. Blind speech separation is applicable to the realization of noise-robust speech recognition, high quality hands-free telecommunication systems and hearing aids.

Various efforts have been devoted to the separation of convolutive mixtures. They can be classified into two major approaches: time-domain BSS [5, 6] and frequency-domain BSS [7]. With time-domain BSS, a cost function is defined for time-domain signals, and optimized with convolutive separation filters. However, the optimization with convolutive separation filters is not as simple as BSS for instantaneous mixtures, and generally computationally expensive. With frequency-domain BSS, time-domain mixed signals observed at microphones are converted into frequency-domain time-series signals by a short-time Fourier transform (STFT). However, choosing the length of STFT has relationship with the length of room impulse response [8]. The merit of these approaches is that the ICA algorithm becomes simple and can be performed separately at each frequency by any complex-valued instantaneous ICA algorithm [9-11]. However, the drawbacks of frequency-domain ICA are the permutation and scaling ambiguities of an ICA solution. In the frequency-domain ICA, different permutations at different frequencies lead to re-mixing of signals in the final output. Also, different scaling at different frequencies leads to distortion of the frequency spectrum of the output signal. For the scaling problem, in one method, the output is filtered by the inverse of the separation filter [12]. For the permutation problem, spatial information, such as the direction-of-arrivals (DOA) of sources, can be estimated and used [13, 14]. Another method utilizes the coherency of the mixing matrices in several

adjacent frequencies [15]. For non-stationary sources such as speech, many methods exploit the dependency of separated signals across frequencies to solve the permutation problem [16, 17]. We propose a method for the permutation problem, by maximizing the correlation of power ratio measure of each bin frequency with the average of previous bin frequencies [18].

This chapter deals with the frequency-domain BSS for convolutive mixtures of speech signals. We begin by formulating the BSS problem for convolutive mixtures in Section 2. Section 3 provides an overview of the frequency-domain BSS. Section 4 discusses Principal Component Analysis (PCA) as a pre-processing step. Fast ICA algorithm for complex-valued signals is discussed in Section 5. We then present several important techniques along with our proposed method for solving the permutation problem in Section 6. Section 7 introduces a common method for the scaling problem. Section 8 considers ways of choosing the STFT length for a better performance of the separation problem. In Section 9, we compare our proposed method in the permutation problem with some other conventional methods by conducting several experiments. Finally, Section 10 concludes this chapter.

2. Mixing process and convolutive BSS

Convolutive mixing arises in acoustic scenarios due to time delays resulting from sound propagation over space and the multipath generated by reflections of sound from different objects, particularly in rooms and other enclosed settings. If we denote by $s_j(t)$ the signal emitted by the *j*-th source $(1 \le j \le N)$, $x_i(t)$ the signal recorded by the *i*-th microphone $(1 \le i \le M)$, and $h_{ij}(t)$ the impulse response from source j to sensor i, we have:

$$x_i(t) = \sum_{j=1}^{N} \sum_{\tau} h_{ij}(\tau) s_j(t-\tau) . \tag{1}$$

We can write this equation into a more elegant form as:

$$\mathbf{x}(t) = \sum_{\tau} \mathbf{h}(\tau) \mathbf{s}(t-\tau), \tag{2}$$

where $\mathbf{h}(t)$ is an unknown $M \times N$ mixing matrix. Now, the goal of a convolutive BSS is to obtain separated signals $y_1(t),...,y_N(t)$, each of which corresponds to each of the source signals. The task should be performed only with M observed mixtures, and without information on the sources and the impulse responses:

$$y_j(t) = \sum_{i=1}^{M} \sum_{\tau} b_{ji}(\tau) x_i(t-\tau), \tag{3}$$

where $b_{ji}(t)$ represents the impulse response of the multichannel separation system. Convolutive BSS as applied to speech signal mixtures involves relatively-long multichannel FIR filters to achieve separation with even moderate amounts of room reverberation. While time-domain algorithms can be developed to perform this task, they can be difficult to code primarily due to the multichannel convolution operations involved [5, 6]. One way to simplify the conceptualization of the convolutive BSS algorithms is to transform the task

into the frequency domain, as convolution in time becomes multiplication in frequency. Ideally, each frequency component of the mixture signal contains an instantaneous mixture of the corresponding frequency components of the underlying source signals. One of the advantages of the frequency-domain BSS is that we can employ any ICA algorithm for instantaneous mixtures, such as the information maximization (Infomax) approach [19] combined with the natural gradient [20], Fast ICA [21], JADE [22], or an algorithm based on non-stationarity of signals [23].

3. Frequency-domain convolutive BSS

This section presents an overview of the frequency-domain BSS approach that we consider in this chapter. First, each of the time-domain microphone observations $x_j(t)$ is converted into frequency-domain time-series signals $X_j(k,f)$ by a short-time Fourier transform (STFT) with a *K*-sample frame and its *S*-sample shift:

$$X_j(k,f) = \sum_t x_j(t)\,\mathbf{win}\left(t - k\frac{S}{f_s}\right) e^{-i2\pi ft}, \tag{4}$$

for all discrete frequencies $f \in \left\{0, \frac{1}{K}f_S, \ldots, \frac{K-1}{K}f_S\right\}$, and for frame index *k*. The analysis window $\mathbf{win}(t)$ is defined as being nonzero only in the *K*-sample interval $\left[-\frac{K}{2}\frac{1}{f_s}, (\frac{K}{2}-1)\frac{1}{f_s}\right]$ and tapers smoothly to zero at each end of the interval, such as a Hanning window $\mathbf{win}(t) = \frac{1}{2}\left(1 + \cos\frac{2\pi f_s}{K}t\right)$.

If the frame size *K* is long enough to cover the main part of the impulse responses h_{ij}, the convolutive model (2) can be approximated as an instantaneous model at each frequency [8, 24]:

$$\mathbf{X}(k,f) = \mathbf{H}(f)\mathbf{S}(k,f), \tag{5}$$

where $\mathbf{H}(f)$ is an $M \times N$ mixing matrix in frequency domain, and $\mathbf{X}(k,f)$ and $\mathbf{S}(k,f)$ are vectors of observations and sources in frequency domain, respectively. Notice that, the convolutive mixture problem is reduced to a complex but instantaneous mixture problem and separation is performed at each frequency bin by:

$$\mathbf{Y}(k,f) = \mathbf{B}(f)\mathbf{X}(k,f), \tag{6}$$

where $\mathbf{B}(f)$ is an $N \times M$ separation matrix. As a basic setup, we assume that the number of sources *N* is no more than the number of the microphones *M*, i.e., $N \le M$. However, in a case with $N > M$ that is referred to as underdetermined BSS, separating all the sources is a rather difficult problem [25].

We can limit the set of frequencies to perform the separations by $\left\{0, \frac{1}{K}f_s, \ldots, \frac{1}{2}f_s\right\}$ due to the relationship of complex conjugate:

$$X_j(k,\frac{m}{K}f_s)=X_j^*\left(k,\frac{K-m}{K}f_s\right),\quad (m=1,...,K/2-1). \tag{7}$$

We employ the complex-valued instantaneous ICA to calculate the separation matrix $\mathbf{B}(f)$. Section 5 describes the detailed procedure for the complex-valued ICA used in our implementation and experiments. However, the ICA solution at each frequency bin has permutation and scaling ambiguity. In order to construct proper separated signals in time domain, frequency-domain separated signals originating from the same source should be grouped together. This is the permutation problem. Also, different scaling at different frequencies leads to distortion of the frequency spectrum of the output signal. This is the scaling problem. There are some methods to solve the permutation and scaling problems [12-18]. After solving the permutation and the scaling problem, the time-domain output signals $y_i(t)$ are calculated with an inverse STFT (ISTFT) of the separated signals $Y_i(k,f)$. The flow of the frequency-domain BSS is shown in Figure 1.

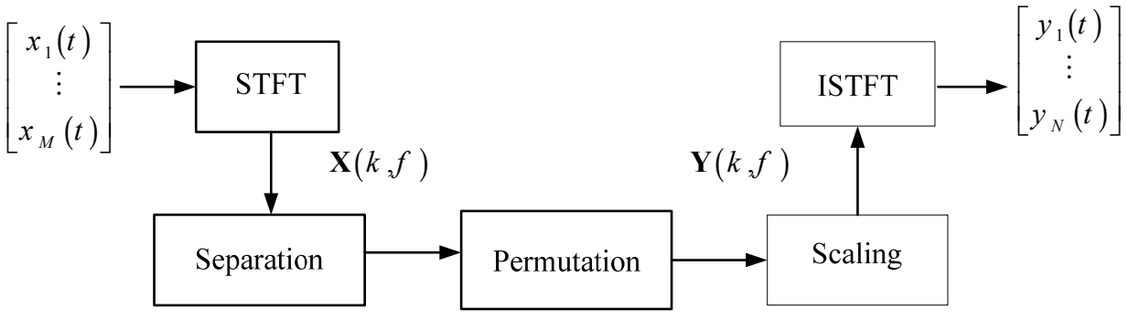


Fig. 1. System structure for the frequency-domain BSS

4. Pre-processing with principal component analysis

It is known that choosing the number of microphones more than the number of sources improves the separation performance. This is termed as the overdetermined case, in which the dimension of the observed signals is greater than the number of sources. Many methods have been proposed to solve the overdetermined problem. In a typical method, the subspace procedure is used as a pre-processing step for ICA in the framework of BSS [15, 26, 27]. The subspace method can be understood as a special case of principal component analysis (PCA) with $M \geq N$, where M and N denote the number of observed signals and source signals, respectively. This technique reduces room reflections and ambient noise [15]. Also, as pre-processing, PCA improves the convergence speed of ICA. Figure 2 shows the use of PCA as pre-processing to reduce the dimension of microphone signals.

