Pseudo-Determined Blind Source Separation for Ad-hoc Microphone Networks

Lin Wang, Andrea Cavallaro

***Abstract*—We propose a pseudo-determined blind source separation framework that exploits the information from a large number of microphones in an ad-hoc network to extract and enhance sound sources in a reverberant scenario. After compensating for the time offsets and sampling rate mismatch between (asynchronous) signals, we interpret as a determined $M \times M$ mixture the over-determined $M \times N$ mixture, where $M > N$ is the number of microphones and N is the number of sources. Next, we propose a pseudo-determined mixture model that can apply an $M \times M$ independent component analysis (ICA) directly to the M-channel recordings. Moreover, we propose a reference-based permutation alignment scheme that aligns the permutation of the ICA outputs and classifies them into target channels, which contain the N sources, and non-target channels, which contain reverberation residuals. Finally, using the signals from non-target channels, we estimate in each target channel the power spectral density of the noise component that we suppress with a spectral post-filter. Interestingly, we also obtain late-reverberation suppression as by-product. Experiments show that each processing block improves incrementally source separation and that the performance of the proposed pseudo-determined separation improves as the number of microphones increases.**

***Index Terms*—Ad-hoc, asynchronous recording, blind source separation, over-determined mixture**

I. Introduction

Smartphones, tablets and body-worn cameras equipped with audio interfaces and wireless communication modules can be used as scalable and flexible ad-hoc microphone networks [1]. An important task when a group of people record the same event with their devices is to enhance the input signals and to localize sound sources [6], [7]. In order to employ traditional microphone array techniques with ad-hoc networks, specific challenges such as device localization [2], [3] and clock synchronization [4], [5] have to be addressed.

Blind source separation (BSS) is suitable for processing signals captured by an ad-hoc microphone network and can extract the speech of an individual from a mixture of speakers talking concurrently, without prior knowledge of the location of the microphones [8]. BSS employs independent component analysis (ICA) to estimate a demixing network to recover the sources from the mixture exploiting the statistical independence of the source signals [9]. For the mixing network to be invertible, ICA typically requires the number of microphones, M, to be equal to the number of sources, N.

Manuscript received: February 3, 2018

This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/K007491/1, and by the ARTEMIS-JU and the UK Technology Strategy Board (Innovate UK) through the COPCAMS Project under grant 332913.

The authors are with Centre for Intelligent Sensing, Queen Mary University of London, London, UK (e-mail: {lin.wang, a.cavallaro}@qmul.ac.uk)

BSS can be *determined* (DBSS: $M{=}N$) [10], *under-determined* (UBSS: $M{<}N$) [11] or *over-determined* (OBSS: $M{>}N$) [12]. Source separation with an ad-hoc network generally leads to an over-determined problem as the microphones outnumber the sources [7], [13]. A typical solution is to convert OBSS to DBSS by selecting a number of sensors equal to the number of sources or by dimensionality reduction [8]. However, dimensionality reduction may discard information that helps the separation task.

In this paper, we present a frequency-domain BSS framework that applies an $M \times M$ ICA directly to the M-channel recordings when $M{>}N$. We interpret the overdetermined $M \times N$ mixture as a determined $M \times M$ mixture, thus grounding the feasibility of an $M \times M$ ICA. In contrast to a regular-determined $N \times N$ mixture, we term this $M \times M$ mixture pseudo-determined mixture and the proposed method *pseudo-determined* BSS (PBSS). Compared to [13], the proposed method includes a new signal model to interpret the pseudo-determined mixture and to classify the ICA outputs into target channels (containing the N sources) and non-target channels (containing reverberant residuals). Based on this model, we derive three insights that are the basis for PBSS in an ad-hoc network with a large number of microphones. Specifically, we discuss (i) the performance improvement of PBSS when the number of microphones increases; (ii) the performance degradation when the reverberation density increases, and show how increasing the number of microphones addresses this problem; and (iii) the benefits of using the signals in the non-target channels as reference to estimate the noise in each target channel, which allows us to further improve the source separation performance with a post-filter. Moreover, we define a new source separation framework cascading PBSS and post-filtering, and propose a reference-based permutation alignment scheme to solve the permutation ambiguity and the target-channel detection problems.

After reviewing related works (Sec. II), we formulate the problem (Sec. III) and present three insights for pseudo-determined BSS (Sec. IV). Next, we introduce the new source separation framework in Sec. V and measures for performance evaluation in Sec. VI. We then test the advantage of PBSS with simulations in Sec. VII and real data in Sec. VIII. Finally, in Sec. IX we draw conclusions.

II. Background

Multiple simulateous sound sources undergo convolutive mixing due to reverberation. The convolutive BSS problem can be addressed using short-time Fourier transform (STFT)

to approximate the convolution in the time domain as linear instantaneous mixing in the frequency domain [8]. *Independent component analysis* (ICA) is then applied at individual frequency bins to separate linear and instantaneous mixtures by adaptively estimating a demixing matrix and maximizing the statistical independence of the output signals [9]. To obtain the estimate of the demixing matrix, ICA typically requires the mixing network to remain stationary for a certain period. Next, *permutation alignment* groups separated components from the same source, which are finally transformed back into the time domain via inverse STFT.

Permutation ambiguity problems have been addressed with inter-frequency dependency, location-based or joint optimization strategies. *Inter-frequency dependency* strategies are the most robust under reverberations, especially for speech signals [10] and exploit the temporal structure of separated signal amplitudes or speech activities. This temporal structure has high correlation, for the same source, between neighboring bins. Clustering-based and region-wise permutation alignment schemes exploit such inter-frequency dependency [10], [11], [15]. *Location*-based strategies exploit spatial information since contributions from the same source are likely to originate from the same direction [16], [17]. *Joint optimization* strategies, e.g. independent vector analysis (IVA), 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 [18], [19].

For the mixing network to be invertible, ICA usually works with an equal number of sources and microphones [9]. To convert the over-determined BSS problem ($M{>}N$) to a determined BSS problem ($M{=}N$), a regular-determined or pseudo-determined strategy can ben used (Table I).

The *regular-determined* strategy converts an over-determined $M{\times}N$ mixture to a regular-determined $N{\times}N$ mixture by subset selection [20] or dimensionality reduction [21]. Subset selection identifies a subset of microphones from the whole set. The selection can be based on geometric information [20] or on selecting the microphone subset with the best outputs [12], [29]. Subspace-based pre-processing (e.g. PCA - principal component analysis) can also be used to extract an equal number of components [21]–[26]. After PCA, the signal-to-noise ratio in the retained components is generally higher than in any individual input signal and the mixing matrix is usually better conditioned. Alternatively, a set of fixed beamformers each pointing at one source can be applied before separation if the location of each source is known [27], [28]. The fixed beamformer can reduce noise and reverberation for each source, thus making the subsequent separation task easier.

The *pseudo-determined strategy* strategy converts the over-determined $M{\times}N$ mixture to a pseudo-determined $M{\times}M$ mixture so that one can apply an $M{\times}M$ ICA, which achieves better separation than a regular $N{\times}N$ ICA. However, with $M > N$, each source may occupy one or more channels at the outputs, leading to inter- and intra-source ambiguities [13]. This is a more challenging problem than the one for a regular $N{\times}N$ ICA, where only inter-source ambiguities exist. While a source merging-based permutation alignment scheme can classify the outputs and merge those belonging to the same source [13], this procedure does not discriminate the noise components, which are therefore merged into the output thus degrading the overall separation performance. To address this problem, in this paper we propose a reference-based permutation alignment scheme.

TABLE I
COMPARISON OF OVER-DETERMINED SOURCE SEPARATION ALGORITHMS. KEY: $\mathsf{R_M}$: MICROPHONE LOCATION; $\mathsf{R_S}$: SOURCE LOCATION; N: NUMBER OF SOURCES

References	Prior knowledge			Approach		Strategy
	$\mathsf{R_M}$	$\mathsf{R_S}$	N			
[21]–[26]			✓	dimensionality reduction	subspace	regular-determined
[27], [28]	✓	✓	✓		fixed beamforming	
[20]	✓		✓	subset selection	geometry-based	
[12], [29]			✓		separation-based	
[13]	✓			source merging		pseudo-determined
Proposed			✓	reference-based		

III. PROBLEM FORMULATION

Let M microphones be distributed at unknown locations in a reverberant acoustic environment. Let these microphones record a known number, $N \leq M$, of sound sources at unknown (fixed) locations. Let $\boldsymbol{s}(n) = [s_1(n), \cdots, s_N(n)]^{\mathrm{T}}$ be the N source signals and $\boldsymbol{x}(n) = [x_1(n), \cdots, x_M(n)]^{\mathrm{T}}$ be the signals received by the M microphones, where n is the sample index and the superscript $(\cdot)^{\mathrm{T}}$ is the transpose operator. Writing $\boldsymbol{s}(n)$ and $\boldsymbol{x}(n)$ in the STFT domain, we get $\boldsymbol{S}(k,l) = [S_1(k,l), \cdots, S_M(k,l)]^{\mathrm{T}}$ and $\boldsymbol{X}(k,l) = [X_1(k,l), \cdots, X_M(k,l)]^{\mathrm{T}}$, where k and l are the frequency and frame indices, respectively[1]. Let K and L denote the total number of frequency bins and time frames, respectively.

If $x_{ij}(n)$ is the component of $s_j(n)$ received by microphone i and $h_{ij}(n)$ is the impulse response between them, then

$$x_{ij}(n) = h_{ij}(n) * s_j(n), \tag{1}$$

where the operator $*$ denotes the convolution. Let $H_{ij}(k)$ be the frequency-domain version of $h_{ij}(n)$. Note that with static microphones and sources, the mixing filter $H_{ij}(k)$ is time-invariant. If the STFT frame length is larger than that of the impulse response, the convolution in Eq. 1 can be written in the STFT domain as

$$X_{ij}(k,l) = H_{ij}(k) S_j(k,l). \tag{2}$$

The microphone signal $\boldsymbol{X}(k,l)$ is obtained by passing $\boldsymbol{S}(k,l)$ through a mixing network $\boldsymbol{H}(k)$:

$$\boldsymbol{X}(k,l) = \underbrace{\boldsymbol{H}(k)}_{M\times N}\underbrace{\boldsymbol{S}(k,l)}_{N\times 1} = \begin{bmatrix} H_{11} & \cdots & H_{1N} \\ \vdots & \ddots & \vdots \\ H_{M1} & \cdots & H_{MN} \end{bmatrix} \begin{bmatrix} S_1 \\ \vdots \\ S_N \end{bmatrix}, \tag{3}$$

which is an over-determined mixture when $M{>}N$.

