Reverberant speech separation with probabilistic time-frequency masking for B-format recordings

Xiaoyi Chen [a,*], Wenwu Wang [b], Yingmin Wang [a],
Xionghu Zhong [c], Atiyeh Alinaghi [b]

[a] *Department of Acoustic Engineering, School of Marine Science and Technology, Northwestern Polytechnical University, China, 710072.*
[b] *Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering, University of Surrey, UK, GU2 7XH.*
[c] *School of Computer Engineering, College of Engineering, Nanyang Technological University, Singapore, 639798.*

Abstract

Existing speech source separation approaches overwhelmingly rely on acoustic pressure information acquired by using a microphone array. Little attention has been devoted to the usage of B-format microphones, by which both acoustic pressure and pressure gradient can be obtained, and therefore the direction of arrival (DOA) cues can be estimated from the received signal. In this paper, such DOA cues, together with the frequency bin-wise mixing vector (MV) cues, are used to evaluate the contribution of a specific source at each time-frequency (T-F) point of the mixtures in order to separate the source from the mixture. Based on the von Mises mixture model and the complex Gaussian mixture model respectively, a source separation algorithm is developed, where the model parameters are estimated via an expectation-maximization (EM) algorithm. A T-F mask is then derived from the model parameters for recovering the sources. Moreover, we further improve the separation performance by choosing only the reliable DOA estimates at the T-F units based on thresholding. The performance of the proposed method is evaluated in both simulated room environments and a real reverberant studio in terms of signal-to-distortion ratio (SDR) and the perceptual evaluation of speech quality (PESQ). The experimental results show its advantage over four baseline algorithms including three T-F mask based approaches and one convolutive independent component analysis (ICA) based method.

Keywords:
B-format signal, acoustic intensity, expectation-maximization (EM) algorithm, blind source separation (BSS), direction of arrival (DOA)

1. Introduction

Blind speech separation (BSS) aims to estimate the desired speech signals in the presence of other speech signals or interfering sounds, without the prior knowledge (or with very little information) about the sources and the mixing process (Pedersen et al., 2007). It offers great potentials in many applications such as automatic speech recognition, teleconferencing and hearing aids.

In the past, independent component analysis (ICA) (Lee, 1998; Stone, 2004; Hyvärinen and Oja, 2000; Comon, 1994; Hyvärinen et al., 2009; Comon and Jutten, 2010) has been widely employed and shown to be promising in BSS problems. Significant contributions have been made in anechoic (i.e. without room reflections) and over-determined/even-determined (i.e. the number of microphones is greater than or equal to the number of sources) situations. However, the performance of ICA is degraded in the reverberant environments (i.e. with room reflections), especially for under-determined (i.e. the number of microphones is smaller than the number of sources) case, since the unmixing process becomes increasingly ambiguous due to the overlap of the reflected sound with the direct sound, and/or the lack of information in the under-determined case.

To separate sources under reverberant environments, two types of methods are often used, namely time-domain (Aichner et al., 2002; Thomas et al., 2006; Nishikawa et al., 2003) and frequency-domain (Sawada et al., 2004; Araki et al., 2001; Saruwatari et al., 2001; Sawada et al., 2005) approaches, respectively. The time-domain methods are often based on the extension of the instantaneous ICA to the convolutive case, and the computational complexity associated with the estimation of the filter coefficients can be high, especially when dealing with the mixtures in a heavily reverberant environment, i.e. large T_{60} (Amari et al., 1997; Buchner et al., 2004).

For approaches in frequency domain (Araki et al., 2003; Parra and Spence, 2000; Wang et al., 2005), the convolutive mixtures are transformed to the complex-valued instantaneous source separation problems by e.g. the short-time Fourier transform (STFT), and then the separated source components in each frequency bin are aligned to remove the permutation ambiguities

before being used to reconstruct the sources in the time-domain using inverse short-time Fourier transform (ISTFT). Due to the use of STFT, the frequency-domain approaches are, in general, computationally more efficient as compared with time-domain methods.

Recently, various methods have been developed to separate the speech mixtures in the underdetermined scenarios. By exploiting the sparseness property of the speech signals in the time-frequency (T-F) domain, different approaches such as T-F masking method (Yilmaz and Rickard, 2004; Sawada et al., 2006; Wang et al., 2009) and maximum *a posterior* (MAP) estimation (D O'Grady and Pearlmutter, 2008) have been proposed. The former method is more attractive due to its lower computational complexity than the latter one (Sawada et al., 2006; Wang et al., 2009). In this paper, we focus on the T-F masking approach.

The T-F masking approach can be divided into two categories. One is based on the binary mask, where the mask value is set as either one or zero to retain or to reject the mixture energy at each T-F unit. For example, in (Araki et al., 2003), a binary mask based source separation method is introduced by clustering the feature of the level ratio and the weighted phase difference with the K-means algorithm. The other category is based on the probabilistic (soft) mask, where the mask value is the probability of each source being active at each T-F point of the mixtures, hence ranging from zero to one. Examples in this category include the model-based method in (Mandel et al., 2010) where binaural cues such as the interaural phase difference (IPD) and interaural level difference (ILD) are estimated from the mixtures to generate the mask, and the method (Sawada et al., 2007, 2011) where the mixing vector (MV) cue is used for estimating the T-F mask. The probabilistic mask can be estimated iteratively using the Expectation-Maximization (EM) algorithm.

Most of the methods discussed above are performed by using a microphone array together with the estimation techniques developed based on acoustic pressure information. Different from these traditional microphone arrays which measure only the acoustic pressure, the soundfield microphone system (Farrar, 1979; Malham and Myatt, 1995), also known as B-format microphone, consists of four closely co-located microphones and is able to measure the full soundfield information, i.e., the pressure gradient at forward, leftward and upward as well as the acoustic pressure information. Another system which is named acoustic vector sensor (AVS) (Nehorai and Paldi, 1994; Hawkes and Nehorai, 2000), can also be used to collect the particle ve-

locity information in three dimensional space as well as the acoustic pressure information. Both the B-format microphone and the AVS have promising advantages over the conventional microphones due to the three bidirectional pick-ups (pressure gradient or the velocity), and show good performance on several applications, such as sound localization (Hawkes and Nehorai, 1998; Zhong and Premkumar, 2012) and speech enhancement (Shujau et al., 2010).

Nevertheless, only few works have been conducted in the literature in dealing with the BSS problem with speech mixtures acquired by the B-format microphone/AVS. Two typical examples are (Gunel et al., 2008; Shujau et al., 2011), where the direction-of-arrival (DOA) information obtained from the B-format microphone/AVS are used to separate the speech sources based on the T-F masking approach.

In (Gunel et al., 2008), the DOA at each T-F unit is estimated based on the intensity vector (Nehorai and Paldi, 1994), by exploiting the T-F representation of the outputs of the B-format microphone. The soft T-F masking approach is employed for the B-format mixtures under reverberant environment, the contribution of a specific source at each T-F point is obtained by fitting the DOA histogram with the von Mises distribution. The von Mises distribution can be characterized by the mean direction (μ) and the concentration parameter (κ). In (Gunel et al., 2008), the mean direction (μ) for each source is estimated by picking the peaks of the DOA histogram. However, the concentration parameter (κ) is searched experimentally over a range of all possible solutions, which is computationally expensive. In (Shujau et al., 2011), a binary T-F masking approach is employed for the mixtures recorded by a single AVS. The peaks of the DOA histogram (which is obtained by the estimation of the intensity vector, the same as in (Gunel et al., 2008)) are estimated and regarded as the directions of the source signals. The binary T-F mask is obtained by comparing the DOAs at each T-F point with the direction of the target speech, with 1 assigned to the T-F unit where the DOA is closer to the target signal than the interferences, and 0 otherwise.

There are two main drawbacks with the methods described above. Firstly, the separation performance of these two methods is strongly dependent on the accuracy of the DOA information, however, as demonstrated in (Levin et al., 2010), the intensity based DOA estimation, which is used in these two methods, produces biased results under reverberant environment, and the angular error becomes larger with the increase of the reverberation level. Secondly, the separation performance of the two algorithms is dependent on the accuracy of the estimation of mean directions, which are identified by the

histogram peaks. The performance deteriorates when the sources are located close to each other, since it is difficult to distinguish the mean directions from the histogram under such a situation.

Several approaches are proposed in this paper to address these problems. Firstly, the T-F bin-wise MV cue is incorporated with the DOA cue to improve the accuracy of each T-F point of the mixture being assigned to a specific source under the reverberant environment. Secondly, different from the above two methods, in which the masks are constructed by the mean directions directly, the mean directions are adopted as the initialization value of the DOA cue in the EM algorithm, and the parameters of the MV and DOA cues are updated iteratively at each frequency bin until convergence. Lastly, the DOA cue is evaluated at each T-F unit and a thresholding method is used to select the reliable DOA estimates and thus further improve the separation performance.

