Vietnam J Comput Sci
DOI 10.1007/s40595-016-0085-x

CrossMark

REGULAR PAPER

Source separation employing beamforming and SRP-PHAT localization in three-speaker room environments

Hai Quang Hong Dam[1] · **Sven Nordholm**[2]

Received: 24 March 2016 / Accepted: 22 September 2016

Abstract This paper presents a new blind speech separation algorithm using beamforming technique that is capable of extracting each individual speech signal from a mixture of three speech sources in a room. The speech separation algorithm utilizes the steered response power phase transform for obtaining a localization estimate for each individual speech source in the frequency domain. Based on those estimates each desired speech signal is extracted from the speech mixture using an optimal beamforming technique. To solve the permutation problem, a permutation alignment algorithm based on the mutual output correlation is employed to group the output signals into the correct sources from each frequency bin. Evaluations using real speech recordings in a room environment show that the proposed blind speech separation algorithm offers high interference suppression level whilst maintaining low distortion level for each desired signal.

Keywords Blind speech separation · SRP-PHAT · Beamformer

1 Introduction

Over the last 10–15 years research in machine interfaces for voice pick-up in reverberant and noisy environments has been very actively conducted using multi-channel systems like microphone arrays [1–4]. Multi-channel techniques have been useful in many applications such as hearing aids, hands-free communication, robotics, audio and video conference systems, and speech recognition [1,2,5,6]. One of the most popular techniques applied to multi-microphone systems is the optimal beamforming technique [1]. Optimal beamformers are formulated to exploit spatial information of desired and undesired signals in such a way that the desired one is extracted and undesired signals are suppressed [1,2]. Many methods have been proposed for determining the location of the desired source such as predefined well-determined array geometry combined with source localization [7,8], a calibration method using training samples of pre-recording desired and undesired sources [9,10]. Based on this information, optimal beamformers are designed using the spatial information to suppress the contribution of all undesired signals while reserving the contribution of the desired signal [1,11,12]. Specifically, the optimal beamformer weights are calculated using knowledge about the location of the target signal and array geometry. It is also possible to obtain estimates of speech and noise correlation matrices. These estimates are then used to form the optimal beamformer weights; for this method to be efficient a priori knowledge about the statistical characteristics of the noise is necessary. When the background noise is stationary over the measurement period either a voice activity detector (VAD) [2] estimate or a relative transfer function (RTF) estimate can be found [13]. Either of these estimates can be used to form optimal beamformers [2]. This leads us to a more general case where the spatial knowledge is not known a priori and the observed mixture signals are the only available information to be used for speech separation and noise suppression. In this case, blind source separation (BSS) techniques can be deployed for separating the different sound sources. Many blind source separation techniques using microphone array have been proposed for speech sep-

✉ Hai Quang Hong Dam
damhai@uit.edu.vn

Sven Nordholm
S.Nordholm@curtin.edu.au

[1] University of Information Technology, Ho Chi Minh City, Vietnam

[2] Curtin University of Technology, Perth, Australia

Published online: 06 October 2016

aration in both time domain and frequency domain. Some prominent BSS techniques for speech separation are independent component analysis (ICA), maximum likelihood, second-order gradient, and kurtosis maximization [14–18]. Most of the BSS techniques are based on either statistical independence or non-stationarity of the different input sources in the observed signal.

Speech separation in cocktail party or multiple-speaker environment is one of the significant problems in speech enhancement research. It occurs when the observed signals are obtained from several speakers in different spatial locations. Here, the spatial separation of speech sources is very important for speech separation due to the fact that all speech signals have the same spectral characteristic. We can categorize two different cases:

1. When the sources' spatial information is available, many separation techniques such as steering beamforming, optimum beamforming, and post-filtering have been proposed [3,4,6,10,19]. In [19], we introduced a post-filtering method which is implemented after an optimum beamformer to extract the desired speech source from a mixture of signals in multiple speakers environments. However, the source spatial information in those studies was obtained using a calibration method.
2. When the sources' spatial information is not available then blind separation techniques in a multiple-speaker environment need to be employed. For this scenario, a number of different BSS techniques have been proposed for the case of two speech sources in both time domain and time–frequency domain [4,18,20–22]. When the number of speech sources is more than two, the blind signal separation is more of a complicated and computational intense problem [23–25]. For this case, popular blind separation techniques are conducted to extract the desired source signal by finding a separating vector that maximizes the deterministic character (such as non-Gaussianity in ICA technique) of the extracted source signals [4,24,26,27].

In this paper, a blind signal separation method is proposed which estimates the source spatial information without having prior knowledge about the spatial location of speech sources in three-speaker environments. Once the source spatial information is estimated, it is used to design optimum beamformers for extracting speech sources from the observed signal. As such, the source spatial information estimation is performed in the frequency domain without having prior knowledge about the spatial location of the speech sources. Here, a spatial localization technique employing steered response power phase transform (SRP-PHAT) is proposed for estimating each source's spatial information based on the observed signal. The SRP-PHAT localization employs cross-correlation and phase transform weighting of the received signals from all microphone pairs in the array [28]. From the SRP-PHAT estimates, the proposed spatial localization technique calculates the spatial information of three speech sources from the observed signal. Based on the spatial information of the three speech sources, an optimum beamformer is proposed for extraction of each individual speech source from the observed signal. A permutation alignment is used for grouping each extracted signal into the correct source output before transforming them into the time domain. The performance of the proposed algorithm shows that the proposed algorithm offers a good interference suppression level while maintaining low speech distortion.

The paper is organized as follows: Sect. 2 outlines the problem formulation and details the signal model. In Sect. 3, the spatial localization method is derived and discussed in detail. Section 4 provides the details and derivation of the optimum beamforming technique. Section 5 discusses the method used for permutation alignment. In Sect. 6, the experimental results are presented and discussed. Finally, Sect. 7 summarizes the paper.