Our objective is to blindly extract the N sources from the recordings of the M microphones. While BSS approaches have been widely used to solve this problem, their performance

[1]To improve readability, n, k and l may be omitted in some equations.

usually degrades considerably when the number of sources and the reverberation density increase. In this paper, we show how to exploit a sufficient number of microphones in an ad-hoc network to tackle this challenge. We will first assume that the signals from the M microphones are synchronously sampled (Sec. IV) and then consider a more general case with unsynchronized signals (Sec. V).

IV. Pseudo-determined mixture model

We aim to build a complete theoretical framework based on pseudo-determined BSS [13], an approach that achieves better source separation in reverberant scenarios by applying an $M\times M$ICA directly to an $M\times N$mixture.

A. Pseudo-determined BSS

Based on the image-source model [30], we approximate the room reverberation as an aggregated contribution from a set of image sources, including an early-reverberant and multiple late-reverberant image sources.

Let for a physical source $s_j(n)$ be R_j image sources, where $s_{j1}(n)$ is the early-reverberant image source and $s_{j2}(n),\cdots,s_{jR_j}(n)$ are the late-reverberant image sources. Let $\tilde{h}_{ijr}(n)$ be the impulse response from the r-th image source $\tilde{s}_{jr}(n)$ to microphone i. The signal $x_{ij}(n)$ in Eq. 1 can therefore be represented as

$$x_{ij}(n)=\sum_{r=1}^{R_j}\tilde{h}_{ijr}(n)*\tilde{s}_{jr}(n). \tag{4}$$

Let $\tilde{S}_{jr}(k,l)$ and $\tilde{H}_{ijr}(k)$ be the frequency-domain version of $\tilde{s}_{jr}(n)$ and $\tilde{h}_{ijr}(n)$, respectively. The convolution in Eq. 4 written in the STFT domain becomes

$$X_{ij}(k,l)=\sum_{r=1}^{R_j}\tilde{H}_{ijr}(k)\tilde{S}_{jr}(k,l). \tag{5}$$

Let $R=\sum_{j=1}^{N}R_j$ virtual image sources generate from the N physical sources, i.e. $\tilde{\boldsymbol{S}}(k,l)=[\tilde{S}_{11}(k,l),\cdots,\tilde{S}_{1R_1}(k,l),\cdots,\tilde{S}_{N1}(k,l),\cdots,\tilde{S}_{NR_N}(k,l)]^{\mathrm{T}}$. The microphone signal $\boldsymbol{X}(k,l)$ can be obtained by passing $\tilde{\boldsymbol{S}}(k,l)$ through a mixing network $\tilde{\boldsymbol{H}}(k)$, i.e.

$$\boldsymbol{X}(k,l)=\underbrace{\tilde{\boldsymbol{H}}(k)}_{M\times R}\underbrace{\tilde{\boldsymbol{S}}(k,l)}_{R\times 1}=\begin{bmatrix}\tilde{H}_{111}&\cdots&\tilde{H}_{1NR_N}\\ \vdots&\ddots&\vdots\\ \tilde{H}_{M11}&\cdots&\tilde{H}_{MNR_N}\end{bmatrix}\begin{bmatrix}\tilde{S}_{11}\\ \vdots\\ \tilde{S}_{NR_N}\end{bmatrix}. \tag{6}$$

The value of R ($>M$) is unknown but proportional to the reverberation density. These image sources originate from different spatial locations (with different delays) and each has higher non-Gaussianity than the microphone signal due to room reverberation (see Fig. 4). ICA usually employs non-Gaussianity to measure the independence of the outputs [9]. When applying an $M\times M$ ICA to Eq. 6, ICA (with M degrees of freedom) can separate from the mixture N early-reverberant plus $M-N$ late-reverberant image sources that originate from different spatial locations and have the maximum non-Gaussianity. Let us represent these M separated image sources as an $M\times 1$ vector

$$\bar{\boldsymbol{S}}_A(k,l)=[\bar{S}_1(k,l),\cdots,\bar{S}_M(k,l)]^{\mathrm{T}} \tag{7}$$

and its corresponding mixing network between these image sources and the microphones as the $M\times M$ matrix $\bar{\boldsymbol{H}}_A(k)$. The demixing matrix $\bar{\boldsymbol{W}}(k)$ estimated by ICA ideally inverses $\bar{\boldsymbol{H}}_A(k)$, i.e.

$$\bar{\boldsymbol{W}}(k)\bar{\boldsymbol{H}}_A(k)=\boldsymbol{I}_M, \tag{8}$$

where $\boldsymbol{I}_M$ is an $M\times M$ identity matrix if we do not consider scaling and permutation ambiguities of ICA. Because the number of sources is still N, we term the BSS approach using this $M\times M$ ICA pseudo-determined BSS (PBSS).

B. Advantages of Pseudo-determined BSS

Let us divide the components in $\tilde{\boldsymbol{S}}(k,l)$ into two subvectors: an $M\times 1$ vector $\bar{\boldsymbol{S}}_A(k,l)$, defined in Eq. 7 and containing N early-reverberant and $M-N$ late-reverberant image sources, and an $(R-M)\times 1$ vector, $\bar{\boldsymbol{S}}_B(k,l)$, which contains the remaining late-reverberant image sources. A new vector is formulated as $\bar{\boldsymbol{S}}(k,l)=[\bar{S}_1(k,l),\cdots,\bar{S}_R(k,l)]^{\mathrm{T}}=[\bar{\boldsymbol{S}}_A^{\mathrm{T}}(k,l)\ \bar{\boldsymbol{S}}_B^{\mathrm{T}}(k,l)]^{\mathrm{T}}$. The model in Eq. 6 is then updated as

$$\boldsymbol{X}(k,l)=\underbrace{\bar{\boldsymbol{H}}(k)}_{M\times R}\underbrace{\bar{\boldsymbol{S}}(k,l)}_{R\times 1}=\begin{bmatrix}\bar{H}_{11}&\cdots&\bar{H}_{1R}\\ \vdots&\ddots&\vdots\\ \bar{H}_{M1}&\cdots&\bar{H}_{MR}\end{bmatrix}\begin{bmatrix}\bar{S}_1\\ \vdots\\ \bar{S}_R\end{bmatrix}, \tag{9}$$

where $\bar{H}_{ir}(k)$ is the transfer function between $\bar{S}_r(k,l)$ and microphone i.

We split $\bar{\boldsymbol{H}}(k)$ into two sub-matrices, $\bar{\boldsymbol{H}}_A(k)$ and $\bar{\boldsymbol{H}}_B(k)$, corresponding to $\bar{\boldsymbol{S}}_A(k,l)$ and $\bar{\boldsymbol{S}}_B(k,l)$, and thus

$$\begin{aligned}\boldsymbol{X}(k,l)&=\begin{bmatrix}\bar{\boldsymbol{H}}_A(k)&\bar{\boldsymbol{H}}_B(k)\end{bmatrix}\begin{bmatrix}\bar{\boldsymbol{S}}_A(k,l)\\ \bar{\boldsymbol{S}}_B(k,l)\end{bmatrix}\\&=\underbrace{\bar{\boldsymbol{H}}_A(k)}_{M\times M}\underbrace{\bar{\boldsymbol{S}}_A(k,l)}_{M\times 1}+\underbrace{\bar{\boldsymbol{H}}_B(k)}_{M\times(R-M)}\underbrace{\bar{\boldsymbol{S}}_B(k,l)}_{(R-M)\times 1},\end{aligned} \tag{10}$$

which is a decomposition of the original mixture into a pseudo-determined mixture plus a residual mixture.

Due to the residual term $\bar{\boldsymbol{H}}_B(k)\bar{\boldsymbol{S}}_B(k,l)$ in Eq. 10 and the fact that $\bar{\boldsymbol{W}}(k)\bar{\boldsymbol{H}}_B(k)=\bar{\boldsymbol{Q}}(k)\neq\boldsymbol{I}_M$, applying $\bar{\boldsymbol{W}}(k)$ to $\boldsymbol{X}(k,l)$ will lead to a noisy output $\bar{\boldsymbol{Y}}(k,l)=[\bar{Y}_1(k,l),\cdots,\bar{Y}_M(k,l)]^{\mathrm{T}}$:

$$\begin{aligned}\bar{\boldsymbol{Y}}(k,l)&=\bar{\boldsymbol{W}}(k)\boldsymbol{X}(k,l)\\&=\bar{\boldsymbol{S}}_A(k,l)+\bar{\boldsymbol{V}}_A(k,l)=\bar{\boldsymbol{S}}_A(k,l)+\bar{\boldsymbol{Q}}(k)\bar{\boldsymbol{S}}_B(k,l)\\&=\begin{bmatrix}\bar{S}_1(k,l)\\ \vdots\\ \bar{S}_M(k,l)\end{bmatrix}+\begin{bmatrix}\sum_{j=1}^{R-M}\bar{q}_{1j}(k)\bar{S}_{j+M}(k,l)\\ \vdots\\ \sum_{j=1}^{R-M}\bar{q}_{Mj}(k)\bar{S}_{j+M}(k,l)\end{bmatrix},\end{aligned} \tag{11}$$

where $\bar{\boldsymbol{S}}_A(k,l)$ and $\bar{\boldsymbol{V}}_A(k,l)$ contain the source and the noise components, respectively.