In the PCA process, the input microphone signals are assumed to be modeled as:

$$\mathbf{X}(k,f)=\mathbf{A}(f)\mathbf{S}(k,f)+\mathbf{n}(k,f), \tag{8}$$

where the (*m*,*n*)-th element of $\mathbf{A}(f)$ is the transfer function from the *n*-th source to the *m*-th microphone as:

$$A_{m,n}(f) = T_{m,n}(f)e^{-i2\pi f\tau_{m,n}}. \tag{9}$$

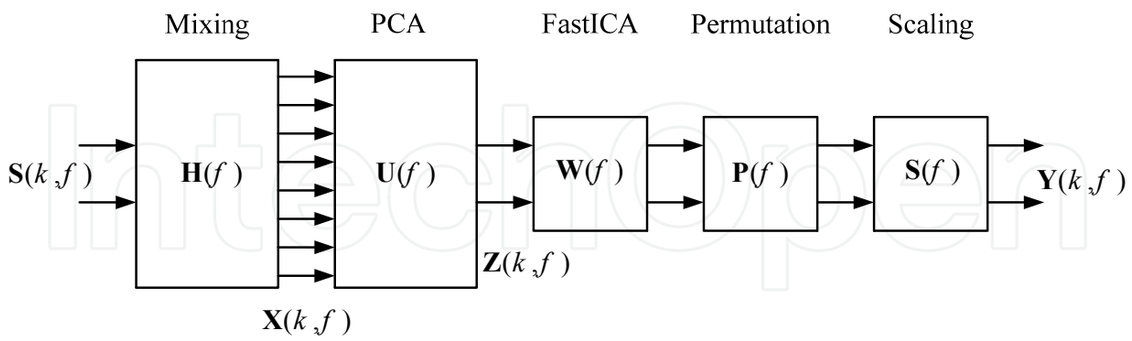


Fig. 2. The use of PCA as a pre-processing step in the frequency-domain BSS

Here, the symbol $T_{m,n}(f)$ is the magnitude of the transfer function. The symbol $\tau_{m,n}$ denotes the propagation time from the n-th source to the m-th microphone. The first term in Eq. (8), $\mathbf{A}(f)\mathbf{S}(k,f)$, expresses the directional components in $\mathbf{X}(k,f)$ and the second term, $\mathbf{n}(k,f)$, is a mixture of less-directional components which includes room reflections and ambient noise.

The spatial correlation matrix $\mathbf{R}(f)$ of $\mathbf{X}(k,f)$ is defined as:

$$\mathbf{R}(f) = E[\mathbf{X}(k,f)\mathbf{X}^H(k,f)]. \tag{10}$$

The eigenvalues of $\mathbf{R}(f)$ are denoted as $\lambda_1(f),...,\lambda_M(f)$ with $\lambda_1(f) \geq,...,\geq \lambda_M(f)$ and the corresponding eigenvectors are denoted as $\mathbf{e}_1(f),...,\mathbf{e}_M(f)$. Assuming that $\mathbf{s}(t)$ and $\mathbf{n}(t)$ are uncorrelated, the energy of the N directional signals $\mathbf{s}(t)$ is concentrated on the N dominant eigenvalues and the energy of $\mathbf{n}(t)$ is equally spread over all eigenvalues. In this case, it is generally satisfied that:

$$\lambda_1(f),...,\lambda_N(f) >> \lambda_{N+1}(f),...,\lambda_M(f). \tag{11}$$

The vectors $\mathbf{e}_1(f),...,\mathbf{e}_N(f)$ and $\mathbf{e}_{N+1}(f),...,\mathbf{e}_M(f)$ are the basis of the signal and noise subspaces, respectively.

In the PCA method, the input signal is processed as:

$$\mathbf{Z}(k,f) = \mathbf{U}(f)\mathbf{X}(k,f), \tag{12}$$

that reduces the energy of $\mathbf{n}(t)$ in the noise subspace, and the PCA filter is defined as:

$$\mathbf{U}(f) = \mathbf{\Lambda}^{-\frac{1}{2}}(f)\,\mathbf{E}^H(f), \tag{13}$$

where

$$\mathbf{\Lambda}(f) = diag\left(\lambda_1(f),...,\lambda_N(f)\right),\ \mathbf{E}(f) = \left[e_1(f),...,e_N(f)\right]. \tag{14}$$

The PCA filtering of $\mathbf{X}(k,f)$ reduces the dimension of input signal to the number of sources N which is equivalent to a spatially whitening operation, i.e., $E\{\mathbf{Z}(k,f)\mathbf{Z}^H(k,f)\} = \mathbf{I}$ where $\mathbf{I}$ is the $N \times N$ identity matrix.

5. Complex-valued fast fixed-point ICA

The ICA algorithm used in this chapter is fast fixed-point ICA (fast ICA). The fast ICA algorithm for the separation of linearly mixed independent source signals was presented in [21]. This algorithm is a computationally efficient and robust fixed-point type algorithm for independent component analysis and blind source separation. However, the algorithm in [21] is not applicable to frequency-domain ICA as these are complex-valued. In [9], the fixed-point ICA algorithm of [21] has been extended to involve complex-valued signals. The fast fixed-point ICA algorithm is based on the assumption that when the non-Gaussian signals get mixed, it becomes more Gaussian and thus its non-Gaussianization can yield independent components. The process of non-Gaussianization consists of two-steps, namely, pre-whitening or sphering and rotation of the observation vector. Sphering is half of the ICA task and gives spatially decorrelated signals. The process of sphering (pre-whitening) is accomplished by the PCA stage as described in the previous section. The task remaining after whitening involves rotating the whitened signal vector $\mathbf{Z}(k,f)$ such that $\mathbf{Y}(k,f) = \mathbf{W}(f)\,\mathbf{Z}(k,f)$ returns independent components. For measuring the non-Gaussianity, we can use the negentropy-based cost function:

$$J = E\left\{G\left(|\,\mathbf{w}^H\mathbf{Z}\,|^2\right)\right\}, \tag{15}$$

where $G(t) = \log(0.01 + t)$ [9].

The elements of the matrix $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_N)$ are obtained in an iterative procedure. The fixed-point iterative algorithm for each column vector $\mathbf{w}$ is as follows (the frequency index f and frame index k are dropped hereafter for clarity):

$$\mathbf{w} = E\left\{\mathbf{y}(\mathbf{w}^H\mathbf{Z})^* g\left(\left|\mathbf{w}^H\mathbf{Z}\right|^2\right)\right\} - E\left\{g\left(\left|\mathbf{w}^H\mathbf{Z}\right|^2\right) + \left|\mathbf{w}^H\mathbf{Z}\right|^2 g'\left(\left|\mathbf{w}^H\mathbf{Z}\right|^2\right)\right\}\mathbf{w}, \tag{16}$$

where $g(.)$ and $g'(.)$ are first- and second-order derivatives of G:

$$g(t) = \frac{1}{(0.01 + t^2)}, \quad g'(t) = \frac{o.5}{(0.01 + t^2)^2}. \tag{17}$$

After each iteration, it is also essential to decorrelate $\mathbf{W}$ to prevent its convergence to the previously converged point. The decorrelation process to obtain $\mathbf{W}$ for the next iteration is obtained as [9]:

$$\mathbf{W} = \mathbf{W}(\mathbf{W}^H\mathbf{W})^{-1/2}. \tag{18}$$

Then, the separation matrix is obtained by the product of $\mathbf{U}(f)$ and $\mathbf{W}(f)$:

$$\mathbf{B}(f) = \mathbf{W}(f)\mathbf{U}(f). \tag{19}$$

6. Solving the permutation problem

In order to get separated signals correctly, the order of separation vectors (position of rows) in $\mathbf{B}(f)$ must be the same at each frequency bin. This is called permutation problem. In this section, we review various methods which have already been proposed to solve permutation problem.

6.1 Solving permutation by Direction of Arrival (DOA) estimation

Some methods for permutation problem use the information of source locations, such as direction of arrival (DOA). In the totally blind setup, DOA cannot be known so it is estimated from the directivity pattern of the separation matrix. In this method, the effect of room reverberation is neglected, and the elements of the mixing matrix in Eq. (9) can be written as the following expression:

$$A_{m,n}(f) = T_{m,n}(f)e^{-i2\pi f\tau_{m,n}}, \qquad (\tau_{m,n} \equiv \frac{1}{c}d_m \sin\theta_n), \tag{20}$$

where $\tau_{m,n}$ is the arriving lag with respect to the n-th source signal from the direction of θ_n, observed at the m-th microphone located at d_m, and c is the velocity of sound. Microphone array and sound sources are shown in Figure 3.

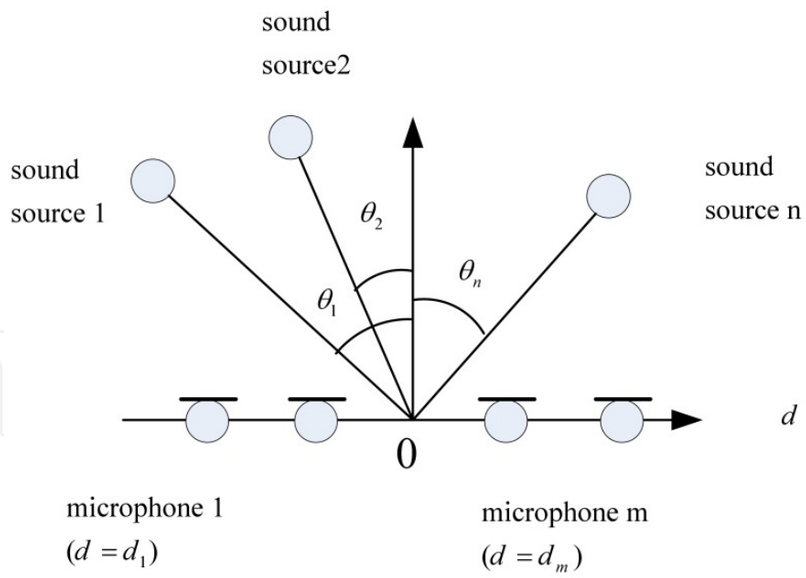


Fig. 3. Configuration of a microphone array and sound sources

From the standpoint of array signal processing, directivity patterns (DP) are produced in the array system. Accordingly, directivity patterns with respect to $B_{nm}(f)$ are obtained at every frequency bin to extract the DOA of the n-th source signal. The directivity pattern $F_n(f,\theta)$ is given by [13]:

$$F_n(f,\theta) = \sum_{m=1}^{M} B_{nm}(f).\exp\left[i2\pi f d_m \sin\theta / c\right]. \tag{21}$$

The DP of the separation matrix contains nulls in each source direction. Figure 4 shows an example of directivity patterns at frequency bins f_1 and f_2 plotted for two sources. As it is observed, the positions of the nulls vary at each frequency bin for the same source direction. Hence, in order to solve the permutation problem and sort out the different sources, the separation matrix at each frequency bin is arranged in accordance with the directions of nulls.

