

Chapter 11

Speech Separation and Extraction by Combining Superdirective Beamforming and Blind Source Separation

Lin Wang, Heping Ding and Fuliang Yin

Abstract Blind source separation (BSS) and beamforming are two well-known multiple microphone techniques for speech separation and extraction in cocktail-party environments. However, both of them perform limitedly in highly reverberant and dynamic scenarios. Emulating human auditory systems, this chapter proposes a combined method for better separation and extraction performance, which uses superdirective beamforming as a preprocessor of frequency-domain BSS. Based on spatial information only, superdirective beamforming presents abilities of dereverberation and noise reduction and performs robustly in time-varying environments. Using it as a preprocessor can mitigate the inherent “circular convolution approximation problem” of the frequency-domain BSS and enhances its robustness in dynamic environments. Meanwhile, utilizing statistical information only, BSS can further reduce the residual interferences after beamforming efficiently. The combined method can exploit both spatial information and statistical information about microphone signals and hence performs better than using either BSS or beamforming alone. The proposed method is applied to two specific challenging tasks, namely a separation task in highly reverberant environments with the positions of all sources known, and a target speech extraction task in highly dynamic cocktail-party environments with

L. Wang (✉) · F. Yin
School of Electronic and Information Engineering,
Dalian University of Technology, Dalian, China
e-mail: wanglin_2k@sina.com; lin.wang@uni-oldenburg.de

F. Yin
e-mail: flyin@dlut.edu.cn

L. Wang
Institute of Physics - Signal Processing Group, University of Oldenburg, Oldenburg, Germany

H. Ding
Information and Communications Technology,
National Research Council, Ottawa, Canada
e-mail: heping.ding@nrc-cnrc.gc.ca

only the position of the target known. Experimental results prove the effectiveness of the proposed method.

11.1 Introduction

Extracting one or several desired speech signals from their corrupted observations is essential for many applications of speech processing and communication. One of the hardest situations to handle is the extraction of desired speech signals in a “cocktail party” condition—from mixtures picked up by microphones placed inside a noisy and reverberant enclosure. In this case, the target speech is immersed in ambient noise and interferences, and distorted by reverberation. Furthermore, the environment may be time varying. Generally, there are two well-known techniques that may achieve the objective: blind source separation (BSS) and beamforming.

With a microphone array, beamforming is a well-known technique for directional signal reception [1, 2]. Depending on how the beamformer weights are chosen, it can be implemented as a data-independent fixed beamforming or data-dependent adaptive one [3–7]. Although an adaptive beamformer generally exhibits better noise reduction abilities, a fixed beamformer is more preferred in complicated environments due to its robustness. By coherently summing signals from multiple sensors based on a model of the wavefront from acoustic sources, a fixed beamformer presents a specified directional response. With abilities of enhancing signals from the desired direction while suppressing ones from other directions, it can be used to perform both noise suppression and dereverberation. The most conventional fixed beamformer is a delay-and-sum one, which however requires a large number of microphones to achieve high performance. Another filter-and-sum beamformer has superdirective response with optimized weights [4]. Assuming the directions of the sources are known, speech separation or extraction can be obtained by forming individual beams at the target sources separately. However, fixed beamforming performs limitedly in real cocktail-party scenarios. First, the performance is closely related to the microphone array size—a large array is usually required to obtain a satisfactory result but may not be practically feasible. Second, beamforming cannot suppress the interfering reverberation coming from the desired direction.

BSS is a technique for recovering the source signals from observed signals with the mixing process unknown. By exploiting the statistical independence of the sources, independent component analysis (ICA)-based algorithms are commonly used to solve the BSS problem [8–13]. While time domain ICA-based techniques are well suited for instantaneous mixing problem, they are not efficient in addressing the convolutive mixture problem encountered in reverberant environments [14–17]. By considering the BSS problem in the frequency domain, the convolutive mixing problem can be transformed into an instantaneous mixing problem for each frequency bin, reducing computation complexity significantly. However, the inherent permutation and amplitude scaling ambiguity problem occurs at each frequency bin in frequency-domain BSS, deteriorating signal reconstruction in the time domain significantly

[18, 19]. The permutation ambiguity problem has been investigated intensively and there are generally three strategies to tackle it. The first is to make the separation filters smooth in the frequency domain by limiting the filter length in the time domain [16, 20, 21]. The second strategy is to exploit the interfrequency dependence of the amplitude of separated signals [22–30]. The third strategy is to exploit the position information about sources such as direction of arrival (DOA). By estimating the arriving delay of sources or analyzing the directivity pattern formed by a separation matrix, source direction can be estimated and permutations aligned [31–36].

The relationship between blind source separation and beamforming has been intensively investigated in recent years, and adaptive beamforming is commonly used to explain the physical principle of convolutive BSS [37, 38]. In addition, many approaches have been presented that combine both techniques. Some of these combined approaches are aimed at resolving the permutation ambiguity inherent in frequency-domain BSS [31, 39], whereas other approaches utilize beamforming to provide a good initialization for BSS or to accelerate its convergence [40–43]. For now, there were no systematically studies on combining the two techniques to improve the separation performance in challenging acoustic scenarios.

Compared with beamforming, which extracts desired speech and suppress interference, BSS aims at separating all the involved desired and interfering sources equally. One advantage with blind source separation is that it does not need to know the direction of arrival of any signals and the array geometry can be arbitrary and unknown to the system. Nevertheless, blind source separation also performs limitedly in real cocktail-party scenarios. First, BSS performs poorly in high reverberation with long mixing filters, due to the “circular convolution approximation problem”. Second, underdetermined situations can result from the fact that there are only a limited number of microphones. Third, the performance of BSS degrades in dynamic environments.

Due to the reasons above, few methods proposed in recent years show good separation/extraction results in a real cocktail-party environment. In contrast, a human has a remarkable ability to focus on a specific speaker in that case. This selective listening capability is partially attributed to binaural hearing. Two ears work as a beamformer, which enables directive listening [44], then the brain analyzes the received signals to extract sources of interest from the background, just as blind source separation does [45–47]. Stimulating this principle, we propose to do speech separation and extraction by combining beamforming and blind source separation. Specifically, the following two issues will be addressed:

- To improve the separation performance in highly reverberant scenarios, a combined method is proposed which uses beamforming as a preprocessor of blind source separation by forming a number of beams each pointing at a source. With beamforming shortening mixing filters, the inherent “circular convolution approximation” problem in the frequency-domain BSS is mitigated and the performance of proposed method can improve significantly especially in high reverberation.
- The combined method is further extended to a special case of target speech extraction problem in noisy cocktail-party environments, where only one source

is of interest. Instead of focusing on all the sources, the proposed method forms just several fixed beams at an area containing the target source. The proposed scheme can enhance the robustness to time-varying environments and make the target source dominant in the output of the beamformer. Consequently, the subsequent extraction task with blind source separation becomes easier and satisfactory extraction results can be obtained even in challenging scenarios.

The rest of the chapter is organized as follows. The principles of frequency-domain blind source separation and superdirective beamforming are reviewed in Sects. 11.2 and 11.3, respectively. Especially, the inherent “circular convolution approximation” problem with frequency-domain BSS is discussed in detail in Sect. 11.2. BSS and beamforming are combined for a better separation results and the performance of the combined method is experimentally evaluated in Sect. 11.4. The combined method is further extended to a target speech extraction problem with some experimental results shown in Sect. 11.5. Finally, conclusions are drawn in Sect. 11.6.

11.2 Frequency-Domain BSS and Its Fundamental Problem

BSS is a powerful tool for solving cocktail-party problems since it aims at recovering the source signals from observed signals with the mixing process unknown. The simplest instantaneous BSS problem can be solved by independent components analysis (ICA), which assumes that all the source signals are independent of each other. One challenge arises when the mixing process is convolutive, i.e., the observations are combinations of filtered versions of sources. The convolutive BSS problem can be solved in the time domain, where the separation network is derived by optimizing a time-domain cost function. However, the task of estimating many parameters simultaneously has to face the challenge of slow convergence and high-computational demand. Alternatively, the convolutive BSS problem can be solved in the frequency domain, where instantaneous BSS is performed at individual frequency bins, reducing the computational complexity significantly. In this section, the principle of the frequency-domain BSS is introduced at first, followed by a discussion of an inherent problem, namely the circular convolution approximation problem, which degrades the performance of the frequency-domain BSS in high reverberation.

