Over-Determined Source Separation and Localization Using Distributed Microphones

Lin Wang, Joshua D. Reiss, and Andrea Cavallaro

***Abstract*—We propose an over-determined source separation and localization method for a set of M microphones distributed around an unknown number, $N < M$, of sources. We reformulate the over-determined acoustic mixing procedure with a new determined mixing model and apply a determined $M \times M$ independent component analysis (ICA) in each frequency bin directly. The reformulated ICA operates without knowing N and also leads to better separation in reverberant scenarios. To solve the challenging permutation ambiguity problem, we first employ a time activity-based clustering approach to cluster the separated frequency components into M channels. We then propose a remixing procedure to detect and merge channels from the same source. The detection is done by analyzing time and frequency activities, spectral likeliness and spatial location. To estimate the spatial location we propose a time-frequency masking-based steered response power algorithm. Simulated and real-data experiments in a very challenging reverberant scenario confirm the effectiveness of the proposed method in obtaining the number of sources, the separated signals, and the location and spatial likelihood of each source.**

***Index Terms*—Blind source separation; over-determined mixture; source localization; permutation alignment**

I. Introduction

Sound source localization and separation are fundamental tasks for acoustic scene analysis [1]. *Source localization* enables visualizing sound directions, while source separation assists auditory information processing, such as speech communication and recognition, by extracting the constituent sources from the mixture signals received by the microphones. *Blind source separation* (BSS) is a well-known technique that can implement these two tasks simultaneously. A widely used approach to blind source separation is independent component analysis (ICA) which estimates a demixing network that recovers the unknown sources from the observed mixture [2]. The demixing network can be interpreted as an inverse of the acoustic mixing network and thus can be used to estimate the source locations if the microphone locations are known [3]. With the increasing flexibility in sensor placement, blind source separation has been investigated intensively in recent decades [4]. Although significant progress has been achieved, several challenges related to blind source separation still remain, including the uncertainty about the number of sound sources, performance degradation in reverberant environments, and acoustical changes due to sound source movement [5].

The BSS problem can be classified as determined (DBSS), under-determined (UBSS) and over-determined (OBSS), corresponding to the number of microphones being equal to, larger than and smaller than the number of sources, respectively [2]. In the most commonly encountered DBSS problem, ICA requires an equal number of sources and microphones to make the mixing network invertible [6]. When the microphones outnumber the sources, subspace-based dimensionality reduction pre-processing is usually applied to get a determined mixture [6]. When the sources outnumber the microphones, the mixing network becomes non-invertible, and nonlinear filtering techniques, such as time-frequency masking, are used instead of ICA [7], [8]. Thus, a prior knowledge of the number of sources is crucial for choosing appropriate BSS algorithms. The performance of BSS degrades significantly in highly reverberant environments, where the acoustic filters are typically very long, thus making the mixing system difficult to invert [9]. Moreover, ICA usually requires the mixing network to remain static for a relatively long period to provide a reasonable estimate of a long demixing filter. This assumption is difficult to fulfil in realistic scenarios where human speakers may turn their heads or move around [10].

Among the various approaches, OBSS has been relatively overlooked compared to DBSS and UBSS, which have attracted the majority of research attention [2], [4]. DBSS and UBSS find wider applicabilities in portable recording devices, where only a limited number of microphones are available. However, due to the small array sizes, most DBSS and UBSS algorithms have limited performance in complex acoustic environments, e.g., with many speakers distributed over a large area. In recent years, distributed microphone networks have become popular [11], [12] as many portable devices, such as smartphones, cameras and laptops, are equipped with wireless communication modules and audio interfaces (e.g., in the case of many people recording the same event with hand-held devices). Utilizing the information from all microphones may lead to better source separation and localization performance. OBSS becomes a common problem in such a network, where the microphones usually outnumber the sources. How to efficiently exploit the redundant information from distributed microphones to tackle the source separation problem, especially in complex acoustic environments, is an important topic.

In this paper we propose an over-determined source separation and localization system for a set of M microphones distributed around an unknown number, $N < M$, of sources. When using distributed arrays, several research

Manuscript received: May 24, 2016

The authors are with the Centre for Intelligent Sensing, Queen Mary University of London, London E1 4NS, U.K. (e-mail: lin.wang@qmul.ac.uk; joshua.reiss@qmul.ac.uk; a.cavallaro@qmul.ac.uk)

This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/K007491/1.

questions arise, such as microphone self-localization [13], [14], asynchronous recording alignment [15], and sampling rate mismatch compensation [16]. The focus in this work is the computation of the spatial filter to separate and localize the sources, hence we assume all the microphones are synchronized and their locations are known. The proposed method employs a frequency-domain BSS framework, where the sources are separated with ICA in each frequency bin and then permutation aligned. By exploiting a sufficient number of microphones, the proposed method can address the challenges of source number uncertainty and reverberation. The novelties are summarized as below.

(a) We formulate the original $M \times N$ mixture as a new $M \times M$ mixing model, which allows us to apply an $M \times M$ ICA directly in each frequency bin without knowing N. The new model considers environmental reverberation and leads to better separation performance in reverberant scenarios.

(b) A fundamental problem with frequency-domain BSS is the unknown and random order of the ICA outputs at each frequency bin, which collapses signal reconstruction in the time domain. Traditional permutation alignment approaches only consider the case with a determined ICA where each source occupies only one output channel. For our case with $M > N$ and N unknown, the employed $M \times M$ ICA results in more ambiguities: each source may randomly occupy an unknown number of output channels. This is a new permutation problem, consisting of both inter-source and intra-source ambiguities. We first solve inter-source ambiguities by clustering separated components with similar time activities into the same channel. We then address the residual intra-source ambiguities with a remixing procedure, which detects and merges channels from the same source. The detection is done by defining and analyzing four measures: time and frequency activities, spectral likeliness and spatial location. The remixing procedure can also estimate the number and locations of the sources.

(c) To estimate the spatial location of each clustered channel, which will be used in the remixing procedure, we propose a time-frequency (T-F) masking-based steered response power (SRP) approach. With the T-F masks that enhance target sources, SRP outperforms traditional approaches based on ICA demixing matrices.

Fig. 1 depicts the block diagram of the proposed method, which consists of three main blocks: blind source separation, source localization and remixing, which are presented in Sections IV, V and VI, respectively.

II. Background

A. Steered Response Power for Source Localization

SRP is a steered-beamforming based algorithm which is suitable for source localization with distributed microphones [17], [18]. This approach steers the beamformer over a predefined set of spatial points and searches for peaks in the steered response power (the output signal). The simplest (delay-and-sum) beamformer computes the propagation delays from the source position to each microphone and compensates for these delays in order to coherently sum the signals arising from the source position. More sophisticated beamformers filter the microphone signals in addition to delaying them [19]. The steered response power can be interpreted as a spatial likelihood map [20] whose peaks correspond to the locations of the sources. SRP is robust to noise and can localize the source reliably in a single-source scenario. However, when multiple sources are present, the peaks may not provide reliable information for all sources, especially when they overlap in time [21], [22]. Multi-path propagation due to reverberation causes additional peaks, which also degrade the localization performance. Another drawback is the high computational cost, which increases with the number of microphones and also with the size of the search area. Some fast searching schemes have been proposed [23], [24].

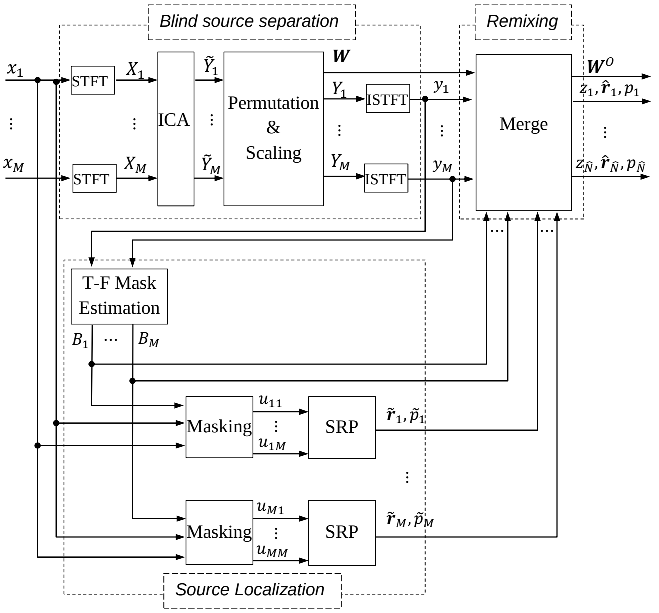


Fig. 1. Block diagram of the proposed over-determined source separation and localization method.

SRP-PHAT, a phase transformed version of SRP, is another popular approach for multi-microphone source localization [17], [25], [26]. Instead of using steered beamformers, SRP-PHAT calculates the spatial likelihood map by summing the phase-transformed generalized cross-correlations for all possible pairs of the set of microphones. The phase-transform may increase the robustness to reverberation [27]; however, the pair-wise evaluation of the steered response power makes the computational cost increase exponentially with the number of microphones. This could be a bottleneck when using SRP-PHAT with a large number of microphones. Both SRP and SRP-PHAT can be seen as phase-based approaches since they rely on delaying the microphone signals appropriately, a procedure which is equivalent to modifying signal phase in the frequency domain.

B. ICA and Permutation Alignment

In the context of BSS, ICA is a well-known tool for the separation of linear and instantaneous mixed signals captured by multiple sensors [6]. ICA adaptively estimates a demixing matrix by maximizing the statistical independence of the output signals. Based on the definition of independence measure, various ICA algorithms have been proposed, including joint approximate diagonalization of eigenmatrices (JADE) [28], Infomax [29] and fixed-point FastICA [6]. ICA usually relies on two prerequisites: the independence between source signals and the invertibility of the mixing matrix. The former condition is satisfied with most audio signals such as speech and music. The latter condition typically requires the mixing matrix to be square, i.e., a mixture with an equal number of sources and observations.

For many real-world problems, the signals undergo a convolutive mixing due to reverberation. Various attempts have been made to solve the convolutive BSS problem [30], including frequency-domain approaches [2]. By transforming the mixture to the frequency domain with the short-time Fourier transform (STFT), convolution in the time domain translates to linear mixing in the frequency domain. Subsequently, ICA can be performed on each frequency bin. However, since ICA is indeterminate of source permutation, further post-processing methods are necessary to align the permutations in each frequency bin.

Most permutation alignment algorithms were proposed under the framework of DBSS, where only inter-source ambiguities occur. Three strategies exist to tackle the permutation ambiguity problem. The *inter-frequency dependency-based* strategy exploits the time structure of separated signal amplitudes or speech activities [33]. This time structure shows high correlation between neighboring bins for the same source. Various approaches, including clustering-based and region-wise permutation alignment schemes, exploit such inter-frequency dependency [8], [31]–[33]. The *location-based* strategy exploits the spatial information since contributions from the same source are likely to come from the same direction [34]–[37]. This approach typically works well only in low-reverberant scenarios and may suffer from spatial aliasing ambiguities if the microphones are far apart [38], [42]. The *joint optimization* strategy, e.g., independent vector analysis, directly incorporates the inter-frequency dependency measure into ICA so that the permutation ambiguity can be minimized by joint optimization across all the frequency bins [39], [40]. However, this approach may get easily stuck in local optima by simultaneously optimizing many parameters across the whole frequency band. Among the three strategies, the inter-frequency dependency-based approach performs most robustly in reverberant scenarios, especially for speech signals [33]. Relying only on signal amplitudes, this approach works independently of the microphone placement and is robust to spatial aliasing problems.

Since the demixing matrix obtained by ICA can be interpreted as an inverse of the mixing matrix, several ICA-based source localization approaches have been proposed [3], [41]–[43]. By blindly identifying the acoustic transfer functions for each separated source, these approaches are suitable for multi-source environments. However, since ICA is performed individually in each frequency bin, ICA-based localization relies on successfully addressing the permutation ambiguity problem.