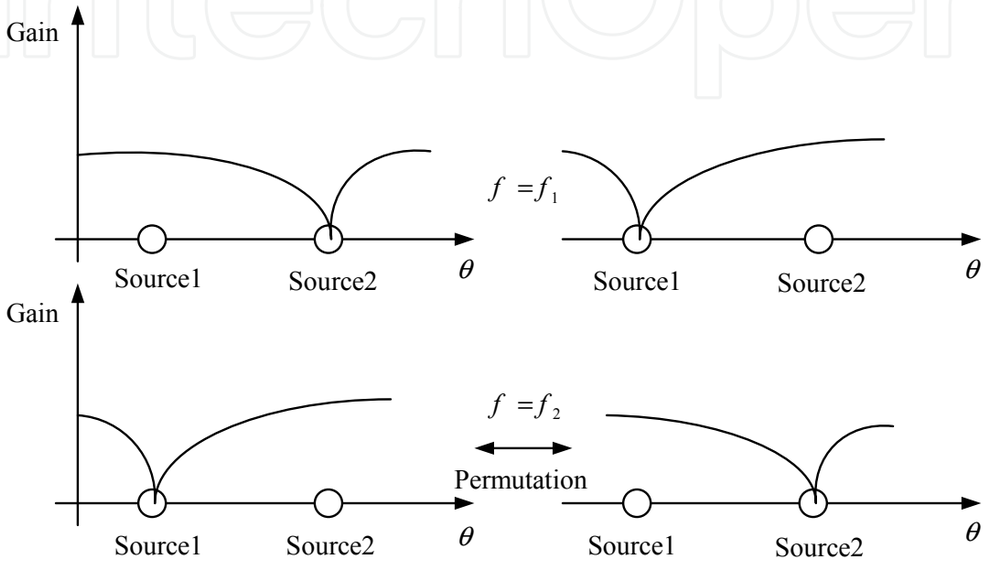


Fig. 4. Examples of directivity patterns

This method is not always effective in the overdetermined case, because the directions giving the nulls of the directivity patterns of the separation matrix $\mathbf{B}(f)$ do not always correspond to the source directions. Figure 5 shows the directivity pattern for the case ($M = 2,\ N = 2$) , and the overdetermined case ($M = 8,\ N = 2$).

6.1.1 Closed-form formula for estimating DOAs

The DOA estimation method by the directivity pattern has three problems, a high computational cost, the difficulty of using it for mixtures of more than two sources, and for overdetermined case in which the number of microphones is more than the number of sources. Instead of plotting directivity patterns and searching for the minimum as a null direction, some propose a closed-form formula for estimating DOAs [16]. In principle, this method can be applied to any number of source signals as well as to the overdetermined case. It can be shown that the DOAs for sources are estimated by the following relation [16]:

$$\theta_k = \arccos \frac{\arg\left(\dfrac{\left[\mathbf{B}^{-1}\right]_{jk}}{\left[\mathbf{B}^{-1}\right]_{j'k}} \right)}{2\pi f c^{-1}\left(d_j - d_{j'}\right)}, \tag{22}$$

where, d_j and $d_{j'}$ are the positions of sensors x_j and $x_{j'}$.

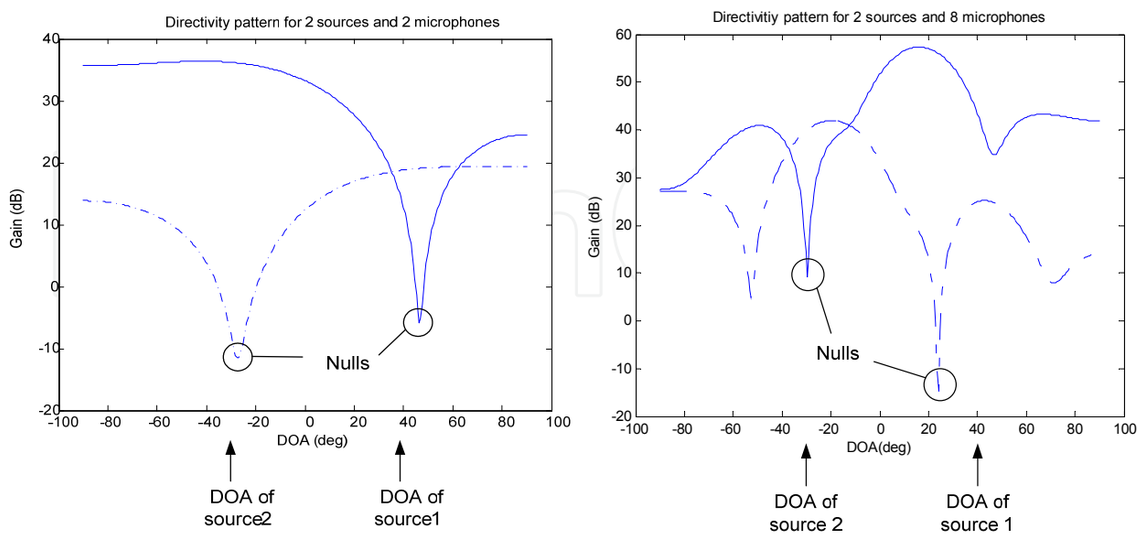


Fig. 5. The directivity patterns for the case ($M = 2, N = 2$), and the overdetermined case ($M = 8, N = 2$)

If the absolute value of the input variable of arccos(.) is larger than 1, θ_k becomes complex and no direction is obtained. In this case, formula (22) can be tested with another pair j and j'.

If $N < M$, the Moore–Penrose pseudoinverse $\mathbf{B}^{+}$ is used instead of $\mathbf{B}^{-1}$. Based on these DOA estimations, the permutation matrix is determined. In this process, no reverberation is assumed for the mixing signals. Therefore, for the reverberant case the method based on DOA estimation is not efficient.

6.2 Permutation by interfrequency coherency of mixing matrix

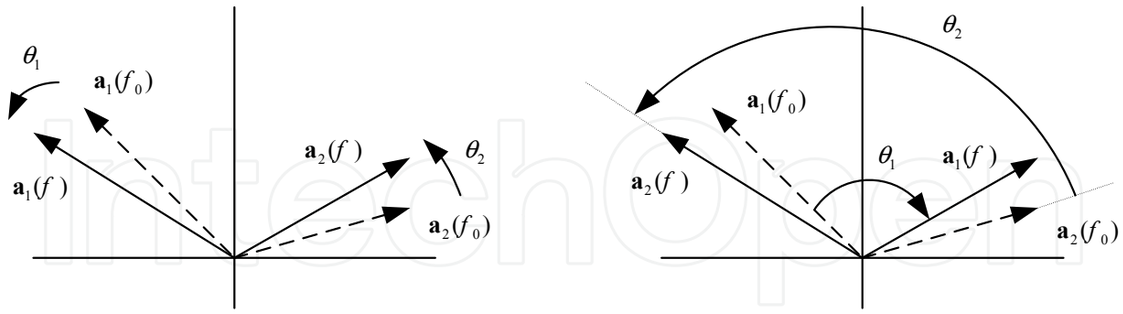
Another method to solve the permutation problem utilizes the coherency of the mixing matrices in several adjacent frequencies [15]. For the mixing matrix $\mathbf{A}(f)$ in the Eq. (8), the n-th column vector (location vector of the n-th source) at frequency f has coherency with that at the adjacent frequency $f_0 = f - \Delta f$. Therefore, the location vector $\mathbf{a}_n(f)$ is $\mathbf{a}_n(f_0)$ which is rotated by the angle θ_n as depicted in Figure 6(a). Accordingly, θ_n is expected to be the smallest for the correct permutation as shown in Figure 6. Based on this assumption, permutation is solved so that the sum of the angles $\{\theta_1, \theta_2, ..., \theta_N\}$ between the location vectors in the adjacent frequencies is minimized. An estimate of the mixing matrix $\hat{\mathbf{A}}(f) = [\hat{\mathbf{a}}_1(f), ..., \hat{\mathbf{a}}_N(f)]$ can be obtained as the pseudoinverse of the separation matrix as:

$$\hat{\mathbf{A}}(f) = \mathbf{B}^{+}(f). \tag{23}$$

For this purpose, we define a cost function as [15]:

$$F(\mathbf{P}) = \frac{1}{N}\sum_{n=1}^{N} \cos\theta_n, \qquad \cos\theta_n = \frac{\hat{\mathbf{a}}_n^H(f)\hat{\mathbf{a}}_n(f_0)}{\|\hat{\mathbf{a}}_n(f)\|.\|\hat{\mathbf{a}}_n(f_0)\|}. \tag{24}$$

This cost function is calculated for all arranges of columns of mixing matrix $\hat{\mathbf{A}}(f) = [\hat{\mathbf{a}}_1(f),...,\hat{\mathbf{a}}_N(f)]$ and the permutation matrix $\mathbf{P}$ is obtained by maximizing it.



(a) Correct Permutation (b) Incorrect Permutation

Fig. 6. The column vectors of the mixing matrix in two adjacent frequencies, with correct and incorrect permutations

To increase the accuracy of this method, the cost function is calculated for a range of frequencies instead of the two adjacent frequencies and a confidence measure is used to determine which permutation is correct [15].

The mixing matrix is defined as the transfer function of direct path from each source to each microphone where the coherency of mixing matrices is used in several adjacent frequencies to obtain the permutation matrix.