The frequency-dependent model parameters for both the DOA and MV cues are evaluated and refined iteratively by the EM algorithm. In the E-step, the von Mises and the complex Gaussian probability distributions are applied respectively to calculate the probability that each source is dominant in each T-F point of the mixture. In the M-step, the parameters of each source model are re-estimated according to the T-F regions of the mixtures that are most likely to be dominated by that source. It is noticed from (Mandel et al., 2010) that the EM algorithm is sensitive to the initialization value because of the non-convex characteristics of the total log likelihood, so the more accurate mean direction used in the initialization has the potential to improve the separation performance. Moreover, due to the exploitation of the DOA information, the permutation problem is solved in the first iteration of the EM algorithm.

Preliminary studies of this work have been presented in (Chen et al., 2013; Zhong et al., 2013). Different from (Chen et al., 2013; Zhong et al., 2013), however, we have made the following improvements in this paper. Firstly, we use the von Mises distribution to model the circular statistics for the DOA cue, as opposed to the use of the Gaussian distribution in (Chen et al., 2013; Zhong et al., 2013). This provides a better fit to the statistics of the DOA cue and more accurate estimate for the source occupation probability at each T-F point in the EM algorithm, especially for the circular case, when the mean DOA is close to the estimated DOA, e.g. the mean DOA at around 0° and the estimated DOA at around 360°. In our previous work (Chen et al., 2013; Zhong et al., 2013), only the semi-circular case, i.e. DOAs from 0° to

180°, was considered. Secondly, we propose a simple but efficient method to improve the separation performance under reverberant environment by selecting only the reliable DOA estimates obtained based on the intensity information and discarding the un-reliable DOAs caused by reverberations. Lastly, the separation performance of the proposed method was evaluated under the over-, even- and under-determined case respectively, as well as under various reverberation times and configurations.

For performance comparison, we choose four baseline methods, namely, the two DOA based T-F masking approaches (Gunel et al., 2008) and (Shujau et al., 2011) as already discussed earlier, the MV cue based T-F clustering method (Sawada et al., 2011), and a conventional second-order statistics based convolutive ICA algorithm (Wang et al., 2005).

The remainder of this paper is organized as follows. In Section 2 the B-format microphone based source separation model and the two DOA-based T-F masking methods are introduced. In Section 3, the T-F masking based source separation approach is presented firstly, then the proposed separation method, which combines the reliability-based DOA classification and the bin-wise classification based on the EM algorithm, is introduced in detail. The experimental setup and the results of the proposed method as compared with the baseline methods are presented in Section 4, and finally Section 5 gives the conclusions.

2. Background

This section first introduces the T-F masking based source separation model in which the mixtures are obtained from the B-format microphone system, and then gives an overview of two previous methods for speech separation based on the B-format/AVS recordings that will be used as baselines in our numerical evaluations.

2.1. B-format Microphone based Source Separation Model

The geometry of the B-format microphone array is made up of four compact microphones which are placed at the four non-adjacent corners of a cube, forming a regular tetrahedron, as shown in Figure 1. The x-, y- and z-coordinates indicate the forward, leftward and upward direction, respectively. The four capsules, which show the information at left-front L_F, left-back L_B, right-front R_F and right-back R_B respectively, are mounted as closely as possible to eliminate the phase aliasing (Farrar, 1979). The B-format outputs

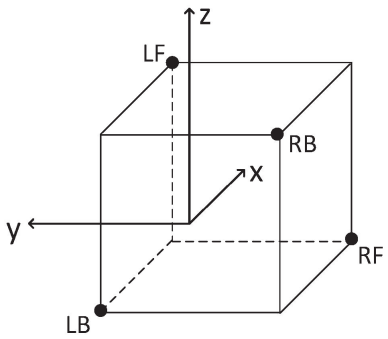


Figure 1: An illustration of the microphone array setup in the B-format microphone.

(Farrar, 1979), which include the pressure (or omnidirectional) component (p_0) and the pressure gradient values corresponding to the x-, y- and z-coordinate (g_x, g_y and g_z), can be obtained from the four raw tetrahedral capsule outputs as

$$\begin{bmatrix} p_0(n) \\ g_x(n) \\ g_y(n) \\ g_z(n) \end{bmatrix} = \begin{bmatrix} L_F(n) + L_B(n) + R_F(n) + R_B(n) \\ L_F(n) - L_B(n) + R_F(n) - R_B(n) \\ L_F(n) + L_B(n) - R_F(n) - R_B(n) \\ L_F(n) - L_B(n) - R_F(n) + R_B(n) \end{bmatrix} \tag{1}$$

where n is the discrete time index.

In this work, we assume that the sources are strictly located at a 2-D ($x - y$) plane, i.e., the elevation angle of the sources are zero. Under this assumption, only the $p_0(n)$, $g_x(n)$ and $g_y(n)$ are considered as the outputs of the B-format microphone.

Assume I different speech signals $s_i(n)$ ($i = 1, \ldots, I$) are presented in a noise-free acoustic room environment, the received mixtures from the B-format microphone array can be written as

$$\mathbf{x}(n) = \begin{bmatrix} p_0(n) \\ g_x(n) \\ g_y(n) \end{bmatrix} = \sum_{i=1}^{I} \begin{bmatrix} h_0^i(n) \\ h_x^i(n) \\ h_y^i(n) \end{bmatrix} \otimes s_i(n) \tag{2}$$

where I is the number of sources, $\otimes$ denotes convolution, $h_0^i(n)$, $h_x^i(n)$ and $h_y^i(n)$ represent the corresponding room impulse response (RIR) from the i-th source to $p_0(n)$, $g_x(n)$ and $g_y(n)$ respectively, cascading the direct path as well as the multipath responses. It should be noted that the RIR here is used for both the acoustic pressure and pressure gradient, representing an

expanded version of the traditional RIR, which is normally related to the acoustic pressure only.

To realize the frequency-domain separation, the mixture observations $\mathbf{x}(n)$ from the B-format microphone are first converted into frequency-domain time-series signals $\mathbf{X}(\omega, t)$ by the STFT. It is known that if the frame size in the STFT approach is long enough to cover the main part of the impulse response, the time-domain convolutive mixture model $\mathbf{x}(n)$ can be approximated as an instantaneous mixture model in the frequency domain (Smaragdis, 1998)

$$\mathbf{X}(\omega, t) = \sum_{i=1}^{I} \mathbf{H}_i(\omega) S_i(\omega, t) \tag{3}$$

where ω and t are the frequency bin and time frame indices, respectively. $\mathbf{X}(\omega, t) = [P_0(\omega, t), G_x(\omega, t), G_y(\omega, t)]^T$, in which $P_0(\omega, t)$, $G_x(\omega, t)$ and $G_y(\omega, t)$ are the STFT of $p_0(n)$, $g_x(n)$ and $g_y(n)$, respectively. $\mathbf{H}_i(\omega) = [h_0^i(\omega), h_x^i(\omega), h_y^i(\omega)]^T$ is the frequency domain representation of the RIR from the i-th source to the three components of the B-format microphone respectively. $S_i(\omega, t)$ is the STFT of the i-th source.

The separated signals in the frequency domain $Y_i(\omega, t)$ can be obtained by the T-F masking as

$$Y_i(\omega, t) = M_i(\omega, t) P_0(\omega, t) \tag{4}$$

where $0 \leq M_i(\omega, t) \leq 1$ is the mask for the i-th separated signal.

After the T-F masking approach, the source signals in the time-domain $y_i(n)$ can then be reconstructed by the inverse STFT.

The goal of blind source separation with the B-format microphone system is to obtain the separated signals $y_i(n)$, $i = 1, \ldots, I$, which corresponds to the source signals $s_i(n)$, $i = 1, \ldots, I$. The separation approach is performed only with the mixtures $\mathbf{x}(n)$, without knowing RIRs, $h_0^i(n)$, $h_x^i(n)$ and $h_y^i(n)$. To achieve this, the DOA based soft and binary T-F masking techniques are adopted (Gunel et al., 2008; Shujau et al., 2011), and a brief introduction of these two approaches is given next.

2.2. DOA based T-F Masking Approaches

The estimation of DOA, which is employed as a cue to estimate the T-F mask in (Gunel et al., 2008; Shujau et al., 2011), is introduced first based on the T-F domain intensity vector estimation. In (Nehorai and Paldi, 1994), it

is assumed that the signal behaves as a plane wave at the sensor. With this assumption, the acoustic particle velocity can be expressed as

$$\mathbf{v}(n) = -\frac{1}{\rho_0 c}\mathbf{g}(n) \odot \vec{\mathbf{u}} \tag{5}$$

where $\mathbf{v}(n) = [v_x(n), v_y(n)]^T$ is the velocity components along x- and y-direction, and $\odot$ denotes the element-wise product, and ρ_0 is the ambient density of the air, and c is the velocity of sound wave in the air, and $\mathbf{g}(n) = [g_x(n), g_y(n)]^T$ is the pressure gradient value corresponding to the x- and y- coordinates, and $\vec{\mathbf{u}}$ is an unit vector denotes the direction in x- and y- coordinates, which points from the sensor towards the source, i.e., $\vec{\mathbf{u}} = [\vec{u}_x, \vec{u}_y]^T$.