11.2.1 Frequency-Domain BSS

Supposing N sources and M sensors in a real-world acoustic scenario, the source vector $\mathbf{s}(n) = [s_1(n), \dots, s_N(n)]^T$, and the observed vector $\mathbf{x}(n) = [x_1(n), \dots, x_M(n)]^T$, the mixing channels can be modeled by FIR filters of length P , the convolutive mixing process is formulated as

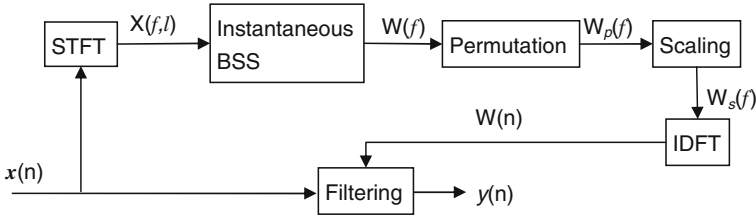


Fig. 11.1 Workflow of frequency-domain blind source separation

$$\mathbf{x}(n) = \mathbf{H}(n) * \mathbf{s}(n) = \sum_{p=0}^{P-1} \mathbf{H}(p) \mathbf{s}(n-p) \quad (11.1)$$

where $\mathbf{H}(n)$ is a sequence of $M \times N$ matrices containing the impulse responses of the mixing channels, n is the time index, and the operator “*” denotes matrix convolution. For separation, FIR filters of length L can be used to estimate the source signals $\mathbf{y}(n) = [y_1(n), \dots, y_N(n)]^T$ by

$$\mathbf{y}(n) = \mathbf{W}(n) * \mathbf{x}(n) = \sum_{l=0}^{L-1} \mathbf{W}(l) \mathbf{x}(n-l) \quad (11.2)$$

where $\mathbf{W}(n)$ is a sequence of $N \times M$ matrices containing the unmixing filters.

The demixing network $\mathbf{W}(n)$ can be estimated in the frequency domain. By using a blockwise Q -point short-time Fourier transform (STFT), the time-domain convolution regarding the mixing process can be converted into frequency-domain multiplications and correspondingly the convolutive BSS problem is converted into multiple instantaneous BSS problem at each frequency bin. This is expressed as

$$\mathbf{x}(f, l) = \mathbf{H}(f) \mathbf{s}(f, l) \quad (11.3)$$

where l is a decimated version of the time index n , f is the frequency index, $\mathbf{H}(f)$ is the Fourier transform of $\mathbf{H}(n)$, and $\mathbf{x}(f, l)$ and $\mathbf{s}(f, l)$ are the STFTs of $\mathbf{x}(n)$ and $\mathbf{s}(n)$, respectively. The remaining task will be to find a demixing matrix $\mathbf{W}_{\text{demix}}(f)$ at individual frequency bins, so that the original signals can be recovered. This is expressed as

$$\mathbf{y}(f, l) = \mathbf{W}_{\text{demix}}(f) \mathbf{x}(f, l) \quad (11.4)$$

Based on the discussion above, the workflow of the frequency-domain BSS is shown in Fig. 11.1. The observed time-domain signals are converted into the time–frequency domain by STFT; then instantaneous BSS is applied to each frequency bin; after solving the inherent permutation and scaling ambiguities, the separated signals of all frequency bins are combined and inverse-transformed to the time domain. The procedure of a frequency-domain BSS mainly consists of three blocks: instantaneous BSS, permutation alignment, and scaling correction.

(1) *Instantaneous BSS*

After decomposing time-domain convolutive mixing into frequency-domain instantaneous mixing, it is possible to perform instantaneous separation at each frequency bin with a complex-valued ICA algorithm. The ICA algorithms for instantaneous BSS have been studied for many years and are considered to be quite mature. For instance, the demixing matrix can be estimated iteratively by using the well-known Infomax algorithm [11, 12], i.e.,

$$\begin{cases} \mathbf{y}(f, l) = \mathbf{W}(f)\mathbf{x}(f, l) \\ \mathbf{W}(f) = \mathbf{W}(f) + \eta(\mathbf{I} - \mathbb{E}[\Phi(\mathbf{y}(f, l))\mathbf{y}^H(f, l)])\mathbf{W}(f) \end{cases} \quad (11.5)$$

where \mathbf{I} is an identity matrix, $\Phi(\cdot)$ is a nonlinear function and $\mathbb{E}[\cdot]$ is the expectation operator.

(2) *Permutation alignment*

Although satisfactory instantaneous separation may be achieved within all frequency bins, combining them to recover the original sources is a challenge because of the unknown permutations associated with individual frequency bins. This permutation ambiguity problem is the main challenge in the frequency-domain BSS and how to solve this problem has been a hot topic in the research community in recent years.

Two kinds of strategies can be used to solve this problem. The first strategy is to exploit the interfrequency dependence of the amplitude of separated signals [22–25]. The second strategy is to exploit the position information of sources such as direction of arrival or the continuity of the phase of separation matrix [31–33]. By analyzing the directivity pattern formed by a separation matrix, source direction can be estimated and permutations aligned. Since the performance of the second approach is generally limited by the reverberation density of the environment and the source positions, we prefer to use the first approach. In [24], a region-growing permutation alignment approach is proposed with good results, which is based on the interfrequency of separated signals. Bin-wise permutation alignment is applied first across all frequency bins, using the correlation of separated signal powers; then the full frequency band is partitioned into small regions based on the bin-wise permutation alignment result. Finally, region-wise permutation alignment is performed, which can prevent the spreading of the misalignment at isolated frequency bins to others and thus improves permutation alignment results. After permutation alignment, we can assume that the separated frequency components from the same source are grouped together.

(3) *Scaling correction*

The scaling indeterminacy can be resolved relatively easily by using the Minimal Distortion Principle [48]:

$$\mathbf{W}_s(f) = \text{diag}(\mathbf{W}_p^{-1}(f)) \cdot \mathbf{W}_p(f) \quad (11.6)$$

where $\mathbf{W}_p(f)$ is $\mathbf{W}(f)$ after permutation correction, $(\cdot)^{-1}$ denotes inversion of a square matrix or pseudo inversion of a rectangular matrix; $\text{diag}(\cdot)$ retains only the

main diagonal components of the matrix. $\mathbf{W}_s(f)$ is the demixing matrix $\mathbf{W}_{\text{demix}}(f)$, which we are looking for.

Finally, the demixing network $\mathbf{W}(n)$ is obtained by inverse Fourier transforming $\mathbf{W}_s(f)$, and the estimated source $\mathbf{y}(n)$ is obtained by filtering $\mathbf{x}(n)$ through $\mathbf{W}(n)$.

11.2.2 Circular Convolution Approximation Problem

Besides the permutation and scaling ambiguity, another problem also affects the performance of frequency-domain BSS: the STFT circular convolution approximation [24, 37, 43]. The convolutive mixture is decomposed into an instantaneous mixture at each frequency bin as shown in (11.3). Equation (11.3) is only an approximation since it implies a circular convolution but not a linear convolution in the time domain. It is correct only when the STFT analysis frame length L is larger than the mixing filter length P . Thus, a large L is required to ensure a sufficient separation performance. However in that case, the instantaneous separation performance is saturated before reaching a sufficient level, because decreased time resolution for STFT and fewer data available in each frequency bin will violate the independence assumption.

To verify the statement above, a simple example is given below. As is well known, non-Gaussianity is an important measure for the independence of signals while kurtosis is an important measure for non-Gaussianity [8]. The kurtosis of a signal s is defined as

$$\text{kurt}(s) = E\{s^4\} - 3(E\{s^2\})^2 \quad (11.7)$$

where the operator $E\{\cdot\}$ denotes expectation. A high kurtosis value indicates strong non-Gaussianity and independence. We compare kurtosis values of the STFT coefficients of a speech signal when different STFT frame sizes (varying from 128 to 16,384) are used. The kurtosis value is calculated for the real and imaginary parts of the complex-valued coefficients, respectively. Since the kurtosis value, which is calculated for the time sequences at each frequency bin, varies with respect to frequency, a median value is chosen from the set of kurtosis values at all frequencies to represent the independence measure of the signal after STFT analysis. Considering the possible influence of insufficient data points at each frequency bin after a long-frame STFT, three speech signals with lengths of 10s, 40s, and 160s, respectively, are tested. The obtained kurtosis for different test signals and different STFT frame sizes are shown in Fig. 11.2. For reference, the kurtosis of a normalized Gaussian white signal is also given. As can be seen in Fig. 11.2, the real and imaginary parts of the STFT coefficients show similar variation trend with respect to the STFT frame size. Additionally, two phenomena can be observed.