This method assumes that the spectrum of microphone signals consists of the directional components and reflection components of sources and employs the subspace method to reduce the reflection components. However, if the reflection components are not reduced by the subspace method, the mixing matrix consists of indirect path components, and the method will not be efficient.

6.3 A new method to solve the permutation problem based on power ratio measure

Another group of permutation methods use the information on the separated signals which are based on the interfrequency correlation of separated signals. Conventionally, the correlation coefficient of separated signal envelopes has been employed to measure the dependency of bin-wise separated signals. Envelopes have high correlations at neighboring frequencies if separated signals correspond to the same source signal. Thus, calculating such correlations helps us to align permutations. A simple approach to the permutation alignment is to maximize the sum of the correlations between neighboring frequencies [16]. The method in [12] assumes high correlations of envelopes even between frequencies that are not close neighbors and so it does not limit the frequency range in which correlations are calculated.

However, this assumption is not satisfied for all pairs of frequencies. Therefore, the use of envelopes for maximizing correlations in this way is not a good choice. Recently, the power ratio between the *i*-th separated signal and the total power sum of all separated signals has been proposed as another type of measure [17]. In this approach, the dependence of bin-wise separated signals can be measured more clearly by calculating correlation coefficients with power ratio values rather than with envelopes. This is shown by comparing Figures 7 and 8.

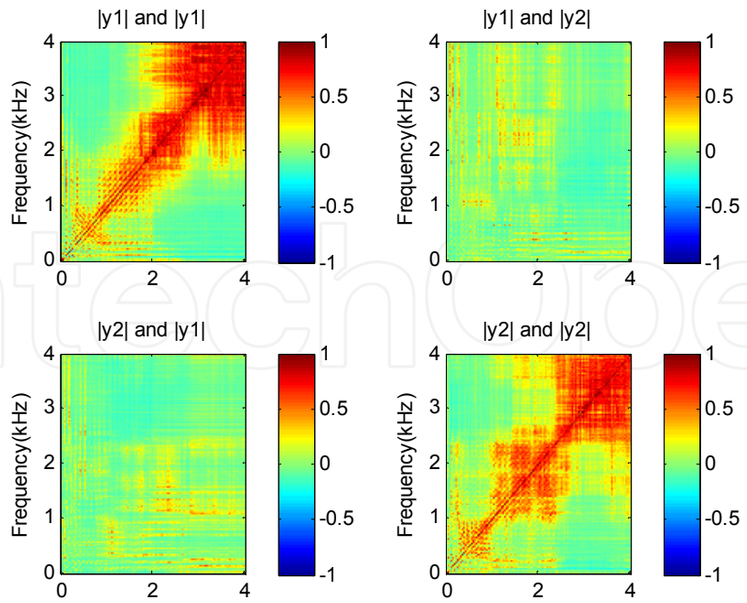


Fig. 7. Correlation coefficients between the separated signal envelopes

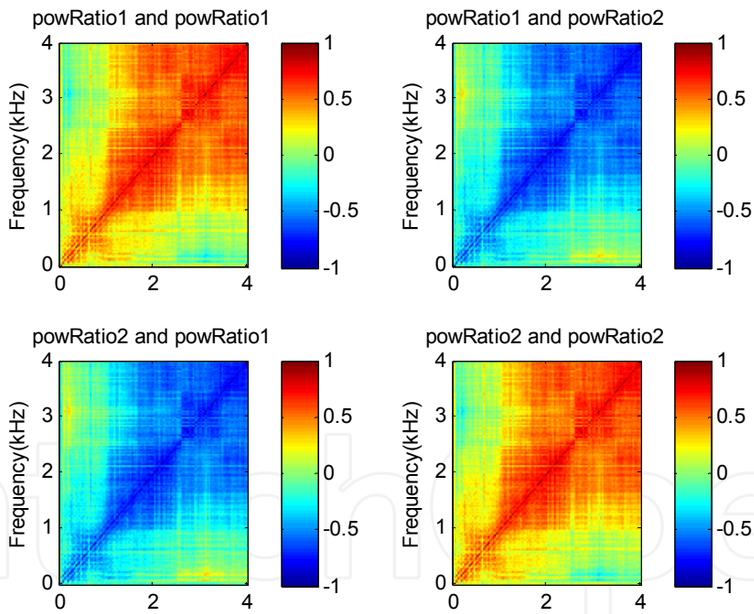


Fig. 8. Correlation coefficients between the power ratios of separated signals

This method uses two optimization techniques for permutation alignment; a rough global optimization and a fine local optimization. In rough global optimization, a centroid is calculated for each source as the average value of power ratio with the current permutation. The permutations are optimized by an iterative maximization between the power ratio measures and the current centroid. In fine local optimization, the permutations are obtained by maximizing the correlation coefficients over a set of frequencies consisting of adjacent frequencies and harmonic frequencies. Here, the experiments show that the fine local optimization alone does not provide good results in permutation alignment. But using both

global and local optimization achieves almost optimal results. This method, however, is somewhat complicated for calculating the permutations.

In our proposed method, we take a rather simple technique to compute the permutation matrices. Here, we assume that the correlation coefficients of power ratios of bin-wise separated signal to be high if they come from the same source for each two frequencies even if they are not close together. Therefore, we extend the frequency range for calculating correlation to all previous frequencies, where the permutation was solved for them. We decide on the permutation by maximizing the correlation of power ratio measure of each bin frequency with the average of power ratio measures of previous bin frequencies, iteratively with increasing frequency. Therefore, this criterion is not based on local information and does not have the drawback of propagation of mistakes by the computation of permutation at each frequency.

If the separation works well, the bin-wise separated signals $\mathbf{Y}_1(k,f),...,\mathbf{Y}_N(k,f)$ are the estimations of the original source signals $\mathbf{S}_1(k,f),...,\mathbf{S}_N(k,f)$ up to the permutation and scaling ambiguity. Thus, the observation vector $\mathbf{X}(k,f)$ can be represented by the linear combination of the separated signals as:

$$\mathbf{X}(k,f) = \mathbf{A}(f)\,\mathbf{Y}(k,f) = \sum_{i=1}^{N} \mathbf{a}_i(f)\,Y_i(k,f), \tag{25}$$

where the mixing matrix $\mathbf{A}(f) = [\mathbf{a}_1(f),...,\mathbf{a}_N(f)]$ is the pseudoinverse of the separation matrix $\mathbf{B}(f)$:

$$\mathbf{A}(f) = \mathbf{B}(f)^{+}. \tag{26}$$

Now, we use the power ratio measure as given by [17]:

$$powRatio_i(k,f) = \frac{\left\| \mathbf{a}_i(f)\,Y_i(k,f) \right\|^2}{\sum_{n=1}^{N} \left\| \mathbf{a}_n(f)\,Y_n(k,f) \right\|^2}. \tag{27}$$

In the following, $v_i^{f_l}(k) = powRatio_i(k,f_l)$ denotes the power ratio measure obtained at frequency $f_l = (l/K)f_s$ $(l = 0,...,K/2)$, where f_s is the sampling rate.

The details of the proposed method are as follows:

1. Obtain $v_i^{f_0}(k)$, $(i = 1,...,N)$, set $l = 1$.
2. Obtain $v_i^{f_l}(k)$ and $c_i^{f_l}(k) = \sum_{g \in T} v_i^{g}(k)$ $(i = 1,...,N)$, where $T = \{f_0,...,f_{l-1}\}$.
3. Obtain all permutation matrices $\mathbf{P}_e$ $(e = 1,2,...,N)$. Permutation matrix is an $N \times N$ matrix where in each row and each column there is one nonzero element of unit value. For example, for a case of 2 sources, the permutation matrices are:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ \mathbf{P}_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{28}$$

4. Obtain $\mathbf{u}^{f_l} = \mathbf{P}_e \mathbf{v}^{f_l}$ for all permutation matrices.

5. Determine the permutation matrix that maximizes the correlation of power ratio measure of current frequency bin with the average of power ratio measures of previous bin frequencies:

$$\mathbf{P} = \arg\max_{\mathbf{P}_e} \sum_{i=1}^{N} \rho\left(u_i^{f_l}, c_i^{f_l}\right). \tag{29}$$

6. Then, process the separated signal $\mathbf{Y}(k, f_l)$ with the permutation matrix at the bin frequency f_l:

$$\mathbf{Y}(k, f_l) \leftarrow \mathbf{P}(f_l)\, \mathbf{Y}(k, f_l). \tag{30}$$

7. Set $l = l + 1$, and return to step 2), if $l < K / 2$.

The steps of the proposed method are shown in the block diagram of Figure 9.