Among the M outputs of $\bar{\boldsymbol{Y}}(k,l)$, we are interested in the first N channels as they contain the early-reverberant components of the N sources. We thus split $\bar{\boldsymbol{S}}_A(k,l)$ into two sub-vectors: $\bar{\boldsymbol{S}}_{A1}(k,l)=[\bar{S}_1(k,l),\cdots,\bar{S}_N(k,l)]^{\mathrm{T}}$, containing

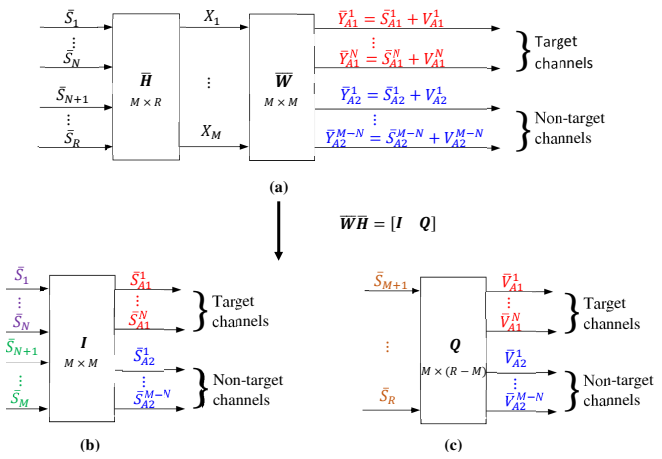


Fig. 1. Pseudo-determined blind source separation: (a) the mixing and demixing procedure; (b) source components in the output channels; (c) noise components in the output channels.

the N early-reverberant image sources; and $\bar{\boldsymbol{S}}_{A2}(k,l) = [\bar{S}_{N+1}(k,l), \cdots, \bar{S}_M(k,l)]^{\mathrm{T}}$, containing the $M-N$ late-reverberant image sources.

Similarly, we split $\bar{\boldsymbol{Y}}(k,l)$ and $\bar{\boldsymbol{V}}_A(k,l)$:

$$\bar{\boldsymbol{Y}}(k,l) = \begin{bmatrix} \bar{\boldsymbol{Y}}_{A1}(k,l) \\ \bar{\boldsymbol{Y}}_{A2}(k,l) \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{S}}_{A1}(k,l) + \bar{\boldsymbol{V}}_{A1}(k,l) \\ \bar{\boldsymbol{S}}_{A2}(k,l) + \bar{\boldsymbol{V}}_{A2}(k,l) \end{bmatrix}, \quad (12)$$

and refer to $\bar{\boldsymbol{Y}}_{A1}(k,l)$ as *target channels*, which contain the target sources $\bar{\boldsymbol{S}}_{A1}(k,l)$; and to $\bar{\boldsymbol{Y}}_{A2}(k,l)$ as *non-target channels*, which contain the non-target sources $\bar{\boldsymbol{S}}_{A2}(k,l)$. Moreover, we refer to $\bar{\boldsymbol{S}}_B(k,l)$ as redundant sources, which contribute to the noise components in $\bar{\boldsymbol{V}}_{A1}(k,l)$ and $\bar{\boldsymbol{V}}_{A2}(k,l)$. These relationships are visualized in Fig. 1.

For each target channel $\bar{Y}^m_{A1}(k,l) = \bar{S}^m_{A1}(k,l) + \bar{V}^m_{A1}(k,l)$, the noise component $\bar{V}^m_{A1}(k,l)$ can be represented as a linear combination of the elements in $\bar{\boldsymbol{S}}_B(k,l)$. Let $\mathbb{S}^m_{A1}$ represent the set of image sounds that originate from the target source $\bar{S}^m_{A1}$, we can decompose $\bar{V}^m_{A1}(k,l)$ as

$$\begin{aligned} \bar{V}^m_{A1}(k,l) &= \sum_{j=M+1}^{R} \bar{q}_{m,j-M}(k)\bar{S}_j(k,l) \\ &= \sum_{j\in\mathbb{S}^m_{A1}} \bar{q}_{m,j-M}\bar{S}_j(k,l) + \sum_{j\notin\mathbb{S}^m_{A1}} \bar{q}_{m,j-M}\bar{S}_j(k,l), \end{aligned} \quad (13)$$

where the first term represents the contribution from the late-reverberant sounds of the target source, while the second term represents the contribution from other interfering sources. Thus, the noise component will introduce not only interferences but also reverberation residuals in the source separation output. The energy of the noise component $\bar{V}^m_{A1}$ is proportional to the overall energy of the $R-M$ components in $\bar{\boldsymbol{S}}_B(k,l)$. The separation performance of PBSS thus mainly depends on two factors: R and M.

Based on Eq. 13, we obtain the following insights on PBSS.

Insight 1: *The separation performance tends to improve as the number of microphones increases.* Let us use as an example $M = N$ and $M = M_1$ ($M_1 > N$). When R is fixed, the noise component in the target channel in the two cases can be represented for $M = M_1$ as

$$\bar{V}^m_{A1}[M_1] = \sum_{j=M_1+1}^{R} \bar{q}_{m,j-M}\bar{S}_j, \quad (14)$$

and for $M = N$ as

$$\bar{V}^m_{A1}[N] = \sum_{j=N+1}^{M_1} \bar{q}_{m,j-M}\bar{S}_j + \sum_{j=M_1+1}^{R} \bar{q}_{m,j-M}\bar{S}_j, \quad (15)$$

with $\bar{V}^m_{A1}[N]$ having a higher energy than $\bar{V}^m_{A1}[M_1]$. When M increases from N to M_1, the redundant sources $\bar{S}_{N+1}, \cdots, \bar{S}_{M_1}$ are extracted from $\bar{\boldsymbol{S}}_B$ to $\bar{\boldsymbol{S}}_{A2}$ and no longer appear in the target channels. These displaced elements contain late-reverberant image sounds from both the target source and interfering sources. Increasing M reduces the energy of the noise component in the target channel, thus increasing the signal-to-interference ratio (SIR) while suppressing artificial reverberation effects, i.e. achieving dereverberation as by-product.

Insight 2: *The separation performance tends to degrade as the reverberation density increases.* Let us use as an example $R = R_1$ and $R = R_2$ ($R_1 < R_2$). When M is fixed, the noise component in the target channel can be represented for $R = R_1$ as

$$\bar{V}^m_{A1}[R_1] = \sum_{j=M+1}^{R_1} \bar{q}_{m,j-M}\bar{S}_j, \quad (16)$$

and for $R = R_2$ as

$$\bar{V}^m_{A1}[R_2] = \sum_{j=M+1}^{R_1} \bar{q}_{m,j-M}\bar{S}_j + \sum_{j=R_1+1}^{R_2} \bar{q}_{m,j-M}\bar{S}_j, \quad (17)$$

with $\bar{V}^m_{A1}[R_2]$ having a higher energy than $\bar{V}^m_{A1}[R_1]$. Increasing R from R_1 to R_2 does not change the target and non-target sources in $\bar{\boldsymbol{S}}_{A1}$ and $\bar{\boldsymbol{S}}_{A2}$, but produces more redundant sources, i.e. $\bar{S}_{R_1+1}, \cdots, \bar{S}_{R_2}$. This raises the energy of the noise component in the target channel, thus decreasing the SIR and introducing artificial reverberation effects. Performance degradation in reverberant scenarios is a general problem of BSS caused by the poor separation performance of ICA for long mixing filters [27], [32]. PBSS instead tackles this problem effectively when increasing the number of microphones: as M increases, more high-energy late-reverberant image sounds are extracted as non-target sources, thus reducing interference and reverberation in the target channels.

Insight 3: *By dividing the outputs into target and non-target channels, PBSS naturally allows a post-filter to enhance the separation output.* Referring to Eq. 11 and Eq. 12, the noise components $\boldsymbol{V}_{A1}$ in the target channel are a linear combination of the elements in $\bar{\boldsymbol{S}}_B$, which consist of late-reverberant images of the N sources. Likewise, the non-target channel $\boldsymbol{Y}_{A2}$ is a linear combination of the elements in $\boldsymbol{S}_{A2}$ and $\bar{\boldsymbol{S}}_B$, which both consists of late-reverberant images of the N sources. The signals in the non-target channels thus provide valuable information to estimate the noise components in the target channels. If we manage to exploit this information to estimate the power spectrum density (PSD) of the noise

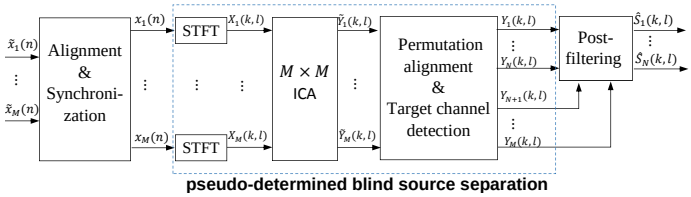


Fig. 2. Block diagram of the proposed pseudo-determined BSS framework.

component, we can design a spectral post-filter to further enhance the separated signals in the target channels.

V. The Proposed Separation Framework

The three insights presented in Sec. IV lead to the proposed pseudo-determined BSS framework (see Figure 2 and Table II) for ad-hoc networks with *asynchronously sampled* signals $\tilde{x}_1(n), \cdots, \tilde{x}_M(n)$ from M independent devices.

A. Synchronization

The fist step towards formulating a unified separation network is to synchronize the signals from independent microphones. The synchronization of these signals requires the estimation of time offset and sampling rate offset.

The *time offset* can be estimated by maximizing the cross-correlation between audio fingerprints in the time-frequency domain [33], [34] or between time-domain sequences [35]. We opt for the latter solution as BSS works robustly even with small misalignments between sequences [4].

A *sampling rate offset* leads to different unit lengths of the digital samples and creates a Doppler effect, i.e. the digital sequence either shrinks or expands along the time axis compared to the original waveform. This generates a time-varying delay between asynchronous recordings, which significantly degrades the performance of BSS [4]. To estimate the sampling rate offset we maximize the correlation of the phase information of the microphone signals [5]. Given the offset, we correct the sampling rate mismatch via resampling.

Let the time offset and sampling rate offset between two sequences $\tilde{x}_1(n)$ and $\tilde{x}_2(n)$ be δ_{12} and ε_{12}, respectively; and f_s be the nominal sampling rate of the first microphone. Then the synchronized sequences can be expressed as

$$\begin{cases} x_1(n) = \tilde{x}_1(n) \\ x_2(n) = \mathscr{R}(\tilde{x}_2(n-\delta_{12}), f_s, f_s+\epsilon_{12}) \end{cases}, \tag{18}$$

where $\mathscr{R}(\cdot)$ is the resampling operator [5] that converts the sampling rate $f_s+\epsilon_{12}$ to f_s.

We synchronize all the signals from the M independent microphones using one of the microphones as reference.

B. Permutation alignment and target channel detection

The $M\times N$ over-determined mixing network obtained after synchronization could undergo an $M\times M$ ICA directly on the signals from the M microphones. This would result in better separation but more challenging permutation ambiguities as, with $M>N$, each source may occupy multiple output channels and thus lead to inter-source and intra-source permutation ambiguities. Since only N target channels are of interest out of these M outputs, the permutation alignment task can be simplified as detecting the N target channels and aligning their permutation.

TABLE II
Algorithms used in the PBSS framework.