- (1) Large kurtosis values can be observed for small STFT frame sizes. The kurtosis value increases slightly with increased STFT frame size and then decreases significantly when the STFT frame size is larger than 1,024. The kurtosis value is close to a Gaussian white signal when the STFT frame size is very large.

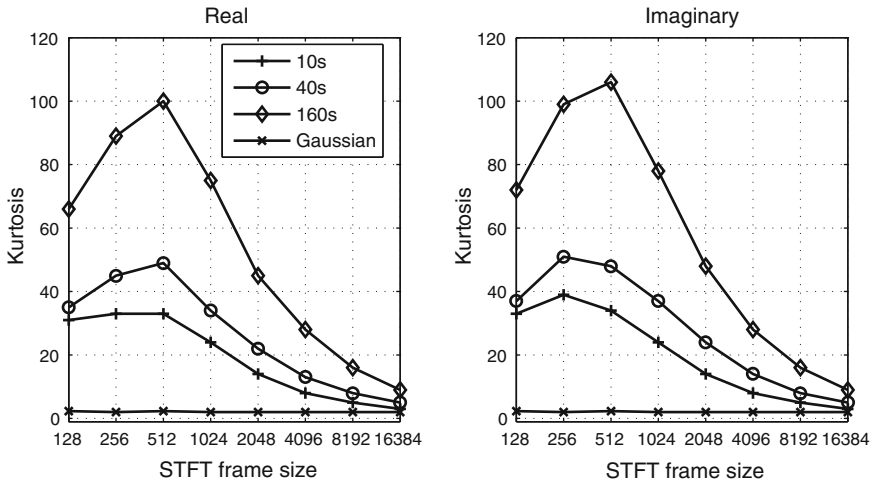


Fig. 11.2 Kurtosis of the STFT coefficients versus STFT frame size (calculated for speech signals of different lengths)

This demonstrates that the STFT coefficients of a speech signal show strong non-Gaussianity for small STFT frame sizes, but tend to be Gaussian for large STFT frame sizes.

- (2) Even for a same STFT frame size, the kurtosis of a speech signal with different length is different. Generally, a long signal shows higher kurtosis than short signals. This may be due to insufficient data points available at each frequency bin with short signals.

The results shown in Fig. 11.2 demonstrate that the independence assumption of sources may collapse when a large STFT frame size is used, hence degrading the separation performance significantly. There is a dilemma in determining the STFT frame size: short frames make the conversion to instantaneous mixture incomplete, while long ones disturb the separation. The conflict becomes severer in highly reverberant environments and leads to the degraded performance. Generally, a frequency-domain BSS which works well in low (100–200 ms) reverberation has degraded performance in medium (200–500 ms) and high (>500 ms) reverberation. Since the problem originates from a processing step, which approximates linear convolutions with circular convolutions in frequency-domain BSS, we call it circular convolution approximation problem.

Furthermore, some separation experiments (2×2 and 4×4) are carried out in a reverberant environment (reverberation time 300 ms) using a frequency-domain BSS algorithm proposed in [24] with different STFT frame sizes. The source speech signals are of 8s length and 8 kHz sampling rate.¹ The resultant signal-interference-ratios (SIR) are shown in Fig. 11.3. In both 2×2 and 4×4 cases, the separation

¹ More details of the experiment can be found in Sect. 11.4.2.

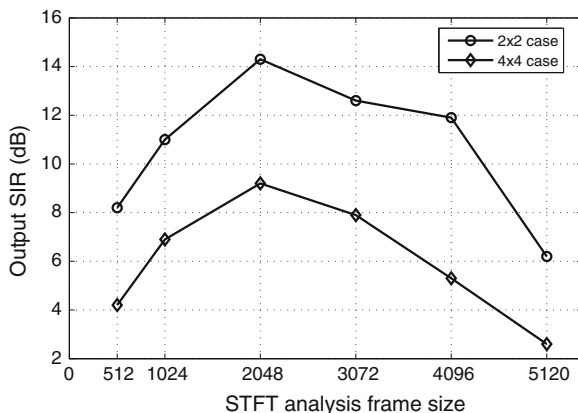


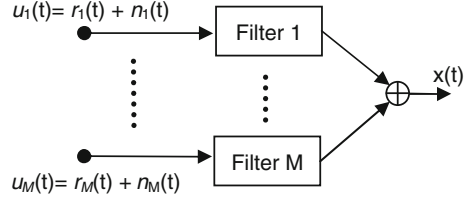
Fig. 11.3 Performance of BSS versus STFT frame size (calculated for speech signals of 8s length, $RT_{60} = 300$ ms)

performance peaks at the STFT frame size of 2,048, while degrading for both shorter and longer frame sizes. This verifies the discussion about the dilemma in determining the STFT frame size. Obviously, an optimal STFT frame size may exist for a specific reverberation. However, due to complex acoustical environments and varieties of source signals, it is difficult to determine this value precisely. Generally, at sampling frequency of 8,000 Hz, 1,024 or 2,048 can be used as a balanced choice for the frame length.

11.3 Superdirective Beamforming

Beamforming is a technique used in sensor arrays for directional signal reception by enhancing target directions and suppressing unwanted ones. Beamforming can be classified as either fixed or adaptive, depending on how the beamformer weights are chosen. An adaptive beamformer obtains directive response mainly by analyzing the statistical information contained in the array data, not by utilizing the spatial information directly. It generally adapts its weights during breaks in the target signal. The challenge to predict signal breaks when several people are talking concurrently limits the feasibility of adaptive beamforming in cocktail-party environments significantly. In contrast, the weights of a fixed beamformer do not depend on array data and are chosen to present a specified response for all scenarios. The directional response is achieved by coherently summing signals from multiple sensors based on a model of the wavefront from acoustic sources. A filter-and-sum beamformer has super directivity response with optimized weights. The superdirective beamformer can be designed in the frequency-domain.

Fig. 11.4 Principle of a filter-and-sum beamformer



The principle of a fixed beamformer is given in Fig. 11.4, where a weighted sum of signals from M sensors is produced to enhance the target direction. Suppose a beamformer model with a target source $r(t)$ and background noise $n(t)$, the components received by the l -th sensor is $u_l(t) = r_l(t) + n_l(t)$ in the time domain. Similarly, in the frequency domain, the l -th sensor output is $u_l(f) = r_l(f) + n_l(f)$. The array output in the frequency domain is

$$\mathbf{x}(f) = \mathbf{b}^H(f)\mathbf{u}(f) \quad (11.8)$$

where $\mathbf{b}(f) = [b_1(f), \dots, b_M(f)]^T$ is the beamforming weight vector composed of beamforming weights from each sensor, and $\mathbf{u}(f) = [u_1(f), \dots, u_M(f)]^T$ is the output vector composed of outputs from each sensor, and $(\cdot)^H$ denotes conjugate transpose. The $\mathbf{b}(f)$ depends on the array geometry and source directivity, as well as the array output optimization criterion such as a signal-to-noise ratio (SNR) gain criterion [4, 49, 50].

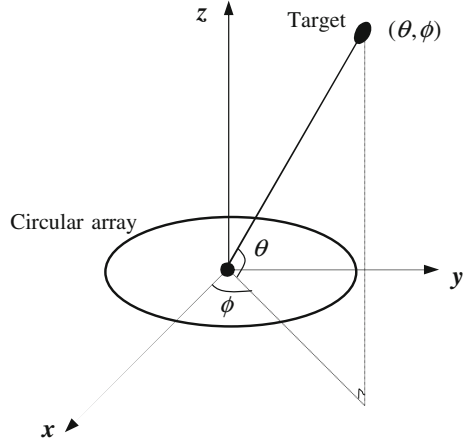
Suppose $\mathbf{r}(f) = [r_1(f), \dots, r_M(f)]^T$ is the source vector, which is composed of the target source signals from the sensors, and $\mathbf{n}(f)$ is the noise vector, which is composed of the spatial diffuse noises from the sensors. The array gain is a measure of the improvement in signal-to-noise ratio. It is defined as the ratio of the SNR at the output of the beamforming array to the SNR at a single reference microphone. For development of the theory, the reference SNR is defined to be the ratio of average signal power spectral densities over the microphone array, $\sigma_r^2(f) = E\{\mathbf{r}^H(f)\mathbf{r}(f)\}/M$, to the average noise power spectral density over the array, $\sigma_n^2(f) = E\{\mathbf{n}^H(f)\mathbf{n}(f)\}/M$. By derivation, the array gain at frequency f is expressed as

$$G(f) = \frac{\mathbf{b}^H(f)\mathbf{R}_{rr}(f)\mathbf{b}(f)}{\mathbf{b}^H(f)\mathbf{R}_{nn}(f)\mathbf{b}(f)} \quad (11.9)$$

where $\mathbf{R}_{rr}(f) = \mathbf{r}(f)\mathbf{r}^H(f)/\sigma_r^2(f)$ is the normalized signal cross-power spectral density matrix, and $\mathbf{R}_{nn}(f) = \mathbf{n}(f)\mathbf{n}^H(f)/\sigma_n^2(f)$ is the normalized noise cross-power spectral density matrix. Provided $\mathbf{R}_{nn}(f)$ is nonsingular, the array gain is maximized with the weight vector

$$\mathbf{b}_{\text{opt}}(f) = \mathbf{R}_{nn}^{-1}(f)\mathbf{r}(f) \quad (11.10)$$