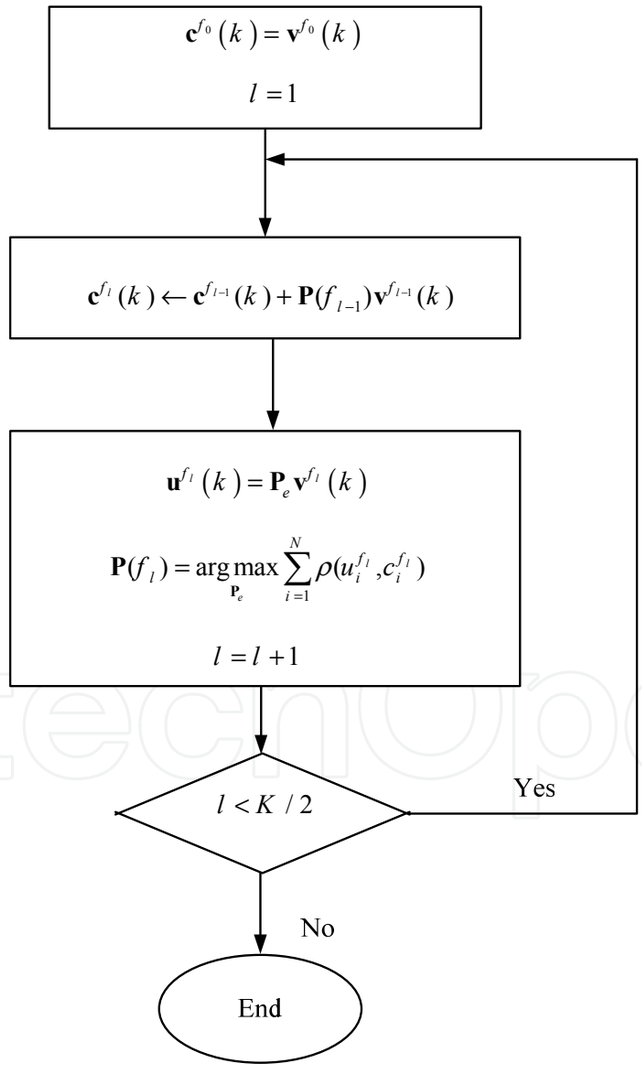


Fig. 9. The block diagram that describes our proposed method for solving the permutation problem

7. Scaling problem

The scaling problem can be solved by filtering individual outputs of the separation filter by the inverse of $\mathbf{B}(f)$ separately [12]. In the overdetermined case, (i.e., $M > N$), the pseudoinverse of $\mathbf{B}(f)$, denoted as $\mathbf{B}(f)^+$, is used instead of the inverse of $\mathbf{B}(f)$. This is due to the fact that in this case, because of employing the subspace method $\mathbf{B}(f)$ is not square. The scaling matrix can be expressed as:

$$\mathbf{S}(f) = diag[B^+_{m,1},...,B^+_{m,N}], \tag{31}$$

where $B^+_{m,n}$ denotes the (m,n)-th element of $\mathbf{B}(f)^+$.

8. Suitable length of STFT for better separation

It is commonly believed that the length of STFT (i.e., frame size), K, must be longer than P to estimate the unmixing matrix for a P-point room impulse response. The reasons for this belief are: 1) A linear convolution can be approximated by a circular convolution if $K > 2P$, and 2) If we want to estimate the inverse system of a system with impulse response of P-taps long, we need an inverse system that is Q-taps long, where $Q > P$. If we assume that the frame size is equal to the length of unmixing filter, then we should have $K > P$. Moreover, when the filter length becomes longer, the number of separation matrices to be estimated increases while the number of samples for learning at each frequency bin decreases. This violates the assumption of independence in the time series at each bin frequency, and the performance of the ICA algorithm becomes poor [8]. Therefore, there is an optimum frame size determined by a trade-off between maintaining the assumption of independence and the length of STFT that should be longer than the room impulse response length in the frequency-domain BSS. Section 9 shows this understanding by some experiments.

9. Experimental results

The experiments are conducted to examine the effectiveness of the proposed permutation method [18]. We use two experimental setups. Setup A is considered to be a basic one, in which there are two sources and two microphones. In setup B, we have two sources and eight microphones, and discuss the effect of a background interference noise on our proposed method. Table 1 summarizes the configurations common to both setups. As the original speech, we use the wave files from 'TIMIT speech database' [28] to test the performance of different BSS algorithms. The lengths of the speech signals are 4 seconds. We have the voice of three male and three female speakers in our experiments and the investigations are carried out for nine different combinations of speakers. The image method

room dimension	L = 3.12 m, W = 5.73 m, H = 2.70 m
direction of arrivals	30° and -40°
window function	Hamming
sample rate	16000 Hz

Table 1. Common Experimental Configuration

has been used to generate multi-channel Room Impulse Responses [29]. Microphone signals are generated by adding the convolutions of source signals with their corresponding room impulse responses. Figure 10 shows the layout of the experimental room for setup *B*. For the setup *A*, we use only two microphones m1 and m2 shown in the figure.

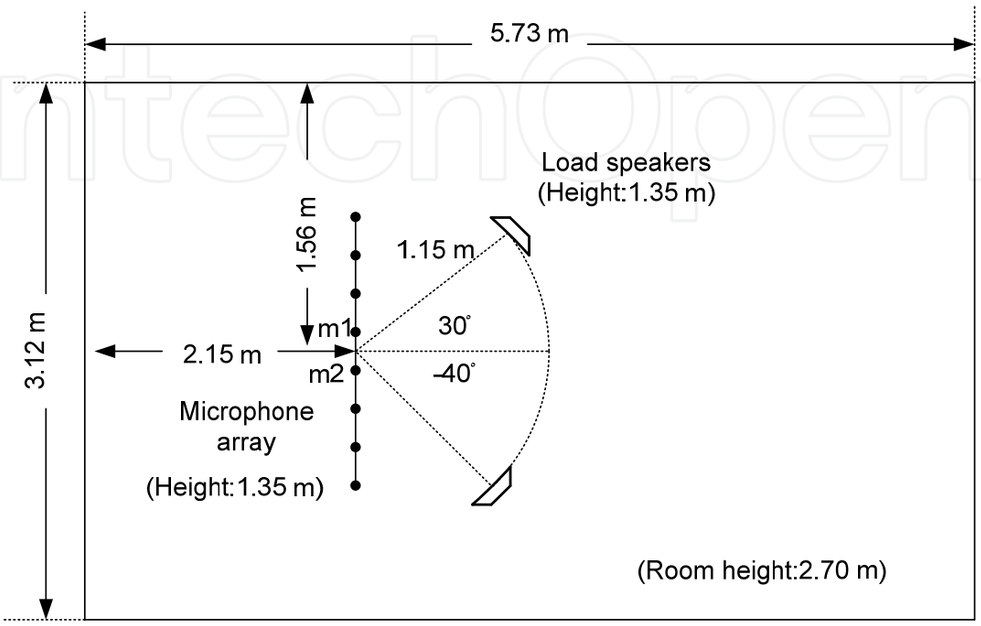


Fig. 10. Experimental Setup *B*

9.1 Evaluation criterion

For the computation of the evaluation criterion, we start by the decomposition of $y_i(t)$ (i.e., the estimation of $s_i(t)$):

$$y_i = s_{\mathbf{target}} + e_{\mathbf{interf}} + e_{\mathbf{noise}}, \tag{32}$$

where $s_{\mathbf{target}}$ is a version of $s_i(t)$ modified by mixing and separating system, and $e_{\mathbf{interf}}$ and $e_{\mathbf{noise}}$ are respectively the interference and noise terms. Figure 11 shows the source, the microphone, and the separated signals.

We use Signal-to-Interference Rate (SIR) as performance criterion by computing energy ratios between the target signal and the interference signal expressed in decibels [30]:

$$\mathrm{SIR}_i = 10 \log_{10} \frac{\left\| s_{\mathbf{target}} \right\|^2}{\left\| e_{\mathbf{interf}} \right\|^2}. \tag{33}$$

To calculate $s_{\mathbf{target}}$, we set the signals of all sources and noises to zero except $s_i(t)$ and measure the output signal. In the same way, to calculate $e_{\mathbf{interf}}$, we set $s_i(t)$ and all noise signals to zero and obtain the output signal.

Setup A: The case of 2-Sources and 2-Microphones

In this experiment, we use only two microphones m1 and m2 in Figure 10. In this case, the reverberation time of the room is set to 130 ms. The frame length and frame shift in the STFT analysis are set to 2048 and 256 samples, respectively. Three different methods for the permutation problem are applied on 9 pairs of speech signals. The results of our simulations are shown in Figure 12. In the MaxSir approach, we select the best permutation by maximizing SIR at each frequency bin for solving perfectly the permutation ambiguity [16]. This gives a rough estimate of the upper bound of the performance.

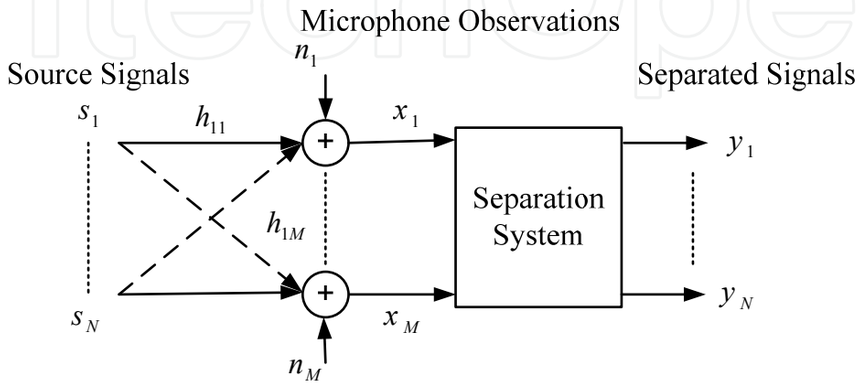


Fig. 11. Block diagram of the separating system

As seen from Figure 12, the results with Murata's method [12] are sometimes very poor, but our proposed method [18] offers almost the same results as that of MaxSir. Figure 13 shows SIRs at each frequency for the 8th pair of speech signals, obtained by the proposed method and Murata's method. The change of signs of SIRs in this figure shows the regions of permutation misalignments. Here, we see the permutation misalignments below 500 Hz obtained by Murata's method, whereas the proposed method has almost perfect permutation alignment. This shows that it is not always true to assume that frequencies not in close proximity have a high correlation of envelopes.

Setup B: The case of 2-Sources and 8-Microphones

In this experiment, we compare the separation performance of our proposed method with those of three other methods, namely, the Interfrequency Coherency method (IFC) [15], the DOA approach with a closed-form formula [17], and MaxSir for the case of 2-Sources and 8-Microphones [17]. To avoid aliasing in the DOA method, we select the distance between the microphones to be 2 cm. All these experiments are performed for three reverberation times T_R = 100 ms, 130 ms, and 200 ms. Before assessing different separation techniques, we first obtain the optimum frame length of STFT at each reverberation time. Then, we evaluate the proposed method in noisy and noise-free cases.

Optimum length of STFT for better separation

To show what frame length of STFT is suitable for better performance of BSS, we perform separation experiments at three reverberation times of T_R = 100 ms, 130 ms, and 200 ms, and by different lengths of STFT. Since the sampling rate is 16 kHz, these reverberation times correspond to P = 1600, 2080, and 3200 taps, respectively.