2 Problem formulation

Consider a linear microphone array, according to Fig. 1, consisting of L microphones and observed mixture signals $\mathbf{x}(n)$. The observed signals are a speech mixture from three speakers sitting in front of the microphones. The observed sampled signal $\mathbf{x}(n)$ at one time instant is an $L \times 1$ vector, which can be expressed as

$$\mathbf{x}(n) = \mathbf{s}_1(n) + \mathbf{s}_2(n) + \mathbf{s}_3(n) \tag{1}$$

where $\mathbf{s}_1(n)$, $\mathbf{s}_2(n)$ and $\mathbf{s}_3(n)$ are the received signals from each respective speech source. In the short-term time–

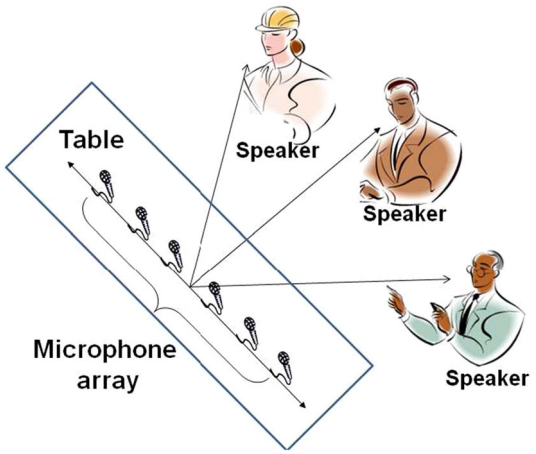

Fig. 1 Position of three speakers and the microphone array in the three-speaker environment

frequency (STFT) domain, the observed signal can be written as

$$\mathbf{x}(\omega, k) = \mathbf{s}_1(\omega, k) + \mathbf{s}_2(\omega, k) + \mathbf{s}_3(\omega, k) \tag{2}$$

where $\mathbf{x}(\omega, k)$, $\mathbf{s}_1(\omega, k)$, $\mathbf{s}_2(\omega, k)$ and $\mathbf{s}_3(\omega, k)$ are the contribution from the observed signal, the first, the second and the third speech sources, respectively. The objective is to separate each individual source signal from the observed signal. As such, one speech source is treated as the desired source while the others become undesired in a round robin fashion. In this case, the VAD cannot be employed to detect the desired source active or inactive periods because all sources can be active at the same time. Thus, a spatial localization technique needs to be employed. In this case, SRP-PHAT is utilized to estimate the spatial information for each speech source based only on the statistics of the observed signal.

3 Spatial localization technique employing SRP-PHAT

For the SRP-PHAT processing, we divide the sequence of observed signal into Q blocks, each consisting of N samples with the index $[(q-1)N+1, qN]$, $1 \le q \le Q$. The estimated correlation matrix $\mathbf{R}(\omega, q)$ of the observed signal in the qth block can be obtained as

$$\mathbf{R}(\omega, q) = \frac{1}{N} \sum_{k=(q-1)N+1}^{qN} \mathbf{x}(\omega, k)\mathbf{x}^H(\omega, k). \tag{3}$$

Denote by $\mathbf{R}(\omega)$ the estimated correlation matrix of the observed signal. This matrix can be obtained based on $\mathbf{R}(\omega, q)$ as

$$\mathbf{R}(\omega) = \frac{1}{QN} \sum_{k=1}^{QN} \mathbf{x}(\omega, k)\mathbf{x}^H(\omega, k) = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{R}(\omega, q). \tag{4}$$

Clearly, during the conversation either speech sources can be active or non-active. Therefore, there exist periods in which all speech sources are inactive. Since, $\mathbf{R}(\omega)$ in (4) is the average of all estimated correlation matrices $\mathbf{R}(\omega, q)$, this matrix can be used as a reference to detect non-speech blocks or blocks with low speech presence. Thus, we propose to use a threshold $\varepsilon\, R(\ell, \ell, \omega)$ to detect the speech presence where ε is a pre-set threshold, $0 < \varepsilon < 1$, and ℓ is a reference microphone. The value $R(\ell, \ell, \omega)$ is the (ℓ, ℓ)th element of the matrix $\mathbf{R}(\omega)$.

Denote by $\mathcal{S}$ the index set of all blocks with at least one active speech source. Based on the proposed threshold, this set can be obtained as

$$\mathcal{S} = \{q,\ 1 \le q \le Q :\ R(\ell, \ell, \omega, q) > \varepsilon\, R(\ell, \ell, \omega)\} \tag{5}$$

where $R(\ell, \ell, \omega, q)$ is the (ℓ, ℓ)th element of the matrix $\mathbf{R}(\omega, q)$. Note that $\mathcal{S}$ is not an empty set since $R(\ell, \ell, \omega)$ is the average of $R(\ell, \ell, \omega, q)$, see (4). For each $q \in S$, denote by $\bar{\mathbf{R}}_x(\omega, q)$ the normalized correlation matrix of the qth block

$$\bar{\mathbf{R}}(\omega, q) = \frac{\mathbf{R}(\omega, q)}{R(\ell, \ell, \omega, q)}. \tag{6}$$

By assuming that the speech signals of three speakers are statistically independent, the matrix $\mathbf{R}(\omega, q)$ can be decomposed as

$$\mathbf{R}(\omega, q) = \mathbf{R}_1(\omega, q) + \mathbf{R}_2(\omega, q) + \mathbf{R}_3(\omega, q) \tag{7}$$

where $\mathbf{R}_1(\omega, q)$, $\mathbf{R}_2(\omega, q)$ and $\mathbf{R}_3(\omega, q)$ are the correlation matrices for the first, the second and the third speech signals, respectively. We have

$$\begin{aligned}\mathbf{R}(\omega, q) = {} & p_1(\omega, q)\bar{\mathbf{R}}_1(\omega) + p_2(\omega, q)\bar{\mathbf{R}}_2(\omega) \\ & + p_3(\omega, q)\bar{\mathbf{R}}_3(\omega)\end{aligned} \tag{8}$$

where $p_1(\omega, q)$, $p_2(\omega, q)$, $p_3(\omega, q)$ and $\bar{\mathbf{R}}_1(\omega)$, $\bar{\mathbf{R}}_2(\omega)$, $\bar{\mathbf{R}}_3(\omega)$ are, respectively, the PSD and the normalized spatial correlation matrices of the first, the second and the third speech signals with (ℓ, ℓ)th elements are 1. Based on the idea of DOA estimation of acoustic signals using Near-field model [29], the spatial correlation matrices of speakers' speech signals are available. Since the (ℓ, ℓ)th elements of the normalized spatial correlation matrices $\bar{\mathbf{R}}_1(\omega)$, $\bar{\mathbf{R}}_2(\omega)$ and $\bar{\mathbf{R}}_3(\omega)$ are one, it follows from (8) that (6) can be rewritten as

$$\begin{aligned}\bar{\mathbf{R}}(\omega, q) = {} & \frac{p_1(\omega, q)}{p_1(\omega, q) + p_2(\omega, q) + p_3(\omega, q)}\bar{\mathbf{R}}_1(\omega) \\ & + \frac{p_2(\omega, q)}{p_1(\omega, q) + p_2(\omega, q) + p_3(\omega, q)}\bar{\mathbf{R}}_2(\omega) \\ & + \frac{p_3(\omega, q)}{p_1(\omega, q) + p_2(\omega, q) + p_3(\omega, q)}\bar{\mathbf{R}}_3(\omega).\end{aligned} \tag{9}$$