TABLE I
Comparison of over-determined blind source separation algorithms. ($\boldsymbol{G}$: Microphone location; $\boldsymbol{R}$: Source location; N: Number of sources)

Strategy	**Approach**	**Prior Knowledge**			**Reference**
		$\boldsymbol{G}$	$\boldsymbol{R}$	N	
dimensionality reduction	subspace	no	no	yes	[44]–[49]
	fixed beamforming	yes	yes	yes	[9], [50]
subset selection	geometry-based	yes	no	yes	[51]
	separation-based	no	no	yes	[52], [53]
separation and remixing		yes	no	no	proposed

C. Over-determined BSS

Table I summarizes the state-of-the-art of OBSS algorithms, which exploit the redundant information available when using more sensors than sources. The *dimensionality reduction* strategy applies dimensionality reduction before separation so that the numbers of input observations and sources become equal. Two approaches have been proposed for dimensionality reduction. The first approach employs subspace-based preprocessing, e.g., principle component analysis (PCA), to extract an equal number of components, and subsequently performs ICA [44]–[49]. After PCA, the signal-to-noise ratio in the retained components is generally higher than in any individual sensor and the mixing matrix is usually better conditioned. Assuming the spatial location of each source to be known, the second approach applies a set of fixed beamformers, each pointing at one source, before separation [9], [50]. The fixed beamformer can reduce noise and reverberation for each source, making the subsequent separation task easier. The *subset selection* strategy selects a subset of microphones from the whole microphone set. The selection can be done based on geometric information, e.g., using wide microphone spacing for doing separation at low frequencies and narrow spacing at higher frequencies [51]. The selection can also be done by trying possible microphone subsets and choosing the one with the best outputs [52], [53].

All the OBSS algorithms discussed above are based on DBSS and require prior knowledge of the number of sources, and some also require to know the locations of the sources. In contrast, our proposed method does not need to know either the number of sources or their locations.

III. Problem Formulation

Consider M microphones and N sources, with $M \geq N$, randomly distributed in a reverberant acoustic scenario. Microphones and sources are physically static. The M microphones are synchronously sampled and their locations $\boldsymbol{G} = [\boldsymbol{g}_1, \cdots, \boldsymbol{g}_M]_{3\times M}$ are known. The number of sources N and their locations $\boldsymbol{R} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_N]_{3\times N}$ are both unknown. The signals received at the microphones, $\boldsymbol{x}(n) = [x_1(n), \cdots, x_M(n)]^{\mathrm{T}}$, are expressed as

$$\boldsymbol{x}(n) = \boldsymbol{H}(n) * \boldsymbol{s}(n) = \sum_{n'=0}^{L_h-1} \boldsymbol{H}(n')\boldsymbol{s}(n-n'), \qquad (1)$$

where n is the time index, $\boldsymbol{s}(n) = [s_1(n), \cdots, s_N(n)]^{\mathrm{T}}$ is the source signal vector, $\boldsymbol{H}(n)$ is a sequence of $M \times N$ matrices containing the impulse responses of mixing channels, L_h is the length of the impulse response, the operator '$*$' denotes the convolution between two sequences of matrices, and the superscript $(\cdot)^{\mathrm{T}}$ denotes transpose.

The task is to estimate the number of sources N, the source locations $\boldsymbol{R}$, and individual sources $\boldsymbol{s}(n)$, given the microphone recordings $\boldsymbol{x}(n)$ and microphone locations $\boldsymbol{G}$.

IV. $M \times M$ Determined Source Separation

In this section we show that, by reformulating the $M \times N$ over-determined acoustic mixing with a new $M \times M$ determined model, it is possible to apply an $M \times M$ source separation directly, which leads to better separation in reverberant scenarios than an $N \times N$ separation.

A. $M \times M$ ICA

Using STFT, the time-domain convolution (1) is converted to instantaneous mixing in the frequency domain:

$$\boldsymbol{X}(k,l) = \boldsymbol{H}(k)\boldsymbol{S}(k,l), \tag{2}$$

where k and l are frequency and frame indices, respectively; $\boldsymbol{H}_{M\times N}(k)$ is the Fourier transform of $\boldsymbol{H}(n)$; $\boldsymbol{X}_{M\times 1}(k,l)$ and $\boldsymbol{S}_{N\times 1}(k,l)$ are the STFTs of $\boldsymbol{x}(n)$ and $\boldsymbol{s}(n)$, respectively. Usually, a subspace-based dimensionality reduction procedure [49] is performed in each frequency bin so that an $N \times N$ ICA can be applied to separate the N sources.

In (1), the sound of the j-th source received by the i-th microphone, $x_{ij}(n)$, equals the convolution result between the original source $s_j(n)$ and the impulse response $h_{ij}(n)$. Based on the image-source theory [59], $x_{ij}(n)$ can also be approximated as a sum of contributions from multiple image sounds emitting from different spatial locations:

$$x_{ij}(n) = h_{ij}(n) * s_j(n) \approx \sum_{r=1}^{R_j} \tilde{h}_{ijr}(n) * \tilde{s}_{jr}(n), \tag{3}$$

where s_j is decomposed into R_j image sounds: $\tilde{s}_{j1}, \cdots, \tilde{s}_{jR_j}$, and $\tilde{h}_{ijr}(n)$ denotes the impulse response between the image sound $\tilde{s}_{jr}$ and the microphone i. Usually, $\tilde{h}_{ijr}(n)$ is shorter than $h_{ij}(n)$. The STFT counterpart of (3) can be written as

$$X_{ij}(k,l) = H_{ij}(k)S_j(k,l) \approx \sum_{r=1}^{R_j} \tilde{H}_{ijr}(k)\tilde{S}_{jr}(k,l), \tag{4}$$

where X_{ij}, S_j, $\tilde{S}_{jr}$, H_{ij} and $\tilde{H}_{ijr}$ are the STFTs of x_{ij}, s_j and $\tilde{s}_{jr}$, h_{ij} and h_{ijr}, respectively.

With the new model (4), the original $M \times N$ over-determined mixing system can be approximated as an $M \times M$ determined mixing system:

$$\begin{aligned}\boldsymbol{X}(k,l) &\approx \widetilde{\boldsymbol{H}}(k)\widetilde{\boldsymbol{S}}(k,l)\\ &= \begin{bmatrix} \tilde{H}_{111}(k) & \cdots & \tilde{H}_{1NR_N}(k) \\ \vdots & \ddots & \vdots \\ \tilde{H}_{M11}(k) & \cdots & \tilde{H}_{MNR_N}(k) \end{bmatrix}_{M\times M} \begin{bmatrix} \tilde{S}_{11}(k,l) \\ \vdots \\ \tilde{S}_{1R_1}(k,l) \\ \vdots \\ \tilde{S}_{N1}(k,l) \\ \vdots \\ \tilde{S}_{NR_N}(k,l) \end{bmatrix}_{M\times 1}, \end{aligned} \tag{5}$$

where $\widetilde{\boldsymbol{S}}$ represents M image sounds from the N sources. Depending on the acoustic scenario, R_j varies with the source j, with $R_j \geqslant 1$ and $\sum_{j=1}^{N} R_j = M$. The new mixing model (5) allows us to apply an $M \times M$ ICA directly to $\boldsymbol{X}(k,l)$. Since the M image sounds usually originate from different spatial locations, the square mixing matrix $\widetilde{\boldsymbol{H}}$ is invertible.

There are two benefits of using such an $M \times M$ mixing model. First, the mixing filter $h_{ijr}(n)$ is usually shorter than the original mixing filter $h_{ij}(n)$, thus making the new $M \times M$ system easier to invert than the original mixing system. Second, the frequency-domain ICA can be interpreted as a set of null-beamformers which extracts a target source by suppressing the sources from other directions [55]. The $M \times M$ ICA allows suppressing at most $M-1$ interferences for each target source. These $M-1$ interferences may include the $N-1$ sources and their associated reverberant image sounds, which are difficult to suppress with a normal $N \times N$ ICA. Thus, the separation performance will improve when M is increased. A possible drawback of using an $M \times M$ ICA is that the size of the mixing matrix also grows with M, thus requiring more data to estimate the demixing matrix.

We choose a widely used ICA algorithm, Infomax, to estimate the demixing matrix using the iteration [29], [33]

$$\begin{cases} \widetilde{\boldsymbol{Y}}(k,l) = \widetilde{\boldsymbol{W}}(k)\boldsymbol{X}(k,l) \\ \widetilde{\boldsymbol{W}}(k) = \widetilde{\boldsymbol{W}}(k) + \eta\left(\boldsymbol{I} - \mathrm{E}\{\Phi(\widetilde{\boldsymbol{Y}}(k,l))\widetilde{\boldsymbol{Y}}^{\mathrm{H}}(k,l)\}\right)\widetilde{\boldsymbol{W}}(k) \end{cases} \tag{6}$$

where the superscript $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose, η is a step-size parameter, $\boldsymbol{I}$ is the identity matrix, $\Phi(\cdot)$ is a nonlinear function, $\mathrm{E}\{\cdot\}$ is the expectation operator, and $\widetilde{\boldsymbol{Y}}(k,l) = [\widetilde{Y}_1(k,l), \cdots, \widetilde{Y}_M(k,l)]^{\mathrm{T}}$ is the separated signal vector.

The demixing matrix can recover the source signals up to scaling and permutation ambiguities [33]:

$$\widetilde{\boldsymbol{Y}}(k,l) = \boldsymbol{\Lambda}(k)\boldsymbol{D}(k)\widetilde{\boldsymbol{S}}(k,l), \tag{7}$$

where $\boldsymbol{D}(k)$ is a permutation matrix and $\boldsymbol{\Lambda}(k)$ a scaling matrix at frequency index k. In our case with $M > N$, each source may occupy one or more ICA outputs. This leads to a more challenging permutation alignment problem, where the permutation ambiguities come not only from different sources (e.g., $\widetilde{S}_{11}$ and $\widetilde{S}_{21}$) but also from the same source (e.g., $\widetilde{S}_{11}$ and $\widetilde{S}_{12}$). This challenge is made even harder by the fact that the number of sources and the number of constituent image sounds of each source are all unknown.

We propose to solve the permutation ambiguity problem in two stages, with a clustering-based permutation alignment approach (Sec. IV-B) addressing inter-source ambiguities and a remixing procedure (Sec. VI) addressing intra-source ambiguities.

B. Clustering-based Permutation Alignment

The time activity of speech signals typically shows strong dependency among frequency components from the same source, and an ability to discriminate among frequency components from different sources. This discriminability usually increases with the duration of the signal. This feature can be exploited to align the permutation of the components from different frequencies. Assuming the number of sources N to be known, a clustering procedure has been proposed to cluster the separated frequency components with similar time-activities into the same group [8], [33], [54]. We use this approach to align the permutation of the $M \times M$ ICA outputs, directly assuming the number of sources to be M.

Given the demixing matrix $\widetilde{\boldsymbol{W}}(k)$, the mixing matrix can be estimated as $\boldsymbol{A}(k) = \boldsymbol{W}^{-1}(k) = [\boldsymbol{a}_1(k), \cdots, \boldsymbol{a}_N(k)]$ with $\boldsymbol{a}_i(k)$ being an $M \times 1$ vector describing the path from the i-th component $\widetilde{Y}_i$ to M microphones. The time activity sequence of $\widetilde{Y}_i$ at frequency k is defined as [54]

$$\boldsymbol{v}_i^k(l) = \frac{\left\| \boldsymbol{a}_i(k)\widetilde{Y}_i(k,l) \right\|^2}{\sum_{j=1}^{M} \left\| \boldsymbol{a}_j(k)\widetilde{Y}_j(k,l) \right\|^2}, \tag{8}$$

where $\|\cdot\|$ denotes the norm-2 operation. The inter-frequency dependency between two time activity sequences $\boldsymbol{v}_i^{k_1}$ and $\boldsymbol{v}_j^{k_2}$ (corresponding, respectively, to the i-th separated signal at frequency k_1 and the j-th separated signal at frequency k_2) is measured by their correlation coefficient

$$\rho(\boldsymbol{v}_i^{k_1}, \boldsymbol{v}_j^{k_2}) = \tfrac{\gamma_{ij}(k_1,k_2) - \mu_i(k_1)\mu_j(k_2)}{\sigma_i(k_1)\sigma_j(k_2)}, \tag{9}$$

where $\gamma_{ij}(k_1, k_2) = \mathrm{E}\{\boldsymbol{v}_i^{k_1}\boldsymbol{v}_j^{k_2}\}$, $\mu_i(k) = \mathrm{E}\{\boldsymbol{v}_i^k\}$, $\sigma_i(k) = \sqrt{\mathrm{E}\{(\boldsymbol{v}_i^k)^2\} - \mu_i^2(k)}$ are the correlation, mean, and standard deviation, respectively.