Functionality	Algorithm
Alignment	correlation maximization-based time offset estimation [35]
Synchronization	correlation maximization-based sampling rate offset estimation [5]
$N \times N$ ICA	Infomax [14]
Blind permutation alignment	clustering-based permutation alignment [10]
$M \times M$ ICA	Infomax [14]
Reference-based permutation alignment	**proposed (Sec. V-B)**
Noise PSD estimation	**proposed (Sec. V-C)**
Spectral post-filter	Wiener filter [38]

If N is known and we pick only N microphones, $N\times N$ ICA would produce worse separation but fewer permutation ambiguities (inter-source only). With an equal number of sources and output channels, the N outputs have a one-to-one correspondence with the N sources. The permutation alignment problem of the determined $N\times N$ ICA has been investigated intensively [10], [19] and we use here the permutation aligned results of the $N\times N$ ICA as reference for the target channel detection and permutation alignment of the $M\times M$ ICA.

The proposed permutation alignment method (Fig. 3(a)) consists of an $M\times M$ ICA step with M unordered outputs at each frequency bin, an $N\times N$ ICA step together with blind permutation alignment providing N permutation aligned outputs at each frequency bin, and a reference-based permutation alignment step that aligns the permutation of the $M\times M$ ICA outputs and classify them as target or non-target channels.

Applying an $M \times M$ ICA to the microphone signal $\boldsymbol{X}_M(k,l) = [X_1(k,l), \cdots, X_M(k,l)]^{\mathrm{T}}$, we obtain the demixing matrix $\tilde{\boldsymbol{W}}_M(k)$ with unordered outputs

$$\tilde{\boldsymbol{Y}}(k,l) = \tilde{\boldsymbol{W}}_M(k)\boldsymbol{X}_M(k,l) = [\tilde{Y}_1(k,l), \cdots, \tilde{Y}_M(k,l)]^{\mathrm{T}}. \tag{19}$$

Applying an $N\times N$ ICA to the microphone signal $\boldsymbol{X}_N(k,l) = [X_1(k,l), \cdots, X_N(k,l)]^{\mathrm{T}}$, we obtain the demixing matrix $\tilde{\boldsymbol{W}}_N(k)$ with unordered outputs

$$\tilde{\boldsymbol{Z}}(k,l) = \tilde{\boldsymbol{W}}_N(k)\boldsymbol{X}_N(k,l) = [\tilde{Z}_1(k,l), \cdots, \tilde{Z}_N(k,l)]^{\mathrm{T}}. \tag{20}$$

We then employ the algorithm [10] to align the permutation of the $N\times N$ ICA outputs as

$$\boldsymbol{Z}(k,l) = [Z_1(k,l), \cdots, Z_N(k,l)]^{\mathrm{T}}, \tag{21}$$

and use $\boldsymbol{Z}(k,l)$ as a reference to detect the target channels in $\tilde{\boldsymbol{Y}}(k,l)$ and align the permutation. This is achieved by computing the similarity between the components in $\boldsymbol{Z}(k,l)$ and in $\tilde{\boldsymbol{Y}}(k,l)$. We measure the similarity between sequence $\tilde{Y}_i(k,l)$ and $Z_j(k,l)$ by the correlation coefficient of their

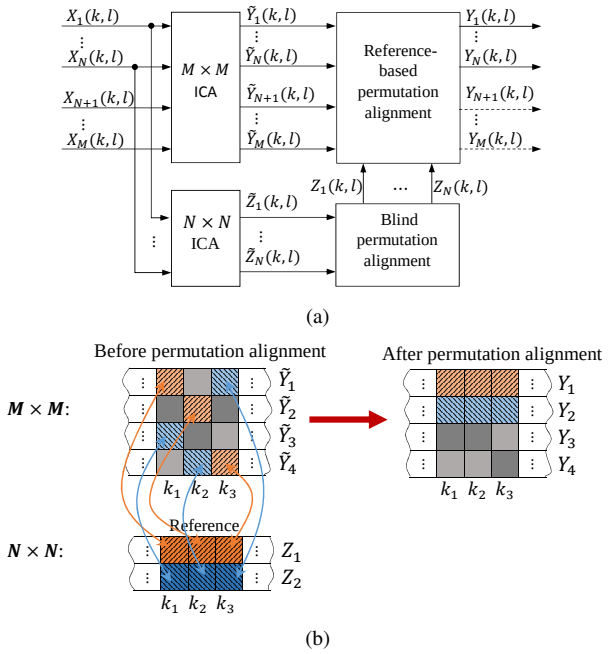


Fig. 3. Using the permutation aligned result from the $N \times N$ ICA as reference for target channel detection and permutation alignment of the $M \times M$ ICA. (a) Block diagram of reference-based permutation alignment algorithm. (b) Illustration of reference-based permutation alignment with $M = 4$ and $N = 2$. The cells with orange and blue shadows belong to target channels while the cells with gray shadows belong to non-target channels.

amplitudes, γ_{ij}, defined as

$$\gamma_{ij}(k) = \frac{\sum_{l=1}^{L} |\tilde{Y}_i(k,l)||Z_j(k,l)|}{\sqrt{\sum_{l=1}^{L} |\tilde{Y}_i(k,l)|^2}\sqrt{\sum_{l=1}^{L} |Z_j(k,l)|^2}}. \quad (22)$$

Let Π_M be the permutation of the M outputs, i.e. the projection from the original order $[1, \cdots, M]$ to a new order $[\Pi_M(1), \cdots, \Pi_M(M)]$, and let $\boldsymbol{\Pi}_M$ be the set of all possible projections. The permutation of the elements in $\tilde{\boldsymbol{Y}}(k,l)$ is then determined as

$$\Pi_M^k = \underset{\Pi_M \in \boldsymbol{\Pi}_M}{\arg\max} \sum_{j=1}^{N} \left\{ \gamma_{ij}(k)|_{i=\Pi_M(j)} \right\}, \quad \forall k \quad (23)$$

where Π_M^k is the permutation at frequency k. By sticking to the N references in $\boldsymbol{Z}(k,l)$, the N target channels can be naturally detected and permutation aligned.

We update the demixing matrix as

$$\hat{\boldsymbol{W}}_M(k) \xleftarrow{\Pi_M^k} \tilde{\boldsymbol{W}}_M(k), \quad (24)$$

and correct the scaling ambiguity with a back projection [36]

$$\boldsymbol{W}_M(k) = \text{diag}\left(\hat{\boldsymbol{W}}_M^{-1}(k)\right) \cdot \hat{\boldsymbol{W}}_M(k), \quad (25)$$

where the operator diag$(\cdot)$ retains only the diagonal elements of a matrix.

Finally, the permutation aligned outputs are represented as

$$\boldsymbol{Y}(k,l) = \boldsymbol{W}_M(k)\boldsymbol{X}_M(k,l) = [Y_1(k,l), \cdots, Y_M(k,l)]^{\mathrm{T}}, \quad (26)$$

where the permutation-aligned target channels are $\boldsymbol{Y}_{A1}(k,l) = [Y_1(k,l), \cdots, Y_N(k,l)]^{\mathrm{T}}$ and the non-target channels are $\boldsymbol{Y}_{A2}(k,l) = [Y_{N+1}(k,l), \cdots, Y_M(k,l)]^{\mathrm{T}}$. Note that the order of non-target channels is irrelevant as the post-filtering will use the average PSD across all the non-target channels as an estimate of the noise PSD in the target channel (Eq. 27).

An example of reference-based permutation alignment is shown in Fig. 3(b). The permutation of the N reference channels is correctly aligned across frequencies, while the permutation of the M input channels is ambiguous. In each frequency bin, we detect N channels that are highly correlated with the reference channels, and align them according to the order of the reference channels. For instance, at frequency k_1, we choose $\Pi_M^{k_1} = [1, 3, 2, 4]$ as the new permutation maximizing the objective function (23). After permutation alignment, the target channels are extracted in the first N output channels with their permutation aligned.

The better separation results of $M \times M$ ICA and the better permutation results of $N \times N$ ICA allow the proposed reference-based alignment scheme to solve the target channel detection and permutation alignment problem simultaneously. The knowledge of the number of sources, N, and a robust permutation alignment algorithm for $N \times N$ ICA are crucial for the success of this scheme.

C. Noise PSD estimation and post-filtering

The signals in the non-target channels can provide a reference to estimate the noise components in the target channels (see Insight 3), because both can be seen as linear combination of late-reverberant image sources. However, these image sources typically undergo different spatial filtering and thus contribute different energy in each target and non-target channel. Deriving the relationship between noise components in the target channel and signals in the non-target channels is therefore a challenging task.

Since the noise components in the target channels and the signals in the non-target channels originate from the same N physical sources, they tend to occupy similar time-frequency bins. We thus propose to approximate the PSD of the noise in the target channel by averaging the PSDs of the signals across all non-target channels.

Let $S_m(k,l)$ and $V_m(k,l)$ be the target and noise components in the m-th target channel, respectively, and $Y_m(k,l) = S_m(k,l) + V_m(k,l)$. We estimate the PSD of V_m as

$$\hat{P}_{V_m}(k,l) = \frac{\sum_{j=N+1}^{M} |Y_j(k,l)|^2}{M - N}, \quad m = 1, \cdots, N. \quad (27)$$

With this noise PSD estimation, we can design a spectral post-filter that further suppresses the noise component in each target channel. For instance, the Wiener filter enhances the target channel as

$$\hat{S}_m(k,l) = G_m(k,l)Y_m(k,l), \quad (28)$$

where the spectral gain is computed from $\hat{P}_{V_m}(k,l)$ and $Y_m(k,l)$ [38]. Applying inverse STFT to $\hat{S}_1(k,l), \cdots, \hat{S}_N(k,l)$, we get the enhanced time-domain signals

$$\hat{\boldsymbol{s}}(n) = [\hat{s}_1(n), \cdots, \hat{s}_N(n)]^{\mathrm{T}}. \quad (29)$$

TABLE III
DECOMPOSITION OF THE MICROPHONE SIGNAL x_i WITH RESPECT TO s_j.