Eq. (9) can then be expressed as

$$\bar{\mathbf{R}}(\omega, q) = \gamma_1(\omega, q)\bar{\mathbf{R}}_1(\omega) + \gamma_2(\omega, q)\bar{\mathbf{R}}_2(\omega) + \gamma_3(\omega, q)\bar{\mathbf{R}}_3(\omega) \tag{10}$$

where the values $\gamma_1(\omega, q)$, $\gamma_2(\omega, q)$ and $\gamma_3(\omega, q)$ represent, respectively, the proportions of the matrices $\bar{\mathbf{R}}_1(\omega)$, $\bar{\mathbf{R}}_2(\omega)$ and $\bar{\mathbf{R}}_3(\omega)$ in the normalized correlation matrix $\bar{\mathbf{R}}(\omega, q)$, i.e.,

$$\gamma_1(\omega, q) = \frac{p_1(\omega, q)}{p_1(\omega, q) + p_2(\omega, q) + p_3(\omega, q)} \tag{11}$$

and

$$\gamma_2(\omega, q) = \frac{p_2(\omega, q)}{p_1(\omega, q) + p_2(\omega, q) + p_3(\omega, q)}. \quad (12)$$

and

$$\gamma_3(\omega, q) = \frac{p_3(\omega, q)}{p_1(\omega, q) + p_2(\omega, q) + p_3(\omega, q)}. \quad (13)$$

Since $p_1(\omega, q) \geq 0$, $p_2(\omega, q) \geq 0$ and $p_3(\omega, q) \geq 0$ we have

$$\gamma_1(\omega, q) \geq 0,\ \gamma_2(\omega, q) \geq 0,\ \gamma_3(\omega, q) \geq 0 \quad (14)$$

and

$$\gamma_1(\omega, q) + \gamma_2(\omega, q) + \gamma_3(\omega, q) = 1. \quad (15)$$

Since $\mathbf{R}(\omega)$ in (4) is the correlation matrix of the observed signal it follows

$$\bar{\mathbf{R}}(\omega) = \gamma_1(\omega)\bar{\mathbf{R}}_1(\omega) + \gamma_2(\omega)\bar{\mathbf{R}}_2(\omega) + \gamma_3(\omega)\bar{\mathbf{R}}_3(\omega) \quad (16)$$

where $\bar{\mathbf{R}}(\omega)$ is the normalized correlation matrix of the observed signal. The values $\gamma_1(\omega)$, $\gamma_2(\omega)$ and $\gamma_3(\omega)$ represent, respectively, the proportions of the matrices $\bar{\mathbf{R}}_1(\omega)$, $\bar{\mathbf{R}}_2(\omega)$ and $\bar{\mathbf{R}}_3(\omega)$ in the matrix $\bar{\mathbf{R}}(\omega)$, also

$$\gamma_1(\omega) \geq 0,\ \gamma_2(\omega) \geq 0,\ \gamma_3(\omega) \geq 0 \quad (17)$$

and

$$\gamma_1(\omega) + \gamma_2(\omega) + \gamma_3(\omega) = 1. \quad (18)$$

In the sequel, a spatial localization technique employing SRP-PHAT is proposed. Here, the (m, m)th element of $\mathbf{R}(\omega, q)$ is the cross-correlation of mth and nth microphone observed signals in the qth block. As such, the SRP-PHAT in block q can be estimated as follows

$$\Psi(\bar{\mathbf{R}}(\omega, q)) = \sum_{m=1}^{L} \sum_{n=m+1}^{L} \bar{R}(m, n, \omega, q) \quad (19)$$

where $\bar{R}(m, n, \omega, q)$ is the (m, n) element of the normalized correlation matrix $\bar{\mathbf{R}}(\omega, q)$. From (19) and (10), we have the following

$$\Psi(\bar{\mathbf{R}}(\omega, q)) = \gamma_1(\omega, q)\Psi(\bar{\mathbf{R}}_1(\omega)) + \gamma_2(\omega, q)\Psi(\bar{\mathbf{R}}_2(\omega)) + \gamma_3(\omega, q)\Psi(\bar{\mathbf{R}}_3(\omega)). \quad (20)$$

Clearly, the Eq. (20) shows the contribution balance of three speech sources in block q. As such, during the conversation, each speech sources can be active and non-active so the correlation matrices of blocks, in which only one speech source is active, are useful for speech spatial estimation. In the block of only one active source, the contribution of this source should be 1 and all contributions of other sources should be 0. In the complex plane, based on (14) (15) (20), the point of $\Psi(\bar{\mathbf{R}}(\omega, q))$ is located inside a triangle, with vertices given by these points $\Psi(\bar{\mathbf{R}}_1(\omega))$, $\Psi(\bar{\mathbf{R}}_2(\omega))$ and $\Psi(\bar{\mathbf{R}}_3(\omega))$. In addition, based on (14), the point of $\Psi(\bar{\mathbf{R}}(\omega))$ is located inside this triangle too, see Fig. 2a. As such, normalized spatial correlation matrices $\bar{\mathbf{R}}_1(\omega)$, $\bar{\mathbf{R}}_2(\omega)$ and $\bar{\mathbf{R}}_3(\omega)$ can be estimated by detecting triangle vertices of blocks' SRP-PHAT of observed signal, see Fig. 2b. Hence, a spatial detection of speech sources is proposed that employs an algorithm for finding triangle vertices, i.e., the blocks of only one source active.