Let us represent the new permutation with respect to the original outputs as $\Pi = \{1, \cdots, M\} \to \{1, \cdots, M\}$ with $\Pi(m)$ being the new order of the m-th output [33]. The clustering is implemented as an expectation maximization (EM) procedure, which maximizes the correlation coefficient between the clustered time-activity sequences and the corresponding centroids. The iterative EM procedure is expressed as

$$\begin{cases} \boldsymbol{c}_m = \dfrac{1}{K}\displaystyle\sum_{k=1}^{K} \boldsymbol{v}_m^k, \quad m = 1, \cdots, M \\ \Pi_k = \arg\max\limits_{\Pi} \displaystyle\sum_{m=1}^{M} \left\{ \rho(\boldsymbol{v}_i^k, \boldsymbol{c}_m) \big|_{i=\Pi(m)} \right\}, \quad \forall k \end{cases} \tag{10}$$

where K is the number of frequency bins; $\boldsymbol{c}_1, \cdots, \boldsymbol{c}_M$ denote the estimated centroids; and Π_k is the permutation at frequency k. Accordingly, the demixing matrix is permutated as

$$\widehat{\boldsymbol{W}}(k) \xleftarrow{\Pi_k} \widetilde{\boldsymbol{W}}(k). \tag{11}$$

We address the scaling ambiguity problem, i.e., the unknown scale of the ICA outputs, by minimal distortion principle-based back projection [56]:

$$\boldsymbol{W}(k) = \mathrm{diag}\left(\widehat{\boldsymbol{W}}^{-1}(k)\right) \cdot \widehat{\boldsymbol{W}}(k), \tag{12}$$

where $(\cdot)^{-1}$ denotes the inversion of a matrix and $\mathrm{diag}(\cdot)$ retains only the diagonal components of a matrix.

Finally, we compute the separated signal as

$$\boldsymbol{Y}(k,l) = \boldsymbol{W}(k)\boldsymbol{X}(k,l), \tag{13}$$

with $\boldsymbol{Y}(k,l) = [Y_1(k,l), \cdots, Y_M(k,l)]^{\mathrm{T}}$ and $\boldsymbol{y}(n) = [y_1(n), \cdots, y_M(n)]^{\mathrm{T}}$ being the inverse STFT of $\boldsymbol{Y}(k,l)$. We refer to $\boldsymbol{y}(n)$ as the *DBSS output*.

C. Discussion

Clustering based on time activity sequences has been successfully applied to permutation alignment with $M = N$ [8], [33]. However, in cases with unknown N and $M > N$, a new challenge arises: the clustering algorithm produces more (M) clusters than (N) sources. In addition, it is observed that the time activity of a speech signal may vary slightly across frequencies, e.g., between high and low frequencies [33]. This leads to two types of clustering results. On the one hand, the clustering algorithm tends to allocate components from different sources to different clusters, thus solving the *inter-source ambiguity* problem. On the other hand, with $M > N$, the clustering algorithm tends to allocate components (with slightly different time activities) from the same source to different clusters. Consequently, the obtained M clusters can be virtually divided into N source sets. Each set may consist of several clusters which all correspond to the same source. The number of sets and the association between clusters and source sets are unknown. Therefore, the *intra-source ambiguity* problem still remains unsolved.

We illustrate the intermediate processing results of the determined $M \times M$ source separation with a realistic acoustic scenario shown in Fig. 2. This scenario is included in an existing dataset [26]. In a room of size 8m×6m×3m and with reverberation time $\mathrm{T}_{60} = 0.45$ s, 10 speakers, dividing into 3 groups, are chatting simultaneously. The speeches of the 10 speakers are recorded by 171 distributed microphones and also by 10 close-talk microphones attached to the speakers. The database provides a 120 s long real recording as well as the locations of all microphones and speakers. For convenience of comparison, we generate the same scenario using the closely recorded speech from each speaker and simulated room impulse responses by the image-source method [59]. We randomly choose 20 microphones (Fig. 2, $M = 20$ and $N = 10$) with signal length 25 s and sampling rate 8 kHz. We use signal-to-interference ratio (SIR), as defined in (36), to measure the source separation performance.

The considered acoustic scenario is very challenging since the input SIRs of the sources at microphones can be as low as -15 dB (Fig. 11). After DBSS, the SIR of each output is shown in Fig. 3(a), where each column of the map represents the SIRs of each source in all the outputs, while each row represents the SIRs of all the sources in each output. The row-wise SIRs

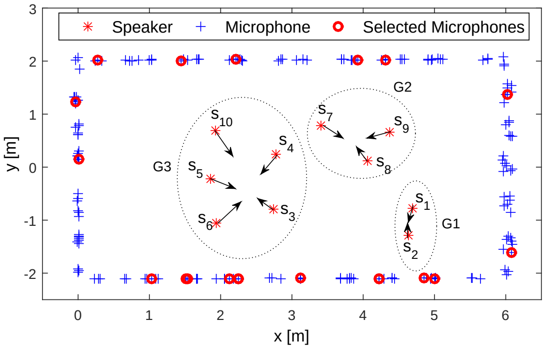


Fig. 2. Configuration of the acoustic scenario. Ten speakers are divided into three groups. An arrow denotes the orientation of the face of each speaker. An example of randomly selected $M = 20$ microphones is indicated with red circles.

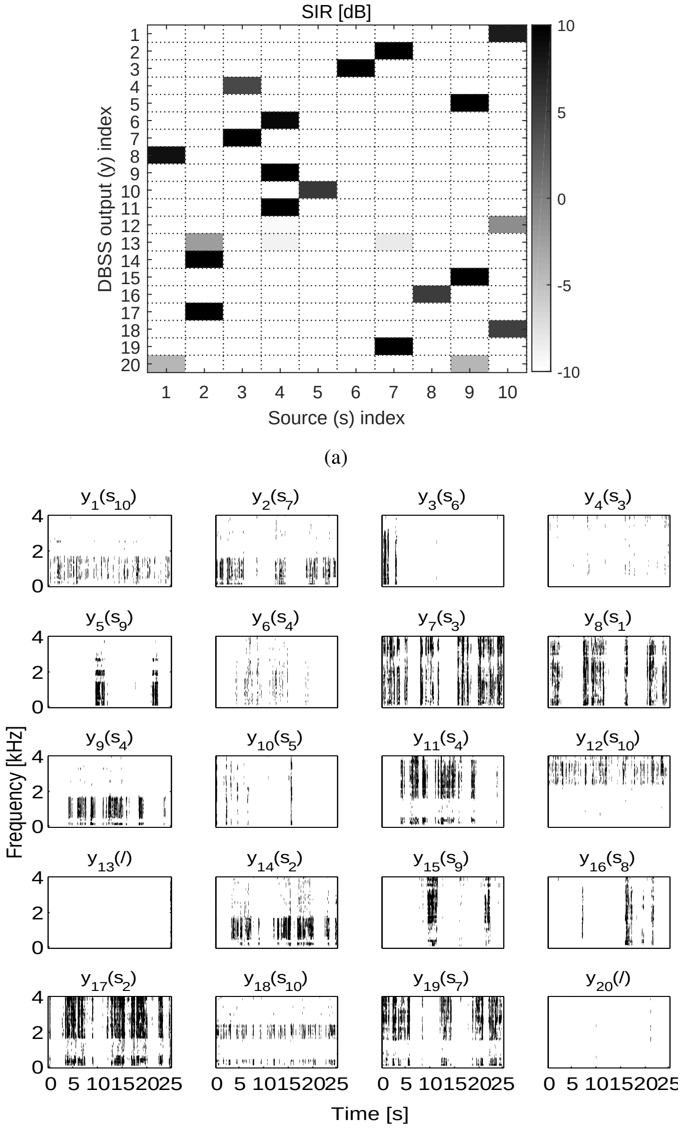
show that in each DBSS output only one source is dominant. This indicates that ICA performs well for separation in each frequency bin and the inter-source ambiguities are well solved by the clustering-based permutation alignment. Exceptions are y_{13} and y_{20} that are both identified as noise later. The column-wise SIRs show that each source can be dominant in one or several outputs. This indicates that the intra-source ambiguity problem is still unsolved.

To better show the intra-source ambiguity problem, we depict in Fig. 3(b) the speech activity of each DBSS output in the form of a binary map, where the speech activity (8) is set to 1 when it is larger than 0.5, and is set to 0 otherwise. In Fig. 3(b) the title of each panel $y_i(s_j)$ represents that the output y_i is associated with the source s_j. This association is inferred from Fig. 3(a), based on the highest SIR in each row. As observed in Fig. 3(b), some sources appear only in one channel, e.g., $y_3(s_6)$ and $y_8(s_1)$, whereas other sources appear in several channels. These channels can be different frequency bands of a source, e.g., $y_{14}(s_2)$ and $y_{17}(s_2)$. These channels can also be different image sounds of a source, e.g., $y_6(s_4)$ represents a reverberant version of s_4, while $y_9(s_4)$ and $y_{11}(s_4)$ constitute the full-band direct sound of s_4. Some channels contain noise only and present no speech activity, e.g., y_{13} and y_{20}.

The above observations provide valuable information to design the remixing algorithm (Sec. VI), which aims to find the association between the sources and DBSS outputs.

V. T-F Masking based Localization

In this section, we estimate the spatial locations of the DBSS outputs, which will be used in the subsequent remixing procedure. Existing approaches perform localization based on the acoustic transfer functions of each separated component, namely steering vectors, estimated from the ICA demixing matrix [42], [43]. These approaches typically rely on a high direct-to-reverberant ratio in the microphone signals so that the phase of the steering vector varies approximately linearly with frequency. Moreover, the intra-source ambiguity may randomly distribute the frequency bin-wise steering vectors of a single source to different channels, degrading the localization performance significantly.



(a)

(b)

Fig. 3. Source separation results by the DBSS algorithm. (a) SIR of each source in each DBSS output. In each row only one source is dominant. In each column one source may dominate several output channels. (b) Binary map of speech activity of each DBSS output. The title of each panel, $y_i(s_j)$, represents the association between y_i and s_j.

SRP is a steered-beamforming based multi-microphone localization algorithm, which is robust to noise but has degraded performance in reverberant and multi-source scenarios. Combining source separation and SRP, we propose a new time-frequency (T-F) masking-based SRP algorithm for source localization.

For each DBSS output y_m, a T-F mask is estimated which indicates the dominance of y_m in each T-F bin in the microphone signal. After applying this T-F mask to all microphone signals, an SRP algorithm is employed to estimate the location of y_m. There are three benefits when using this T-F mask. First, the T-F mask can effectively improve the SIR of the target source by suppressing interferences. Second, applying a T-F mask will not change the phase information embedded in the microphone signals. This allows us to apply the phase-based SRP algorithm for localization. Finally, by directly working on the microphone signals, this approach is robust to intra-source ambiguities. The algorithm is summarized below, using y_m as an example.

The (k, l)-th element of the T-F mask B_m is estimated as

$$B_m(k,l) = \begin{cases} 1, & \lambda_m(k,l) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where $\lambda_m(k,l) = \frac{|Y_m(k,l)|}{\sum_{i=1}^{M}|Y_i(k,l)|}$ indicates the proportion of $Y_m(k,l)$ among all the outputs $Y_1(k,l),\cdots,Y_M(k,l)$. Applying B_m to the M microphone signals, the masked signal at the i-th microphone is

$$U_{mi}(k,l) = B_m(k,l)X_i(k,l), \quad i = 1,\cdots,M. \tag{15}$$

Applying the inverse STFT to $U_{mi}(k,l)$, we obtain the time-domain signal $u_{mi}(n)$.