Fig. 11.5 Circular array geometry



The terms $\mathbf{R}_{nn}(f)$ and $\mathbf{r}(f)$ in (11.10) depend on the array geometry and the target source direction. For instance, given a circular array, $\mathbf{R}_{nn}(f)$ and $\mathbf{r}(f)$ can be calculated as below [45].

Figure 11.5 shows an M -element circular array with a radius of r and a target source coming from the direction (θ, ϕ) . The elements are equally spaced around the circumference, and their positions, which are determined from the layout of array, are given in a matrix form as

$$\mathbf{v} = \begin{bmatrix} v_{x_1} & v_{y_1} \\ \vdots & \vdots \\ v_{x_M} & v_{y_M} \end{bmatrix} \tag{11.11}$$

The source vector $\mathbf{r}(f)$ can be derived as

$$\mathbf{r}(f) = \begin{bmatrix} \exp(-jk(\sin \theta \cdot \cos \phi \cdot v_{x_1} + \sin \theta \cdot \sin \phi \cdot v_{y_1})) \\ \vdots \\ \exp(-jk(\sin \theta \cdot \cos \phi \cdot v_{x_M} + \sin \theta \cdot \sin \phi \cdot v_{y_M})) \end{bmatrix} \tag{11.12}$$

where $k = 2\pi f/c$ is the wave number, and c is the sound velocity. The normalized noise cross-power spectral density matrix $\mathbf{R}_{nn}(f)$ is expressed as

$$(\mathbf{R}_{nn}(f))_{m_1 m_2} = \begin{cases} \frac{\sin(kd_{m_1 m_2})}{kd_{m_1 m_2}}, & m_1 \neq m_2 \\ 1, & m_1 = m_2 \end{cases} \tag{11.13}$$

where $(\mathbf{R}_{nn}(f))_{m_1 m_2}$ is the (m_1, m_2) entry of the matrix $\mathbf{R}_{nn}(f)$, $m_1, m_2 = 1, \dots, M$, k is the wave number, $d_{m_1 m_2}$ is the distance between two microphones m_1 and m_2 .

After calculating the beamforming vector by (11.10), (11.12) and (11.13) at each frequency bin, the time-domain beamforming filter $\mathbf{b}(n)$ is obtained by inverse Fourier transforming $\mathbf{b}_{\text{opt}}(f)$.

The procedure above is to design a beamformer with only one target direction. In a speech separation or extraction system, each source signal may be separately obtained using the directivity of the array, if the directions of sources are known. However, beamforming in principle performs limitedly in highly reverberant conditions because it cannot suppress the interfering reverberation coming from the target direction.

11.4 Enhanced Separation in Reverberant Environments by Combining Beamforming and BSS

Due to the circular convolution approximation problem, the performance of a frequency-domain BSS algorithm degrades seriously when the mixing filters are long, e.g., in high reverberation environments. Thus, the problem may be mitigated if the mixing filters become shorter. With directive response enhancing desired direction and suppressing unwanted ones, a superdirective beamforming can deflate the reflected paths and hence equivalently shorten the mixing filter. It thus may help compensate for the deficiency of blind source separation. If we use beamforming as a preprocessor for blind source separation, at least three advantages can be achieved:

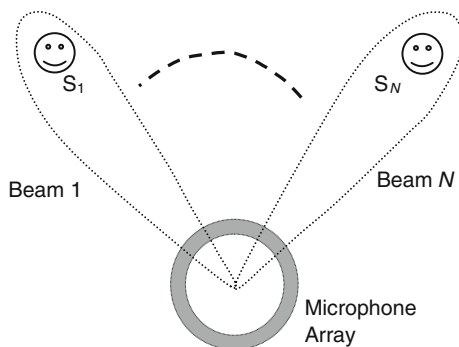
- (1) The interfering residuals due to reverberation after beamforming are further reduced by blind source separation;
- (2) The poor separation performance of blind source separation in reverberant environments is compensated for by beamforming, which suppresses the reflected paths and shortens the mixing filters;
- (3) Beamformer enhances the source in its path and suppresses the ones outside. It thus enhances signal-to-noise ratio and provides a cleaner output for blind source separation to process.

From another point of view, beamforming makes primary use of spatial information while blind source separation utilizes statistical information contained in signals. Integrating both pieces of information should help to get better separation results, just like the way our ears separate audio signals. The details of the combined method are given below, followed by experimental results and analysis.

11.4.1 Workflow of the Combined Method

The illustration of the combined method is shown in Fig. 11.6. For N sources received by an array of M microphones, N beams are formed toward them, respectively,

Fig. 11.6 Illustration of the proposed method combining beamforming and blind source separation



assuming the directions of all sources are known. Then the N beamformed outputs are fed to blind separation to recover the N sources. The signal flow of the proposed method is shown in Fig. 11.7, which mainly consists of three stages: acoustic mixing, beamforming, and separation.

The mixing stage results in the observed vector

$$\mathbf{u}(n) = \mathbf{H}(n) * \mathbf{s}(n) \quad (11.14)$$

where $\mathbf{u}(n) = [u_1(n), \dots, u_M(n)]^T$ and $\mathbf{s}(n) = [s_1(n), \dots, s_N(n)]^T$ are the observed and the source vectors, respectively, $\mathbf{H}(n)$ is a sequence of $M \times N$ matrices containing the impulse responses of the mixing channels, and the operator ‘*’ denotes matrix convolution.

The beamforming stage is expressed as

$$\mathbf{x}(n) = \mathbf{B}(n) * \mathbf{u}(n) = \mathbf{B}(n) * \mathbf{H}(n) * \mathbf{s}(n) = \mathbf{F}(n) * \mathbf{s}(n) \quad (11.15)$$

where $\mathbf{x}(n) = [x_1(n), \dots, x_N(n)]^T$ is the beamforming output vector, $\mathbf{B}(n)$ is a sequence of $N \times M$ matrices containing the impulse responses of beamformer, $\mathbf{F}(n)$ is the global impulse response by combining $\mathbf{H}(n)$ and $\mathbf{B}(n)$.

The blind source separation stage is expressed as

$$\mathbf{y}(n) = \mathbf{W}(n) * \mathbf{x}(n) = \mathbf{W}(n) * \mathbf{F}(n) * \mathbf{s}(n) \quad (11.16)$$

where $\mathbf{y}(n) = [y_1(n), \dots, y_N(n)]^T$ is the estimated source signal vector, and $\mathbf{W}(n)$ is a sequence of $N \times N$ matrices containing the unmixing filters.

It can be seen from (11.14)–(11.16) that with beamforming reducing reverberation and enhancing signal-to-noise ratio, the combined method is able to replace the original mixing network $\mathbf{H}(n)$, which results from the room impulse response, with a new mixing network $\mathbf{F}(n)$, which is easier to separate.

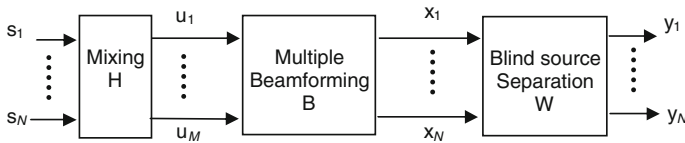


Fig. 11.7 Signal flow of the proposed method combining beamforming and blind source separation

The blind source separation and beamforming algorithms introduced in Sects. 11.2 and 11.3 can be used directly for the combined method. However, the following two issues should be clarified when implementing the combined method.