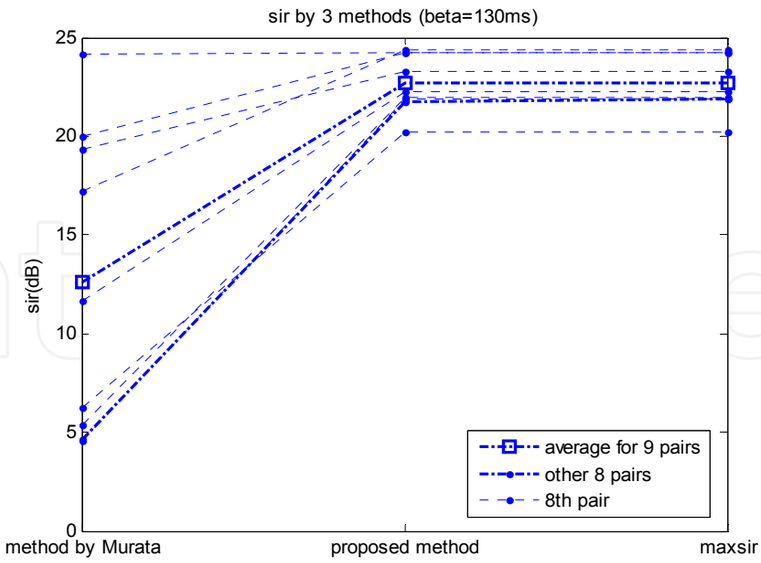


Fig. 12. The separation results of 9 pairs of speech signals for three different methods of permutation problem: Murata's method, proposed method, and MaxSir

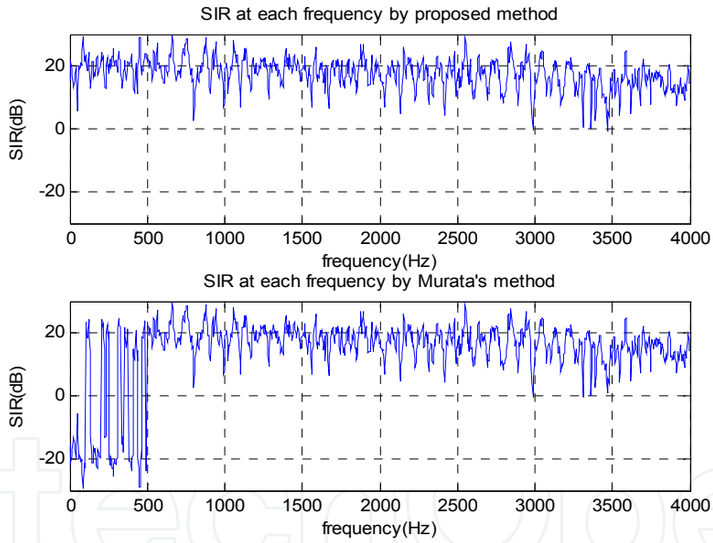


Fig. 13. SIRs measured at different frequencies for the proposed method and the Murata's method

Figure 14 shows the room impulse responses h_{11} for $T_R = 100\,\text{ms}, 130\,\text{ms}, \text{and } 200\,\text{ms}$. We vary the length of STFT by $K = 512$, 1024, 2048, 4096, and 8192 with corresponding frame shifts of $S = 64$, 128, 256, 512, and 1024, respectively. The best permutation is selected by maximizing SIR at each frequency bin. In this way, the results are ideal under the condition that the permutation problem is solved perfectly. The experimental results of SIR for different lengths of STFT are shown in Figure 15. These values are averaged over all nine combinations of speakers to obtain average values of SIR_1 and SIR_2. As it is observed from this figure, in the case of $T_R = 100\,\text{ms}$ we obtain the best performance with $K = 1024$. For

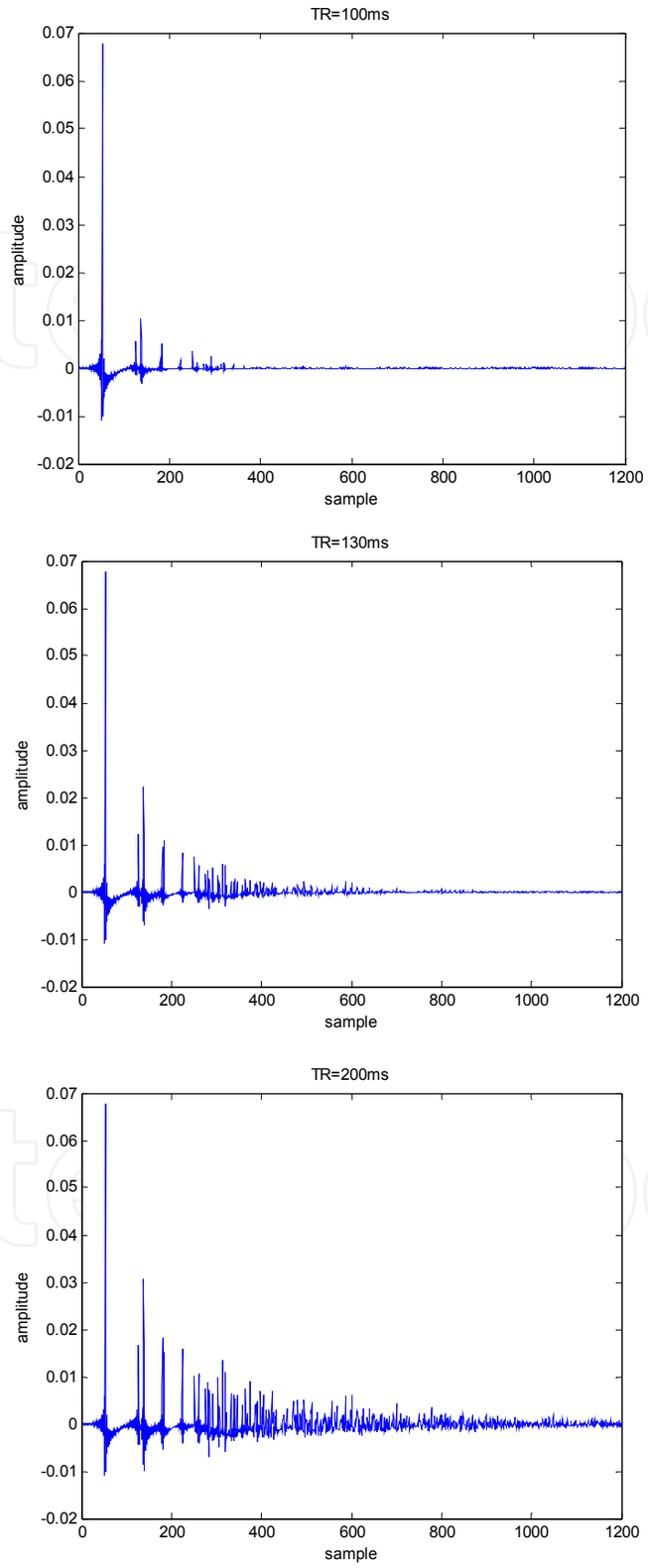


Fig. 14. The room impulse responses h_{11} for $T_R = 100\,\text{ms}, 130\,\text{ms}, \text{and } 200\,\text{ms}$

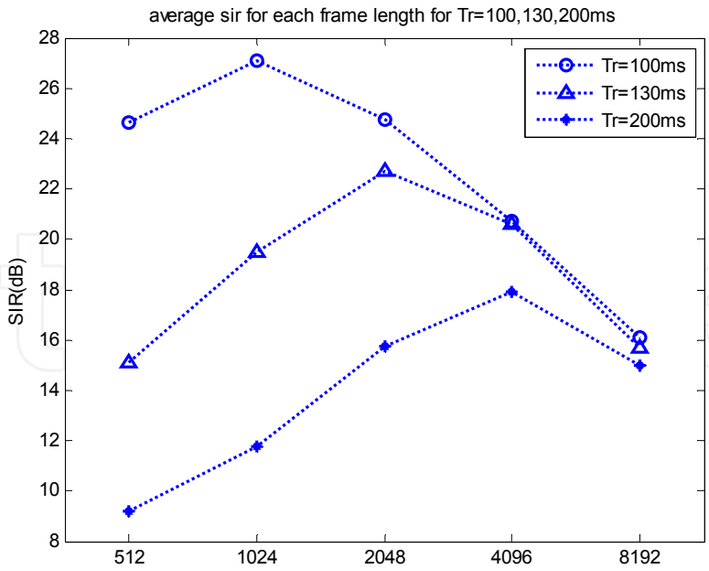


Fig. 15. The experimental results of SIR for different lengths of STFT

the reverberant conditions with $T_R = 130, \text{and } 200 \text{ ms}$, the best performance is realized with $K = 2048$, and 4096, respectively. Figure 16 shows the average of neg-entropy (Eq. 15) as a measurement of independence. We see that by longer lengths of STFT the independence is smaller, and the performance of the fixed-point ICA is poorer [8].