The location of y_m is estimated by applying the SRP algorithm to the M masked signals $u_{m1}(n),\cdots,u_{mM}(n)$ in the time-domain. Here we use the simplest delay-and-sum beamformer for the SRP algorithm. The algorithm calculates an SRP map in a pre-defined spatial space $\mathbb{R}$, where the SRP for a candidate position $\boldsymbol{r} \in \mathbb{R}$ is defined as

$$\text{SRP}_m(\boldsymbol{r}) = \sum_{n=1}^{L_s} \tilde{u}_m^2(n), \tag{16}$$

where L_s is the signal length in samples, $\tilde{u}_m(n) = \sum_{i=1}^{M} u_{mi}(n - \tau_i(\boldsymbol{r}))$, with $\tau_i(\boldsymbol{r})$ being the delay of the i-th channel with respect to the first channel for the location $\boldsymbol{r}$. The location of y_m is estimated by detecting the highest peak in the map, i.e.,

$$\tilde{\boldsymbol{r}}_m = \arg\max_{\boldsymbol{r}\in\mathbb{R}} \text{SRP}_m(\boldsymbol{r}). \tag{17}$$

We also calculate the spatial likelihood map, which indicates the confidence that y_m originates from a certain location. After removing a floor value with $\widetilde{\text{SRP}}_m(\boldsymbol{r}) = \text{SRP}_m(\boldsymbol{r}) - \text{SRP}_{\text{floor}}$, the spatial likelihood (SL) is defined as

$$\text{SL}_m(\boldsymbol{r}) = \frac{\sum_{\boldsymbol{r}\in\mathbb{N}} \widetilde{\text{SRP}}_m(\boldsymbol{r})}{\sum_{\boldsymbol{r}\in\mathbb{R}} \widetilde{\text{SRP}}_m(\boldsymbol{r})}, \tag{18}$$

where the floor $\text{SRP}_{\text{floor}}$ is the mean SRP value in the space $\mathbb{R}$, and $\mathbb{N}$ denotes a predefined neighbourhood surrounding $\boldsymbol{r}$. Accordingly, the spatial likelihood of $\tilde{\boldsymbol{r}}_m$ is defined as

$$p_m = \text{SL}_m(\tilde{\boldsymbol{r}}_m). \tag{19}$$

Spatial aliasing may occur at frequencies where the corresponding sound wavelength is shorter than twice the inter-microphone distance. Spatial aliasing introduces phase ambiguities, which lead to ghost locations when localizing a sound source [57]. The aliasing is usually severe at high frequencies and less pronounced at lower frequencies. After clustering-based permutation alignment, each DBSS output contains the broadband signal of a single source. Source localization with a broadband signal increases the robustness to spatial aliasing, i.e., a true location will present a higher spatial likelihood than the spurious locations. In some exceptional cases when the DBSS output only contains the high-frequency band of the source signal, spatial aliasing can still be observed with ghost locations, thus degrading the localization performance. This problem will be addressed in the remixing stage by merging different frequency bands of the same source together.

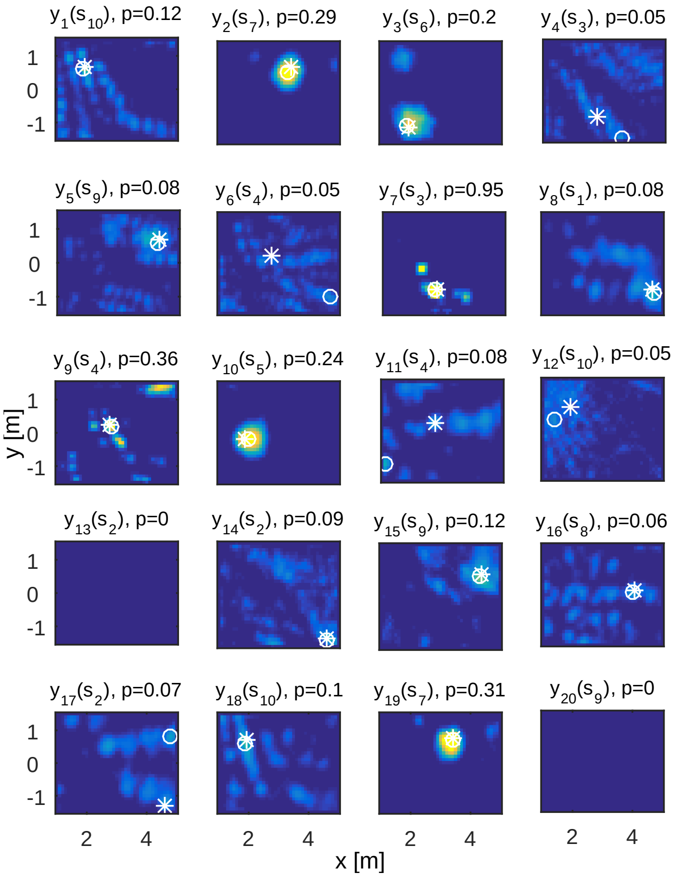


Fig. 4. Spatial likelihood map for each DBSS output. In the title of each panel $y_i(s_j)$ denotes the association between y_i and s_j, whereas p denotes the spatial likelihood of the estimated location. The true and estimated locations are indicated with white stars and white circles, respectively.

As an example, Fig. 4 depicts the spatial likelihood map, the estimated and true locations of each DBSS output obtained in Sec. IV-C. As observed in Fig. 4, the spatial likelihood can provide information regarding estimation accuracy. A high spatial likelihood (e.g. $p \geq 0.2$) is usually associated with an evident peak in the spatial likelihood map and correct localization, e.g., $y_2(s_7)$ and $y_7(s_3)$. A low spatial likelihood (e.g. $p \leq 0.05$) is usually associated with multiple peaks in the spatial likelihood map, leading to inaccurate or even wrong localization, e.g., $y_4(s_3)$ and $y_6(s_4)$. However, it may happen that a low spatial likelihood is associated with a correct localization, e.g., $y_8(s_1)$. In addition, y_{13} and y_{20} contain only noise with $p = 0$. The two channels $y_{12}(s_{10})$ and $y_{17}(s_2)$ contain signals only in the high-frequency band. The wrong localization at these two channels is possibly caused by the ghost locations that are introduced by the spatial aliasing at high frequencies.

VI. Remixing

The intra-source ambiguity problem is dealt with using a remixing procedure, which merges any two DBSS output channels if they are detected to be from the same source. As shown in Fig. 3(b), the relationship between two channels can be classified into three types as below.

- `Inter`: from different sources. Two channels have different activities along time and frequency.
- `Intra-B`: from different frequency *bands* of the same source. Two channels have similar activities along time but different activities along frequency.
- `Intra-I`: from different sound *images* of the same source. Two channels are both full-band signals but represent different image sounds of a source, e.g., direct and reverberant sounds. The two channels may have different time activities, as measured by the dominance in each time-frequency bin. However, the direct and reverberant sounds have similar spectral contents.

In addition to spectral information, all channels from the same source are supposed to come from the same spatial location. However, as shown in Fig. 4, the estimated location may deviate from the true value if the channel contains a reverberant sound.

Based on the above analysis, we define time and frequency activity measures to detect channels from different frequency bands of the same source, and define a spectral likeliness measure to detect channels containing direct and reverberant image sounds of the same source. For channels from the same source but not detected by the above spectral measures, we further as complement define a spatial distance measure.

A. Remixing Measures

1) Speech Activity: The time- and frequency-activity sequences of y_m are calculated from the T-F mask B_m as

$$a_m(l) = \sum_{k=1}^{K} B_m(k,l); \quad b_m(k) = \sum_{l=1}^{L} B_m(k,l), \tag{20}$$

where K and L denote the numbers of frequency bins and time frames, respectively. The time-activity correlation coefficient between two channels y_{m_1} and y_{m_2} is defined as

$$R_a(m_1,m_2) = \frac{\sum_{l=1}^{L} a_{m_1}(l) a_{m_2}(l)}{\sqrt{\sum_{l=1}^{L}(a_{m_1}(l))^2}\sqrt{\sum_{l=1}^{L}(a_{m_2}(l))^2}}. \tag{21}$$

Similarly, the frequency-activity correlation coefficient between y_{m_1} and y_{m_2} is defined as

$$R_b(m_1,m_2) = \frac{\sum_{k=1}^{K} b_{m_1}(k) b_{m_2}(k)}{\sqrt{\sum_{k=1}^{K}(b_{m_1}(k))^2}\sqrt{\sum_{k=1}^{K}(b_{m_2}(k))^2}}. \tag{22}$$

A high value of $R_a(m_1,m_2)$ indicates that y_{m_1} and y_{m_2} tend to have similar time activities. A low value of R_b indicates that the two signals tend to have different frequency activities and thus occupy different frequency bands.

Combining both time and frequency activities, we define a global speech activity correlation coefficient measure

$$R_{ab}(m_1,m_2) = R_a(m_1,m_2) - R_b(m_1,m_2). \tag{23}$$

Two channels y_{m_1} and y_{m_2} are detected to be from the same source (`Intra-B`) if this measure satisfies

$$R_{ab}(m_1,m_2) > T_{ab}, \tag{24}$$

where T_{ab} is a predefined threshold.

2) Spectral Likeliness: For two channels $Y_{m_1}(k,l)$ and $Y_{m_2}(k,l)$ in the time-frequency domain, the likeliness of their spectral magnitudes is defined as

$$R_s(m_1,m_2) = \frac{\sum\limits_{k=1}^{K}\sum\limits_{l=1}^{L} |Y_{m_1}(k,l) Y_{m_2}(k,l)|}{\sqrt{\sum\limits_{k=1}^{K}\sum\limits_{l=1}^{L}|Y_{m_1}(k,l)|^2}\sqrt{\sum\limits_{k=1}^{K}\sum\limits_{l=1}^{L}|Y_{m_2}(k,l)|^2}}. \tag{25}$$

A high $R_s(m_1,m_2)$ indicates that y_{m_1} and y_{m_2} tend to have similar spectral contents. Thus, y_{m_1} and y_{m_2} are detected to be from the same source (`Intra-I`) if their spectral likeliness satisfies

$$R_s(m_1,m_2) > T_s, \tag{26}$$

where T_s is a predefined threshold.

3) Spatial Location: The spatial distance between y_{m_1} and y_{m_2} is defined as

$$D(m_1,m_2) = ||\tilde{\boldsymbol{r}}_{m_1} - \tilde{\boldsymbol{r}}_{m_2}||, \tag{27}$$

where the locations $\tilde{\boldsymbol{r}}_{m_1}$ and $\tilde{\boldsymbol{r}}_{m_2}$ are estimated with (17). The two channels are regarded as originated from the same source if their spatial distance is sufficiently small, i.e.,

$$D(m_1,m_2) < T_d, \tag{28}$$

where T_d is a predefined threshold.

4) Outlier Measure: A channel y_m is detected to be uncorrelated or diffuse noise if its spatial likelihood p_m is sufficiently small, i.e.,

$$p_m < T_p, \tag{29}$$

where T_p is a predefined threshold.

Because of the non-Gaussianity of speech signals, y_m can be detected as noise if its kurtosis [6] is sufficiently small, i.e.,

$$\text{kurt}(y_m) < T_k, \tag{30}$$

where T_k is a predefined threshold.

B. Remixing Procedure

Given the above measures, the remixing procedure consists of five stages (Fig. 5). Each stage generates a new set of channels by either removing or merging channels inherited from its preceding stage. The first stage removes noise outliers based on the kurtosis measure (30). The second, third and fourth stages merge channels based on the speech activity measure (24), the spectral likeliness measure (26), and the distance measure (28), respectively. The fifth stage removes the residual outliers, which satisfy either the spatial likelihood measure (29) or the kurtosis measure (30).