(1) *The choice of a beamformer*

Beamformer can be implemented as a fixed one or an adaptive one. As mentioned before, comparing to fixed beamforming, an adaptive method is not appropriate for the combined method. First, an adaptive beamformer obtains directive response mainly by analyzing the statistical information contained in the array data, not by utilizing the spatial information directly. Its essence is similar to that of convolutive blind source separation [37, 38]. Cascading them together is equivalent to using the same techniques repeatedly, hence contributing little to performance improvement. Second, An adaptive beamformer generally adapts its weights during breaks in the target signal. However, it is a challenge to predict signal breaks when several people are talking concurrently. This significantly limits the applicability of adaptive beamforming to source separation. In contrast, a fixed beamformer, which relies mainly on the spatial information, does not have such disadvantages. It is data independent and more robust. With its directive response fixed in all acoustic scenarios, a superdirective beamformer is preferred in the combined method.

(2) *The permutation ambiguity problem in BSS*

Permutation ambiguity inherent in frequency-domain BSS is always a challenging problem. Generally, there are two approaches to solve it. One is to exploit the dependence of separated signals across frequencies, and the other is to exploit the position information of sources: the directivity pattern of the mixing/unmixing matrix provides a good reference for permutation alignment. However, in the proposed method, the directivity information contained in the mixing matrix does not exist any longer after beamforming. Even if the source positions are known, they are not much helpful to permutation alignment in the subsequent blind source separation. Consequently, what we can use for permutation is merely the first reference: the interfrequency dependence of separated signals.

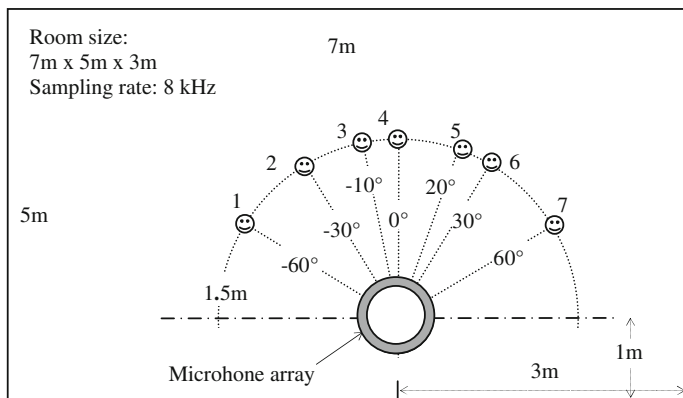


Fig. 11.8 Simulated room environment for speech separation

11.4.2 Experimental Results and Analysis

We evaluate the performance of the proposed method in simulated experiments from two aspects. The first experiment verifies the advantage of the beamforming preprocessing, i.e., dereverberation and noise reduction; the second one investigates the performance of the proposed method in various reverberant conditions, and compares it with a BSS-only method and a beamforming-only one.

The simulation environment is shown in Fig. 11.8, the room size is $7\text{m} \times 5\text{m} \times 3\text{m}$, all sources and microphones are 1.5 m high. The room impulse response was obtained by using the image method [51], and the reverberation time was controlled by varying the absorption coefficient of the wall. The sampling rate is 8 kHz. For BSS, a STFT frame size of 2,048 is used. For beamforming, a circular microphone array is used to design the beamformer with the filter length 2,048. A commonly used objective measure, signal-to-interference ratio (SIR), is employed to evaluate the separation performance [45].

11.4.2.1 Influence of Beamforming Preprocessing

The proposed algorithm is used for separating three sources in the environment shown in Fig. 11.8, using a 16-element circular microphone array with a radius of 0.2 m. The simulated room reverberation time is $RT_{60} = 300$ ms, where RT_{60} is the time required for the sound level to decrease by 60 dB. This is a medium reverberant condition. Three source locations (2, 4, 6) are used, and the sources are two male speeches and one female speech of 8s each. Three beams are formed by the microphone array pointing at the three sources, respectively. Impulse responses associated with the global transfer function of beamforming is shown in Fig. 11.10, which are calculated from the impulse responses of mixing filters and beamforming filters using

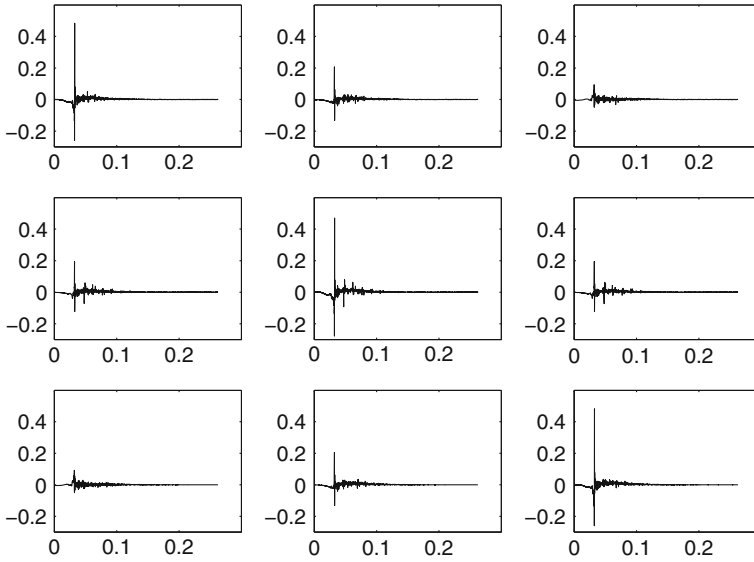


Fig. 11.9 Global impulse responses after beamforming

$$\mathbf{F}(n) = \mathbf{B}(n) * \mathbf{H}(n) \quad (11.17)$$

It can be seen that the diagonal components in Fig. 11.9 are superior to off-diagonal ones. This implies that the target sources are dominant in the outputs. To demonstrate the dereverberation performance of beamforming, the top left panel in Fig. 11.9 is enlarged in Fig. 11.10 and compared with the original impulse response. Obviously, the mixing filter becomes shorter after beamforming, and the reverberation becomes smaller. This indicates that dereverberation is achieved. So far, the two advantages of beamforming, dereverberation and noise reduction, are observed as expected. Thus, the new mixing network (11.17) should be easier to separate than the original mixing network. In this experiment, the average input SIR is -2.8 dB, and the output one, enhanced by beamforming, is 3.3 dB. Applying BSS to the beamformed signals, we get an average output SIR of the combined method of 16.3 dB, a 19.1 dB improvement over the input: 6.1 dB improvement at the beamforming stage, and 13 dB further improvement at the BSS stage.

11.4.2.2 Performance in Reverberant Environments

The performances of the combined method, the BSS-only method and the beamforming-only method, are compared in the simulated environment shown in Fig. 11.8 with different reverberation times. The beamforming-only method is just the first processing stage of the combined method. For the combined method,

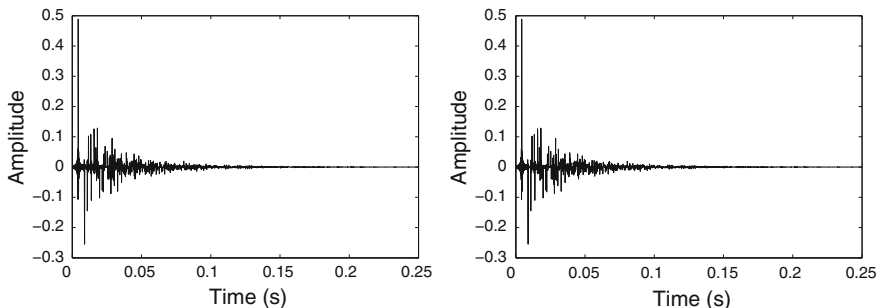


Fig. 11.10 Comparison of the impulse responses before and after beamforming: the left panel is simulated room impulse response for $RT_{60} = 300$ ms; the right panel is the resultant impulse response after beamforming