$x_i = x_{ij}^e + x_{ij}^l + x_{ij}^u = x_{ij}^d + x_{ij}^v$	the i-th microphone signal
$x_{ij} = x_{ij}^e + x_{ij}^l$	source component (early- and late-reverberant components)
$x_{ij}^u = \sum_{j' \neq j} x_{ij'}$	interference component
$x_{ij}^d = x_{ij}^e$	target component
$x_{ij}^v = x_{ij}^l + x_{ij}^u$	noise component

While Eq. 27 can only approximate the noise PSD in the target channel, it is useful for noise reduction. First, the noise components in the target channels are usually non-stationary and their energy is sparsely concentrated in the time-frequency domain. The knowledge of the locations of these dominant time-frequency bins would be valuable for noise suppression, even if their magnitudes are not accurately known. Second, this approximation tends to overestimate the noise PSD due to the inclusion of non-target sources into the averaging operation. The energy of non-target sources is usually higher than that of the noise components in the target channels, thus leading to an overestimate.

This overestimate leads to better noise reduction but might also lead to target signal cancellation, especially when the dominant time-frequency bins of the estimated noise are overlapped with those of the target sources. Thus, the trade-off between noise reduction and target signal cancellation depends on the energy of these non-target sources. For instance, when $M \gg N$ and most late-reverberant image sources extracted into non-target channels, a post-filter might be unnecessary.

VI. PERFORMANCE MEASURES

We evaluate the *source separation* performance in terms of SIR and the *dereverberation* effect in terms of early-late reverberation ratio (ELR). Moreover, we evaluate the *signal distortion* and the *global sound enhancement* in terms of Perceptual Evaluation of Speech Quality (PESQ). To this end, we first decompose the microphone signal into early-reverberant, late-reverberant, and interference components.

A. Signal decomposition

Assuming the original source, $s_j(n)$, and its corresponding components received by the microphones, $x_{ij}(n)$, to be known, we decompose the microphone signal $x_i(n) = \sum_{j'=1}^{N} x_{ij'}(n)$ into an early-reverberant component $x_{ij}^e(n)$, a late-reverberant component $x_{ij}^l(n)$ and an interference component $x_{ij}^u(n)$, with respect to each source s_j, i.e.

$$\begin{aligned} x_i(n) &= x_{ij}^e(n) + x_{ij}^l(n) + x_{ij}^u(n) \\ &= x_{ij}(n) + x_{ij}^u(n) = x_{ij}^d(n) + x_{ij}^v(n), \end{aligned} \tag{30}$$

where $x_{ij}(n) = x_{ij}^e(n) + x_{ij}^l(n)$, $x_{ij}^u(n) = \sum_{j' \neq j} x_{ij'}(n)$, and $x_i(n)$ can be decomposed into target component $x_{ij}^d(n) = x_{ij}^e(n)$ and noise component $x_{ij}^v(n) = x_{ij}^l(n) + x_{ij}^u(n)$ (see the summary in Table III).

We aim to extract the early-reverberant component of each source, $x_{ij}^e(n)$, which can be calculated by convolving the original source, $s_j(n)$, with an early-reverberant filter $\boldsymbol{h}_{ij}^e = [h_{ij}^e(1), \cdots, h_{ij}^e(L_e)]$, i.e.

$$x_{ij}^e(n) = h_{ij}^e(n) * s_j(n), \tag{31}$$

where the length of early reverberation L_e is chosen to be 64 ms (i.e. 1024 at the sampling rate 16 kHz). Usually, the early part of the reverberant signal (the first 50-100 ms after the direct sound) helps improve speech intelligibility [39]. The filter $\boldsymbol{h}_{ij}^e$ is computed via a projection procedure between $x_{ij}(n)$ and $s_j(n)$, which can be represented as [37]

$$\boldsymbol{h}_{ij}^e = \arg\min_{\boldsymbol{h}} \sum_n \left(x_{ij}(n) - h(n) * s_j(n) \right)^2. \tag{32}$$

Given an $M \times M$ demixing network $\boldsymbol{W}$, the i-th output channel is represented as $y_i(n) = \sum_{j=1}^{N} y_{ij}(n)$, where $y_{ij}(n) = \sum_{m=1}^{M} W_{im}(n) * x_{mj}(n)$ is the component of the source j in the output channel i. Similarly, the i-th output for a post-filter $\boldsymbol{G}$ is represented as $\hat{s}_i(n) = \sum_{j=1}^{N} \hat{s}_{ij}(n)$ with $\hat{s}_{ij}(n)$ being the component of the source j in the output channel i. Similarly to $x_i(n)$, the source separation output $y_i(n)$ and the post-filtering output $\hat{s}_i(n)$ can also be decomposed into early-reverberant, late-reverberant and interference components, i.e.

$$y_i(n) = y_{ij}^e(n) + y_{ij}^l(n) + y_{ij}^u(n), \tag{33}$$

$$\hat{s}_i(n) = \hat{s}_{ij}^e(n) + \hat{s}_{ij}^l(n) + \hat{s}_{ij}^u(n). \tag{34}$$

B. The measures

We use SIR to evaluate the source separation performance. Let $\mathcal{P}\{y_{ij}\} = \sum_n y_{ij}^2(n)$ be the energy of a sequence $y_{ij}(n)$. For $\boldsymbol{W}$, the SIR of the source j in the output channel i is

$$\text{SIR}_{ij}(\boldsymbol{W}) = \frac{\mathcal{P}\{y_{ij}\}}{\sum_{j' \neq j} \mathcal{P}\{y_{ij'}\}}. \tag{35}$$

The SIR of source j is then the maximum SIR among all the output channels:

$$\text{SIR}_j(\boldsymbol{W}) = \text{SIR}_{\mathcal{I}_j j}(\boldsymbol{W}), \tag{36}$$

where $\mathcal{I}_j = \max_{i \in [1,M]} \{\text{SIR}_{ij}(\boldsymbol{W})\}$ is the index of the channel where the source j is dominant. The overall SIR obtained by $\boldsymbol{W}$ is defined as the average SIR among all the sources: $\text{SIR}(\boldsymbol{W}) = \frac{1}{N} \sum_{j=1}^{N} \text{SIR}_j(\boldsymbol{W})$.

We use ELR to evaluate the dereverberation performance. For $\boldsymbol{W}$, the ELR of the source j is defined as

$$\text{ELR}_j(\boldsymbol{W}) = \frac{\mathcal{P}\{y_{\mathcal{I}_j}^e\}}{\mathcal{P}\{y_{\mathcal{I}_j}^l\}}. \tag{37}$$

The overall ELR obtained by $\boldsymbol{W}$ is defined as the average among all the sources, i.e. $\text{ELR}(\boldsymbol{W}) = \frac{1}{N} \sum_{j=1}^{N} \text{ELR}_j(\boldsymbol{W})$.

We use PESQ to evaluate the signal distortion (i.e. DPESQ) and the global sound enhancement (i.e. GPESQ). PESQ $\in [0, 4.5]$ is a widely used measure to assess the overall quality of the processed speech, $s_e(n)$, relative to the referenced clean speech, $s_o(n)$ [40]. The higher PESQ, the better the speech quality. We represent PESQ as $\mathcal{Q}\{s_e, s_o\}$.

Let source j have its early-reverberant component in the first channel as $x_{1j}^e(n)$, and is extracted in the $\mathcal{I}_j$-th channel

$y_{\mathcal{I}_j}(n)$ with the corresponding component being $y_{\mathcal{I}_j j}(n)$. The distortion measure DPESQ is defined as

$$\text{DPESQ}_j(\boldsymbol{W}) = \mathcal{Q}\{y_{\mathcal{I}_j j}, x^e_{1j}\}, \tag{38}$$

and the overall DPESQ obtained by $\boldsymbol{W}$ is defined as the average DPESQ value among all the sources, i.e. $\text{DPESQ}(\boldsymbol{W}) = \frac{1}{N}\sum_{j=1}^{N}\text{DPESQ}_j(\boldsymbol{W})$. The global sound enhancement measure GPESQ is defined as

$$\text{GPESQ}_j(\boldsymbol{W}) = \mathcal{Q}\{y_{\mathcal{I}_j}, x^e_{1j}\}, \tag{39}$$

and the overall GPESQ value obtained by $\boldsymbol{W}$ is defined as $\text{GPESQ}(\boldsymbol{W}) = \frac{1}{N}\sum_{j=1}^{N}\text{GPESQ}_j(\boldsymbol{W})$.

For a post-filter $\boldsymbol{G}$, the SIR and ELR can be calculated similarly as Eq. 36 and Eq. 37. DPESQ is calculated by comparing the early-reverberant component in the spatial filter output, $y^e_{\mathcal{I}_j j}(n)$, and the target source component in the post-filter output, $\hat{s}_{\mathcal{I}_j j}(n)$:

$$\text{DPESQ}_j(\boldsymbol{G}) = \mathcal{Q}\{\hat{s}_{\mathcal{I}_j j}, y^e_{\mathcal{I}_j j}\}. \tag{40}$$

GDESQ is calculated by comparing $y^e_{\mathcal{I}_j j}(n)$ with the post-filter output, $\hat{s}_{\mathcal{I}_j}(n)$:

$$\text{GPESQ}_j(\boldsymbol{G}) = \mathcal{Q}\{\hat{s}_{\mathcal{I}_j}, y^e_{\mathcal{I}_j j}\}. \tag{41}$$

VII. The Advantages of PBSS: Validation

In this section we verify the independence of the image sources of a reverberant sound and the three insights of PBSS presented in Sec. IV. The evaluation data is simulated with the image-source model [30] in a 7×7×4m enclosure. Four sound sources (20 s by 2 male and 2 female speakers with sampling rate 16 kHz) are placed in the center of the room, equally distributed along a circle with 0.5 m radius. Sixteen microphones are placed around the sources, equally distributed along a circle with radius 2 m. The reverberation time (RT) varies from 400 to 1000 ms, with 200 ms step. The microphone signals are obtained by convolving the sound sources with the room impulse responses from the source location to the microphones. We assume that the signals are synchronously sampled and the permutation ambiguity are solved by referring to clean source signals [10]. The STFT frame lengths are $N_{F1} = 4096$ for spatial filtering and $N_{F2} = 512$ for post-filtering, both with half overlap. To bridge these two STFT lengths, we transform the spatial filtering outputs, N_{F1}, into the time domain and then reanalyze them into the STFT domain, N_{F2}, as the input to the post-filter.