To elaborate on the remixing procedure, we denote each newly merged channel as a source set, i.e., $\mathbb{S}_m = \left\{\left(\boldsymbol{I}_m, \hat{\boldsymbol{r}}_m, q_m, z_m(n), \bar{B}_m(k,l)\right)\right\}$, which contains five

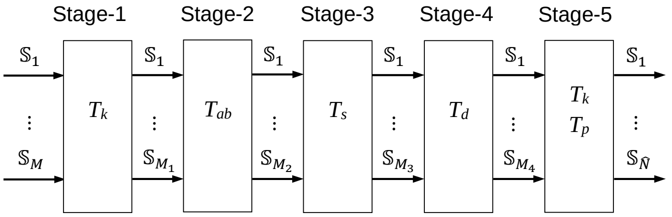


Fig. 5. Block diagram of the remixing procedure. M_1-M_4 denote the resulting number of source sets in Stages 1-4.

TABLE II
PARAMETERS USED IN THE REMIXING PROCEDURE.

Parameter	T_{ab}	T_s	T_p	T_k	T_d
Value	0.4	0.5	0.03	5	0.25 m

elements denoting the constituent channels, spatial location, spatial likelihood, sound signal and binary mask, respectively.

Before remixing, we initialize M source sets as $\boldsymbol{I}_m = \{m\}$, $q_m = p_m$, $\hat{\boldsymbol{r}}_m = \tilde{\boldsymbol{r}}_m$, $z_m(n) = y_m(n)$, and $\bar{B}_m(k,l) = B_m(k,l)$, for $m = 1, \cdots, M$. In each stage, if m_G sets $\{\mathbb{S}_{m_1}, \cdots, \mathbb{S}_{m_G}\}$ are detected to be from the same source, we merge them into the first set $\mathbb{S}_{m_1}$ and then remove others. The merging procedure is performed as below.

$$\begin{cases} \boldsymbol{I}_{m_1} = \boldsymbol{I}_{m_1} \cup \boldsymbol{I}_{m_2} \cup \cdots \cup \boldsymbol{I}_{m_G}, \\ q_{m_1} = q_{\bar{m}},\ \hat{\boldsymbol{r}}_{m_1} = \hat{\boldsymbol{r}}_{\bar{m}}, \\ z_{m_1}(n) = z_{m_1}(n) + z_{m_2}(n) + \cdots + z_{m_G}(n), \\ \bar{B}_{m_1}(k,l) = \bar{B}_{m_1}(k,l) | \bar{B}_{m_2}(k,l) | \cdots | \bar{B}_{m_G}(k,l), \end{cases} \tag{31}$$

where $\bar{m} = \arg\max_{m \in \{m_1, \cdots, m_G\}} q_m$ represents the index of the channel with the highest spatial likelihood, the operator '$\cup$' denotes union of two sets, and '$|$' the binary 'OR' operator.

Finally, we obtain $\hat{N}$ source sets $\left(\mathbb{S}_1, \cdots, \mathbb{S}_{\hat{N}}\right)$. We denote the corresponding output as $\boldsymbol{z}(n) = [z_1(n), \cdots, z_{\hat{N}}(n)]^{\mathrm{T}}$, the locations $\hat{\boldsymbol{r}}_1, \cdots, \hat{\boldsymbol{r}}_{\hat{N}}$, and the spatial likelihoods $q_1, \cdots, q_{\hat{N}}$. The OBSS demixing matrix $\boldsymbol{W}^O$ is

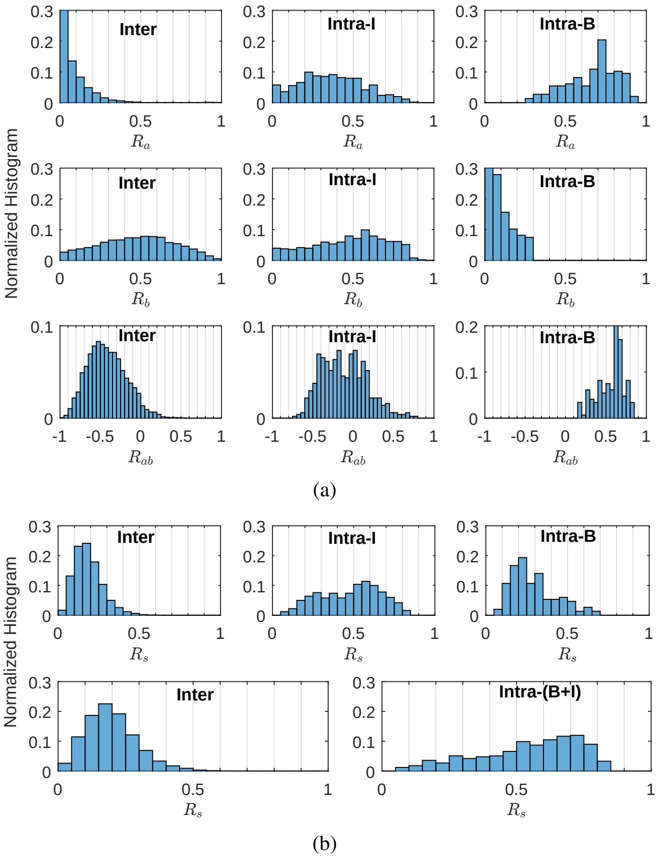
$$\boldsymbol{W}^O_m(k) = \sum_{m' \in \boldsymbol{I}_m} \boldsymbol{W}_{m'}(k), \quad m = 1, \cdots, \hat{N}; \tag{32}$$

where $\boldsymbol{W}_{m'}$ denotes the m'-th row of $\boldsymbol{W}$, and the same for $\boldsymbol{W}^O_m$. We refer to $\boldsymbol{z}(n)$ as the *OBSS output.*

C. Parameter Selection

The remixing procedure uses 5 thresholds: T_{ab}, T_s, T_d, T_p and T_k (Table II). Among them, T_k and T_d can be easily determined. Since the kurtosis of babble noise is around 5 [58], we choose $T_k = 5$ to distinguish speech from noise. With prior knowledge of the acoustic environment, we choose the minimum distance between speakers as $T_d = 0.25$ m.

The thresholds T_{ab} and T_s play important roles in the remixing procedure since they determine whether two channels should be merged. We examine the distribution of the parameters R_{ab} and R_s in different classes (`Inter`, `Intra-B` and `Intra-I`) in order to choose optimal threshold values. We generate 64 testing cases with the simulated dataset in Sec. IV-C, including different numbers of microphones M, signal lengths L_t, and realizations, where $M \in \{10, 20, 30, 40\}$, $L_t \in \{10, 20, 30, 40\}$ s, and each (M, L_t) configuration has four realizations with the start time of source signal set as the $\{5,\ 25,\ 45,\ 65\}$-th second in the original speech. For each pair of DBSS output channels, we hand-labelled its classification and calculated the speech activity measures R_a, R_b and R_{ab}, and the spectral measure R_s. By repeating this procedure across all testing cases, we obtain the distribution of these parameters in the three classes.



(a)

(b)

Fig. 6. Distribution of the parameters in three classes: `Inter`, `Intra-I`, `Intra-B`. (a) Speech activity measures R_a, R_b and R_{ab}. (b) Spectral likeliness measure R_s.

Fig. 6(a) depicts the normalized histogram of R_a, R_b and R_{ab}. In the first row, the time-activity measure R_a behaves differently for `Inter` and `Intra-B`. For `Inter`, R_a is distributed between 0 and 0.5 and tends to show small values, being close to 0. For `Intra-B`, R_a is distributed between 0.2 and 1 and tends to show large values, centroiding around 0.7. For `Intra-I`, R_a is distributed almost uniformly between 0 and 0.8. In the second row of Fig. 6(a), the frequency-activity measure R_b behaves uniquely for `Intra-B`, where it is distributed between 0 and 0.3 and tends to show small values, being close to 0. For both `Inter` and `Intra-I`, R_b is distributed almost uniformly between 0 and 1. The different behaviours of R_a and R_b across three classes allow us to use them jointly, i.e., $R_{ab} = R_a - R_b$, to distinguish between `Inter` and `Intra-B`. As shown in the third row of Fig. 6(a), R_{ab} has clear difference in these two classes. For `Inter`, R_{ab} is distributed between -1 and 0.5 and tends to show small values, centroiding around -0.5. For `Intra-B`, R_{ab} is distributed between 0 and 1 and tends to show large values, centroiding around 0.5. For `Intra-I`, R_{ab} is distributed almost uniformly between -0.5 and 0.5.

Fig. 6(b) depicts the normalized histogram of R_s. In the first row, R_s tends to show low values for both `Inter` and `Intra-B`. A pair of channels from `Intra-B` occupy different frequency bands of the same source and thus present low spectral likeliness. To better show the discriminability of R_s for intra- and inter-source channels, we merge all the pairs in `Intra-B` into full-band signals. We then reclassify all the channel pairs as `Inter` and `Intra-(B+I)` and recalculate R_s. The second row of Fig. 6(b) depicts the distribution of the new R_s. For `Inter`, R_s is distributed between 0 and 0.5 and tends to show small values, centroiding around 0.2. For `Intra-(B+I)`, R_s is distributed between 0 and 0.9, centroiding around 0.7. Comparing with the first row, R_s in the second row shows clearer difference between intra- and inter-source channels. This also explains why we merge channels based on the speech activity measure R_{ab} at first (Stage-2) and then the spectral likeliness measure R_s (Stage-3).

When performing channel merging in Stage-2 and Stage-3, two types of errors may occur. The first error, namely miss detection, denotes two channels which are supposed to be merged but not detected. This error typically leads to incompletely reconstructed source signals and also an overestimation of the number of sources. The second error, namely false alarm, occurs when two channels from different sources are erroneously merged. This error is usually irreversible as two different sources are mixed. We therefore give higher importance to the task of minimizing false alarms by choosing

$$T_{ab} = 0.4, \quad T_s = 0.5, \tag{33}$$

as given in Table II. From Fig. 6(a), some channel pairs in `Intra-I`, with $R_{ab} > T_{ab}$, may be detected as `Intra-B` and merged in Stage-2. This will not affect the final remixing result. However, miss detection occurs when some channel pairs in `Intra-B` have $R_{ab} < T_{ab}$ (Stage-2), or when some channel pairs in `Intra-(B+I)` have $R_s < T_s$ (Stage-3). The influence of miss detections can be reduced by using an additional location-based measure, T_d, in Stage-4.

The threshold T_p is used to remove an outlier channel based on the spatial likelihood. We thus examine the distribution of the spatial likelihood for two classes of channels, with correct and incorrect peak locations, respectively, in the spatial likelihood maps. Fig. 7 depicts the normalized histograms of the spatial likelihood in all testing cases. The spatial likelihood shows evident difference between the two classes. For correct localization, around half of the spatial likelihoods are larger than 0.2, while the other half is distributed between 0.03 and 0.2. For incorrect localization, around 40% of the spatial likelihoods are smaller than 0.01, while the others are distributed mainly between 0.02 and 0.1. In order to maximize localization accuracy while removing outliers we choose

$$T_p = 0.03, \tag{34}$$

as given in Table II. Since T_p takes effect only in the last stage, it does not have a large effect on the remixing.

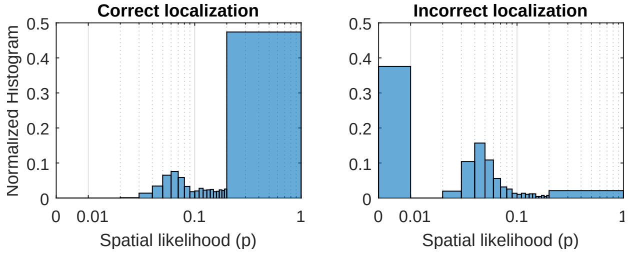


Fig. 7. Distribution of the spatial likelihood, p, in two classes: correct and incorrect localization.