The block of only first source active is detected as block q_1 as follows:

$$q_1 = \arg\max_q |\Psi(\bar{\mathbf{R}}(\omega, q)) - \Psi(\bar{\mathbf{R}}(\omega))| \quad (21)$$

here $|\cdot|$ is the absolute operation. The block of only second source active is detected as block q_2 as follows:

$$q_2 = \arg\max_q |\Psi(\bar{\mathbf{R}}(\omega, q)) - \Psi(\bar{R}(\omega, q_1))|. \quad (22)$$

The block of only third source active is detected as block q_3 as follows:

$$q_3 = \arg\max_q \{|\Psi(\bar{\mathbf{R}}(\omega, q)) - \Psi(\bar{R}(\omega, q_1))| + |\Psi(\bar{\mathbf{R}}(\omega, q)) - \Psi(\bar{R}(\omega, q_2))|\}. \quad (23)$$

Here, the correlation matrix of the observed signal in the block of only one active source contains only spatial characteristic of the active source. As such, the normalized spatial correlation matrix for the active source can be estimated as normalized correlation matrix in the block of only this source active. To reduce the correlation mismatch, we propose to estimate the normalized spatial correlation matrices for the speech sources by taking the average of the estimated normalized correlation matrices corresponding to I blocks which SRP-PHAT are nearest to estimated triangle vertices.

The average is employed to reduce the estimation error which can occur due to a limited number of samples in each block. Then, $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$ are proposed to be subsets of $\mathcal{S}$ and each subset has I block's indexes of blocks which SRP-PHAT are nearest to SRP-PHAT of blocks q_1, q_2, and q_3, respectively. In practice, the value I can be chosen smaller than 5 % of the number of elements in $\mathcal{S}$. The normalized spatial correlation matrix $\hat{\bar{\mathbf{R}}}_1(\omega)$ for the first source can be estimated as follows:

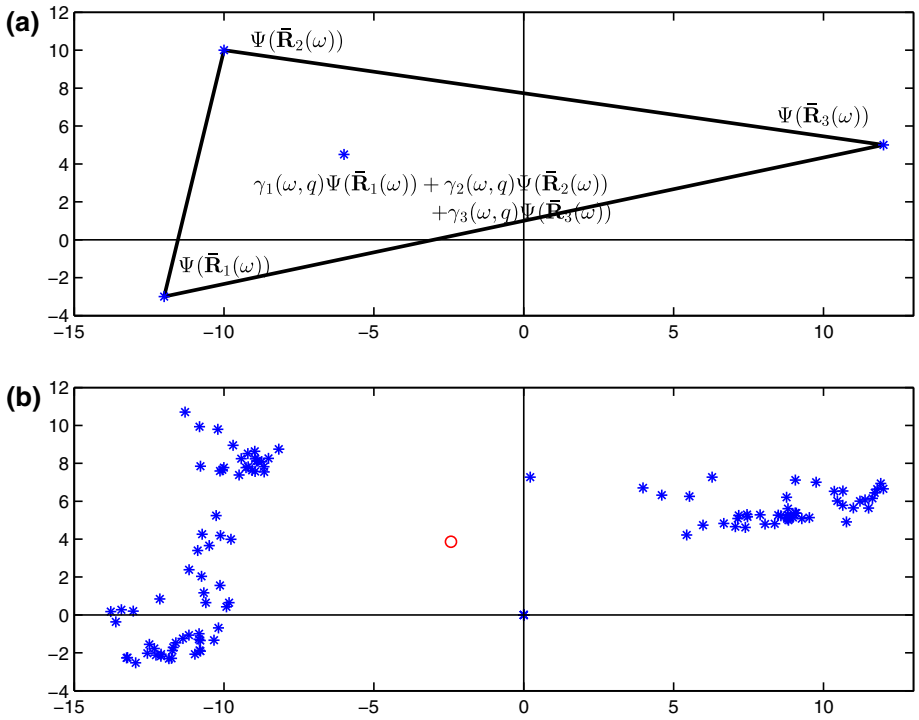

Fig. 2 **a** The triangle with SRP-PHAT vertices in the complex plane; **b** SRP-PHAT values of the observed signal for frequency of 2100 Hz from the simulation in Sect. 6

$$\hat{\bar{\mathbf{R}}}_1(\omega) = \frac{1}{I}\sum_{i\subseteq\mathcal{S}_1}\bar{\mathbf{R}}(\omega, q_{1,i}). \tag{24}$$

The normalized spatial correlation matrix $\hat{\bar{\mathbf{R}}}_2(\omega)$ for the second source can be estimated as follows:

$$\hat{\bar{\mathbf{R}}}_2(\omega) = \frac{1}{I}\sum_{i\subseteq\mathcal{S}_2}\bar{\mathbf{R}}(\omega, q_{2,i}). \tag{25}$$

The normalized spatial correlation matrix $\hat{\bar{\mathbf{R}}}_3(\omega)$ for the second source can be estimated as follows:

$$\hat{\bar{\mathbf{R}}}_3(\omega) = \frac{1}{I}\sum_{i\subseteq\mathcal{S}_3}\bar{\mathbf{R}}(\omega, q_{3,i}). \tag{26}$$

Due to the small value of I, the proportion of non-desired sources in the matrices $\hat{\bar{\mathbf{R}}}_1(\omega)$, $\hat{\bar{\mathbf{R}}}_2(\omega)$, and $\hat{\bar{\mathbf{R}}}_3(\omega)$ is approximately close to zero and their contribution can be neglected. These matrices are now used to estimate the optimum beamformer in each frequency bin.