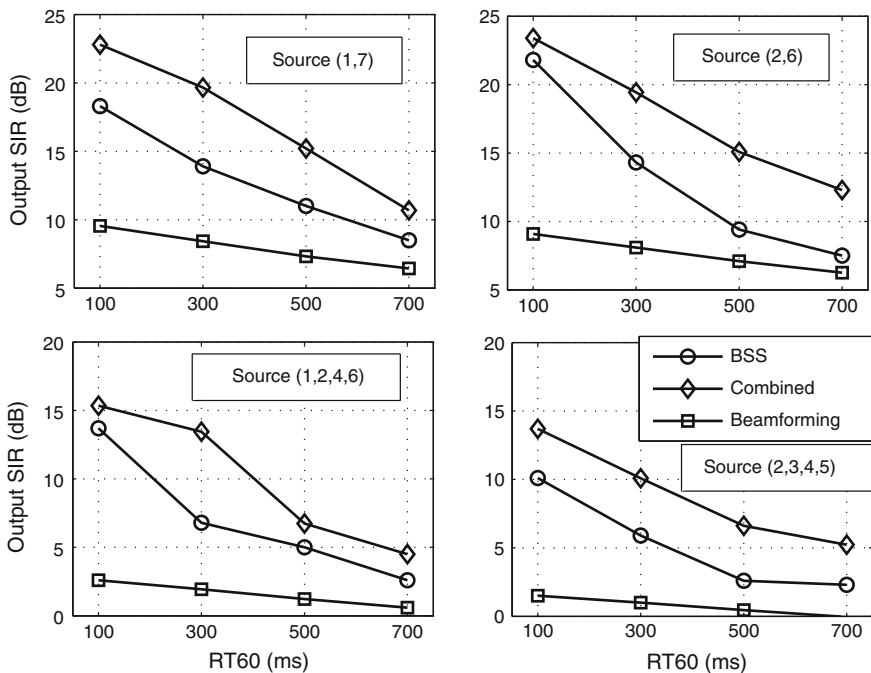


Fig. 11.11 Performance comparison between the combined method, the BSS-only method and the beamforming-only method in different reverberant conditions

a 16-element microphone array with a radius of 0.2 m is used. For the BSS-only method, a linear array consisting of four microphones (inter-space of 6 cm) is used instead of the circular array. Various combinations of source locations are tested (2 sources and 4 sources). The sources are two male speeches and two female speeches of 8s each. RT_{60} ranges from 100 to 700 ms in increments of 200 ms. The average

input SIR does not vary significantly with the reverberation time: it is about 0 dB for two-source cases, and -5 dB for four-source cases. For all three methods, the STFT frame size is set at 2,048. The separation results are shown in Fig. 11.11, with each panel depicting the output SIRs of the three methods for one source combination. It is observed in Fig. 11.11 that for each source configuration, the output SIRs of all methods decrease with increasing reverberation; however, the combined method always outperforms the other two. Beamforming performs worst among the three methods; however, it provides a good preprocessing result, and hence the combined method works better than the BSS-only method.

It is interesting to investigate how big an improvement one can obtain by the use of beamforming preprocessing in different reverberation values. To measure the contribution of this preprocessing, we define the relative improvement of the combined method over the BSS-only method as

$$I_R = \frac{I_c - I_b}{I_b} \times 100\% \quad (11.18)$$

where I is the obtained SIR improvement with the subscripts $(\cdot)_b$ and $(\cdot)_c$ standing for the BSS-only method and the combined method, respectively. We calculate the relative performance improvement for the four separation scenarios listed in Fig. 11.11 and show the average result in Fig. 11.12. As discussed previously, the performance is improved by the combined method for all reverberant conditions. However, it is also observed in Fig. 11.12 that the improvement in low reverberation is not as large as in medium and high reverberation. That is, the use of beamforming in low reverberation is not as beneficial as it would be for high reverberation. The reason is that BSS can work well alone when the circular convolution approximation problem is not evident in low reverberation, and thus the contribution of preprocessing is small. On the other hand, when the circular convolution approximation problem becomes severe in high reverberation, the contribution of preprocessing becomes crucial and hence the separation performance is improved significantly.

Based on the two experiments above, a conclusion can be drawn: With beamforming shortening mixing filters and reducing noise before blind source separation, the combined method performs better than using beamforming or blind source separation alone in highly reverberant environments. A disadvantage of the proposed method is that it requires the knowledge of source locations for beamforming. Generally, the source locations may be estimated with an array sound source localization algorithm [52–54].

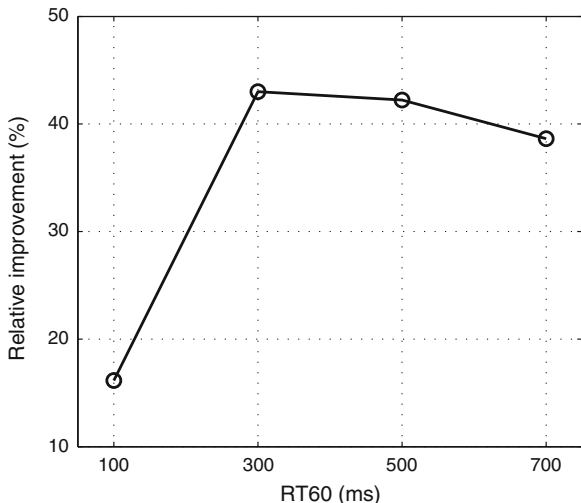
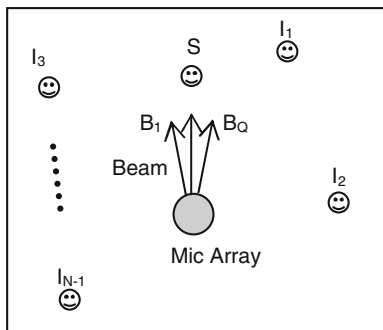


Fig. 11.12 Relative performance improvement of the combined method over the BSS-only method in different reverberant environments

Fig. 11.13 Illustration of the proposed method combining beamforming and blind source separation for target speech extraction



11.5 Target Speech Extraction in a Cocktail-Party Environment

11.5.1 Target Speech Extraction by Combining Beamforming and BSS

In this section, the combined method is extended to a special application of target speech extraction where only the position of the target speaker is known. In real cocktail-party environments, each speaker may move and talk freely. This is very difficult to handle with blind source separation or beamforming alone. Fortunately, it is often in such a case that the target speaker stays in a position or moves slowly and the noisy environment around it is time-varying, *e.g.*, moving interfering speakers and the ambient noise. For this specific situation, a target speech extraction method

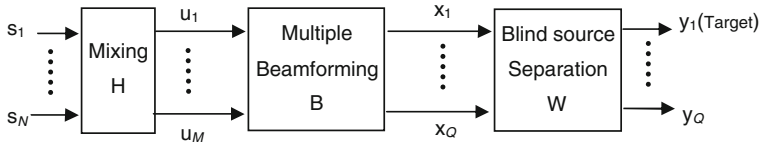


Fig. 11.14 Signal flow of the proposed method combining beamforming and blind source separation for target speech extraction

by combining beamforming and blind source separation is proposed. The principle of the proposed method is illustrated in Fig. 11.13, where the target source S and $N - 1$ interfering sources I_1, \dots, I_{N-1} , are convolutively mixed and observed at an array of M microphones. To extract the target, Q beams ($Q \leq N$) are formed at an area containing it, with a small separation angle between adjacent beams; then the Q beamformed outputs are fed to a blind separation scheme. Using beamforming as a preprocessor for BSS, the target signal becomes dominant in the output of the beamformer and is hence easier to extract. Furthermore, as seen in Fig. 11.13, the beams are pointing at an area containing the target, as opposed to the interfering sources. This is very important for operation under a time-varying condition, because of the following reasons:

- (1) When the target speaker remains in a constant position while others move, it is impractical to know all speakers' positions and steer a beam at each of them;
- (2) There is no need to steer the beams at individual speakers since only the target speaker is of interest;
- (3) The target signal is likely to become dominant in at least one of the beamformed output channels if the beams point at an area containing the target speaker. Thus, it is possible to extract it as an independent source even if the number of beams is less than the sources [55]. This feature is very important for the proper operation of the proposed method;
- (4) A seamless beam area will be formed by several beams with each covering some beamwidth. It is possible to extract the target signal even if it moves slightly inside this area. This feature may improve the robustness of the proposed method; and
- (5) The fact that there are fewer beams than sources reduces the dimensionality of the problem and saves computation.

The signal flow of the proposed method is shown in Fig. 11.14, which is similar to the one shown in Fig. 11.7. The same implementation of beamforming and blind source separation is also employed.