The instantaneous intensity vector can then be denoted as the product of the acoustic pressure and the particle velocity, as follows,

$$\mathbf{i}(n) = p_0(n) \odot \mathbf{v}(n) \tag{6}$$

By taking the STFT, the T-F representation of the intensity vector $\mathbf{I} = [I_x(\omega, t), I_y(\omega, t)]^T$can be given as

$$I_x(\omega, t) = -\frac{1}{\rho_0 c}\big[\Re\{P_0^*(\omega, t)G_x(\omega, t)\}\vec{u}_x\big] \tag{7}$$

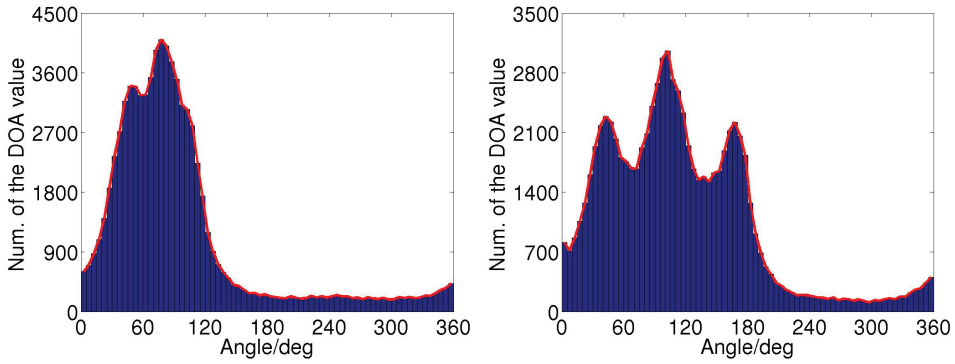
$$I_y(\omega, t) = -\frac{1}{\rho_0 c}\big[\Re\{P_0^*(\omega, t)G_y(\omega, t)\}\vec{u}_y\big] \tag{8}$$

where the superscript $*$ denotes conjugation, $\Re\{\cdot\}$ means taking the real part of its argument. The direction of the intensity can thus be obtained by

$$\theta(\omega, t) = \arctan\left[\frac{\Re\{P_0^*(\omega, t)G_y(\omega, t)\}}{\Re\{P_0^*(\omega, t)G_x(\omega, t)\}}\right] \tag{9}$$

Based on the estimation of $\theta(\omega, t)$ over an entire spectrogram, the algorithm in (Gunel et al., 2008), which we refer to as Gunel, creates a histogram of all the direction value $\theta(\omega, t)$ first. Then, the von Mises density function is utilized to fit the direction histogram and to evaluate the contribution of a specific source at each T-F point of the mixtures, the probability density function of the von Mises distribution is given as

$$f(\theta|\mu, \kappa) = \frac{exp(\kappa \cos(\theta - \mu))}{2\pi I_0(\kappa)} \tag{10}$$



(a) (b)

Figure 2: The direction histogram of three speech sources which are located at (a) 40°, 70° and 100° (b) 40°, 100° and 160° respectively under 0.6 s reverberation.

where $0 \leq \mu < 2\pi$ is the mean direction, $\kappa > 0$ is the concentration parameter, and $I_0(\kappa)$ is the modified Bessel function of order zero. The probability that each T-F point of the mixtures corresponds to the i-th source is obtained as

$$p_i^g(\omega, t) = \sigma_i \frac{exp(\kappa_i(t)\cos(\theta(\omega, t) - \mu_i))}{2\pi I_0(\kappa_i(t))} \tag{11}$$

where $\sigma_i = 1/(I+1)$ is the component weight corresponding to source i, the superscript g is used to identify the probability estimated in Gunel's method. The mean value μ_i is identified as the direction corresponding to the i-th largest peak of the DOA histogram. The concentration parameter κ_i is estimated by the 6-dB beamwidth θ_i^{BW} as

$$\kappa_i = \frac{1}{1 - \cos(\theta_i^{BW}/2)} \tag{12}$$

For each source, θ_i^{BW} is spanned linearly from 10° to 180° with 10° intervals and the related κ_i is calculated by Equation (12). The κ_i which best fits the direction histogram is finally chosen as the concentration parameter. The final mask value of the Gunel's method M_i^g is obtained by normalizing p_i^g across the sources as

$$M_i^g(\omega, t) = \frac{p_i^g(\omega, t)}{\sum_l p_l^g(\omega, t)}, \qquad (l = 1, ..., I) \tag{13}$$

In the algorithm of (Shujau et al., 2011), which we refer to as Shujau, I largest peaks of the histogram of $\theta(\omega, t)$ are found and identified as the DOAs

corresponding to the I sources. Let δ_i, for $i = 1, \ldots, I$ denote the estimated DOAs. The angular difference $\Delta\theta_i$ is calculated by the DOA at each T-F point $\theta(\omega, t)$ with the direction of each source δ_i as

$$\Delta\theta_i(\omega, t) = \begin{cases} |\theta(\omega, t) - \delta_i| - 180^\circ, & |\theta(\omega, t) - \delta_i| > 180^\circ \\ |\theta(\omega, t) - \delta_i|, & otherwise \end{cases} \quad (i = 1, \cdots, I) \tag{14}$$

A binary T-F mask is then obtained to separate the sources as

$$M_i^s(\omega, t) = \begin{cases} 1, & \Delta\theta_i(\omega, t) < \Delta\theta_j(\omega, t) \\ 0, & otherwise \end{cases} \quad (j = 1, \cdots, I, j \neq i) \tag{15}$$

where M_i^s is the mask used to recover the source i and superscript s denotes the mask obtained by Shujau's method.

3. Proposed Method

Using only the DOA cue based source separation such as the method in (Gunel et al., 2008; Shujau et al., 2011), the performance deteriorates when the sources are located close to each other, since the peaks of the DOA histogram considered as the direction of the sources are blurred, as shown in Figure 2. The DOA values in Figure 2 were calculated by Equation (9) with three speech sources mixed together in the same studio as described in Section 4. It has been observed recently in (Alinaghi et al., 2011) that adding the mixing vector (MV) cue can improve the accuracy of the T-F assignment. In this paper, to address the above limitation, the MV cue is incorporated with the DOA cue to improve the estimation of the source occupation likelihood at each T-F point based on a maximum likelihood framework. The proposed system is shown in Figure 3. The T-F masking approach is proposed by combining the DOA classification with the bin-wise classification based on the EM algorithm, in which the DOA values are estimated from the intensity information. The DOA based classification process has already been described in Section 2.2 and therefore is not elaborated any more. In this section, we present a thresholding approach to reduce the errors of the intensity-based DOA estimation caused by reverberation, and to further improve the reliability of the DOA cues and hence the separation performance. The details of the reliability based DOA classification are given later in Section 3.4. Next, we first present the bin-wise based classification, followed by the EM algorithm and its initialization.

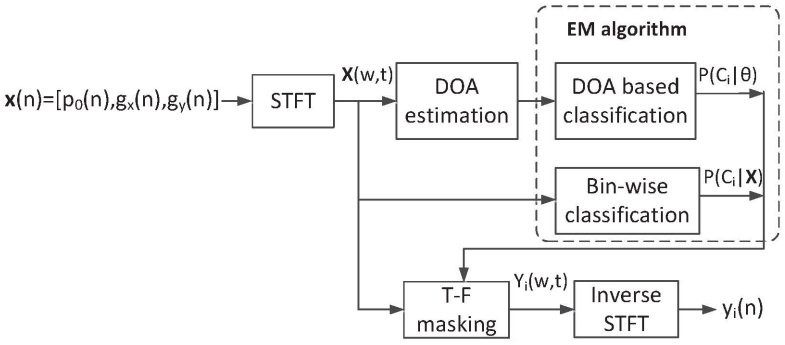


Figure 3: Processing flow for the proposed BSS algorithm with T-F masking.

3.1. Bin-wise Classification

In frequency bin-wise classification, only the x- and y- gradient components of the B-format outputs are used to model the mixing vectors, since it was found experimentally that the performance will degrade when p_0 is employed, the similar phenomenon also found in (Shujau et al., 2010). Assuming that only one source is dominant at each T-F unit, according to Equation (3), the STFT of the observations of the gradient components at the t-th frame can be represented as

$$\begin{aligned}\hat{\mathbf{X}}(\omega,t) &= \sum_{i=1}^{I} \hat{\mathbf{H}}_i(\omega) S_i(\omega,t) \\ &\approx \hat{\mathbf{H}}_i(\omega) S_i(\omega,t), \forall i \in [1,\ldots,I] \end{aligned} \tag{16}$$

where $\hat{\mathbf{X}}(\omega,t) = [G_x(\omega,t), G_y(\omega,t)]^T$, $\hat{\mathbf{H}}_i(\omega) = [H_x^i(\omega), H_y^i(\omega)]^T$. Each observation vector is then normalized to remove the effect of the source amplitude. The mixing filter coefficients, $\hat{\mathbf{H}}_i$, are modeled, similar to (Sawada et al., 2007), by a complex Gaussian density (CGD) function, given as

$$\begin{aligned} p_i^m(\hat{\mathbf{X}}(\omega,t)|\mathbf{a}_i(\omega),\gamma_i^2(\omega)) &= \frac{1}{\left(\pi\gamma_i^2(\omega)\right)^2} \\ &\times \exp\left(-\frac{||\hat{\mathbf{X}}(\omega,t) - (\mathbf{a}_i^H(\omega)\hat{\mathbf{X}}(\omega,t))\mathbf{a}_i(\omega)||^2}{\gamma_i^2(\omega)}\right) \end{aligned} \tag{17}$$