4 Optimum beamformer using spatial information

Based on the estimated normalized spatial correlation matrices $\hat{\bar{\mathbf{R}}}_1(\omega)$, $\hat{\bar{\mathbf{R}}}_2(\omega)$, and $\hat{\bar{\mathbf{R}}}_3(\omega)$, an optimum beamformer is proposed for each desired source in the frequency bin ω. For extracting one speech source from the observed signal, an optimum beamformer is desired to suppress all undesired sources whilst preserving the desired one. Then, the first source is assumed to be the desired source so two other sources are undesired and denote by $\mathbf{w}_1(\omega)$ the filter weight for the first source in the frequency bin ω. The filter weight $\mathbf{w}_1(\omega)$ is designed to minimize two weighted cost functions $\mathbf{w}_1(\omega)^H\hat{\bar{\mathbf{R}}}_2(\omega)\mathbf{w}_1(\omega)$ and $\mathbf{w}_1(\omega)^H\hat{\bar{\mathbf{R}}}_3(\omega)\mathbf{w}_1(\omega)$ while maintaining the source direction as follows:

$$\begin{cases} \min\limits_{\mathbf{w}_1(\omega)} \mathbf{w}_1(\omega)^H\hat{\bar{\mathbf{R}}}_2(\omega)\mathbf{w}_1(\omega)\ ,\ \mathbf{w}_1(\omega)^H\hat{\bar{\mathbf{R}}}_3(\omega)\mathbf{w}_1(\omega) \\ \text{subject to}\ \ \mathbf{w}(\omega)^H\hat{\bar{\mathbf{d}}}_1(\omega) = 1. \end{cases} \tag{27}$$

where $\hat{\bar{\mathbf{d}}}_1(\omega)$ is the estimated cross-correlation vector between the first source at a ℓth reference microphone. This vector is also the ℓth column of the matrix $\hat{\bar{\mathbf{R}}}_1(\omega)$. Thus, from (27) we propose to minimize the following weighted cost function $\mathbf{w}_1(\omega)^H\left[\hat{\bar{\mathbf{R}}}_2(\omega) + \hat{\bar{\mathbf{R}}}_3(\omega)\right]\mathbf{w}_1(\omega)$ and the filter weight $\mathbf{w}_1(\omega)$ can be obtained by solving the optimization problem

$$\begin{cases} \min\ \mathbf{w}_1^H(\omega)\left[\hat{\bar{\mathbf{R}}}_2(\omega) + \hat{\bar{\mathbf{R}}}_3(\omega)\right]\mathbf{w}_1(\omega) \\ \text{subject to}\ \mathbf{w}_1^H\hat{\bar{\mathbf{d}}}_1(\omega) = 1 \end{cases} \tag{28}$$

Similarly, the beamformer weight $\mathbf{w}_2(\omega)$ for the second source can be obtained as the solution to the optimization problem

$$\begin{cases} \min\ \mathbf{w}_2^H(\omega)\left[\hat{\hat{\mathbf{R}}}_1(\omega)+\hat{\hat{\mathbf{R}}}_3(\omega)\right]\mathbf{w}_2(\omega) \\ \text{subject to } \mathbf{w}_2^H\hat{\hat{\mathbf{d}}}_2(\omega)=1 \end{cases} \tag{29}$$

where $\hat{\hat{\mathbf{d}}}_2(\omega)$ is the ℓth column of the matrix $\hat{\hat{\mathbf{R}}}_2(\omega)$. The beamformer weight $\mathbf{w}_3(\omega)$ for the third source can be obtained as the solution to the optimization problem

$$\begin{cases} \min\ \mathbf{w}_3^H(\omega)\left[\hat{\hat{\mathbf{R}}}_1(\omega)+\hat{\hat{\mathbf{R}}}_2(\omega)\right]\mathbf{w}_3(\omega) \\ \text{subject to } \mathbf{w}_3^H\hat{\hat{\mathbf{d}}}_3(\omega)=1 \end{cases} \tag{30}$$

where $\hat{\hat{\mathbf{d}}}_3(\omega)$ is the ℓth column of the matrix $\hat{\hat{\mathbf{R}}}_3(\omega)$. The solutions to three optimization problems can be expressed as

$$\mathbf{w}_1(\omega)=\frac{\left[\hat{\hat{\mathbf{R}}}_2(\omega)+\hat{\hat{\mathbf{R}}}_3(\omega)\right]^{-1}\hat{\hat{\mathbf{d}}}_1(\omega)}{\hat{\hat{\mathbf{d}}}_1^{\,H}(\omega)\left[\hat{\hat{\mathbf{R}}}_2(\omega)+\hat{\hat{\mathbf{R}}}_3(\omega)\right]^{-1}\hat{\hat{\mathbf{d}}}_1(\omega)} \tag{31}$$

and

$$\mathbf{w}_2(\omega)=\frac{\left[\hat{\hat{\mathbf{R}}}_1(\omega)+\hat{\hat{\mathbf{R}}}_3(\omega)\right]^{-1}\hat{\hat{\mathbf{d}}}_2(\omega)}{\hat{\hat{\mathbf{d}}}_2^{\,H}(\omega)\left[\hat{\hat{\mathbf{R}}}_1(\omega)+\hat{\hat{\mathbf{R}}}_3(\omega)\right]^{-1}\hat{\hat{\mathbf{d}}}_2(\omega)} \tag{32}$$

and

$$\mathbf{w}_3(\omega)=\frac{\left[\hat{\hat{\mathbf{R}}}_1(\omega)+\hat{\hat{\mathbf{R}}}_2(\omega)\right]^{-1}\hat{\hat{\mathbf{d}}}_3(\omega)}{\hat{\hat{\mathbf{d}}}_3^{\,H}(\omega)\left[\hat{\hat{\mathbf{R}}}_1(\omega)+\hat{\hat{\mathbf{R}}}_2(\omega)\right]^{-1}\hat{\hat{\mathbf{d}}}_3(\omega)} \tag{33}$$

The beamformer outputs for the three sources are calculated as

$$y_1(\omega,k)=\mathbf{w}_1^H(\omega)\mathbf{x}(\omega,k) \tag{34}$$

and

$$y_2(\omega,k)=\mathbf{w}_2^H(\omega)\mathbf{x}(\omega,k). \tag{35}$$

and

$$y_3(\omega,k)=\mathbf{w}_3^H(\omega)\mathbf{x}(\omega,k). \tag{36}$$

The remaining problem is to align the beamformer output in different frequency bins to the same source. In the sequel, the correlation between the beamformer outputs in neighboring frequencies is employed to overcome the permutation problem.