D. Validation

To demonstrate the effectiveness of the remixing procedure (Fig. 5) as well as the chosen thresholds (Table II), we calculate the detection error in each stage. Let us denote the number of wrongly merged channels in each stage as N_w, and the number of undetected channels as $N_u = |\tilde{M} - N|$, where $\tilde{M}$ is the resulting number of source sets in this stage, and N, as already defined, is the number of sources. We define the miss detection ratio, the false alarm ratio, and the total error ratio, respectively, as

$$E_{\text{miss}} = \frac{N_u}{M}, \; E_{\text{false}} = \frac{N_w}{M}, \; E_{\text{total}} = E_{\text{miss}} + E_{\text{false}}. \tag{35}$$

Fig. 8 depicts the average error ratios for all testing cases in each stage. With more microphones than sources, we obtain a large E_{miss} around 0.5 and $E_{\text{false}} = 0$ at the input stage. The noise outlier removal at Stage-1 reduces E_{miss} without changing E_{false}. The channel merging at Stages 2-4 reduces E_{miss} significantly but also increases E_{false}. Specifically, spectrum-based Stage-2 and Stage-3 only introduce minor false alarms, while location-based Stage-4 introduces evident false alarms, mainly due to inaccurate localization in some DBSS output channels. Removing the outliers generated during channel merging, Stage-5 can reduce both E_{miss} and E_{false}. As a result, the total error E_{total} decreases monotonically across all processing stages, finally reaching a value below 0.1. Since the merging error mainly arises from Stage-4, the remixing performance could be improved if a better localization algorithm was employed. The observations made in Fig. 8 confirm the effectiveness of the remixing procedure and the chosen thresholds.

We investigate the robustness of the remixing procedure in scenarios with a varying number of sources N and reverberation time R_{60}. In the first scenario, we use R_{60} = 450 ms but a varying number of sources $N \in \{5, 7, 8, 10\}$. Referring to Fig. 2, we only consider the speakers from the groups G1 and G2 for $N = 5$; the groups G1 and G3 for $N = 7$; the groups G2 and G3 for $N = 8$; and all three groups for $N = 10$. In the second scenario, we use $N = 10$ but varying $\text{R}_{60} \in \{200, 450, 700\}$ ms. As in Sec. VI-C, we generate 64 testing cases for each configuration. Fig. 9 depicts the total detection error E_{total} in each remixing stage for different scenarios. In Fig. 9(a), E_{total} at the input stage decreases with N, because the number of undetected channels N_u drops as N is increased. It is also observed that, for each N, E_{total} decreases monotonically across all processing stages.

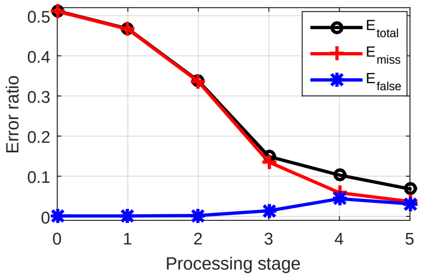


Fig. 8. Detection errors, in terms of miss detection ratio, E_{miss}, false alarm ratio, E_{false}, and total error ratio, E_{total}, in each stage of the remixing procedure. Stages 0-5 denote, respectively, input, outlier removal (T_k), merge (T_{ab}), merge (T_s), merge (T_d), and outlier removal (T_k, T_p).

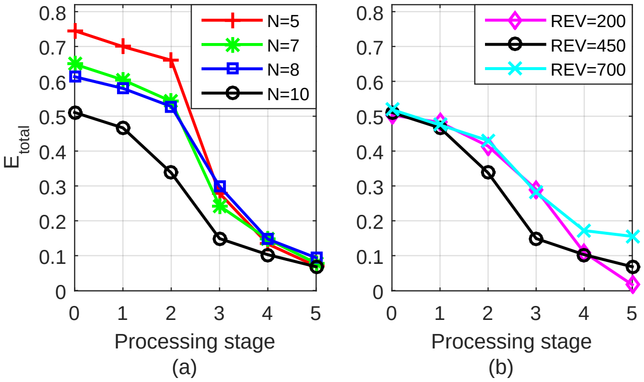


Fig. 9. Detection error E_{total} in each stage of the remixing procedure for (a) different number of sources and (b) different reverberation times. Stages 0-5 denote, respectively, input, outlier removal (T_k), merge (T_{ab}), merge (T_s), merge (T_d), and outlier removal (T_k, T_p).

The final E_{total} are all close to 0.1 for different N. In Fig. 9(b), E_{total} also decreases in all processing stages for each R$_{60}$. The final E_{total} increases with the reverberation time, being close to 0, 0.1 and 0.15 for R$_{60}$ = 200 ms, 450 ms and 700 ms, respectively. The above observations made in Fig. 9 confirm the robustness of the remixing procedure and thresholds in various scenarios.

Finally, we apply the remixing procedure to the example of Sec. IV-C. After channel merging and outlier removal, we obtain 10 source sets, equalling the true number of sources. The SIR of each source in each OBSS output and the locations of sources are depicted in Fig. 10(a) and (b), respectively. As observed in Fig. 10(a), in each row and column of the SIR map only one source is dominant, showing that both the inter-source and intra-source permutation ambiguities have been solved. As shown in Fig. 10(b), the estimated locations are consistent with the true locations. Finally, the SIR of each source in the microphones and in the OBSS outputs is shown in Fig. 11. It is possible to notice that the OBSS algorithm can significantly improve the SIR (about 20 dB) of each source.

VII. Computational Complexity

The proposed algorithm mainly consists of three blocks: DBSS, source localization and remixing. The source localization block dominates the whole computation of the algorithm and its importance grows with M. For each one of the M DBSS outputs, an SRP algorithm is applied which exhaustively search in the candidate space. The computational complexity of the source localization is proportional to M^2L_t. The computational complexity of the DBSS block is typically proportional to ML_t. The computational complexity of the remixing block is dominated by the spectral likeliness calculation, which is applied to each pair of the DBSS outputs. Thus the computational complexity of the remixing block is proportional to M^2L_t.

We run Matlab code of the proposed algorithm on an Intel CPU i7@3.2 GHz with 16 GB RAM, using the same simulated data (20 s) from Sec. IV-C. Fig. 12 depicts the computation time of each block for a varying number of microphones.

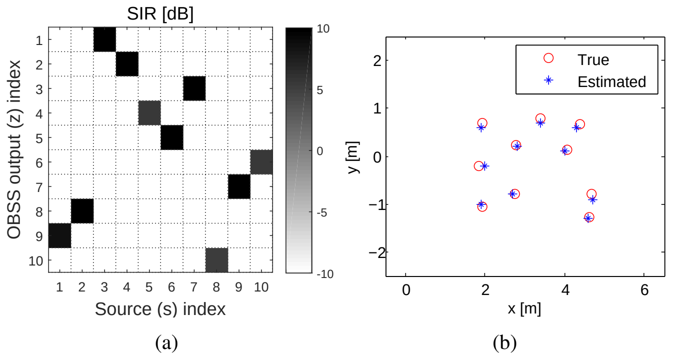


Fig. 10. Source separation and localization results by the proposed algorithm. (a) SIR of each source in each OBSS output. Each row and column is dominated by only one source. (b) True and estimated source locations.

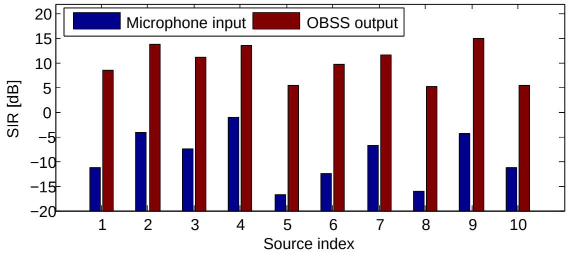


Fig. 11. SIR of each source at microphone inputs and OBSS outputs. The OBSS algorithm can improve the SIR of each source by about 20 dB.

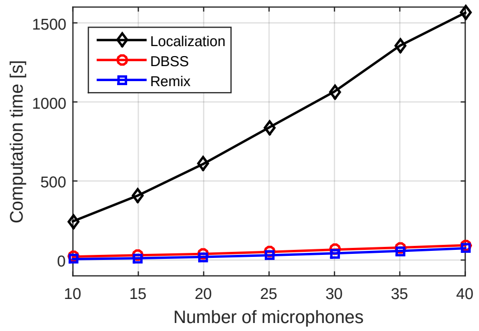


Fig. 12. Computation time of the proposed OBSS algorithm (DBSS, source localization and remixing) for a data-length of 20 s and a varying number of microphones.

VIII. Experimental Results

We evaluate the performance of the proposed algorithm in terms of source separation and source localization.

A. Experiment Setup

We use the same simulated dataset as in Sec. IV-C to evaluate the performance of the proposed OBSS algorithm with a varying number of microphones M and length of signal L_t, where M is increased from 10 to 40 with an interval of 5 and L_t is chosen from {6, 10, 15, 20, 30, 40} s. For each (M, L_t) configuration, we implement four realizations where the start time of the source signal is set as the {5, 25, 45, 65}-th second in the original recording.

For source separation, we compare the performance of the proposed OBSS algorithm (`Proposed`) with four existing algorithms: the subspace-based noise reduction algorithm (`SS`) [46], the subspace-based dimensionality reduction followed by determined BSS (`SS+BSS`) [46], the fixed delay-and-sum beamforming which assumes the locations of all sources are known (`BF`) [50], and the delay-and-sum beamforming followed by determined BSS (`BF+BSS`) [50].

For source localization, we first compare the accuracy of the peak of the spatial likelihood map obtained by the steering vector-based [42] and the proposed T-F masking-based approaches. We then compare the multi-source localization performance of the proposed OBSS algorithm (`Proposed`) with two existing algorithms including `SRP` and `SRP-PHAT`. To adapt these two algorithms for multi-source localization, we localize the most dominant source in each time segment (1 s long) and merge the results across the whole signal duration.

The thresholds used in `Proposed` are listed in Table II. The blind source separation and clustering-based permutation alignment algorithms are implemented as in [33]. We choose 2048 as the STFT frame size at sampling rate 8 kHz. With approximate knowledge of the acoustic environment, we define the search space $\mathbb{R}$ in (17) to be a box of size 5m×3m×2m enclosing all the sources, and define the neighbourhood $\mathbb{N}$ in (18) to be a sphere with radius 0.2 m. The search step is set as 0.1 m in all three dimensions.

B. Performance Measures

We evaluate the *source separation performance* with signal-to-interference ratio (SIR). Denote the mixing system $\boldsymbol{H}(n) = [\boldsymbol{H}_1(n), \cdots, \boldsymbol{H}_N(n)]$ and the demixing system $\boldsymbol{W}(n) = \left[\boldsymbol{W}_1^{\mathrm{T}}(n), \cdots, \boldsymbol{W}_N^{\mathrm{T}}(n)\right]^{\mathrm{T}}$, the SIR of the j-th source s_j in the i-th output y_i is defined as

$$\mathrm{SIR}_{ij} = 10 \log_{10} \frac{\sum_{n=1}^{L_s} (y_{ij}(n))^2}{\sum_{j' \neq j} \sum_{n=1}^{L_s} (y_{ij'}(n))^2}, \tag{36}$$

where $y_{ij}(n) = \boldsymbol{W}_i(n) * \boldsymbol{H}_j(n) * s_j(n)$. The SIR of s_j among all the outputs is defined as the maximum value:

$$\mathrm{SIR}_j = \max_i \{\mathrm{SIR}_{ij}\}. \tag{37}$$

Representing the input and output SIR of s_j as $\mathrm{SIR}_j^{\text{in}}$ and $\mathrm{SIR}_j^{\text{out}}$, respectively, the global SIR improvement by $\boldsymbol{W}$ is defined as

$$\mathrm{SIR}_{\text{imp}} = \frac{1}{N} \sum_{j=1}^{N} \left(\mathrm{SIR}_j^{\text{out}} - \mathrm{SIR}_j^{\text{in}} \right). \tag{38}$$

Given SIR_{ij}, the index of the source associated with the i-th output y_i is estimated as

$$J_i = \arg\max_j \{\mathrm{SIR}_{ij}\}. \tag{39}$$

The SIR regarding the OBSS filter $\boldsymbol{W}^O$ can be calculated in a similar way.

To evaluate the *accuracy of the spatial likelihood peaks* calculated by different approaches, we define an objective measure of peak error rate. Suppose the location of the spatial likelihood peak of the DBSS output y_i is $\tilde{\boldsymbol{r}}_i$ (17) while the true location is $\boldsymbol{r}_{J_i}$ (39), the peak can be seen as a correct estimation if $||\tilde{\boldsymbol{r}}_i - \boldsymbol{r}_{J_i}|| < T_d$. For the M peaks of the M DBSS outputs the peak error rate is defined as

$$R_{\text{pe}} = \frac{M_e}{M}, \tag{40}$$

where M_e denotes the number of incorrect peaks.