where $\mathbf{a}_i(\omega)$ is the centroid with a unit Frobenius norm $||\mathbf{a}_i(\omega)||^2 = 1$, and $\gamma_i^2(\omega)$ is the variance corresponding to the i-th source. The CGD function is evaluated for each observed T-F unit. The orthogonal projection of each

observation $\hat{\mathbf{X}}(\omega,t)$ onto the subspace spanned by $\mathbf{a}_i(\omega)$ can be estimated by $(\mathbf{a}_i^H(\omega)\hat{\mathbf{X}}(\omega,t))\mathbf{a}_i(\omega)$, where the superscript H denotes Hermitian. The minimum distance between the T-F unit $\hat{\mathbf{X}}(\omega,t)$ and the subspace is thus $||\hat{\mathbf{X}}(\omega,t) - (\mathbf{a}_i^H(\omega)\hat{\mathbf{X}}(\omega,t))\mathbf{a}_i(\omega)||$ and represents the probability of that T-F point of the mixture belonging to the $i-$th source. The probability of each T-F unit of the mixture coming from source i can thus be estimated by the normalization across the sources as $\hat{p}_i^m(\omega,t) = p_i^m(\omega,t)/\sum_l(p_l^m(\omega,t))$ where $\hat{p}_i^m(\omega,t)$ is estimated by Equation (17).

3.2. EM Algorithm

As mentioned before, the DOA distribution is blurred when the sources are close to each other, whereas the MV cue is more distinct under the same situation, as demonstrated by (Alinaghi et al., 2013). To improve the reliability of allocating each T-F unit to a specific source, we propose to combine the DOA cue $\theta(\omega,t)$ with the MV cue observed from $\hat{\mathbf{X}}(\omega,t)$, similar in spirit to (Alinaghi et al., 2011). The EM algorithm is employed to find the model parameters that best fit the observations $\{\theta(\omega,t), \hat{\mathbf{X}}(\omega,t)\}$. The parameter set Θ is given by

$$\Theta = \{\mu_i(\omega), k_i(\omega), \mathbf{a}_i(\omega), \gamma_i^2(\omega), \psi_i(\omega)\}$$

where $\mu_i(\omega)$ and $k_i(\omega)$ are the mean and concentration parameter of the DOAs, and $\mathbf{a}_i(\omega)$ and $\gamma_i^2(\omega)$ are the mean and variance of the mixing vector, and $\psi_i(\omega)$ is the mixing weight corresponding to the i-th source. Given an observation set, assuming the statistical independence between the two cues (Alinaghi et al., 2011), the parameters that maximize the log likelihood

$$\begin{aligned} L(\Theta) &= \max_{\Theta} \sum_{\omega,t} \log p(\theta(\omega,t), \hat{\mathbf{X}}(\omega,t)|\Theta) \\ &= \max_{\Theta} \sum_{\omega,t} \log \sum_i [\psi_i(\omega)\mathcal{V}(\theta(\omega,t)|\mu_i(\omega), k_i(\omega)) \\ &\qquad \times \mathcal{N}(\hat{\mathbf{X}}(\omega,t)|\mathbf{a}_i(\omega), \gamma_i^2(\omega))] \end{aligned} \tag{18}$$

can be estimated using the EM algorithm (Mandel et al., 2010) by iterating between the E-step and the M-step until convergence. In Equation (18), $\mathcal{V}\{*\}$ and $\mathcal{N}\{*\}$ represent the von Mises distribution and the complex Gaussian distribution, respectively.

In the E-step, given the parameters, Θ estimated at the M-step, and the observations, the posterior probability that the i-th source presents at each T-F unit of the mixture is calculated as

$$\begin{aligned}\nu_i(\omega,t) \propto \psi_i(\omega)\mathcal{V}(\theta(\omega,t)|\mu_i(\omega),k_i(\omega)) \\ \times\mathcal{N}(\hat{\mathbf{X}}(\omega,t)|\mathbf{a}_i(\omega),\gamma_i^2(\omega))\end{aligned} \tag{19}$$

where the symbol ' $\propto$' means combing the probabilities obtained by the two cues followed by the normalization across the sources.

In the M-step, the DOA parameters $(\mu_i(\omega), k_i(\omega))$ and the MV parameters $(\mathbf{a}_i(\omega), \gamma_i^2(\omega))$ are re-estimated for each source using the normalized probability $\nu_i(\omega,t)$ estimated in the E-step and the observations. As there is usually no prior information about the mixing filters, for the first iteration, we set $\mathcal{N}(\hat{\mathbf{X}}(\omega,t)|\mathbf{a}_i(\omega),\gamma_i^2(\omega)) = 1$ in (19) to remove the effect of the mixing vector contribution. Once the occupation probability $\nu_i(\omega,t)$ is obtained after one iteration based on only the information of DOA cue, the parameters of the mixing vectors, $(\mathbf{a}_i(\omega), \gamma_i^2(\omega))$, can be estimated from the next M-step as follows (Sawada et al., 2007)

$$\mathbf{R}_i(\omega) = \sum_t \nu_i(\omega,t)\hat{\mathbf{X}}(\omega,t)\hat{\mathbf{X}}^H(\omega,t) \tag{20}$$

$$\gamma_i^2(\omega) = \frac{\sum_t \nu_i(\omega,t)||\hat{\mathbf{X}}(\omega,t) - (\mathbf{a}_i^H(\omega)\hat{\mathbf{X}}(\omega,t))\mathbf{a}_i(\omega)||^2}{\sum_t \nu_i(\omega,t)} \tag{21}$$

the optimum $\mathbf{a}_i(\omega)$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{R}_i(\omega)$.

The parameters of the DOA can be updated by the DOAs which are belong to the set Ω in the M-step as (Hung et al., 2012)

$$\mu_i(\omega) = \tan^{-1}\left(\frac{\sum_t \nu_i(\omega,t)\sin(\hat{\theta}(\omega,t))}{\sum_t \nu_i(\omega,t)\cos(\hat{\theta}(\omega,t))}\right) \tag{22}$$

$$k_i(\omega) = A^{-1}\left(\frac{\sum_t \nu_i(\omega,t)\cos(\hat{\theta}(\omega,t) - \mu_i(\omega))}{\sum_t \nu_i(\omega,t)}\right) \tag{23}$$

$$\psi_i(\omega) = \frac{1}{T}\sum_t \nu_i(\omega,t) \tag{24}$$

where $\hat{\theta}(\omega,t)$ represents the reliable DOA values which are included in the set Ω, as calculated by Equation (26). In the current work, it was found

that the best results are obtained when the threshold is set as $\beta = 30°$, i.e. the DOAs which are more than 30° away from all the mean directions are excluded in the estimation of the DOA parameters. A^{-1} is a function that can be computed from Batschelet's table (Batschelet, 1981; Fisher, 1995) and T is the number of all time frames. After the convergence of the EM algorithm, the mask is finally obtained as

$$M_i(\omega, t) \equiv \nu_i(\omega, t) \tag{25}$$

3.3. Initialization and Dealing with the Permutation Problem

The EM algorithm can be initialized either from the E-step or the M-step. As there is usually no prior information about the MVs, similar to (Alinaghi et al., 2011), we initialize the mask with only the DOA cue. The parameters of the DOAs, $\mu_i(\omega)$ and $\kappa_i(\omega)$, are initialized as the peaks of the DOA histograms and 30° respectively. By using these accurate values in the initialization approach, the local optimality problem associated with the EM algorithm can be mitigated.

It should be mentioned that the probabilistic classification in this BSS method is performed for each frequency bin separately and thus the permutation alignment over all the frequency bins is still required. Rather than using a posteriori probability based approach as in (Sawada et al., 2007), due to its high computational cost, we use the information from the DOA cue to solve the permutation alignment problem in the first iteration of the EM algorithm, similar to (Alinaghi et al., 2011). As a result, the remaining iterations of the EM algorithm will not be affected by the permutation problem.

3.4. Reliability-based DOA Classification

It is noticed in (Levin et al., 2010) that the intensity-based DOA estimation method produces biased results under reverberant environment. To address this problem, a new approach based on thresholding is proposed next.

Under reverberant environments, the direction value at each T-F unit $\theta(\omega, t)$ via Equation (9), may contain the information of the sources or the reverberation. Obviously, the tail of the histogram of the DOAs will become broader with the increase of the reverberation level. To mitigate the reverberation effect, the un-reliable DOA estimates should be eliminated or play a less important role for T-F mask estimation.

The mean directions at each frequency $\mu_i(\omega)$, $i = 1, \cdots, I$ are estimated from the peak-finding approach in the first iteration, or from the M-step in the following iterations of the EM algorithm (as explained in Section 3.2). The angular difference between $\theta(\omega, t)$ and each mean direction $\mu_i(\omega)$ is calculated at each frequency bin, the directions which are close to any one of the mean directions are considered as the reliable ones, otherwise, they will be deemed as the points belonging to the reverberation. A set Ω is identified to collect all the reliable direction values at each frequency bin as

$$\Omega = \{\theta(\omega, t) | \cos(\theta(\omega, t) - \mu_i(\omega)) > \cos(\beta), \exists i\} \tag{26}$$

where β is the threshold of the angular difference between the estimated DOAs and the mean directions, which is found empirically in our experiments.