To test the independence of the image sources of a reverberant sound, we select a speech source recorded by four microphones at reverberation time 800 ms. We apply an 4×4 ICA at each frequency bin of the signal transformed into the STFT domain, generating four outputs. Fig. 4(a) shows the amplitudes of the original signal and a microphone signal at 600 Hz, which show that the microphone signal can be interpreted as sum of delayed versions of the original signal. Fig. 4(b) shows the amplitudes of four ICA outputs, which resemble the original source signal but with different delays. These ICA outputs contribute to the microphone signal via the mixing matrix estimated by ICA, and thus can be interpreted as virtual sound sources emitting sounds from different spatial locations, e.g. the first ICA output represents the early-reverberant component of the original sound source and the remaining three represent late-reverberant components. While these virtual sources originate from the same physical source, they each present higher non-Gaussianity than the microphone signal and thus can be separated from the microphone signal with ICA, as observed in Fig. 4(b). For instance, the kurtosis values (a measure of non-Gaussianity [9]) are 15.3 and 8.6 for the original and the microphone signals, and are 17.1, 15.4, 12.9, 13.4 for the four ICA outputs, respectively.

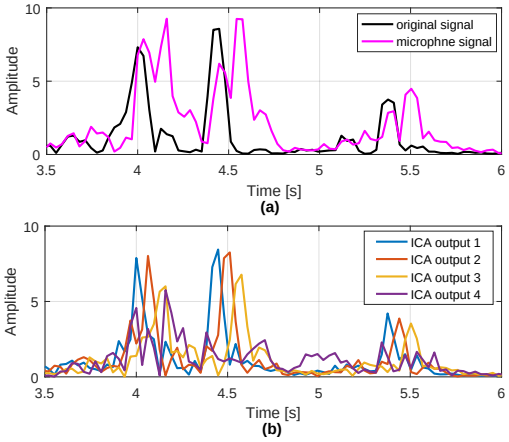


Fig. 4. Applying a 4×4 ICA to one sound source recorded at four microphones in a reverberant environment. (a) The amplitudes of the original signal and the reverberant microphone signal at 600 Hz. (b) The amplitudes of the four ICA outputs at 600 Hz.

Next, we validate the performance degradation with reverberation, the performance improvement (in terms of both separation and late reverberation suppression) with the number of microphones, and the effectiveness of the post-filter. The source separation (SIR), late reverberation suppression (ELR), and global performance (GPESQ) obtained by the PBSS spatial filter are shown in Fig. 5(a). The input SIRs in different reverberant scenarios are all around -4.5 dB. When $M = 4$, PBSS improves the SIR but the performance degrades as the reverberation density increases. As M increases, the SIR performance improves quickly and monotonically for $4 \leq M \leq 8$, then improves slowly for $M > 8$ before becoming saturated at $M = 14$. When the number of microphones increases, the SIR improves considerably, from 6 dB with $M = 4$ to 16 dB with $M = 14$ when RT = 800 ms. The ELR of the input microphone signal drops, as expected, when the reverberation density increases. When $M = 4$, PBSS improves ELR only slightly. As M increases, ELR rises quickly and monotonically for $4 \leq M \leq 8$, and then rises slowly before saturating at $M = 14$. At RT = 800 ms, PBSS improves ELR by up to 2 dB. The variation of GPESQ with respect to RT and M is similar to that of SIR. The GPESQs of the input microphone signal in different reverberant scenarios are all below 1.5. When $M = 4$, PBSS improves GPESQ but the performance degrades as RT increases. As M increases, GPESQ rises quickly and monotonically for $4 \leq M \leq 8$, then rises slowly before saturating at $M = 14$. At RT = 800 ms,

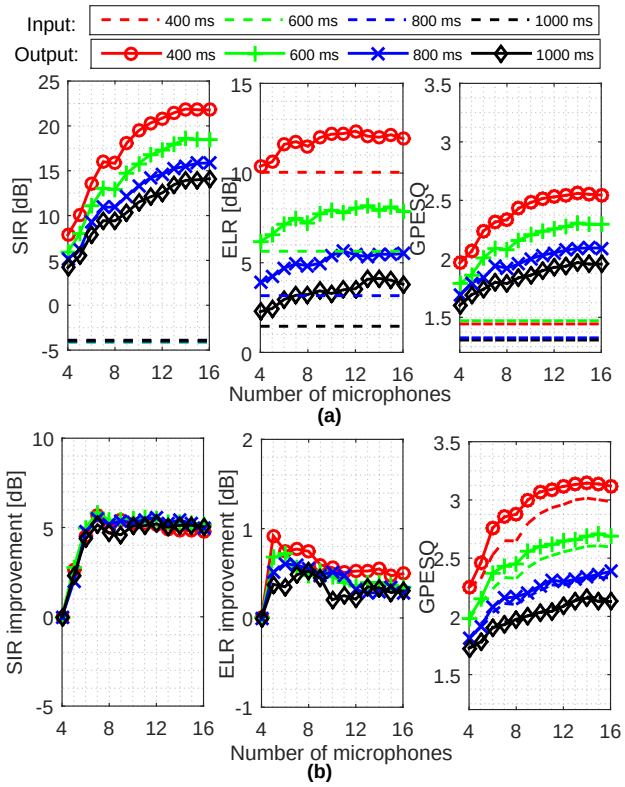


Fig. 5. Performance evaluation of pseudo-determined BSS and the post-filter for 4 sources recorded with a varying number of microphones from 4 to 16 in a scenario with a varying reverberant time from 400 ms to 1000 ms. (a) SIR, ELR and GPESQ obtained by the source separation filter. (b) SIR improvement, ELR improvement, and GPESQ obtained by applying a post-filter to the source separation output.

PBSS improves GPESQ from 1.6 with $M = 4$ to 2.1 with $M = 14$. In summary, the performance of PBSS improves in various reverberant scenarios as M increases, achieving both source separation and late-reverberation suppression.

The performance improvement in terms of SIR, ELR and GPESQ obtained by applying the post-filter to the spatial filtering output are shown in Fig. 5(b). The improvement of the post-filter separation output in terms of SIR remains similar in all reverberant scenarios. The amount of improvement rises quickly from 0 to 5 dB for $4 \leq M \leq 8$, and then saturates afterwards. The post-filter also improves the ELR of the separation output. As M increases, the amount of ELR improvement rises quickly when $4 \leq M \leq 8$, but then drops slowly afterwards. The post-filter improves the ELR more effectively at lower reverberation densities, e.g. by up to 1 dB for RT = 400 ms and up to 0.5 dB for RT = 1000 ms. The GPESQ values of the spatial filtering output and the post-filtering output both improve with M, rising quickly for $4 \leq M \leq 8$ and then slowly before saturation at $M = 14$. The post-filter improves the GPESQ of the spatial filter slightly (by up to 0.1) when RT $\leq$ 600 ms, but performs similarly to the latter when RT $\geq$ 800 ms.

In summary, the post-filter can improve the SIR of the separation output effectively and can also improve the ELR as M increases. The turning point at around $M = 8$ is possibly due to the influence of non-target sources. As M increases

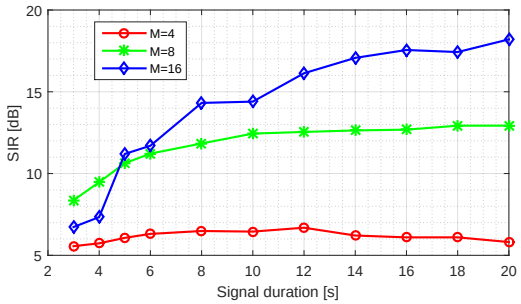


Fig. 6. SIR performance versus signal duration for pseudo-determined BSS with different number of microphones. The reverberation time is 600 s.

from 4 to 8, some high-energy late-reverberant components are sequentially extracted into non-target channels. Using these signals as a reference may help suppress the interference and reverberation residuals in the target channels effectively. As M further increases, more late-reverberant components are extracted as non-target sources, and correspondingly, the energy of the noise in the target channels becomes smaller. The additional noise reduction achieved by increasing M thus becomes less pronounced.

Finally, Fig. 6 shows the impact on PBSS (in terms of SIR) of the signal duration for a varying $M \in \{4, 8, 16\}$ with reverberation time 600 ms. When $M = 4$, SIR does not vary much when the signal duration exceeds 6 s. When $M = 8$, SIR improves with the increase of the signal duration, and saturates with signal durations longer than 10 s. When $M = 16$, SIR improves until the signal duration reaches 16 s. When the signal duration is shorter than 6 s, SIR for $M = 16$ is even lower than that for $M = 8$. This shows that as M increases, the $M \times M$ ICA requires longer data to converge. However, for the same signal duration, the larger M, the higher SIR.

VIII. Real-data Experiments

To evaluate and compare the performance of source separation algorithms we use the data of SISEC 2015 [42]. The development dataset of "asynchronous recordings of speech mixtures" contains eight-channel recording by four independent portable voice recorders (each with two microphones). The sampling rate mismatch of the recording devices is within 1 Hz at the nominal sampling rate 16 kHz. The speech sounds from four loudspeakers are individually recorded by the recording devices and then added together to get the mixed signal. The duration of the signal is 20 s. The reverberation time is around 800 ms. The loudspeakers are set around a table, on which the recorders are set. The locations of the loudspeakers and recorders are unknown.

A. Methods Under Analysis

We compare the proposed $M \times M$ ICA with reference-based permutation alignment (ROBSS) with the following source separation algorithms: NDBSS: $N \times N$ ICA with clustering-based permutation alignment [10]; MDBSS: $M \times M$ ICA with clustering-based permutation alignment [13]; BFBSS: fixed delay-and-sum beamformer followed by NDBSS [27];

SSBSS: subspace based dimensionality reduction followed by NDBSS [24]; and MOBSS: $M \times M$ ICA with source merging-based permutation alignment [13]. We also consider three post-filters applied to the ROBSS outputs, namely Post, the proposed noise PSD estimation based on the signal from non-target channels; UMMSE, a state-of-art single-channel noise PSD estimator [41]); and Benchmark, noise PSD estimation assuming the interference signals to be known (i.e. known $P_{y_{ij}^u}$). These algorithms are applied to microphone signals synchronized as in Eq. 18. We also apply source separation to the original microphone signals which are asynchronously sampled, namely applying NDBSS to the original microphone signals (AsyBSS).

All the spatial filtering algorithms use a STFT frame length of $N_{F1} = 4096$ with half overlap. All the spectral post-filtering algorithms use a STFT frame length of $N_{F1} = 512$, with half overlap and set the minimum gain to $G_{\min} = 0.3$. NDBSS uses a number of microphones equal to that of the sources from all the microphones. We choose a combination that has the highest average SIR. For BFBSS, we estimate the delays from each source to the microphones using the individual recording of each source, i.e. x_{ij}. For MOBSS, as the microphone locations are unknown, we only use the sparseness measure, the time activity measure and the spectral likeliness measure to detect the association between the ICA outputs [13]. After source merging, we retain as output the N channels with the highest energy.