We evaluate the *multi-source localization performance* with recall rate and precision rate. Let the location estimated for the i-th OBSS output z_i be $\hat{\boldsymbol{r}}_i$ (17) and the true location be $\boldsymbol{r}_{J_i}$ (39). The localization is regarded as correct if $||\hat{\boldsymbol{r}}_i - \boldsymbol{r}_{J_i}|| < T_d$. Suppose that the true number of sources is N, the estimated number of sources is $\hat{N}$, and the number of correct estimation is $\hat{N}_c$. Then the recall rate and the precision rate are

$$R_{\text{recall}} = \frac{\hat{N}_c}{N}, \quad R_{\text{prec}} = \frac{\hat{N}_c}{\hat{N}}. \tag{41}$$

C. Validation on Simulated Data

We evaluate the *source separation performance* in terms of SIR improvement by the considered algorithms (`Proposed`, `SS`, `SS+BSS`, `BF`, `BF+BSS`) for various M and L_t. For each (M, L_t) configuration the SIR improvement results are averaged across four realizations. The SIR performance of `Proposed` for various M and L_t is depicted in Fig. 13. The performance increases with L_t, though the improvement slows as L_t increases. ICA typically requires enough data to estimate the demixing matrix, leading to improved separation performance with increased signal length. In most cases, the performance of `Proposed` improves significantly as M is increased from 10 to 25. However, when $M \geq 25$, the performance improves slowly with M when $L_t > 15$ s, and even decreases with M when $L_t \leq 15$ s. This is because the demixing matrix estimation task becomes more challenging. The performance of ICA degrades if there are not enough data, leading to decreased SIR in case of large M but small L_t.

Fig. 14 depicts the SIR performance of the five considered algorithms for various M and L_t. `Proposed` performs best when $L_t > 10$ s. `Proposed` performs similarly to `BF+BSS` when $L_t = 10$ s, but performs worse than `BF+BSS` when $L_t = 6$ s. The rank of the other four algorithms can be `BF+BSS>SS+BSS>BF>SS`. The performance of `SS` and `BF` is improved significantly when combined with BSS. For both

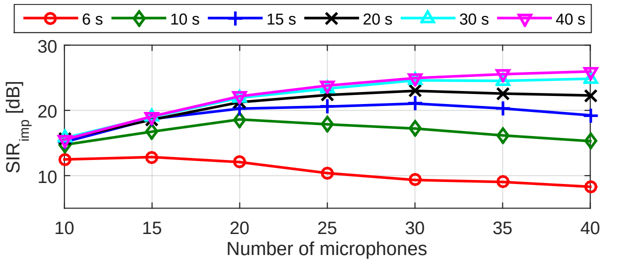


Fig. 13. SIR improvement by `Proposed` for a varying number of microphones (M) and length of signal (L_t). In most cases, the performance improves with M. For $L_t \leq 10$ and $M \geq 20$ the performance decreases with increasing M due to lack of enough data.

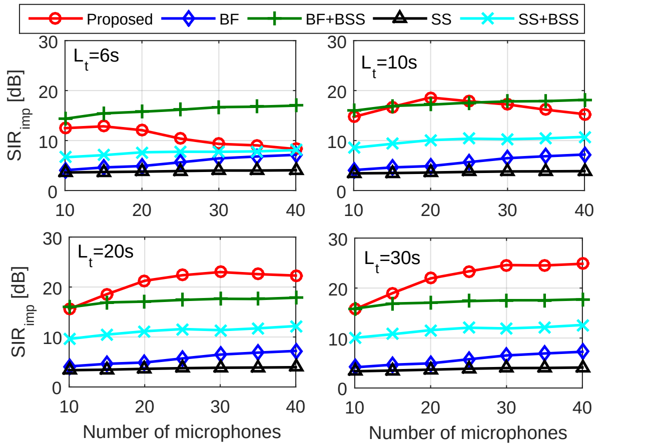


Fig. 14. SIR improvement by the considered algorithms for a varying number of microphones (M) and length of signal (L_t). In most cases, the performance of the `Proposed` algorithm improves with M. For $L_t \leq 10$ and $M \geq 20$ the performance decreases with increasing M due to lack of enough data.

`SS` and `SS+BSS` the SIR performance remains almost constant with respect to M and L_t. For `BF` and `BF+BSS` the SIR performance improves with increasing M but independent of L_t. Note that both `BF` and `BF+BSS` require the number and locations of the sources to be known.

We compare the *accuracy of the spatial likelihood peaks* obtained by the steering vector-based and the T-F masking-based approaches. Fig. 15 depicts the peak error rate (R_{pe}) of the two approaches for various M and L_t. The result for each (M, L_t) configuration is obtained by averaging the four realizations. It is observed that the masking-based approach performs slightly worse than the steering vector-based approach when $M < 20$, but outperforms it significantly especially when $M \geq 30$. The performance degradation of the steering vector-based approach for large M is mainly due to the intra-source ambiguity, which more sever with increasing M. In contrast, the masking-based approach is not affected by intra-source ambiguities, with peak error rate varying slightly with M. Since the performance of the two approaches is complimentary for large and small M, combining the two could lead to better localization results.

We compare the *multi-source localization performance* of the `Proposed`, `SRP`, `SRP-PHAT` algorithms. Fig. 16 depicts the recall rate (R_{recall}) and precision rate (R_{prec}) of the three algorithms for various M and L_t. The result for each (M, L_t) configuration is obtained by averaging the four realizations. In global, the performance improves when increasing L_t. `SRP` performs worst in terms of both R_{recall} and R_{prec}. `SRP-PHAT` performs best in terms of R_{prec}, which remains 1 for all testing cases. However, `SRP-PHAT` achieves a rather low R_{prec}, around 0.6, 0.7 and 0.8 for $L_t = 10$ s, 20 s and 30 s, respectively. `SRP-PHAT` detects multiple sources by merging the localization results across multiple time segments. Increasing signal length can increase the possibility of detecting all the sources. `Proposed` performs best in terms of R_{recall}. Its performance depends on both M and L_t. For $L_t \geq 20$ s, R_{recall} of `Proposed` improves when increasing M. For $L_t = 10$ s, R_{recall} improves when increasing M for $10 \leq M \leq 25$, and then decreases when increasing M for $M > 25$. The decrease of R_{recall} is due to degraded ICA performance in case of large M but small L_t. The remixing procedure of `Proposed` chooses the thresholds that can minimize false alarms, and thus tends to overestimate the number of sources. This is confirmed by the observation that, when $M \geq 20$ and $L_t \geq 20$, R_{recall} of `Proposed` is close to 1 but R_{prec} lies between 0.9 and 1. This drawback can be mitigated by considering the spatial likelihood: a high spatial likelihood is usually associated with correct localization while a low spatial likelihood leads to incorrect localization. Combining spatial likelihoods with particle filtering [60] may

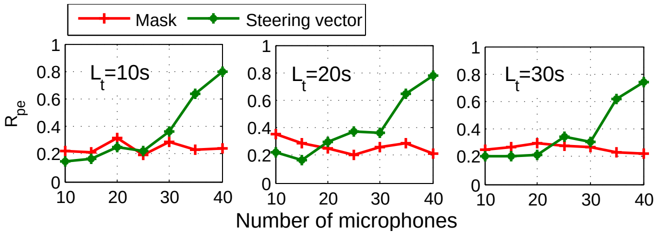


Fig. 15. Peak error rate (R_{pe}) by the T-F masking-based and steering vector-based localization approaches.

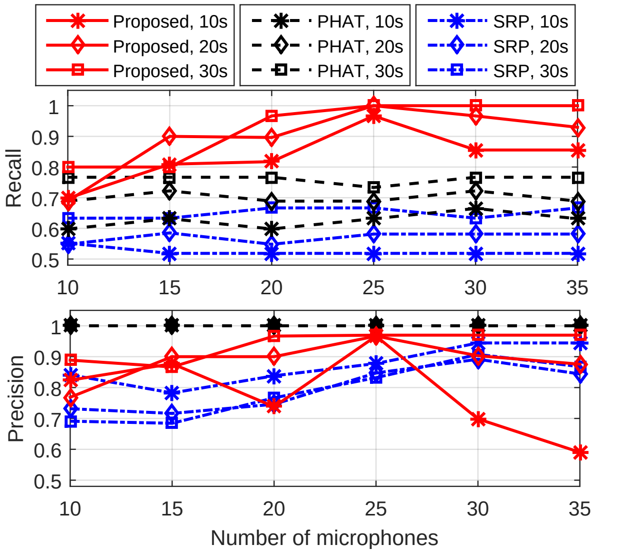


Fig. 16. Multi-source localization performance in terms of recall rate and precision rate for a varying number of microphones and length of signal.

improve the precision rate.

D. *Experiment with Real-data*

We use a database [26] with real recording of the scenario in Fig. 2. The data contain environmental noise, the directivity of the speakers and head movements. We use 30 microphones with signal length 30 s and sampling rate 8 kHz. Since the sound of individual speakers at the microphones is not available, we use the BSS Evaluation Toolbox [61] to calculate the SIR, using the close-recording of each speaker as reference. We compare five algorithms: `Proposed`, `SS`, `SS+BSS`, `BF`, `BF+BSS`.

Fig. 17 depicts the SIRs of each source in the microphone inputs and OBSS outputs. `Proposed` performs best in most cases, except when all the algorithms fail to extract the source s_5. This failure is possibly due to large head movement. The observation that the SIR of `BF` is even lower than the input SIR implies that s_5 has already deviated from its original location. As a result, s_5 is even not extracted in the determined BSS stage (which is not shown here) of `Proposed`.

Fig. 18 depicts the localization result by `Proposed` including the true and estimated source locations and the spatial likelihood of each estimate. It can be observed that `Proposed` can accurately localize 7 out of 10 sources. s_5 is not extracted in the output, s_4 is extracted as z_4 and z_8, and s_7 is extracted as z_{10} and z_{11}. However, these falsely detected sources usually show low spatial likelihoods, which could be used to overcome this problem.

IX. Conclusion

We proposed an over-determined source separation and localization method that can estimate the number and locations of the sources, and separate individual sources in a reverberant and multi-source environment. The proposed method exploits the redundant information of a sufficient number of microphones and performs well in highly reverberant scenarios. Experiments in a very challenging acoustic scenario show the effectiveness of the method, which improves when the number of microphones or the duration of the signal increases.

The separation performance tends to saturate when the number of microphones is large (cf. Fig. 13). Since the computational cost of the proposed method grows quickly with the number of microphones, it would be desirable to decompose the microphone network into several subsets and then perform ICA on each subset. The proposed method requires the sources to be static for a sufficiently long time interval so that the parameters of the demixing filter can be estimated. Extending the proposed method to dynamic acoustic scenarios will be an interesting future research direction. Moreover, the employed parameter selection scheme determines the threshold values based on a limited amount of data and thus may not work optimally in all real-world applications. An intelligent thresholding scheme that determines the threshold values adaptively would be more desirable. A weighted combination of the various threshold parameters could also increase robustness. In addition, there are some existing approaches which are able to estimate the number of sources directly from the microphone signals [62], [63]. Combining these approaches with our proposed remixing procedure may help better address the intra-source ambiguity problem.

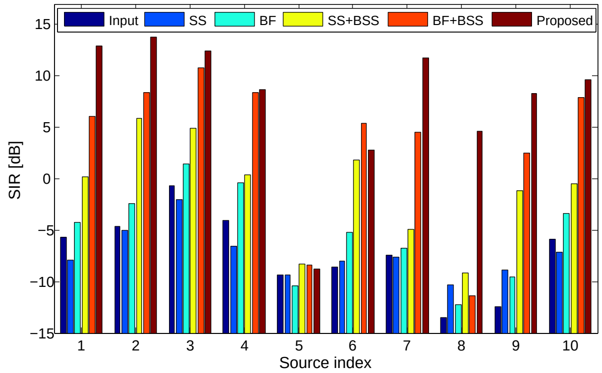


Fig. 17. SIR performance for real data using 30 microphones and 30 s data.