Then, the von Mises distribution is employed to model the DOAs which belong to Ω. For the DOA points which are excluded from Ω, the probability of the DOA cue is set identical and will be determined by the MV cue only, given as

$$p_i^d(\theta(\omega, t)|\mu_i(\omega), \kappa_i(\omega)) = \begin{cases} \dfrac{exp(\kappa_i(\omega)\cos(\theta(\omega, t) - \mu_i(\omega)))}{2\pi I_0(\kappa_i(\omega))}, & \theta(\omega, t) \in \Omega \\ 1/I, & otherwise \end{cases} \tag{27}$$

where $\mu_i(\omega)$ and $\kappa_i(\omega)$ represent the mean direction and the concentration parameter at each frequency corresponding to the i-th source, respectively.

The proposed algorithm is summarized in Algorithm 1.

4. Experiments and Results

To verify the effectiveness of the proposed method, we evaluate its performance with speech mixtures of a varying number of sources. As discussed in Section 2, although the B-format microphone is composed of four microphones, only three outputs (e.g. p_0, g_x, g_y) are used in our tests, and the output of g_z which carries the pressure gradient information at the vertical direction is discarded since in our experiment the sources and the microphone are placed in the same plane (i.e. with the same height in a three dimensional space). Thus, in this work, two, three, and four speech sources

Algorithm 1 soft T-F masking based source separation

Input: $p_0(n)$, $g_x(n)$, $g_y(n)$
Output: $y_i(n)$, $i = 1, \cdots, I$
T-F representation: $P_0(\omega,t) = \mathrm{STFT}(p_0(n))$, $G_x(\omega,t) = \mathrm{STFT}(g_x(n))$,
$G_y(\omega,t) = \mathrm{STFT}(g_y(n))$
calculate $\theta(\omega,t)$ {Equation (9)}
$\hat{\mathbf{X}}(\omega,t) = [G_x(\omega,t), G_y(\omega,t)]^T$
$\hat{\mathbf{X}} = \hat{\mathbf{X}}/||\hat{\mathbf{X}}||$ {normalization}
$\hat{\mathbf{X}} = \mathrm{PreWhitening}(\hat{\mathbf{X}})$
Initialization: $\mu_i = \mathrm{Peaks}(\theta(\omega,t))$, $\omega = 1, \cdots, round(length(\omega)/2)$,
$\kappa_i = 30^\circ$, $\psi_i(\omega) = 1/I$, $\beta = 30^\circ$
for $rep = 1 \rightarrow 16$ **do**
  for $i = 1 \rightarrow I$ **do**
    $p_i^d(\omega,t) = p(\theta(\omega,t)|\mu_i(\omega), k_i(\omega))$ {Equation (27).}
    $\hat{p}_i^d(\omega,t) = \frac{p_i^d(\omega,t)}{\sum_l p_l^d(\omega,t)}$, $l = 1, \cdots, I$ {normalization}
    if $rep < 2$ **then**
      $p_i^m(\omega,t) = 1$
    else
      $p_i^m(\omega,t) = p(\hat{\mathbf{X}}(\omega,t)|\mathbf{a}_i(\omega), \gamma_i^2(\omega))$ {Equation (17).}
    end if
    $\hat{p}_i^m(\omega,t) = \frac{p_i^m(\omega,t)}{\sum_l p_l^m(\omega,t)}$, $l = 1, \cdots, I$ {normalization}
    $\hat{\nu}_i(\omega,t) = \psi_i(\omega)\hat{p}_i^d(\omega,t)\hat{p}_i^m(\omega,t)$
    $\nu_i(\omega,t) = \frac{\hat{\nu}_i(\omega,t)}{\sum_l \hat{\nu}_l(\omega,t)}$ {normalization}
    Update $\mu_i(\omega), k_i(\omega)$ {Equation (22) and (23).}
    if $rep \geq 2$ **then**
      Update $\mathbf{a}_i(\omega), \gamma_i^2(\omega)$ {Equations (20) and (21).})
    end if
    Update $\psi_i(\omega)$ {Equation (24).}
  end for
  $M_i(\omega,t) = \nu_i(\omega,t)$
  $Y_i(\omega,t) = M_i(\omega,t)P_0(\omega,t)$
  $y_i(n) = \mathrm{ISTFT}(Y_i(\omega,t))$
end for

are considered for the over-, even- and under-determined source separation scenarios, respectively.

As mentioned in Section 1, four methods are implemented and used as baselines for performance comparison with the proposed method. First, the two DOA-based separation algorithms (Shujau et al., 2011; Gunel et al., 2008), denoted as 'Gunel' and 'Shujau', respectively, which we have discussed in Section 2.2, are employed to show the performance of the DOA cue based source separation. Then, the bin-wise clustering method (Sawada et al., 2011), referred to as 'Sawada', is adopted to demonstrate the separation performance based only on the mixing vector cue. Finally, the convolutive ICA method (Wang et al., 2005) by exploiting the second-order statistics in the frequency domain is included, which we refer to as 'Wang'. The results by comparing the mixtures with the original sources are also calculated as ref-

erences, which we denote as 'Mixture'. It should be noted that the methods of 'Gunel', 'Shujau', as well as the proposed method, are evaluated based on the outputs of the B-format microphone (p_0, g_x, g_y) directly. However, for the methods 'Sawada' and 'Wang', we considered both the B-format microphone recordings, denoted as 'Sawada-B' and 'Wang-B', respectively, and the recordings with a standard 4-microphone tetrahedral array (L_F, L_B, R_F, R_B) obtained by inverting Equation (1), denoted as 'Sawada-O' and 'Wang-O', respectively..

The experimental setup and the evaluation metrics are introduced first, followed by the separation results for both the synthetic data obtained using a simulated room model and the real room recordings collected in a reverberant studio.

4.1. Experimental Setup

To study the effect of room reverberation, we first test the behavior of the proposed and the baseline methods under various reverberation levels using a simulated room model. As shown in Figure 4 (a), a shoe-box room with a dimension of $9 \times 5 \times 4m^3$ was employed. The B-format microphone was located at the center of the room, as illustrated in Figure 1. The L_F, R_F, L_B, R_B of the B-format microphone were collocated at $(0.005, 0.005, 0.005)$, $(0.005, -0.005, -0.005)$, $(-0.005, 0.005, -0.005)$, $(-0.005, -0.005, 0.005)$, respectively, where the coordinate unit is in meter. The speech sources were fixed at a horizontal distance of 1.5 m to the origin $(0, 0, 0)$ of the microphone. 15 utterances, each with a length of approximately 3 s were randomly chosen from the TIMIT dataset[1] and then shortened to 2.5 s to avoid the silence at the end. Note that the utterances selected contain both male and female speech. Moreover, all the speech signals were normalized before convolving with the room models which were simulated by using the imaging method (Allen and Berkley, 1979) with the reverberation time varied from 0 s to 0.6 s with 0.1 s intervals. 15 pairs of mixtures were chosen randomly from the 15 utterances. In each experimental condition, the first signal (s_1) was fixed at 0°, and other sources were located 50° away with the neighboring source, the position of each source is shown in Figure 4 (a).

[1]TIMIT dataset, widely used by the speech separation and recognition community, is generally considered as a dataset of wideband signals and therefore chosen for the performance evaluation in our work.

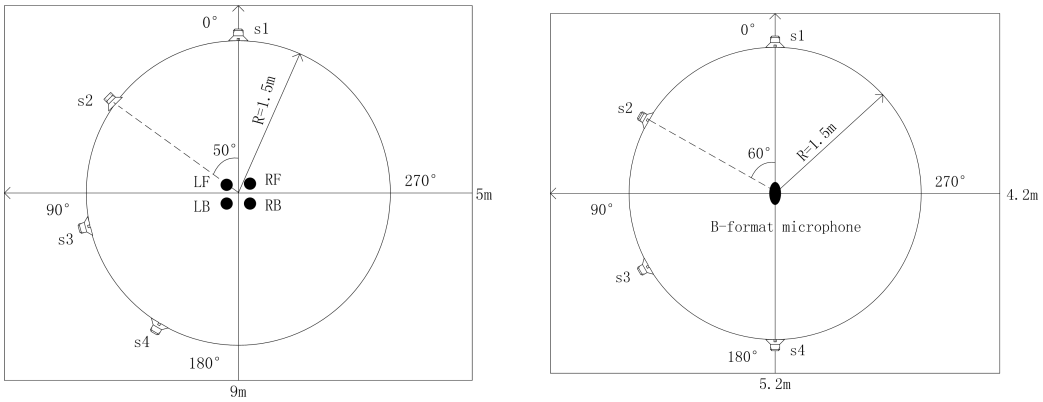


Figure 4: Experimental setup for the B-format recordings in (a) the simulated room model, (b) the studio with a reverberation time of approximately 0.6 s.