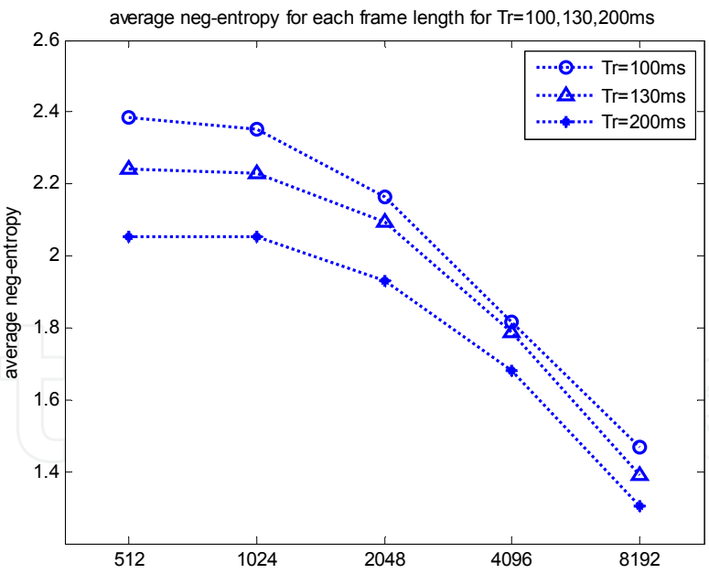
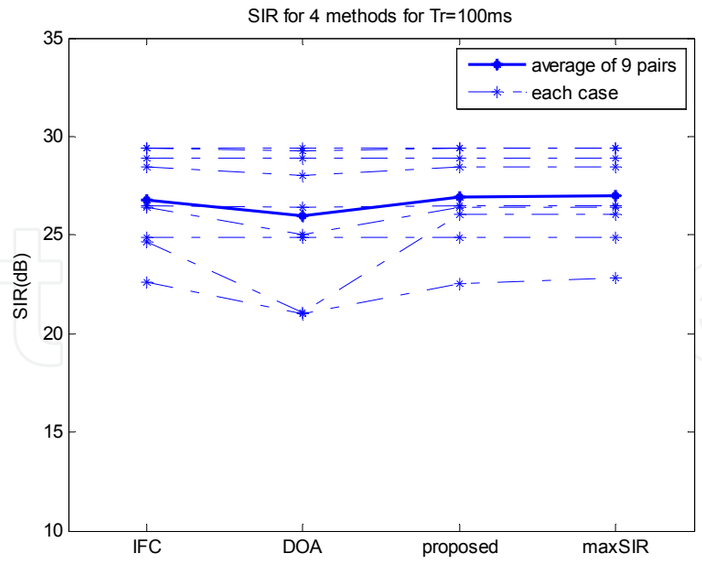


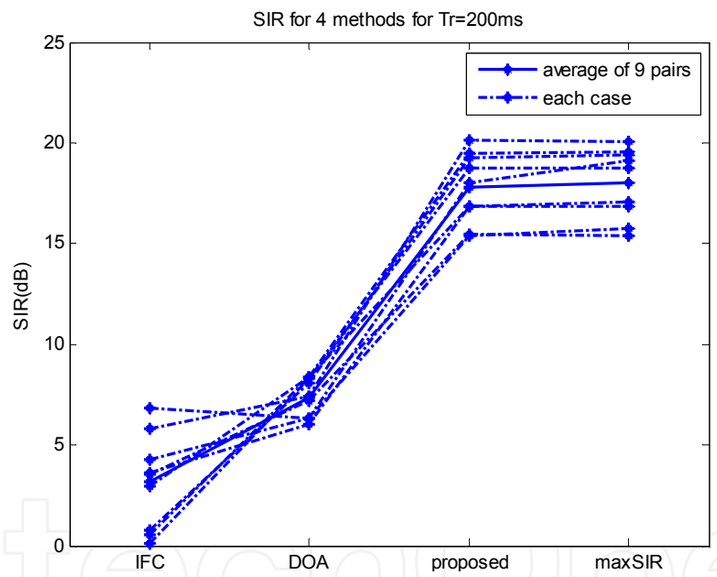
Fig. 16. The average of neg-entropy as a measurement of independence

Evaluation results without background noise

In this section, we compare our proposed method with three methods, namely, IFC, DOA, and MaxSir in the case of 2-Sources and 8-Microphones without the background noise. We select the optimum length of STFT obtained in the previous experiment for each of the three reverberation times. Figure 17 shows the separation results for nine pairs of speech signals



(a)



(b)

Fig. 17. The separation results of 9 pairs of speech signals (a) with $T_R = 100$ ms and (b) with $T_R = 200$ ms as the reverberation times, for four different methods of the permutation problem: the Interfrequency Coherency method (IFC), the DOA method, the proposed method, and the MaxSir method

in the cases where the reverberation times of the room are $T_R = 100$, and 200 ms, respectively. We observe that, when the reverberation time is 100 ms, the separation results for each of the three methods, i.e., IFC, DOA, and proposed methods, are close to the perfect solution obtained by MaxSir. For the reverberant case of $T_R = 200$ ms, the separation performances of IFC and DOA are not good, but the results of SIR for the proposed method are close to the MaxSir approach.

In the IFC method, to use the coherency of the mixing matrices in adjacent frequencies, the mixing matrix should have the form of the transfer function of direct path from each source to each microphone. However, this condition can hold, if the subspace filter reduces the energy of the reflection terms. The performance of the subspace method depends on both the array configuration and the sound environment. In our experiments, the subspace method could not reduce the reflection components, and the performance of the IFC method is poor for the reverberant case. However, in the case of $T_R = 100\,\text{ms}$ the energy of the reflection components is low and the IFC method has good performance. The SIRs at each frequency for four methods in the case of $T_R = 200\,\text{ms}$ are shown in Figure 18. We see a large

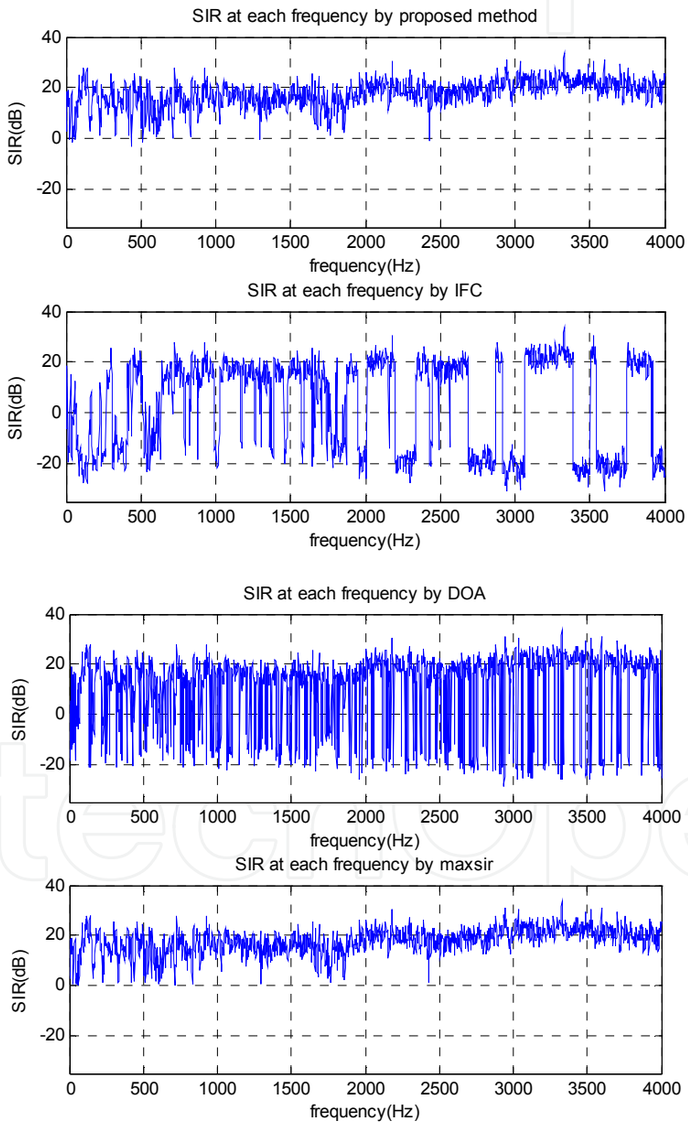


Fig. 18. SIRs measured at each bin frequency for 4 methods: the proposed method, the Interfrequency Coherency method (IFC), the DOA method, and the MaxSir method for the case of $T_R = 200\,\text{ms}$

number of frequencies with permutation misalignments for the IFC and DOA methods. As observed from the simulation results, the proposed approach outperforms the IFC and the DOA methods, where we achieve the best performance in the sense of SIR-improvement.

Evaluation results with background noise

In this part of experiments, we add the restaurant noise from the Noisex-92 database [31] with input SNRs of 5 dB, and 20 dB to the microphone signals. Here, again the optimum window length for the STFT analysis is chosen for each three reverberation times. Figures 19 and 20 show the average SIRs obtained for the proposed, IFC, DOA, and MaxSir methods for the reverberation times of $T_R = 100$ ms, 130 ms, and 200 ms , respectively with input SNRs of 5 dB and 20 dB. It is observed that under the experimental conditions of input SNR = 20 dB and reverberation time of 100 ms, all of the methods, i.e., the proposed, IFC, and DOA give the same separation results. However, as the reverberation time increases, the performance of IFC and DOA decreases. At the reverberation time of 200 ms, the average SIR of the proposed method is slightly reduced. Also, as it is expected, the comparison of Figures 19 and 20 shows that in lower values of input SNRs, the performance of source separation methods decreases. This shows that the ICA-based methods have in general poor separation results in noisy conditions.