5 Permutation alignment

Since the optimum beamformers are performed in each frequency bin, the permutation alignment is needed before transforming the signals to the time domain. Here, the correlation approach is chosen for the permutation alignment and permutation decision is based on inter-frequency correlation of the output signal amplitudes based on the assumption that the amplitudes of the output signals from the one speech signal are correlated with adjoining frequencies. The permutation alignment can be performed continuously with a reference frequency in the middle of the frequency range. In this case, permutation correlation is performed in two directions, with increasing and decreasing frequency indexes until the end of the frequency range. For two neighboring frequencies ω_m and ω_{m+1}, the following correlations between the ith beamformer output of frequencies ω_m and jth beamformer output of frequencies ω_{m+1} are obtained as follows:

$$\mathrm{cor}_{i,j}=\frac{\mu(|y_i(\omega_m,k)y_j(\omega_{m+1},k)|)-\mu(|y_i(\omega_m,k)|)\mu(|y_j(\omega_{m+1},k)|)}{\sigma(|y_i(\omega_m,k)|)\sigma(|y_j(\omega_{m+1},k)|)} \tag{37}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are, respectively, the mean and the standard deviation of $(\cdot)$. Permutation decision Π is made with permutation alignment Π as follows

$$\Pi=\arg\max_{\Pi}\sum_{i,j\in\Pi}\mathrm{cor}_{i,j}, \tag{38}$$

After permutation alignment, three output signals in all frequencies are passed through the synthesis filters for obtaining the output signals with three speech sources in the time domain.

6 Experimental results

For performance evaluations of the proposed blind speech separation algorithm, a simulation is performed in a real room environment using a linear microphone array consisting of 6 microphones. Here, the distance between two adjacent microphones is 6 cm and the positions of three speakers are shown in Fig. 1. The distances between the array and speakers are about 1–1.5 m. The duration of the observed signal is 150 s and the value N was chosen as the number of samples in 0.5 s period while I and ε were chosen as 10 and 0.1, respectively. With the chosen N and I, the evaluation time of each speech source is about 5 s. Based on our experience, the evaluation time 5 s is enough for evaluation of the spatial characteristic of the speech source. We conducted our numerical experiments on HP Laptop with Intel Core i7 and 16GB RAM, using Matlab (R2013b).

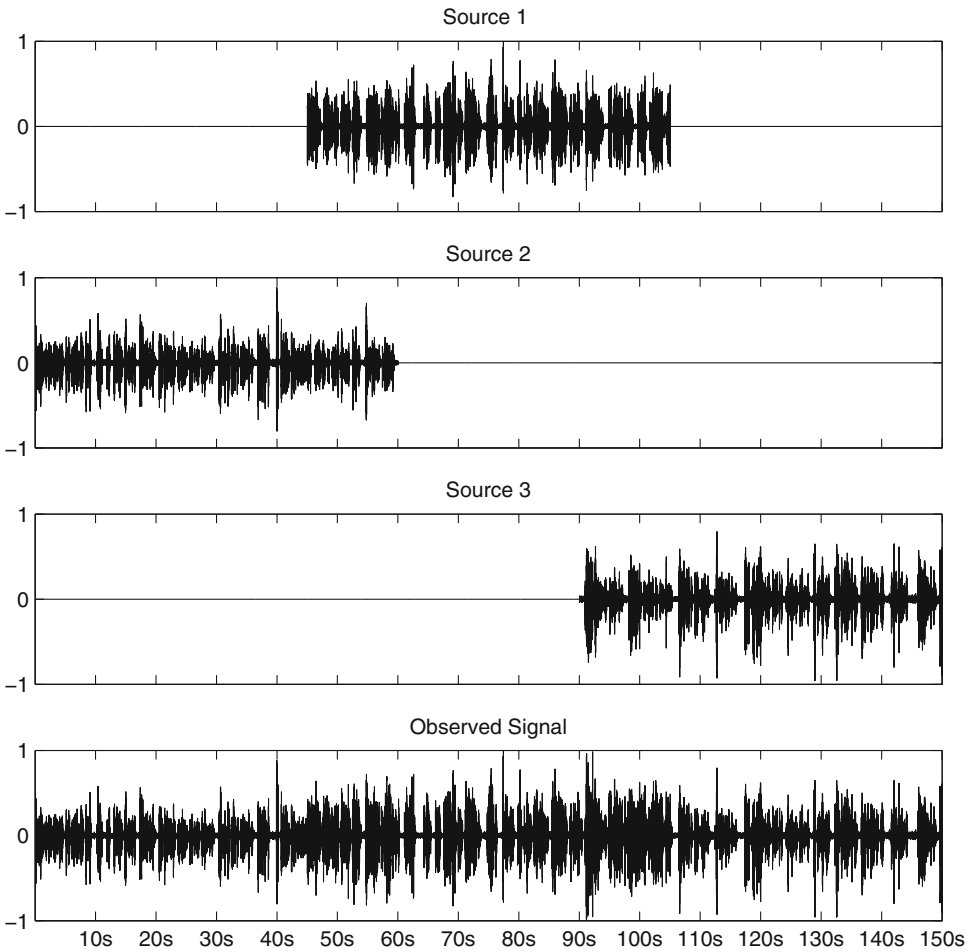

Fig. 3 Time domain plots of the original speech signals and the observed signal at the fourth microphone