B. Discussion

Fig. 7 depicts the SIR maps obtained by various source separation algorithms (MDBSS, NDBSS, MOBSS, ROBSS, and Post) for an 8×3 mixture (M=8 and N=3). Due to the challenging permutation ambiguities in the case of $M>N$, MDBSS can only partly recover the permutation of the separated signals. In the M outputs of MDBSS s_1 and s_2 each dominates only one channel, i.e. $y_{\text{MDBSS-1}}$ and $y_{\text{MDBSS-8}}$, respectively; s_3 dominates two channels $y_{\text{MDBSS-4}}$ and $y_{\text{MDBSS-7}}$, which occupy the low and high frequency bands of s_3, respectively (as shown Fig. 8). MOBSS solves this problem by detecting the association between the M outputs and merge the channels that come from the same source, e.g. merging $y_{\text{MDBSS-4}}$ and $y_{\text{MDBSS-7}}$ into a new channel $y_{\text{MOBSS-3}}$. However, while the merging procedure can reconstruct s_3 properly, it also merges the noise components contained in $y_{\text{MDBSS-4}}$ and $y_{\text{MDBSS-7}}$ into $y_{\text{MOBSS-3}}$, resulting in a lower SIR. With less challenging permutation ambiguities in the case of M=N, NDBSS can recover the permutation of the separated signals. In the N outputs of NDBSS, each source dominates only one channel but with a much lower SIR than MDBSS. Using the NDBSS outputs as a reference, ROBSS realigns the permutation of the MDBSS outputs, extracting the target sources into the first N channels and leaving the residual noise to the remaining $M-N$ channels. This results in a higher SIR at the first N channels than NDBSS and MOBSS. Using the remaining channels $y_{\text{ROBSS-4}} - y_{\text{ROBSS-8}}$ as a reference, Post estimates the noise PSD in $y_{\text{ROBSS-1}} - y_{\text{ROBSS-3}}$ and then implements a spectral filter which further improves the SIR in these channels.

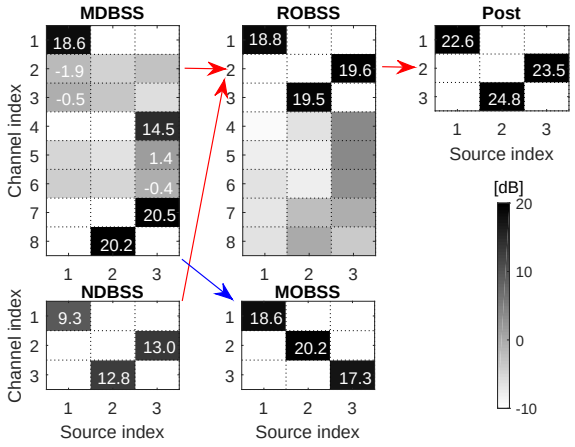


Fig. 7. SIR maps (in dB) obtained by various source separation algorithms (MDBSS, NDBSS, MOBSS, ROBSS, Post) for an 8×3 mixture (M=8, N=3). In each output channel only the highest SIR is indicated.

Fig. 8 depicts the time-frequency spectra of the output signals by MDBSS, NDBSS and ROBSS. For convenience of display, only the signals during 5-10 s are shown. In the first row, the permutation ambiguities are not completely solved by MDBSS, where s_1 is extracted into $y_{\text{MDBSS-1}}$, s_2 is extracted into $y_{\text{MDBSS-3}}$ and $y_{\text{MDBSS-8}}$, and s_3 is exacted into $y_{\text{MDBSS-4}}$ and $y_{\text{MDBSS-7}}$. In the second row, the permutation ambiguities are well solved by NDBSS, where the three sources are extracted into three output channels, respectively. In the third row, the permutation ambiguities are also solved by ROBSS, where the first three output channels contain the three sources and the remaining five channels contain only noise. It is additionally observed that the first three ROBSS outputs contain less residual noise than the corresponding NDBSS outputs.

Fig. 9 depicts the time-frequency PSDs of the intermediate results obtained by two post-filters Post and UMMSE, using $y_{\text{ROBSS-3}}$ (which is dominated by s_2) as an example. Similarly to Eq. 30, $y_{\text{ROBSS-3}}$ can be decomposed into interference y_{32}^u, late reverberation y_{32}^l and early reverberation y_{32}^e, as shown in Fig. 9(b)-(d), respectively. We aim to extract y_{32}^e as a target by suppressing the noise from y_{32}^l and y_{32}^v. Fig. 9(e) depicts the estimated noise PSD by applying a single-channel estimator UMMSE to $y_{\text{ROBSS-3}}$ directly. Since the noise components y_{32}^l and y_{32}^v are both nonstantionary, UMMSE performs poorly in distinguishing them from the target component y_{32}^e. It can be clearly observed that the estimated PSD deviates from the true value. Fig. 9(f) depicts the estimated noise PSD by Post. For convenience of comparison, we decompose the estimated noise PSD into the interference component P_{vv} and the source component P_{vs} (Fig. 9(g)-(h)), corresponding to y_{32}^v and y_{32}^l, respectively. Comparing Fig. 9(b) and Fig. 9(g), P_{vv} can well capture the locations of the most dominant time-frequency bins in y_{32}^v. Similarly in Fig. 9(c) and Fig. 9(h), P_{vs} can well capture the locations of the most dominant time-frequency bins in y_{32}^l. Fig. 9(i) and Fig. 9(j) depict the noise reduction results by Post and UMMSE, respectively. Post achieves a much better noise reduction performance than UMMSE, as supported by their SIR values 24.8 dB and 21.4 dB, respectively. Post and UMMSE achieve similar signal distortion, with DPESQ

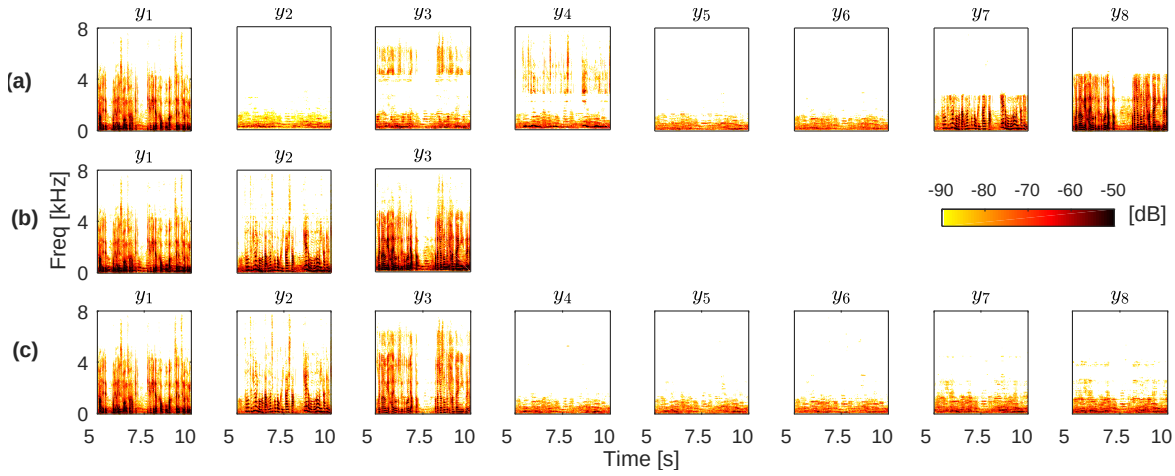


Fig. 8. Time-frequency plots of the output signals by (a) MDBSS, (b) NDBSS, and (c) ROBSS for an 8×3 mixture (M=8, N=3).

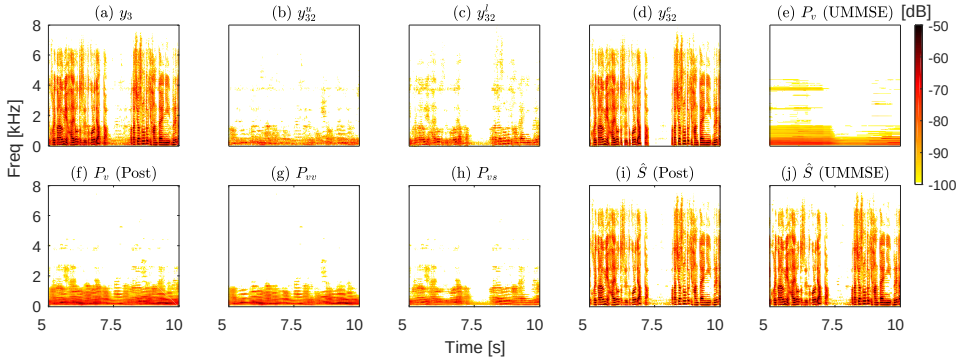


Fig. 9. Time-frequency plots of the intermediate processing results by two post-filters Post and UMMSE for an 8×3 mixture (M=8, N=3). We use the third ROBSS output y_3 as an example, which is dominated by s_2. (a) ROBSS output y_3; (b)-(d) The interference y^u_{32}, late reverberation y^l_{32}, and early reverberation y^e_{32} for the source s_2; (f)-(i) The estimated noise PSD and its interference component P_{vv} and late-reverberant component P_{vs}, and the noise reduction result by Post; (e)(j) The estimated noise PSD and the noise reduction result by UMMSE.

values 2.64 and 2.65, respectively.

We compare the source separation (SIR), signal distortion (DPESQ), and global performance (GPESQ) by the considered algorithms for asynchronous recordings with a varying number of sources $N \in \{2, 3, 4\}$. Fig. 10 depicts the SIR and PESQ values achieved by various algorithms including the input signal (Input), DBSS before and after synchronization (AsyBSS and NDBSS), and four OBSS algorithms (BFBSS, SSBSS, MOBSS and the proposed ROBSS).