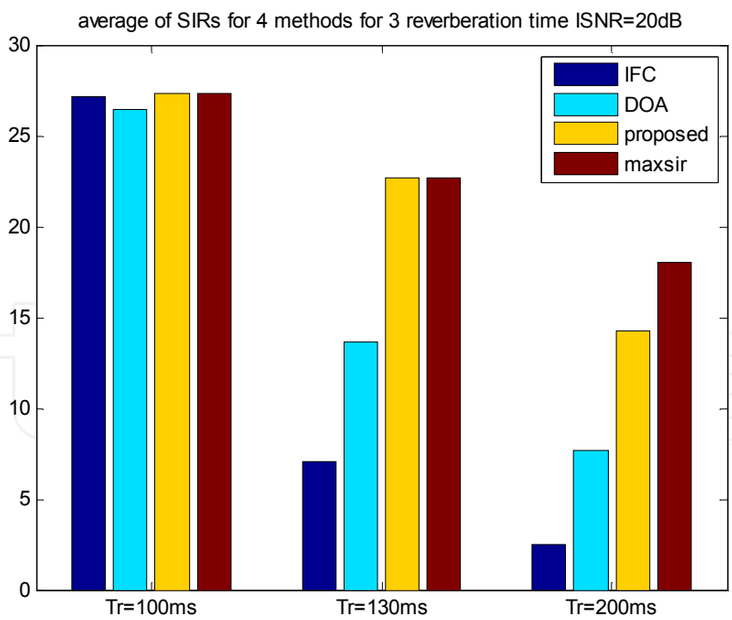


Fig. 19. Average of SIRs for the proposed, IFC, DOA, and MaxSir methods for three reverberation times of $T_R = 100$ ms, 130 ms, and 200 ms , respectively, obtained at the input SNR of 20 dB

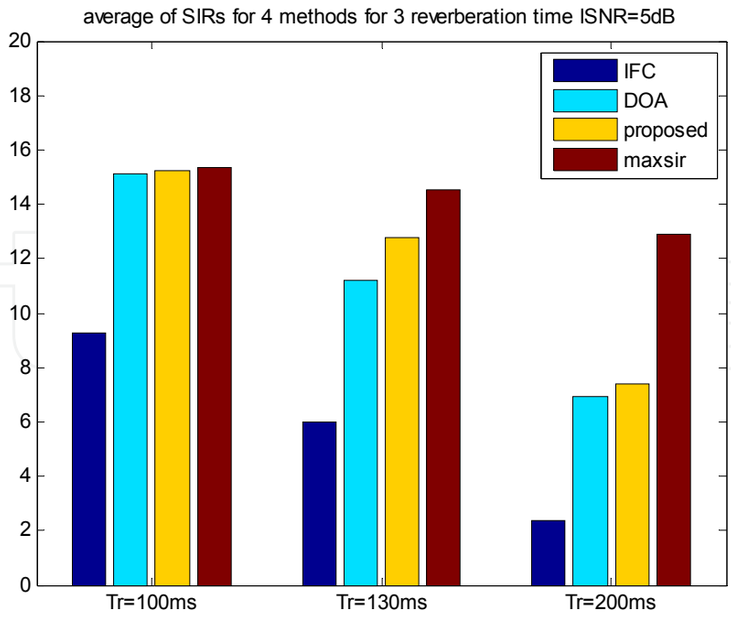


Fig. 20. Average of SIRs for the proposed, IFC, DOA, and MaxSir methods for three reverberation times of $T_R = 100$ ms, 130 ms, and 200 ms , respectively, obtained at the input SNR of 5 dB

10. Conclusion

This chapter presents a comprehensive description of frequency-domain approaches to the blind separation of convolutive mixtures. In frequency-domain approach, the short-time Fourier transform (STFT) is used to convert the convolutive mixtures in time domain to instantaneous mixtures at each frequency. In this way, we can use each of the complex-valued ICA at each frequency bin. We use the fast ICA algorithm for complex-valued signals. The key feature of this algorithm is that it converges faster than other algorithms, like natural gradient-based algorithms, with almost the same separation quality. We employ PCA as pre-processing for the purpose of decreasing the noise effect and dimension reduction. Also, we see that the length of STFT affects the performance of frequency-domain BSS. If the length of STFT becomes longer, the number of coefficients to be estimated increases while the number of samples for learning at each frequency bin decreases. This causes that the assumption of independence in the time series at each bin frequency to collapse, and the performance of the ICA algorithm to become poor. As a result, we select for the frame size an optimum value which is obtained by a trade-off between maintaining the assumption of independence and the length of STFT in the frequency-domain BSS.

We focus on the permutation alignment methods and introduce some conventional methods along with our proposed method to solve this problem. In the proposed method, we maximize the correlation of power ratio measure of each bin frequency with the average of power ratio measures of previous bin frequencies, iteratively with increasing frequency. In the case of 2-sources and 2-microphones, by conducting source separation experiments, we compare the performance of our proposed method with Murata's method which is based on envelope correlation. The results of this comparison show that it is not always true to

assume that frequencies not in close proximity have a high correlation of envelopes. In another overdetermined case of experiment, the proposed method is compared with the DOA, IFC and MaxSir methods. Here, we see that in the reverberant room with high SNR values, the proposed method outperforms other methods. Finally, even though the performance of our proposed method degrades under reverberant conditions with high background noise (low SNRs), the experiments show that the separation results of the proposed method are still satisfactory.

11. Future directions

In this chapter, we have used PCA as a pre-processing technique for the purpose of decreasing the effect of background noise and dimension reduction. This approach assumes that the noise and the signal components are uncorrelated and the noise component is spatially white. Practically, the performance of PCA depends on both the array configuration and the sound environment.

From the results of the experiments, it is clear that two factors affect the performance of BSS methods; background noise and room reverberation. These factors are those that significantly influence the enhancement of audio signals. Therefore, as a future work, we should consider other pre-processing techniques in ICA-based BSS that besides performing dimension reduction also help to decrease the effect of colored noise as well as room reverberation.

12. References

[1] Lee T. W (1998) Independent Component Analysis - Theory and Applications. Norwell, MA: Kluwer.
[2] Comon P (1994) Independent Component Analysis, A New Concept? Signal Processing vol. 36 no. 3: 287-314.
[3] Benesty J, Makino S, Chen J (2005) Speech Enhancement. Springer-Verlag, Berlin, Heidelberg.
[4] Makino S, Lee T. W, Sawada H (2007) Blind Speech Separation. Springer.
[5] Douglas S. C, Sun X (2003) Convolutive Blind Separation of Speech Mixtures Using the Natural Gradient. Speech Communication, vol. 39: 65-78.
[6] Aichner R, Buchner H, Yan F, Kellermann W (2006) A Real-Time Blind Source Separation Scheme and its Application to Reverberant and Noisy Acoustic Environments. Signal Processing, vol. 86, no. 6: 1260-1277.
[7] Smaragdis P (1998) Blind Separation of Convolved Mixtures in the Frequency Domain. Neurocomputing, vol. 22: 21-34.
[8] Araki S, Mukai R, Makino S, Nishikawa T, Saruwatari H (2003) The Fundamental Limitation of Frequency-Domain Blind Source Separation for Convolutive Mixtures of Speech. IEEE Trans. on Speech and Audio Processing, vol. 11, no. 2.
[9] Bingham E, et al. (2000) A Fast Fixed-Point Algorithm for Independent Component Analysis of Complex-Valued Signal. Int. Journal of Neural Systems, vol. 10 (1) 1:8.
[10] Sawada H, Mukai R, Araki S, Makino S (2003) Polar Coordinate-Based Nonlinear Function for Frequency-Domain Blind Source Separation. IEICE Trans. Fundamentals, vol. E86-A, no. 3.

[11] Prasad R, Saruwatari H, Shikano K (2007) An ICA Algorithm for Separation of Convolutive Mixture of Speech Signals. International Journal of Information Technology, vol. 2, no. 4.
[12] Murata N, Ikeda S, Ziehe A (2001) An Approach to Blind Source Separation Based on Temporal Structure of Speech Signals. Neurocomput., vol. 41: 1-24.
[13] Kurita S, Saruwatari H, Kajita S, Takeda K, Itakkura F (2000) Evaluation of Blind Signal Separation Method Using Directivity Pattern Under Reverberant Conditions. ICASSP2000: 3140-3143.
[14] Saruwatari H, Kurita S, Takeda K, Itakura F, Nishikawa T, Shikano K (2003) Blind Source Separation Combining Independent Component Analysis and Beamforming. EURASIP2003: 1135-1146.
[15] Asano F, Ikeda S, Ogawa M, Asoh H, Kitawaki N (2003) Combined Approach of Array Processing and Independent Component Analysis for Blind Separation of Acoustic Signals. IEEE Trans. on Speech and Audio Processing, vol. 11, no. 3: 204-215.
[16] Sawada H, Mukai R, Araki S, Makino S (2004) A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation. IEEE Trans. on Speech and Audio Processing, vol. 12: 530-538.
[17] Sawada H, Araki S, Makino S (2007) Measuring Dependence of Bin-Wise Separated Signals for Permutation Alignment in Frequency-domain BSS. in Proc. ISCAS2007: 3247-3250.
[18] Hesam M, Geravanchizadeh M (2010) A New Solution for the Permutation Problem in the Frequency-Domain BSS Using Power-Ratio Correlation. IEEE Int. Symp. on telecommunications (IST 2010).
[19] Bell A. J, Sejnowski T. J (1995) An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Computation, vol. 7, no. 6: 1129-1159.
[20] Amari S (1998) Natural Gradient Works Efficiently in Learning. Neural Computation, vol. 10: 251-76.
[21] Hyvärinen A (1999) Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Trans. on Neural Networks, vol. 10: 626-634.
[22] Cardoso J.-F (1993) Blind Beamforming for Non-Gaussian Signals. IEE Proceedings-F, vol.140: 362-370.
[23] Matsuoka K, Ohya M, Kawamoto M (1995) A Neural Net for Blind Separation of Nonstationary Signals. Neural Networks, vol. 8: 411-419.
[24] Oppenhaim A. V, Schafer R. W, Buck J. R (1999) Discrete-Time Signal Processing. Prentice Hall, 1999.
[25] Araki S, Makino S, Blin A, Mukai R, Sawada H (2004) Underdetermined Blind Separation for Speech in Real Environments With Sparseness and ICA. in Proc. ICASSP 2004, vol. III: 881-884.
[26] Joho M, Mathis H, Lambert R. H (2000) Overdetermined Blind Source Separation: Using More Sensors Than Source Signals in a Noisy Mixture. Proceedings of ICA 2000: 81-86.
[27] Asano F, Motomura Y, Asoh H, Matsui T (2000) Effect of PCA Filter in Blind Source Separation. Proc. of Int. conf. on Independent Component Analysis (ICA2000).
[28] Allen J. B, Berkley D. A (1979) Image Method for Efficiently Simulating Small Room Acoustics. J. Acoust. Soc. Amer., vol. 65, no. 4: 943-950.
[29] http://www.ldc.upenn.edu/

[30] Vincent E, Gribonval R, Févotte C (2006) Performance Measurement in Blind Audio Source Separation. IEEE Trans. On Audio, Speech, and Language Processing, vol. 14, no. 4.
[31] http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.

IntechOpen

IntechOpen