The B-format signals were also collected in a real studio ($5.2\times4.2\times2.1m^3$) in University of Surrey with the reverberation time of approximately 0.6 s depicted in Figure 4 (b). The B-format microphone was kept at the center of the studio. Similar to the system setup for the synthetic data, the loudspeaker was 1.5 m away from the microphone, and also, both the loudspeakers and the microphone were 1.2 m above the floor to ensure that the recordings would not be affected by the vertical direction. 15 utterances (include both the male and female speakers) were chosen randomly from the same dataset as for the synthetic data, and the first 2.5 s were selected and played by a loudspeaker (Genelec 1030A). The recordings were collected at 44.1 kHz by a SoundField B-format microphone system (SPS422B), and then down-sampled to 16 kHz before being processed. Based on the linearity and time-invariance assumption, the convolutive mixtures were obtained by collecting all the recordings at 0° to 350° with 10° intervals separately, and then summing several (i.e. two, three, or four) recordings at different directions together. Before the collection of each recording, all the utterances were normalized to have the same root mean square energy.

To investigate the effect of source configuration, the speech sources were located with various azimuths for generating the mixtures. When collecting the mixtures in the real studio, the first source s_1 was fixed at 0° for all the experimental cases, other sources were arranged counter clockwise with the same angular difference between the neighboring sources, as shown in Table

1. The angular difference $\Delta\theta$ is varied from 10° to 90° with 10° increasing intervals for the two (i.e. s1, s2), three (i.e. s1, s2 and s3) and four (i.e. s1, s2, s3 and s4) sources case. In Figure 4 (b), an example of the arrangement of four sources at 60° angular difference is shown.

s1	0°	0°	0°	0°	0°	0°	0°	0°	0°
s2	10°	20°	30°	40°	50°	60°	70°	80°	90°
s3	20°	40°	60°	80°	100°	120°	140°	160°	180°
s4	30°	60°	90°	120°	150°	180°	210°	240°	270°
$\Delta\theta$	10°	20°	30°	40°	50°	60°	70°	80°	90°

Table 1: All the orientations of the sources with different angular difference ($\Delta\theta$).

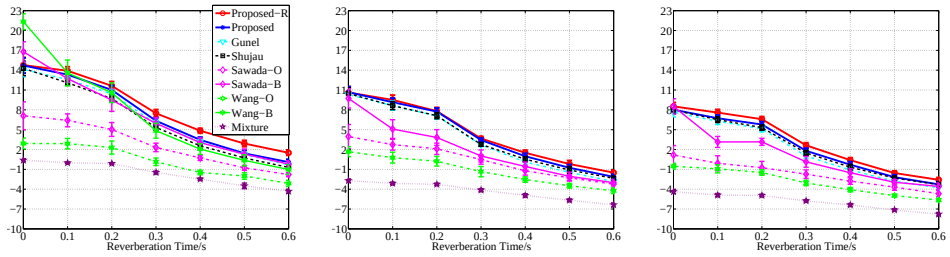
We implemented the baseline methods ourselves and tested them with the same mixtures as for the proposed method. The frame size of the STFT of the mixtures is 1024, with 75% overlap between the neighboring frames. The iteration number of the EM algorithm is chosen as 16 in the Sawada's method and the proposed method.

In Sawada's method, the parameters of the mean value α_i and the variance γ_i^2 are initialized as $1/I$ and 0.1 respectively, the same as in (Sawada et al., 2011). For Gunel's algorithm, following the work in (Gunel et al., 2008), the 6-dB beamwidth is spanned from 10° to 180° with 10° intervals to calculate the related concentration parameters κ.

4.2. Evaluation Metrics

In this work, to quantify the quality of the separated sources, both the signal-to-distortion ratio (SDR) (Vincent et al., 2006) and the perceptual evaluation of speech quality (PESQ) (Loizou, 2007; Di Persia et al., 2008) are evaluated.

The SDR is defined as the ratio of the energy in the original signal to the energy in the interference from other signals and artifacts (i.e. reverberation). The energy of the target signal can be obtained by the energy in the estimated signal y_i which can be considered as a linear combination of delayed version of the original signal s_i. The remaining energy in the estimated signal which does not belong to the target is considered as the distortion energy, including the interference and artifact energy.



(a) (b) (c)

Figure 5: The SDR results in dB for the simulated mixture of (a) two sources, (b) three sources and (c) four sources versus various reverberation times.

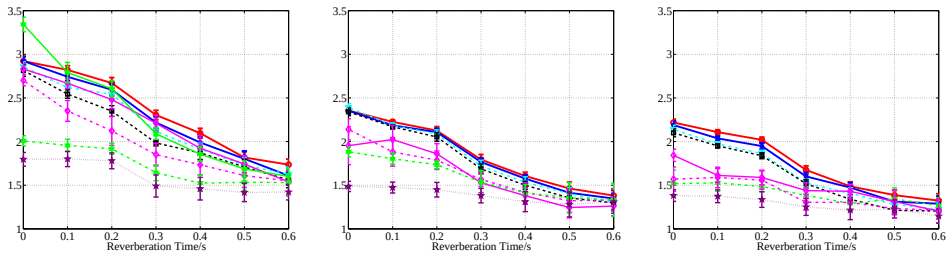
The SDR is calculated as the averaged value for each source

$$SDR = \frac{1}{I}\sum_{i=1}^{I} 10\log_{10}\left(\frac{E\{(s_i)^2\}}{E\{(y_i - s_i)^2\}}\right) \tag{28}$$

where I is the number of the sources.

We also evaluate the PESQ by using the ITU-PESQ software (Thiede et al., 2000). The separated signal is compared with the original clean signal to evaluate the perceptual quality of the separated speech using the Mean Opinion Score (MOS). As noted in (Mandel et al., 2010), the MOS has the range from -0.5 to 4.5, with -0.5 and 4.5 indicating the worst and the best quality of the separated speech, respectively. It is worth noting that PESQ was originally proposed to quantify the perceptual speech quality of telephone networks and speech coding. For example, it is often used to measure the impairment of a speech codec. However, due to its popularity in predicting subjective quality of a speech signal, PESQ has also been widely used in speech separation community for perceptual quality evaluation of separated speech sources.

In order to investigate whether the proposed method shows significant improvements compared with the baseline methods, the one-way ANOVA test (Hoel et al., 1960) is also performed with the significance level set at 5%, and the p-values are calculated to determine whether the performance difference between the methods is statistically significant.



(a) (b) (c)

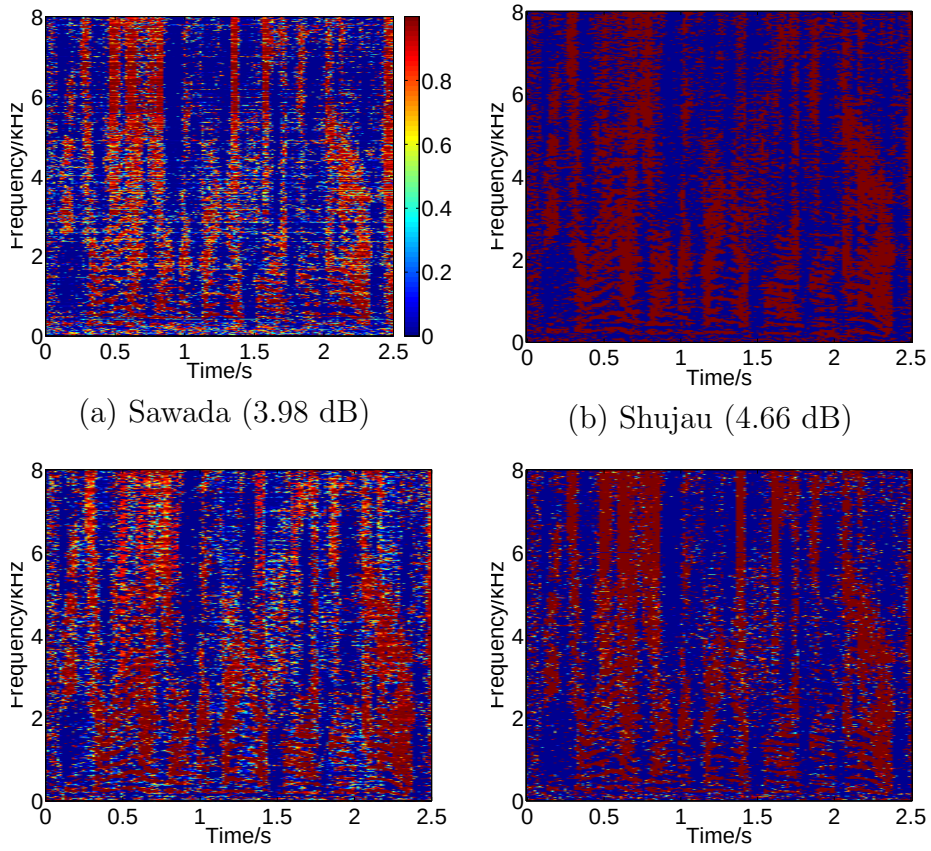
Figure 6: The PESQ results for the simulated mixture of (a) two sources, (b) three sources and (c) four sources versus various reverberation times.