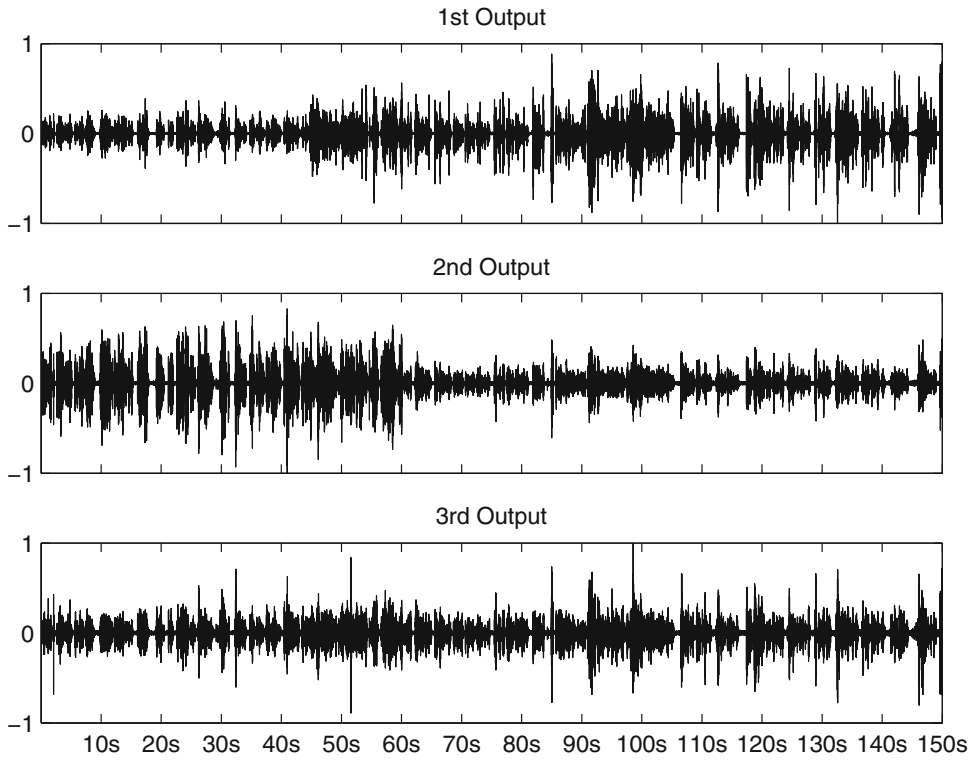

Fig. 4 Time domain plots of the second-order BSS algorithm outputs

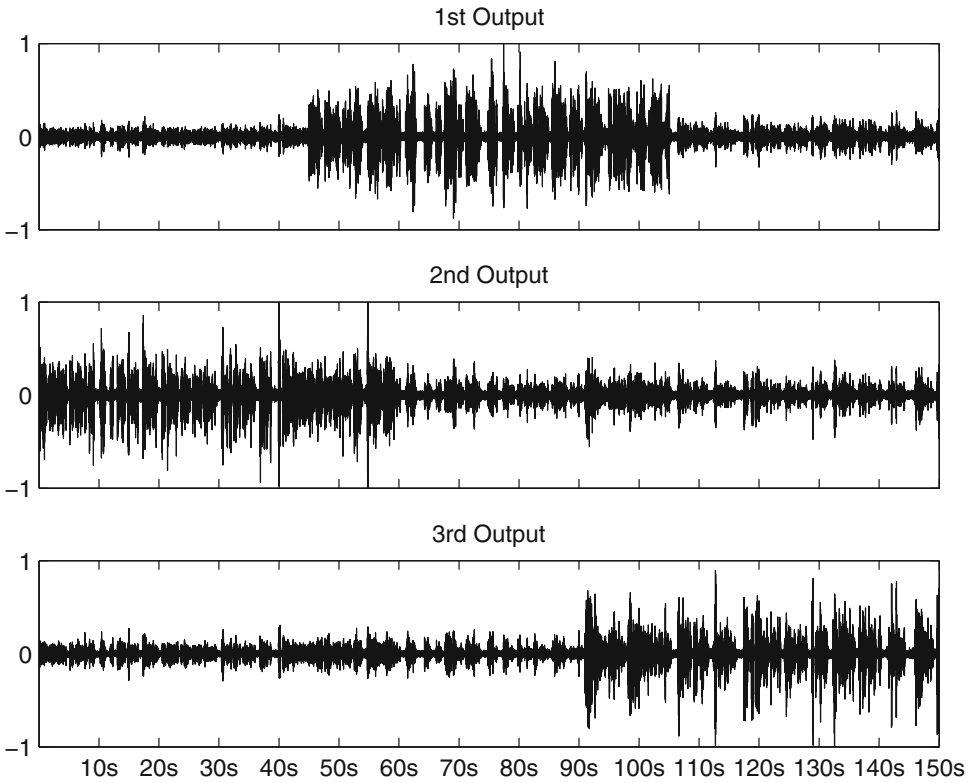

Fig. 5 Time domain plots of the proposed algorithm outputs

The observed signals are decomposed into sub-bands using an oversampled analysis filter bank. Here, an oversampling factor of two is chosen to reduce the aliasing effects between the adjacent sub-bands [30]. After the decomposition, the implementation of the proposed algorithm is performed in sub-bands. Figure 3 shows time domain plots of three speech signals and the observed signal. The speech signals from three speakers occur at different times and can overlap with each other in the observed signal. The overlapping signals simulate simultaneous conversation.

We have compared a second-order BSS algorithm with the suggested method. In Fig. 4, the results for when the second-order blind signal separation (BSS) algorithm is used for separating the observed signal are given. This second-order BBS algorithm was used in [22] for speech separation in two speaker environment. Figure 4 depicts time domain plots of the three outputs of the second-order BSS algorithm. The three outputs are speech signals extracted for three speakers from the observed signal. Hence, Fig. 4 shows a little difference between three output signals and the separation did not have a good result.

Figure 5 depicts time domain plots of the three outputs of the proposed separation algorithm when the proposed blind separation algorithm is used for separating the observed signal. The three outputs are speech signals extracted for three speakers from the observed signal. Thus, Fig. 5 shows that the proposed algorithm can separate the three speech signals from the observed mixture. Informal listening tests suggest the good listening quality of signal outputs from the proposed algorithm. From the Table 1, it is clear that the computation time of proposed algorithm is lower than computation time of the second-order BSS algorithm.

To quantify the performance of the second-order BSS algorithm and the proposed algorithm, the interference suppression (IS) and source distortion (SD) measures as presented in [31] are employed. As such, the speech signal from one speaker is viewed as the desired signal and other speech signals are interferences. Table 1 shows the IS and SD levels

Table 1 The interference suppression and the source distortion levels in the outputs of the proposed blind speech separation algorithm

Methods	First output		Second output		Third output		Computation time (s)
	IS (dB)	SD (dB)	IS (dB)	SD (dB)	IS (dB)	SD (dB)	
Second-order BSS algorithm	1.8	−25.1	2.9	−24.3	2.1	−23.4	42
Proposed algorithm	6.8	−29.2	5.7	−26.6	6.3	−26	27

for the three outputs of the second-order BSS algorithm and the proposed algorithm; the proposed algorithm has a better performance. In addition, the proposed blind speech separation algorithm offers a good interference suppression level (5–7 dB) whilst maintaining a low distortion level (−26 to −29 dB) for the desired source.

7 Summary

In this paper, a new blind speech separation algorithm in the frequency domain was developed for the three-speaker environment. Since, the position of the sources are unknown, the SRP-PHAT localization is used for estimating the spatial location of all speakers in each frequency bin. Based on that information, an optimum beamformer is designed for each speech source to extract the desired signal. The permutation alignment is used before transforming the signals to the time domain. Simulation results show that the proposed blind speech separation algorithm offers a good interference suppression level whilst maintaining a low distortion level for the desired source.