Regarding source separation in Fig. 10(a), the performance of the considered algorithms can be obviously ranked as Input $<$ AsyBSS $<$ BFBSS $<$ NDBSS $<$ SSBSS $<$ MOBSS $<$ ROBSS. Based on (36) the SIR of each source is determined as the maximum value among all the output channels. The observation that the average SIR of Input is higher than 0 dB implies that for each sound source there is a recording device placed closer to it than other devices. AsyBSS can improve the SIR of the input signal even in the case of sampling rate mismatch. After synchronizing the sampling of independent recordings, NDBSS achieves a higher SIR than AsyBSS especially when N is large. BFDBSS does not outperform NDBSS as expected, possibly because the delay-and-sum beamformer does not enhance the source signals effectively in the case of non-uniform response of each recording device. ROBSS, MOBSS and SSBSS can all improve the SIR performance of NDBSS remarkably. ROBSS performs the best, followed by MOBSS and SSBSS. Overall, ROBSS can improve the SIR of Input by around 20 dB and improve NDBSS by around 10 dB in all evaluation scenarios.

Regarding the signal distortion performance (DPESQ) in Fig. 10(b), all the algorithms except SSBSS perform similarly. ROBSS achieves a higher DPESQ value than MOBSS in all evaluation scenarios. SSBSS achieves the lowest DPESQ value, because the subspace-based dimensionality reduction may distort the source signals significantly. Regarding the global performance (GPESQ) in Fig. 10(c), ROBSS performs the best among all the algorithms. NDBSS outperforms AsyBSS especially when $N \geq 3$. ROBSS achieves a higher GPESQ value than MOBSS. Over all, ROBSS improves the GPESQ of NDBSS by around 0.3, and improves the GPESQ of Input by around 1 in all evaluation scenarios.

Fig. 11 depicts the evaluation results achieved by applying three post-filters Post, UMMSE and Benchmark to the ROBSS outputs. In Fig. 11(a), Post achieves a higher SIR than UMMSE because it can estimate the PSD of the interference more accurately. UMMSE underestimates the PSD

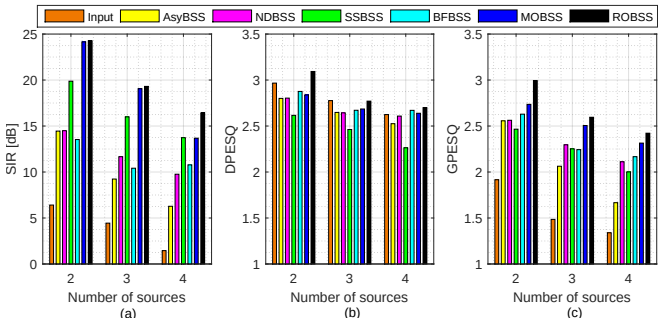


Fig. 10. Performance comparison: source separation (SIR), signal distortion (DPESQ), and global performance (GPESQ) by the considered source separation algorithms.

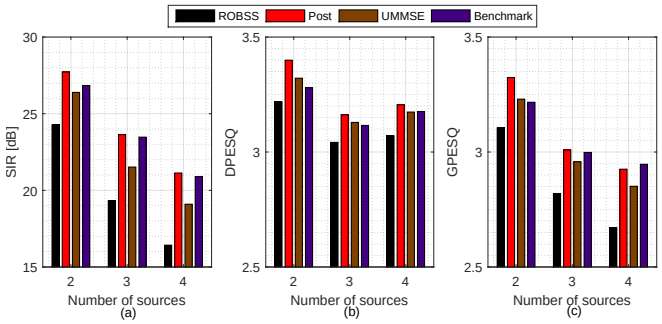


Fig. 11. Performance comparison: source separation (SIR), signal distortion (DPESQ), and global performance (GPESQ) by three post-filtering algorithms (`Post`, `UMMSE` and `Benchmark`). A demo with the audio signals corresponding to Fig. 10 and Fig. 11 is available [47].

of the interference, and thus performs worse than `Post`. `Post` performs similarly to `Benchmark`, which assumes the interference to be known. `Post` can improve the SIR of `ROBSS` by around 5 dB in all evaluation scenarios. In Fig. 11(b), `Post` achieves the highest DPESQ value among all the algorithms for $N = 2$, and achieves similar DPESQ values as another two post-filters for $N \geq 3$. `Post` achieves a higher DPESQ value than `ROBSS` due to its dereverberation effect. For the global measure GPESQ in Fig. 11(c), `Post` performs the best for $N = 2$ and performs similarly to `Benchmark` when $N \geq 3$. `Post` outperforms `UMMSE`, and can improve the GPESQ of `ROBSS` by around 0.2 in all evaluation scenarios.

Finally, we compare our SISEC processing results with the ones obtained from another research group, who performed dimensionality reduction first and then applied IVA to the SISEC data [19], [35]. We evaluate the submitted results (Development2 - asynchrec realmix), which are downloaded from the SISEC website [42], with our own object measures. As shown in Table IV, the propose method clearly outperforms the competing method in terms of SIR and GPESQ.

C. Computation time

Table V lists the computation time of each algorithm block when processing a sequence with 8 microphones and 4 sources. The signal duration is 20 s with sampling rate 16 kHz. We run Matlab code of the proposed algorithm on an Intel CPU i7@3.2 GHz with 16 GB RAM.

TABLE IV
PERFORMANCE COMPARISON OF TWO SISEC SUBMISSIONS

Method	Ref	N	SIR (dB)	GPESQ
ROBSS + Post	proposed	3	20.2	2.6
		4	18.8	2.5
Dimensionality reduction + IVA	[19], [35]	3	13.0	2.2
		4	1.2	1.7

TABLE V
COMPUTATION TIME (SECOND) OF THE PROPOSED METHOD WITH 8 MICROPHONES AND 4 SOURCES. THE SIGNAL DURATION IS 20 S WITH SAMPLING RATE 16 KHZ. KEY: PA - PERMUTATION ALIGNMENT.

alignment & sync	$N \times N$ ICA	blind PA	$M \times M$ ICA	reference-based PA	post-filter
15.1	25.9	14.3	35.1	15.4	2.2

IX. CONCLUSION

We proposed a pseudo-determined mixture model that makes it possible to apply an $M \times M$ ICA directly to an $M \times N$ mixture. We also developed an over-determined BSS system that can be applied to asynchronous recordings from independent devices of an ad-hoc network, such as crowdsourced audio data collected during an event. The proposed approach includes synchronization, pseudo-determined BSS, and post-filtering. Synchronization allows the inclusion of additional independent recording devices for an over-determined separation. The pseudo-determined BSS improves performance when the number of microphones increases. The permutation ambiguity problem is solved with a reference-based permutation alignment scheme. The post-filtering exploits the abundant information from the sensors to further enhance the separated signals. Experimental results show that these steps incrementally improve the source separation performance of the input signals and that dereverberation is obtained as by-product.

There are several directions for future research. The reference-based permutation alignment scheme requires the number of sources N to be known in order to apply a regular $N \times N$ DBSS. When the value of N unavailable, it could be estimated with a source enumeration method (e.g. [43], [44]). The permutation alignment result of the regular DBSS is crucial to the reference-based scheme and could be improved with two strategies: exploiting the information from more sensors, as done by some OBSS algorithms [12], [29]; or considering a time-domain DBSS algorithm, which usually has worse separation performance but is free from permutation ambiguities [45]. Finally, the noise PSD estimation in the post-filtering block employs a simple averaging scheme: exploiting the demixing filter coefficients could further improve the noise PSD estimation performance [38], [46].

REFERENCES

[1] 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.

[2] 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.

[3] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Process. Mag.*, vol. 33, no. 4, pp. 14-29, Apr. 2016.

[4] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, China, 2003, pp. 840-843.

[5] 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.

[6] M. Kim and P. Smaragdis, "Collaborative audio enhancement: crowdsourced audio recording," in *Proc. Neural Inf. Process. Sys.*, Montreal, Canada, 2014, pp. 1-9.

[7] K. Ochi, N. Ono, S. Miyabe, and S. Makino, "Multi-talker speech recognition based on blind source separation with ad hoc microphone array using smartphones and cloud storage," in *Proc. Interspeech*, San Francisco, USA, 2016, pp. 3369-3373.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] L. Wang, J. Reiss, and A. Cavallaro, "Over-determined source separation and localization using distributed microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1573-1588, Sep. 2016.

[14] S. C. Douglas, M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation", in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Honolulu, USA, 2007, pp. 637-640.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.*, vol. 2010, pp. 1-13, 2010.

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

[29] 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.

[30] 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.

[31] 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. 2013, pp. 1157-1166, 2003.

[32] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech, *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 109-116, Mar. 2003.

[33] N. Q. K. Duong, C. Howson, and Y. Legallais, "Fast second screen TV synchronization combining audio fingerprint technique and generalized cross correlation," in *Proc. IEEE Int. Conf. Consum. Electron.*, Berlin, Germany, 2012, pp. 241-244.

[34] 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.

[35] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Process.*, vol. 107, pp. 185-196, Feb. 2015.

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

[37] 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.

[38] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using MaxNSR blocking matrix," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1493-1508, Sep. 2015.

[39] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, no.6, pp. 3233-3244, Jun. 2003.

[40] H. Yi and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229-238, Jan. 2008.

[41] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383-1393, May 2012.

[42] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Variable Analysis Signal Separation*, Liberec, Czech, 2015, pp. 387-395.

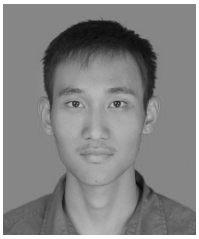
[43] Z. Lu, and A. M. Zoubir, "Flexible detection criterion for source enumeration in array processing," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1303-1314, Mar. 2013.

[44] 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.

[45] S. C. Douglas, M. Gupta, H. Sawada, and S. Makino, "Spatio-temporal FastICA algorithms for the blind separation of convolutive mixtures," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 5, pp. 1511-1520, Jul. 2007.

[46] Y. Zheng, K. Reindl, and W. Kellermann, "Analysis of dual-channel ICA-based blocking matrix for improved noise estimation," *EURASIP J. Adv. Signal Process.*, vol. 2014, pp. 1-24, 2014.

[47] http://www.eecs.qmul.ac.uk/∼andrea/robss.html



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.

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 IEEE/ACM ICDSC 2009, BMVC 2009, M2SFA2 2008, SSPE 2007, and IEEE AVSS 2007. He was Technical Program Chair of IEEE AVSS 2011, EUSIPCO 2008, and WIAMIS 2010. He received the Royal Academy of Engineering Teaching Prize in 2007, three Student Paper Awards at IEEE ICASSP in 2005, 2007, and 2009, respectively, and the Best Paper Award at IEEE AVSS 2009.