4.3. Experimental Results

4.3.1. Results for the synthetic data

Figure 5 shows the SDRs versus T_{60}s for the mixtures of two, three and four sources respectively, with the confidence intervals shown as bars surrounding the means in the plots. As expected, the SDR values decrease when the reverberation level increases. The proposed method ('Proposed') performs better than the baseline methods, giving an improvement of 0.47/0.91 dB, 0.43/0.65 dB, and 0.22/0.60 dB, averaged over all the reverberation levels, as compared with 'Gunel'/'Shujau' under the two, three and four sources cases, respectively. The proposed method based on the reliability information ('Proposed-R') can further improve the separation performance, on average, giving 1.42/1.87 dB, 0.77/0.98 dB, and 0.94/1.32 dB improvements as compared to 'Gunel'/'Shujau', respectively.

As shown in Figure 5, with the same methods, the separation results based on B-format microphone recordings ('Sawada-B' and 'Wang-B') appear to be better than those based on omnidirectional microphone recordings ('Sawada-O' and 'Wang-O'). Note that the omnidirectional microphone recordings are obtained virtually based on the B-format recordings as discussed earlier in this section. It can be seen that under anechoic condition, the ICA method ('Wang-B') outperforms the T-F masking based approaches for B-format recordings. However, with the increase in room reverberation, the methods of 'Proposed'/'Proposed-R' show on average 1.18/1.86 dB improvements as compared with 'Wang-B' for the reverberant cases, and giving an improvement of 0.67/1.35 dB, as compared with 'Sawada-B'. The corresponding improvements are 4.1/4.6 dB and 5.9/6.5 dB, as compared with 'Sawada-O' and 'Wang-O', respectively.



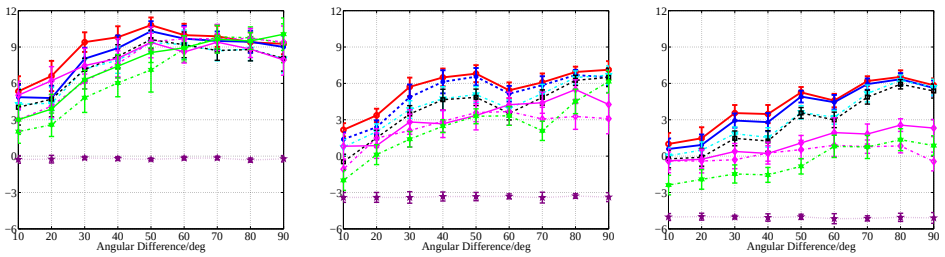
(a) Sawada (3.98 dB) (b) Shujau (4.66 dB)

(c) Gunel (5.62 dB) (d) Proposed with reliability (6.53 dB)

Figure 7: The example masks obtained from B-format recordings by the different algorithms (a) Sawada (i.e. 'Sawada-B'), (b) Shujau, (c) Gunel and (d) proposed method with reliability information, with three speakers located at 0°, 50° and 100° under 0.6 s reverberation. The SDR results in dB corresponding to each method are also shown.

The PESQ results follow the similar trend to the SDR results, as shown in Figure 6. The avarage improvements of 'Proposed'/'Proposed-R' are approximately 0.05/0.1, 0.1/0.15, and 0.18/0.22,, as compared with 'Gunel', 'Shujau', and 'Sawada-B', respectively.

Furthermore, the p-value is estimated by the one-way ANOVA test to determine whether the proposed method gives significant improvements compared with the baseline methods. For the significance level at 5%, the results are considered as statistically significant if the p-value is smaller than 0.05. The p-value of the SDR results (number of mixtures=315) are 1.42×10^{-8},



(a) (b) (c)

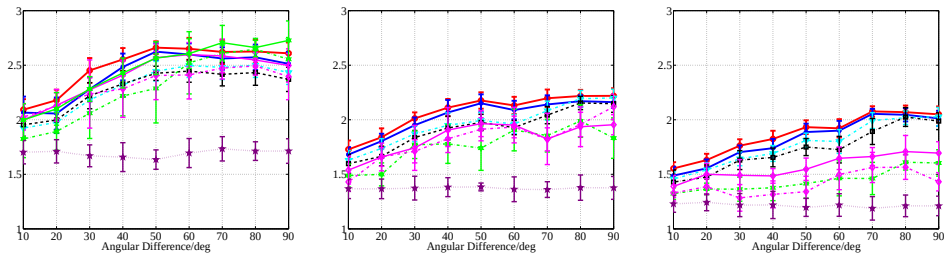
Figure 8: The SDR results in dB for the real collected mixture of (a) two sources, (b) three sources and (c) four sources versus different angular difference ($\Delta\theta$).

2.14×10^{-10}, and 1.48×10^{-22}, by comparing the proposed method with 'Gunel', 'Shujau', and 'Sawada-B', respectively. Thus the improvements by the proposed method are statistically significant as compared with the baseline methods.

It is worth noting that the results of the baseline methods of 'Sawada-B' and 'Wang-B' are obtained based on the $x-$ and $y-$ gradient components of the B-format outputs (g_x, g_y), as we found that the separation performance would degrade when the component p_0 is included. To show this, we present a comparison of the SDR results between discarding and including the pressure component, denoted as 'Sawada-B'/'Sawada-B-3input', 'Wang-B'/'Wang-B-3input' respectively, which were obtained by 15 pairs of mixtures with two sources located at $(40°, 70°)$, and three sources located at $(40°, 70°, 100°)$ and $(40°, 100°, 160°)$ respectively (see Figure 2). The results are shown in Table 2. Due to the common limitation of the ICA algorithms, the separation results of 'Wang-B' are only shown for two sources case, and hence for the three-source case, no results (denoted by '-') are shown in this table.

Direction of sources	Sawada-B/Sawada-B-3input	Wang-B/Wang-B-3input
$40°, 70°$	7.58/5.74 dB	5.88/4.93 dB
$40°, 70°, 100°$	2.92/2.01 dB	−/1.88 dB
$40°, 100°, 160°$	5.10/4.96 dB	−/3.92 dB

Table 2: The SDR results in dB of two baseline methods ('Sawada', 'Wang') by discarding and including the pressure component of the B-format microphone recordings, respectively.



(a) (b) (c)

Figure 9: The PESQ results for the real collected mixture of (a) two sources, (b) three sources and (c) four sources versus different angular difference ($\Delta\theta$).

4.3.2. Results for the real data

In Figure 7, an example is given to show the T-F mask obtained by the proposed method with reliability information based on the DOA values in set Ω, and three baseline methods, respectively. The SDR results corresponding to each mask are also shown in the brackets for comparison.

In Figures 8 and 9, the SDR and PESQ results, which are obtained by averaging over 15 pairs of mixtures at each angular difference, are plotted against the angular difference between the two, three and four sources, respectively. As can be observed from the SDR and PESQ results, the performance gradually deteriorates with the increase in the number of sources.

Almost for all angular differences, the proposed method shows better separation performance than the competing methods. It is because the two DOA-based methods ('Gunel', 'Shujau') rely on the mean directions estimated, which become less accurate and reliable when the sources are located close to each other, especially in highly reverberant environments.

In the proposed method, however, the mean directions are only used at the initialization stage, the parameters of DOA and mixing vector cues are updated iteratively at each frequency bin to improve the estimates towards the true value. The averaged SDR improvements of the proposed method (without the reliability measure) over all the angle differences are about 0.87/0.80/0.53 dB, 0.76/1.05/1.84 dB, and 0.74/1.05/2.76 dB under two, three and four sources cases, compared with the methods of 'Gunel', 'Shujau', and 'Sawada-B", respectively.

The reliability-based approach can further improve the separation performance by removing the un-reliable direction information which is caused by the reverberation. The corresponding SDR improvements are around

1.33/1.41/1.14 dB, 1.27/1.66/2.36 dB, and 1.12/1.42/3.14 dB compared with 'Gunel'/'Shujau'/'Sawada-B', for the mixture of two, three, and four sources, respectively. The p-value of the SDR results (number of mixtures=405) are 4.09×10^{-22}, 7.02×10^{-24}, and 7.20×10^{-30}, by comparing the proposed method with 'Gunel', 'Shujau', and 'Sawada-B', respectively.

The PESQ results follow the trend of the SDR results quite closely. Compared with 'Gunel', 'Shujau', and 'Sawada-B', the proposed method (without the reliability measure) shows approximately 0.08, 0.11,and 0.23 improvements, under two, three, and four sources case respectively, the corresponding improvements are 0.13, 0.17, and 0.29 for the reliability-based method.

For the two sources case, the SDR improvements of 'Proposed'/'Proposed-R' are 0.94/1.55 dB, and the corresponding PESQ results are 0.02/0.05, compared with the method of 'Wang-B'.

In addition, we have also added the step of reliability based DOA classification to the methods of 'Gunel' and 'Shujau', and the results are denoted by 'Gunel-R' and 'Shujau-R', respectively. The SDR results are tested under the same situation with Table 2. As shown in Table 3, similar to the proposed method, the performance of both baseline methods has been improved using the reliability based DOA classification.