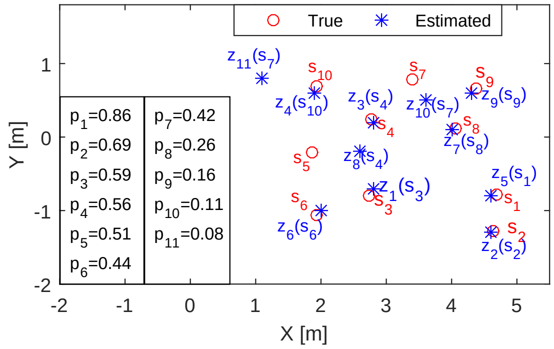


Fig. 18. True and estimated source locations by the proposed OBSS method for real data, as well as the spatial likelihood of each estimated source. $z_i(s_j)$ denotes the association between output z_i and source s_j.

References

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Berlin, Germany: Springer-Verlag, 2001.

[2] S. Makino, T. W. Lee, and H. Sawada, Eds. *Blind Speech Separation*, Berlin, Germany: Springer-Verlag, 2007.

[3] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1592-1604, Jul. 2007.

[4] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Cambridge, USA: Academic Press, 2010.

[5] T. Otsuka, K. Ishiguro, T. Yoshioka, H. Sawada, and H. G. Okuno, "Multichannel sound source dereverberation and separation for arbitrary number of sources based on Bayesian nonparametrics," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2218-2232, Dec. 2014.

[6] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York, USA: John Wiley & Sons, 2004.

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

[8] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516-527, Mar. 2011.

[9] L. Wang, H. Ding, and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," *EURASIP J. Audio, Speech, Music Process.*, pp. 1-13, 2010, Article ID 797962.

[10] S. M. Naqvi, M. Yu, and J. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE J. Selected Topics Signal Process.*, vol. 4, no. 5, pp. 895-910, Oct. 2010.

[11] S. Haykin and K. J. R. Liu, Eds. *Handbook on Array Processing and Sensor Networks*, Hoboken, USA: John Wiley & Sons, 2010.

[12] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," in *Proc. IEEE Symp. Commun. Veh, Technol. Benelux*, Ghent, Belgium, 2011, pp. 1-6.

[13] T. K. Hon, L. Wang, J. D. Reiss, and A. Cavallaro, "Audio fingerprinting for multi-device self-localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1623-1636, Oct. 2015.

[14] L. Wang, T. K. Hon, J. D. Reiss, and A. Cavallaro, "Self-localization of Ad-hoc arrays using time difference of arrivals," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 1018-1033, Feb. 2016.

[15] P. Pertila, M. S. Hamalainen and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 11, pp. 2393-2402, 2013.

[16] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 571-582, Mar. 2016.

[17] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds., Berlin, Germany: Springer-Verlag, 2001, pp. 157-180.

[18] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510-2526, 2007.

[19] J. P. Dmochowski and J. Benesty, "Steered beamforming approaches for acoustic source localization," in *Speech Processing in Modern Communication*, I. Cohen, J. Benesty and S. Gannot, Eds., Berline, Germany: Springer-Verlag, 2010, pp. 307-337.

[20] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP J. Advances in Signal Process.*, vol. 2003, pp. 1-10, 2003.

[21] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP J. Audio, Speech, Music Process.*, pp. 1-17, 2010, Article ID 147495.

[22] Y. Oualil, F. Faubel, and D. Klakow, "Joint detection and localization of multiple speakers using a probabilistic interpretation of the steered response power," in *Proc. Statistical Perceptual Audition Workshop*, Portland, USA, 2012, pp. 1-6.

[23] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 499-508, 2004.

[24] L. O. Nunes, W. A. Martins, and M. V. Lima, et al., "A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays," *IEEE Trans. Sig. Process.*, vol. 62, no. 19, pp. 5171-5183, 2014.

[25] A. Marti, M. Cobos, and J. J. Lopez, "A real-time sound source localization and enhancement system using distributed microphones," in *Proc. AES Convention 130*, London, UK, 2011, pp. 73-84.

[26] H. Do and H. F. Silverman, "Robust cross-correlation-based techniques for detecting and locating simultaneous, multiple sound sources," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Kyoto, Japan, 2012, pp. 201-204.

[27] C. Zhang, D. Florencio and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Process.*, Las Vegas, USA, 2008, pp. 2565-2568.

[28] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proc.-F*, vol. 140, no. 6, pp. 362-370, Dec. 1993.

[29] A. J. Bell and T. J. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129-1159, 1995.

[30] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook on Speech Processing and Speech Communication*, J. Benesty, M. M Sondhi, and Y. Huang, Eds., Berlin, Germany: Springer-Verlag, 2007, pp. 1-34.

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

[32] L. Wang, H. Ding, and F. Yin "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, vol. 3, pp. 549-557, Mar. 2011.

[33] L. Wang, "Multi-band multi-centroid clustering based permutation alignment for frequency-domain blind speech separation," *Digital Signal Process.*, vol. 31, pp. 79-92, 2014.

[34] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Process.*, vol. 12, no.5, pp. 530-538, Sep. 2004.

[35] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: investigation and solutions," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 1-13, Jan. 2005.

[36] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 666-678, Feb. 2006.

[37] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive BSS of short mixtures by ICA recursively regularized across frequencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 624-639, Mar. 2011.

[38] H. Sawada, S. Araki, R. Mukai, and Makino, "Solving the permutation problem of frequency-domain BSS when spatial aliasing occurs with wide sensor spacing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Toulouse, France, 2006, pp. 77-80.

[39] T. Kim, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies, *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70-79, Jan. 2007.

[40] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New York, USA, 2011, pp. 189-192.

[41] A. Lombard, Y. Zheng, H. Buchner, and W. Kallermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1490-1503, Aug. 2011.

[42] F. Nesta and O. Maurizio, "Generalized state coherence transform for multidimensional TDOA estimation of multiple sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 246-260, Jan. 2012.

[43] A. Masnadi-Shirazi and B. Rao, "An ICA-SCT-PHD filter approach for tracking and separation of unknown time-varying number of sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 828-841, Apr. 2013.

[44] A. Westner and V. M. Bove, "Blind separation of real world audio signals using overdetermined mixtures," in *Proc. Int. Workshop Independent Component Analysis and Blind Signal Separation*, Aussois, France, 1999, pp. 11-15.

[45] A. Koutras, E. Dermatas, and G. K. Kokkinakis, "Improving simultaneous speech recognition in real room environments using overdetermined blind source separation," in *Proc. InterSpeech*, Aalborg, Denmark, 2001, pp. 1009-1012.

[46] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 204-215, Jul. 2003.

[47] E. Robledo-Arnuncio and B. H. Juang, "Blind source separation of acoustic mixtures with distributed microphones," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Honolulu, USA, 2007, pp. 949-952.

[48] M. Joho, H. Mathis, and R. H. Lambert, "Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture," in *Proc. Int. Workshop Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland, 2000, pp. 81-86.

[49] S. Winter, H. Sawada, and S. Makino, "Geometrical interpretation of the PCA subspace approach for overdetermined blind source separation," *EURASIP J. Applied Signal Process.*, vol. 2006, pp. 1-11, 2006.

[50] L. Wang, H. Ding, and F. Yin, "Target speech extraction in cocktail party by combining beamforming and blind source separation," *Acoustics Australia*, vol. 39, no. 2, pp. 64-68, 2011.

[51] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind source separation with different sensor spacing and filter length for each frequency range," in *Proc. IEEE Workshop Neural Networks Signal Process.*, Martigny, Switzerland, 2002, pp. 465-474.

[52] Y. Zhang and J. A. Chambers, "Exploiting all combinations of microphone sensors in overdetermined frequency domain blind

separation of speech signals," *Int. J. Adaptive Control Signal Process.*, vol. 25, no. 1, pp. 88-94, 2011.

[53] C. Osterwise and S. L. Grant, "On over-determined frequency domain BSS," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 5, pp. 956-966, May 2014.

[54] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of binwise separated signals for permutation alignment in frequency-domain BSS," in *Proc. IEEE Int. Symp. Circuits Systems*, New Orleans, USA, 2007, pp. 3247-3250.

[55] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP J. Applied Signal Process.*, vol. 2003, pp. 1157-1166, 2003.

[56] K. Matsuoka, "Minimal distortion principle for blind source separation," in *Proc. SICE Annual Conf.*, Osaka, Japan, 2002, pp. 2138-2143.

[57] L. Wang, T. K. Hon, J. D. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1079-1093, Jun. 2016.

[58] G. Li and M. E. Lutman, "Sparseness and speech perception in noise," in *Proc. Interspeech*, Pittsburg, USA, 2006, pp. 1466-1469.

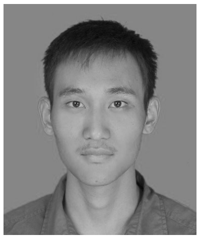
[59] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943-950, 1979.

[60] O. Cappe, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proc. IEEE*, vol. 95, no. 5, pp. 899-924, May 2007.

[61] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.

[62] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36-47, Jul. 2004.

[63] Z. Lu, and A. M. Zoubir, "Flexible detection criterion for source enumeration in array processing," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1303-1314, Mar. 2013.



Lin Wang received the B.S. degree in electronic engineering from Tianjin University, China, in 2003; and the Ph.D degree in signal processing from Dalian University of Technology, China, in 2010. From 2011 to 2013, he was an Alexander von Humboldt Fellow at the University of Oldenburg, Germany. Since 2014, he has been a postdoctoral researcher in the Centre for Intelligent Sensing at Queen Mary University of London. His research interests include video and audio compression, microphone array, blind source separation, and 3D audio processing.

Joshua D. Reiss is a Reader in Audio Engineering with the Centre for Digital Music in the School of Electronic Engineering and Computer Science at Queen Mary University of London. He has bachelors degrees in both physics and mathematics, and earned his Ph.D. in physics from the Georgia Institute of Technology. He is a member of the Board of Governors of the Audio Engineering Society, and co-founder of the company MixGenius, now known as LandR. Dr. Reiss has published more than 100 scientific papers and serves on several steering and technical committees. He has investigated sound synthesis, time scaling and pitch shifting, source separation, polyphonic music transcription, loudspeaker design, automatic mixing for live sound, and digital audio effects. His primary focus of research, which ties together many of the above topics, is on the use of state-of-the-art signal processing techniques for professional sound engineering.

Andrea Cavallaro received the Ph.D. degree in electrical engineering from Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2002. He was a Research Fellow with British Telecommunications in 2004. He is a Professor of Multimedia Signal Processing and the Director of the Centre for Intelligent Sensing at Queen Mary University of London. He has authored more than 150 journal and conference papers, one monograph on Video Tracking (Wiley, 2011), and three edited books, Multi-Camera Networks (Elsevier, 2009), Analysis, Retrieval and Delivery of Multimedia Content (Springer, 2012), and Intelligent Multimedia Surveillance (Springer, 2013). Prof. Cavallaro is Senior Area Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING and Associate Editor of the IEEE MultiMedia Magazine. He is an elected member of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee, and is the Chair of its Awards Committee, and an elected member of the IEEE Circuits and Systems Society Visual Communications and Signal Processing Technical Committee. He is a former elected member of the IEEE Signal Processing Society Multimedia Signal Processing Technical Committee, Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON SIGNAL PROCESSING and IEEE TRANSACTIONS ON IMAGE PROCESSING, and Associate Editor and Area Editor of IEEE Signal Processing Magazine, and Guest Editor of eleven special issues of international journals. He was General Chair for the IEEE/ACM ICDSC 2009, BMVC 2009, M2SFA2 2008, SSPE 2007, and IEEE AVSS 2007. He was Technical Program Chair of the IEEE AVSS 2011, the European Signal Processing Conference in 2008, and WIAMIS 2010. He received the Royal Academy of Engineering Teaching Prize in 2007, three Student Paper Awards on target tracking and perceptually sensitive coding at the IEEE ICASSP in 2005, 2007, and 2009, respectively, and the Best Paper Award at IEEE AVSS 2009.