Acknowledgments This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under Grant Number C2014-26-01.

References

1. Nordholm, S., Dam, H., Lai, C., Lehmann, E.: Broadband beamforming and optimization. Signal processing: array and statistical signal processing, vol 3, pp. 553–598. Academic Press Library (2014)
2. Doclo, S., Kellermann, W., Makino, S., Nordholm, S.E.: Multichannel signal enhancement algorithms for assisted listening devices: exploiting spatial diversity using multiple microphones. IEEE Signal Process. Mag. **32**(2), 18–30 (2015)
3. Cohen, I., Benesty, J., Gannot, S. (eds.): Speech Processing in Modern Communication: Challenges and Perspectives. Springer, Berlin, Heidelberg (2010). ISBN 978-3642111297
4. Benesty, J., Makino, S., Chen, J.: Speech Enhancement. Springer, Berlin, Heidelberg (2005). ISBN 978-3540240396
5. Bai, M.R., Ih, J.-G., Benesty, J.: Acoustic Array Systems: Theory, Implementation, and Application. Wiley-IEEE Press, Singapore (2013). ISBN 978-0470827239
6. Benesty, J., Chen, J., Huang, Y.: Microphone Array Signal Processing Springer, Berlin, Heidelberg (2008). ISBN 978-3540786115
7. Nordebo, S., Claesson, I., Nordholm, S.: Adaptive beamforming: spatial filter designed blocking matrix. IEEE J. Ocean. Eng. **19**, 583–590 (1994)
8. Nagata, Y., Abe, M.: Two-channel adaptive microphone array with target tracking. Electron. Commun. Jpn. **83**(12), 860–866 (2000)
9. Nakadai, K., Nakamura, K., Ince, G.: Real-time super-resolution sound source localization for robots. In: Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012), pp. 694–699. IEEE, Vilamoura (2012)
10. Grbić, N., Nordholm, S., Cantoni, A.: Optimal fir subband beamforming for speech enhancement in multipath environments. IEEE Signal Process. Lett. **10**(11), 335–338 (2003)
11. Brandstein, M., Ward, D. (eds.): Microphone Arrays: Signal Processing Techniques and Applications. Springer, Berlin, Heidelberg (2001). ISBN 978-3540419532
12. Fallon, M., Godsill, S.: Acoustic source localization and tracking of a time-varying number of speakers. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1409–1415 (2012)
13. Gannot, S., Burshtein, D., Weinstein, E.: Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**, 1614–1626 (2001)
14. Low, S.Y., Nordholm, S., Togneri, R.: Convolutive blind signal separation with post-processing. IEEE Trans. Speech Audio Process. **12**(5), 539–548 (2004)
15. Grbić, N., Tao, X.J., Nordholm, S., Claesson, I.: Blind signal separation using overcomplete subband representation. IEEE Trans. Speech Audio Process. **9**(5), 524–533 (2001)
16. Parra, L., Spence, C.: Convolutive blind separation of non-stationary sources. IEEE Trans. Speech Audio Process. **8**(3), 320–327 (2000)
17. Dam, H.H., Nordholm, S., Low, S.Y., Cantoni, A.: Blind signal separation using steepest descent method. IEEE Trans. Signal Process. **55**(8), 4198–4207 (2007)
18. Sawada, H., Araki, S., Makino, S.: Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans. Audio Speech Lang. Process. **19**(3), 516–527 (2011)
19. Dam, H.Q., Nordholm, S., Dam, H.H., Low, S.Y.: Postfiltering using multichannel spectral estimation in multispeaker environments. EURASIP J. Adv. Signal Process ID **860360**, 1–10 (2008)
20. Krishnamoorthy, P., Prasanna, S.R.M.: Two speaker speech separation by lp residual weighting and harmonics enhancement. Int. J. Speech Technol. **13**(3), 117–139 (2010)
21. Dam, H.Q.: Blind multi-channel speech separation using spatial estimation in two-speaker environments. J. Sci. Technol. Spec. Issue Theor. Appl. Comput. Sci. **48**(4), 109–119 (2010)
22. Dam, H.Q., Nordholm, S.: Sound source localization for subband-based two speech separation in room environment. In: 2013 International Conference on Control, Automation and Information Sciences (ICCAIS), pp. 223–227. IEEE, Nha Trang City (2013)
23. Tariqullah, J., Wenwu, W., DeLiang, W.: A multistage approach to blind separation of convolutive speech mixtures. Speech Commun. **53**, 524–539 (2011)
24. Minhas, S.F., Gaydecki, P.: A hybrid algorithm for blind source separation of a convolutive mixture of three speech sources. EURASIP J. Adv. Signal Process. **1**(92), 1–15 (2014)
25. Araki, S., Mukai, R., Makino, S., Nishikawa, T., Saruwatari, H.: The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. IEEE Trans. Speech Audio Process. **11**(2), 109–116 (2003)
26. Makino, H.S.S., Lee, T.-W., Sawada, H. (eds.): Blind Speech Separation. Springer, Netherlands (2007). ISBN 978-1402064784
27. Naik, G.R., Wang, W. (eds.): Blind Source Separation: Advances in Theory. Algorithms and Applications. Springer, Berlin, Heidelberg (2014). ISBN 978-3642550157
28. Cobos, M., Marti, A., Lopez, J.J.: A modified srp-phat functional for robust real-time sound source localization with scalable spatial sampling. IEEE Signal Process. Lett. **18**(1), 71–74 (2010)

29. Sawada, H., Mukai, R., Araki, S., Makino, S.: Frequency-domain blind source separation. In: Speech Enhancement. Signals and Communication Technology, pp. 299–327. Springer, Berlin, Heidelberg (2005). ISBN: 978-3540240396
30. Vaidyanathan, P.P.: Multirate Systems and Filter Banks. Prentice Hall, Englewood Cliffs (1993). ISBN 978-0136057185
31. Dam, H.Q., Nordholm, S., Dam, H.H., Low, S.Y.: Adaptive beamformer for hands-free communication system in noisy environments. IEEE Int. Symp. Circuits Syst. **2**, 856–859 (2005)