Direction of sources	Proposed-R/Proposed	Gunel-R/Gunel	Shujau-R/Shujau
$40°, 70°$	10.18/8.23 dB	8.06/7.62 dB	7.98/7.51 dB
$40°, 70°, 100°$	4.54/3.36 dB	3.13/2.81 dB	3.07/2.71 dB
$40°, 100°, 160°$	6.70/6.45 dB	5.58/5.31 dB	5.57/5.22 dB

Table 3: The SDR results in dB of the proposed method and two baseline methods with and without the step of reliability-based DOA classification, respectively.

5. Conclusions

We have presented a new algorithm for the separation of convolutive mixtures by incorporating the intensity vector of the acoustic field with probabilistic time-frequency masking. The DOA and mixing vector cues are then modeled by the von Mises mixture model and complex Gaussian mixture model respectively, the parameters of which are updated iteratively via the EM algorithm to estimate and refine the probability of each T-F unit of the mixture belonging to each source. Based on this, a reliability-based method is also introduced to improve the performance of source separation in which

the points that are far away from all the mean directions are considered as the outliers due to the effect of room reverberation.

The proposed method has been tested extensively for the mixture of two, three and four speech sources respectively under the simulated room model with different reverberation level, and also for real recordings acquired in a reverberant studio with the reverberation time of approximately 0.6 s with various angular intervals. The proposed method shows better separation performance in SDR and PESQ as compared with the baseline methods under almost all the situations tested.

Acknowledgment

This work was conducted during Xiaoyi Chen's visit at the Centre for Vision Speech and Signal Processing at University of Surrey. The authors wish to thank the anonymous reviewers and the associate editor for their contributions in improving the quality of the paper.

References

Aichner, R., Araki, S., Makino, S., Nishikawa, T., Saruwatari, H., 2002. Time domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming, in: 12th IEEE Workshop on Neural Networks for Signal Processing, pp. 445–454.

Alinaghi, A., Wang, W., Jackson, P.J., 2011. Integrating binaural cues and blind source separation method for separating reverberant speech mixtures, in: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 209–212.

Alinaghi, A., Wang, W., Jackson, P.J., 2013. Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation, in: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 684–688.

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. The Journal of the Acoustical Society of America 65, 943–950.

Amari, S.I., Chen, T.P., Cichocki, A., 1997. Stability analysis of learning algorithms for blind source separation. Neural Networks 10, 1345–1351.

Araki, S., Makino, S., Murai, R., Saruwatari, H., 2001. Equivalence between frequency domain blind source separation and frequency domain adaptive null beamformers, in: the 7th European Conf. on Speech Communication and Technology, pp. 2595–2598.

Araki, S., Mukai, R., Makino, S., Nishikawa, T., Saruwatari, H., 2003. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. IEEE Trans. Speech and Audio Processing 11, 109–116.

Batschelet, 1981. Circular statistics in biology. Academic Press.

Buchner, H., Aichner, R., Kellermann, W., 2004. Trinicon: A versatile framework for multichannel blind signal processing, in: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing.

Chen, X., Alinaghi, A., Zhong, X., Wang, W., 2013. Acoustic vector sensor based speech source separation with mixed gaussian-laplacian distributions, in: Proc. IEEE Int. Conf. on Digital Signal Processing (DSP), pp. 1–5.

Comon, P., 1994. Independent component analysis, a new concept? Signal processing 36, 287–314.

Comon, P., Jutten, C., 2010. Handbook of Blind Source Separation: Independent component analysis and applications. Access Online via Elsevier.

D O'Grady, P., Pearlmutter, B.A., 2008. The lost algorithm: finding lines and separating speech mixtures. EURASIP on Advances in Signal Processing 2008, 1–17.

Di Persia, L., Milone, D., Rufiner, H.L., Yanagida, M., 2008. Perceptual evaluation of blind source separation for robust speech recognition. Signal Processing 88, 2578–2583.

Farrar, K., 1979. Soundfield microphone. Wireless World 85, 48–50.

Fisher, N.I., 1995. Statistical analysis of circular data. Cambridge University Press.

Gunel, B., Hachabiboglu, H., Kondoz, A.M., 2008. Acoustic source separation of convolutive mixtures based on intensity vector statistics. IEEE Trans. Audio, Speech, and Language Processing 16, 748–756.

Hawkes, M., Nehorai, A., 1998. Acoustic vector-sensor beamforming and capon direction estimation. IEEE Trans. Signal Processing 46, 2291–2304.

Hawkes, M., Nehorai, A., 2000. Acoustic vector-sensor processing in the presence of a reflecting boundary. IEEE Trans. Signal Processing 48, 2981–2993.

Hoel, P.G., et al., 1960. Elementary statistics. Elementary statistics .

Hung, W.L., Chang-Chien, S.J., Yang, M.S., 2012. Self-updating clustering algorithm for estimating the parameters in mixtures of von mises distributions. Journal of Applied Statistics 39, 2259–2274.

Hyvärinen, A., Hurri, J., Hoyer, P.O., 2009. Independent component analysis, in: Natural Image Statistics. Springer, pp. 151–175.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Networks 13, 411–430.

Lee, T.W., 1998. Independent component analysis. Springer.

Levin, D., Habets, E.A., Gannot, S., 2010. On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields. The Journal of the Acoustical Society of America 128, 1800–1811.

Loizou, P., 2007. Speech enhancement: theory and practice. CRC, Boca Raton, FL .

Malham, D.G., Myatt, A., 1995. 3-D sound spatialization using ambisonic techniques. Computer Music Journal 19, 58–70.

Mandel, M.I., Weiss, R.J., Ellis, D., 2010. Model-based expectation-maximization source separation and localization. IEEE Trans. Audio, Speech, and Language Processing 18, 382–394.

Nehorai, A., Paldi, E., 1994. Acoustic vector-sensor array processing. IEEE Trans. Signal Processing 42, 2481–2491.

Nishikawa, T., Saruwatari, H., Shikano, K., 2003. Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA. IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences 86, 846–858.

Parra, L., Spence, C., 2000. Convolutive blind separation of non-stationary sources. IEEE Trans. Speech and Audio Processing 8, 320–327.

Pedersen, M.S., Larsen, J., Kjems, U., Parra, L.C., 2007. A survey of convolutive blind source separation methods. Multichannel Speech Processing Handbook , 1065–1084.

Saruwatari, H., Kurita, S., Takeda, K., 2001. Blind source separation combining frequency-domain ICA and beamforming, in: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 2733–2736.

Sawada, H., Araki, S., Makino, S., 2007. A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 139–142.

Sawada, H., Araki, S., Makino, S., 2011. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans. Audio, Speech, and Language Processing 19, 516–527.

Sawada, H., Araki, S., Mukai, R., Makino, S., 2006. Blind extraction of dominant target sources using ICA and time-frequency masking. IEEE Trans. Audio, Speech, and Language Processing 14, 2165–2173.

Sawada, H., Mukai, R., Araki, S., Makino, S., 2004. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. Speech and Audio Processing 12, 530–538.

Sawada, H., Mukai, R., Araki, S., Makino, S., 2005. Frequency-domain blind source separation, in: Speech Enhancement. Springer, pp. 299–327.

Shujau, M., Ritz, C.H., Burnett, I.S., 2010. Speech enhancement via separation of sources from co-located microphone recordings, in: IEEE Int. Conf. on Acoustics Speech and Signal Processing, pp. 137–140.

Shujau, M., Ritz, C.H., Burnett, I.S., 2011. Separation of speech sources using an acoustic vector sensor, in: IEEE Workshop on Multimedia Signal Processing, pp. 1–6.

Smaragdis, P., 1998. Blind separation of convolved mixtures in the frequency domain. Neurocomputing 22, 21–34.

Stone, J.V., 2004. Independent component analysis. Wiley Online Library.

Thiede, T., Treurniet, W.C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J.G., Colomes, C., 2000. PEAQ-the ITU standard for objective measurement of perceived audio quality. Journal of the Audio Engineering Society 48, 3–29.

Thomas, J., Deville, Y., Hosseini, S., 2006. Time-domain fast fixed-point algorithms for convolutive ICA. IEEE Signal Processing Letters 13, 228–231.

Vincent, E., Gribonval, R., Févotte, C., 2006. Performance measurement in blind audio source separation. IEEE Trans. Audio, Speech, and Language Processing 14, 1462–1469.

Wang, D., Kjems, U., Pedersen, M.S., Boldt, J.B., Lunner, T., 2009. Speech intelligibility in background noise with ideal binary time-frequency masking. The Journal of the Acoustical Society of America 125, 23–36.

Wang, W., Sanei, S., Chambers, J.A., 2005. Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources. IEEE Trans. Signal Processing 53, 1654–1669.

Yilmaz, O., Rickard, S., 2004. Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Processing 52, 1830–1847.

Zhong, X., Chen, X., Wang, W., Alinaghi, A., 2013. Acoustic vector sensor based reverberant speech separation with probabilistic time-frequency masking, in: the 21th European Signal Processing Conference (EUSIPCO).

Zhong, X., Premkumar, A.B., 2012. Particle filtering approaches for multiple acoustic source detection and 2-d direction of arrival estimation using a single acoustic vector sensor. IEEE Trans. Signal Processing 60, 4719–4733.