11.5.2 Experimental Results and Analysis

We evaluate the performance of the proposed method in simulated conditions. A typical cocktail-party environment with moving speakers and ambient noises is shown

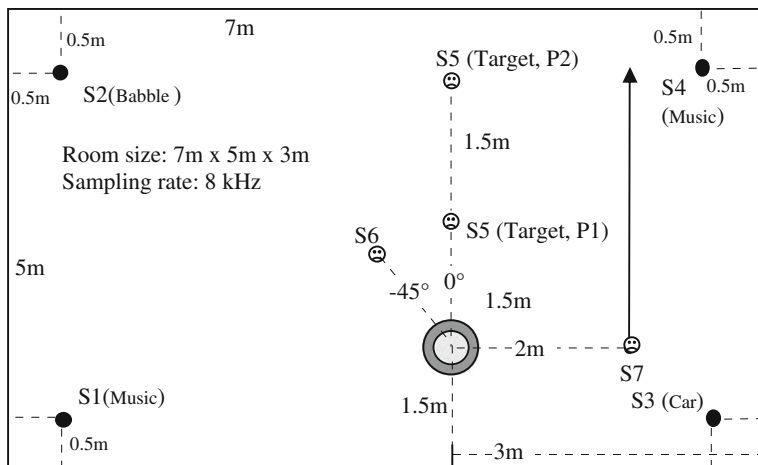


Fig. 11.15 Simulated room environment for target speech extraction

in Fig. 11.15. The room size is $7\text{m} \times 5\text{m} \times 3\text{m}$, and all sources and microphones are 1.5 m high. Four loudspeakers S1–S4 placed near the corners of the room play various interfering sources. Loudspeakers S5, S6, and S7 play speech signals concurrently. S5 and S6 remain in fixed positions, while S7 moves back and forth at a speed of 0.5 m/s. As the target, S5 is placed at either position P1 or P2. S5 simulates a female speaker, while S6 and S7 simulate male speakers. An 8-element circular microphone array with a radius of 0.1 m is placed as shown.

Three beams are formed toward S5, with the separation angle between two adjacent beams being 20° . The room impulse responses are obtained by using the image method, with the reverberation time controlled by varying the absorption coefficient of walls [51]. The test signals last 8s with a sampling rate of 8 kHz. The extraction performance is evaluated in terms of SIR where the signal is the target speech.

With so many speakers present in such a time-varying environment, BSS alone fails to work. Now, we compare the performance of beamforming alone and the proposed method with reverberation RT_{60} of 130 and 300 ms, respectively. The results are given in Table 11.1. As an example, for the close target case (P1) under $RT_{60} = 300$ ms, the input SIR is around -9 dB—the target is almost completely buried in noises and interference. The enhancement by beamforming alone is moderate. On the other hand, the proposed two-stage method improves the SIR by 15.1 dB. In the far target case (P2) of $RT_{60} = 300$ ms, the target signal received at the microphones is much weaker with an input SIR around only -11 dB. The proposed method is still able to extract the target signal with an output SIR of 3.3 dB and a total SIR improvement of 13.5 dB.

For the close target case (P1) under $RT_{60} = 300$ ms, Fig. 11.16 shows the waveforms at various processing stages: sources, microphone signals, beamformer outputs, and finally the BSS outputs. It can be seen that the target signal S5 is totally

Table 11.1 Comparison of beamforming and the proposed method in terms of signal-to-interference ratio (SIR)

Target S5	P1 (close)		P2 (far)	
RT ₆₀	130 ms	300 ms	130 ms	300 ms
Input SIR	-8.2 dB	-9.1 dB	-10.7 dB	-10.8 dB
Beamforming	4.6 dB	0.6 dB	2.5 dB	-2.3 dB
Proposed method	11.9 dB	6.0 dB	9.1 dB	3.3 dB
SIR improvement	20.1 dB	15.1 dB	19.8 dB	13.5 dB

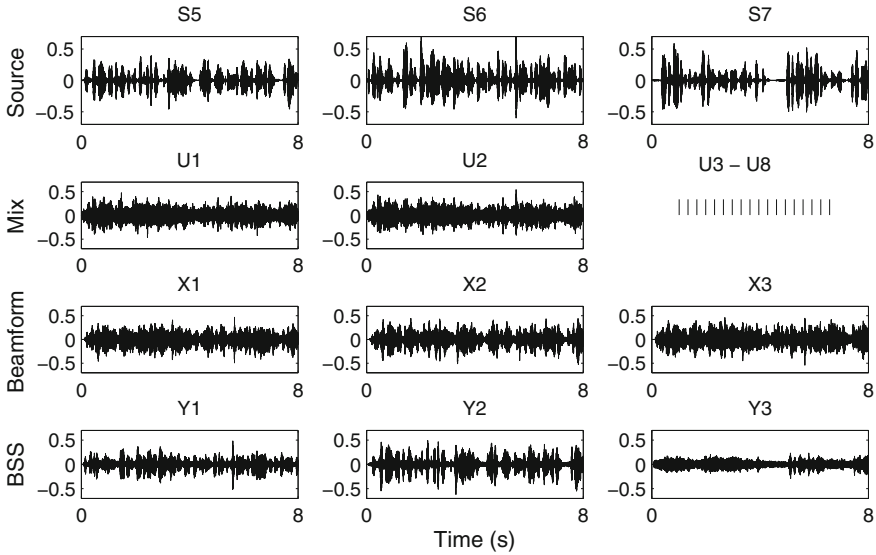


Fig. 11.16 Waveforms at various processing stages

buried in noises and interference in the mixture signals; it is enhanced to a certain degree after beamforming but is still difficult to tell from the background; and after blind source separation, the target signal is clearly exhibited at the channel Y2. In addition, an interference signal (S6) is observed at the output channel Y1, and the noise-like output Y3 is mainly composed of the interfering speech S7 and other noises. The extraction result verifies the validity of the proposed method in noisy cocktail-party environments. Some audio demos can be found at [56].

The good performance of the proposed method in such time-varying environments is due to two reasons. First, fixed beamforming can enhance target signals even in time-varying environments. Second, the spectral components of the target and (moving or static) interfering signals are still independent after beamforming; besides, the target signal becomes dominant in the output of the beamformer. This helps the subsequent blind source separation.

11.6 Conclusions and Prospects

Given the poor performance of blind source separation and beamforming alone in real cocktail-party environments, the chapter proposes a combined method using superdirective beamforming as a preprocessing step of blind source separation. Superdirective beamforming shortens mixing filters and reduces noise for blind source separation, which further reduces the residual interferences. By exploiting both spatial and statistical information, the proposed method can integrate the advantages of beamforming and blind source separation and complement the weakness of them alone. Good results can be obtained when applying the proposed method for speech separation in highly reverberant environments and target speech extraction in dynamic cocktail-party environments.

Although great potentials of the proposed method have been shown, there are still some open problems that need to be addressed. Specifically, beamforming requires the speaker location information to form the beam, but the proposed method in its current form is not capable of identifying the locations, especially with moving speakers. In addition, the separation performance is still limited by the microphone array size, making it a challenge to apply the proposed method to pocket-size applications. These will be investigated in our future research.

Acknowledgments This work is partly supported by the Alexander von Humboldt Foundation.

References

1. Van Veen, B.D., Buckley, K.M.: Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine* **5**, 4–24 (1988)
2. Van Trees, H.L.: *Optimum Array Processing - Part IV of Detection, Estimation, and Modulation Theory*, Chapter 4, pp. 231–331, Wiley-Interscience (2002)
3. Griffiths, L.J., Jim, C.W.: An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Propag.* **30**(1), 27–34 (1982)
4. Cox, H., Zeskind, R.M., Kooij, T.: Practical supergain. *IEEE Trans. Speech Audio Processing*, *ASSP-34*(3), 393–398 (1986)
5. Doclo, S., Moonen, M.: Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics. *IEEE Trans. Signal Process.* **51**(10), 2511–2526 (2003)
6. Doclo, S., Moonen, M.: GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Trans. Signal Process.* **50**(9), 2230–2244 (2002)
7. Doclo, S., Spriet, A., Wouters, J., Moonen, M.: Frequency-domain criterion for the speechdistortion weighted multichannel Wiener filter for robust noise reduction. *Speech Commun.* **49**(7–8), 636–656 (2007)
8. Hyvarinen, A., Karhunen, J., Oja, E.: *Independent Component Analysis*. Wiley, New York (2001)
9. Cardoso, J.: Blind signal separation: statistical principles. *Proc. IEEE* **86**(10), 2009–2025 (1998)
10. Bingham, E., Hyvarinen, A.: A fast fixed-point algorithm for independent component analysis of complex valued signals. *Int. J. Neural Syst.* **10**, 1–8 (2000)
11. Bell, A.J., Sejnowski, T.J.: An information maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7**(6), 1129–1159 (1995)

12. Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. *Adv. Neural Inf. Process. Sys.* **8**, 757–763 (1996)
13. Wang, W., Sanei, S., Chambers, J.A.: Penalty function based joint diagonalisation approach for convolutive blind separation of nonstationary sources. *IEEE Trans. Signal Process.* **53**(5), 1654–1669 (2005)
14. Pedersen, M.S., Larsen, J., Kjems, U., Parra, L.C.: A survey of convolutive blind source separation methods. In: *Handbook on Speech Processing and Speech Communication*, pp. 1–34, Springer (2007)
15. Douglas, S.C., Sun, X.: Convolutive blind separation of speech mixtures using the natural gradient. *Speech Commun.* **39**, 65–78 (2003)
16. Aichner, R., Buchner, H., Yan, F., Kellermann, W.: A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments. *Sig. Process.* **86**(6), 1260–1277 (2006)
17. Douglas, S.C., Gupta, M., Sawada, H., Makino, S.: Spatio-temporal FastICA algorithms for the blind separation of convolutive mixtures. *IEEE Trans. Audio Speech Lang. Process.* **15**(5), 1511–1520 (2007)
18. Sawada, H., Araki, S., Makino, S.: Frequency-domain blind source separation. In: *Blind Speech Separation*, pp. 47–78, Springer (2007)
19. Smaragdis, P.: Blind separation of convolved mixtures in the frequency domain. *Neurocomputing* **22**, 21–34 (1998)
20. Parra, L., Spence, C.: Convolutive blind separation of non-stationary sources. *IEEE Trans. Speech Audio Process.* **8**(3), 320–327 (2000)
21. Mei, T., Mertins, A., Yin, F., Xi, J., Chicharo, J.F.: Blind source separation for convolutive mixtures based on the joint diagonalization of power spectral density matrices. *Sig. Process.* **88**(8), 1990–2007 (2008)
22. Murata, N., Ikeda, S., Ziehe, A.: An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing* **41**(1-4), 1–24 (2001)
23. Sawada, H., Araki, S., Makino, S.: Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS. In: *2007 IEEE International Symposium on Circuits and Systems*, pp. 3247–3250 (2007)
24. Wang, L., Ding, H., Yin, F.: A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures. *IEEE Trans. Audio Speech Lang. Process.* **19**(3), 549–557 (2011)
25. Wang, L., Ding, H., Yin, F.: An improved method for permutation correction in convolutive blind source separation. *Arch. Acoust.* **35**(4), 493–504 (2010)
26. Kim, T., Attias, H.T., Lee, S.Y., Lee, T.W.: Blind source separation exploiting higher-order frequency dependencies. *IEEE Trans. Audio Speech Lang. Process.* **15**(1), 70–79 (2007)
27. Mazur, R., Mertins, A.: An approach for solving the permutation problem of convolutive blind source separation based on statistical signal models. *IEEE Trans. Speech Audio Process.* **17**(1), 117–126 (2009)
28. Serviere, C., Pham, D.T.: Permutation correction in the frequency domain in blind separation of speech mixtures. *EURASIP J. Appl. Sig. Process.* **2006**(1), 177–193 (2006)
29. Ono, N.: Stable and fast update rules for independent vector analysis based on auxiliary function technique. In: *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 189–192, New Paltz (2011)
30. Sawada, H., Araki, S., Makino, S.: Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Trans. Audio Speech Lang. Process.* **19**(3), 516–527 (2011)
31. Saruwatari, H., Kurita, S., Takeda, K.: Blind source separation combining independent component analysis and beamforming. *EURASIP J. Appl. Sig. Process.* **2003**(11), 1135–1146 (2003)
32. Ikram, M.Z., Morgan, D.R.: Permutation inconsistency in blind speech separation: investigation and solutions. *IEEE Trans. Speech Audio Process.* **13**(1), 1–13 (2005)
33. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. Speech Audio Process.* **12**(5), 530–538 (2004)

34. Nesta, F., Svaizer, P., Omologo, M.: Convolutional BSS of short mixtures by ICA recursively regularized across frequencies. *IEEE Trans. Audio Speech Lang. Process.* **19**(3), 624–639 (2011)
35. Nesta, F., Wada, T.S., Juang, B.: Coherent spectral estimation for a robust solution of the permutation problem. In: 2009 IEEE Workshop on Application of Signal Processing to Audio and Acoustics, pp. 1–4, New Paltz, New York (2009)
36. Liu, Q., Wang, W., Jackson, P.: Use of bimodal coherence to resolve the permutation problem in convolutional BSS. *Sig. Process.* **92**(8), 1916–1927 (2012)
37. Araki, S., Mukai, R., Makino, S., Nishikawa, T., Saruwatari, H.: The fundamental limitation of frequency domain blind source separation for convolutional mixtures of speech. *IEEE Trans. Speech Audio Process.* **11**(2), 109–116 (2003)
38. Parra, L., Fancourt, C.: An adaptive beamforming perspective on convolutional blind source separation. In: Davis, G.M. (ed.) *Noise Reduction in Speech Applications*, pp. 361–376. CRC Press (2002)
39. Ikram, M.Z., Morgan, D.R.: A beamforming approach to permutation alignment for multi-channel frequency-domain blind speech separation. In: 2002 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 881–884 (2002)
40. Parra, L.C., Alvino, C.V.: Geometric source separation: Merging convolutional source separation with geometric beamforming. *IEEE Trans. Speech Audio Process.* **10**(6), 352–362 (2002)
41. Saruwatari, H., Kawamura, T., Nishikawa, T., Lee, A., Shikano, K.: Blind source separation based on a fast-convergence algorithm combining ICA and beamforming. *IEEE Trans. Audio Speech Lang. Process.* **14**(2), 666–678 (2006)
42. Gupta, M., Douglas, S.C.: Beamforming initialization and data prewhitening in natural gradient convolutional blind source separation of speech mixtures. In: *Independent Component Analysis and Signal Separation*, vol. 4666, pp. 512–519, Springer, Berlin (2007)
43. Nishikawa, T., Saruwatari, H., Shikano, K.: Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **E86-A**(4), 846–858 (2003)
44. Chen, J., Van Veen, B.D., Hecox, K.E.: External ear transfer function modeling: a beamforming approach. *J. Acoust. Soc. Am.* **92**(4), 1933–1944 (1992)
45. Wang, L., Ding, H., Yin, F.: Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals. *EURASIP J. Audio Speech Music Process.* **2010**, 1–13 (2010). (Article ID 797962)
46. Wang, L., Ding, H., Yin, F.: Target speech extraction in cocktail party by combining beamforming and blind source separation. *IEEE Trans. Audio Speech Lang. Process.* **39**(2), 64–67 (2011)
47. Pan, Q., Aboulnasr, T.: Combined spatial/beamforming and time/frequency processing for blind source separation. In: *European Signal Processing Conference 2005, Antalya, Turkey*, pp. 1–4 (2005)
48. Matsuoka, K., Nakashima, S.: Minimal distortion principle for blind source separation. In: 2001 International Workshop on Independent Component, pp. 722–727 (2001)
49. Ryan, J.G., Goubran, R.A.: Array optimization applied in the near field of a microphone array. *IEEE Trans. Speech Audio Process.* **8**(2), 173–176 (2000)
50. Bouchard, C., Havelock, D.I.: Beamforming with microphone arrays for directional sources. *J. Acoust. Soc. Am.* **125**(4), 2098–2104 (2008)
51. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small room acoustics. *J. Acoust. Soc. Am.* **65**, 943–950 (1979)
52. Silverman, H.F., Yu, Y., Sachar, J.M., Patterson, W.R.: Performance of real-time source-location estimators for a large-aperture microphone array. *IEEE Trans. Speech Audio Process.* **13**(4) (2005)
53. Madhu, N., Martin, R.: A scalable framework for multiple speaker localisation and tracking. In: 2008 International Workshop on Acoustic Echo and Noise Control, Seattle, Washington, pp. 1–4, (2008)

54. Maazaoui, M., Abed-Meraim, K., Grenier, Y.: Blind source separation for robot audition using fixed HRTF beamforming. *EURASIP J. Audio Speech Music Process.* **2012**,1–18 (2012)
55. Sawada, H., Araki, S., Mukai, R., Makino, S.: Blind extraction of dominant target sources using ICA and time-frequency masking. *IEEE Trans. Audio Speech Lang. Process.* **16**(6), 2165–2173 (2006)
56. <https://sites.google.com/site/linwangsig/extraction>