Joint Mixing Vector and Binaural Model Based Stereo Source Separation

Atiyeh Alinaghi, *Member, IEEE*, Philip JB Jackson, *Member, IEEE*, Qingju Liu, *Member, IEEE*, and Wenwu Wang, *Senior Member, IEEE*

***Abstract*—In this paper the mixing vector (MV) in the statistical mixing model is compared to the binaural cues represented by interaural level and phase differences (ILD and IPD). It is shown that the MV distributions are quite distinct while binaural models overlap when the sources are close to each other. On the other hand, the binaural cues are more robust to high reverberation than MV models. According to this complementary behavior we introduce a new robust algorithm for stereo speech separation which considers both additive and convolutive noise signals to model the MV and binaural cues in parallel and estimate probabilistic time-frequency masks. The contribution of each cue to the final decision is also adjusted by weighting the log-likelihoods of the cues empirically. Furthermore, the permutation problem of the frequency domain blind source separation (BSS) is addressed by initializing the MVs based on binaural cues. Experiments are performed systematically on determined and underdetermined speech mixtures in five rooms with various acoustic properties including anechoic, highly reverberant, and spatially-diffuse noise conditions. The results in terms of signal-to-distortion-ratio (SDR) confirm the benefits of integrating the MV and binaural cues, as compared with two state-of-the-art baseline algorithms which only use MV or the binaural cues.**

***Index Terms*—Blind source separation, computational auditory scene analysis, reverberation, time-frequency masking.**

I. Introduction

HEARING aids, automatic speech recognition (ASR) and many other communication systems work reasonably well when there is just one source with almost no echo, but their performance degrades in situations where there are more speakers talking simultaneously or the reverberation is high. Therefore, it is highly desirable to localize and separate the source signals as an auditory front-end especially when the source signals and the mixing process are unknown, introducing a blind estimation problem.

There have been various methods suggested to perform blind source separation (BSS) such as *independent component analysis* (ICA) [1]–[3] and beamforming [4] which need as many mixtures available as the number of sources. To deal with *underdetermined* cases, when the number of mixtures is smaller than that of sources, the signals are transformed into the time-frequency (T-F) domain where the signals become sparse and the sources can be segregated using T-F masks [5].

In [6], probabilistic (soft) masks are generated based on the posterior probability of each source at each T-F unit. The algorithm starts with the statistical model of the mixture signals with additive noise. For T-F units, dominated by one source, the mixing matrix can be replaced with a mixing vector. Both the mixing vector and the active source are latent variables which are estimated by clustering the observation vectors at each frequency bin. Although this BSS technique offers good performance based on SiSEC 2008 Data [7], it degrades as the reverberation time increases.

On the other hand, the human auditory system with just two ears has shown great performance in source separation [8], [9], which has been studied and modelled under the name of *computational auditory scene analysis* (CASA) [10], [11]. In CASA, there are two groups of monaural and binaural cues associated with the features extracted from one or a pair of mixtures, respectively. Among monaural cues, fundamental frequency is the most studied feature which is only effective for voiced speech [12]. In some approaches such as [13] and [14], pitch information is integrated with spatial cues to improve the results. However, their performance depends on accurate pitch estimation which is difficult when there are multiple sources with overlapping frequency components. Here, we focus on binaural cues which contain spatial information.

In [15], the two main binaural cues, namely interaural level difference (ILD) and interaural phase difference (IPD), are applied in a probabilistic context which shows significant improvement over existing algorithms including [5] and [16]. However, it performs poorly when the sources are close to each other with small angular displacement. In [17] the monaural cues are integrated with binaural cues for reverberant speech segregation. Albeit they reported better results compared to [15], they exploited a large training set with known azimuth of the sources which is not always available. Moreover, their method only recovers one (the target) source, while in [15] all the sources can be estimated.

In this paper, we study the method based on mixing vector (MV) estimation [6] and the technique using binaural cues [15] to investigate the strengths and weaknesses of these two approaches. We found that the MV models seem to be more distinct compared to ILD and IPD models for sources that are close to each other. On the other hand, for spatially separated sources the binaural cues become easily distinguishable while MV models may overlap. Moreover, we examined the effect of

Manuscript received September 10, 2013; revised January 31, 2014; accepted April 17, 2014. Date of publication April 29, 2014; date of current version July 18, 2014. This work was supported in part by the EPSRC of the U.K. under Grants EP/H012842/1 and EP/H050000/1 and in part by EPSRC (U.K.) Grant EP/L000539/1. The work of A. Alinaghi was supported by CVSSP. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yunxin Zhao.

The authors are with the Centre for Vision, Speech and Signal Processing (CVSSP), Department of Electronic Engineering, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: a.alinaghi@surrey.ac.uk; p.jackson@surrey.ac.uk; q.liu@surrey.ac.uk; w.wang@surrey.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2320637

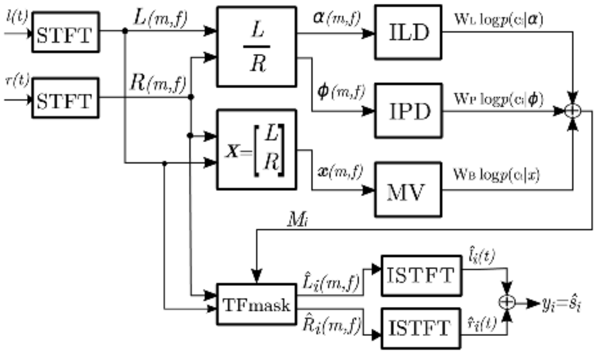


Fig. 1. Block diagram of the proposed algorithm. The two recorded mixtures, $l(t)$, and $r(t)$, are transformed to the time-frequency domain and three different features, ILD, IPD and MV are extracted from each T-F unit. The likelihood of each source being dominant based on each cue is estimated and weighted to contribute to the final soft mask. The mask is then applied to the mixture spectrograms to extract the sources.

two different types of noise on these models and found that MV models deviate due to convolutive noise but are robust to additive noise. On the contrary, the binaural cues, and especially IPD, are robust to convolutive noise introduced by reverberation and degrade in the presence of additive noise. These observations confirm the complementary role of spatial cues in statistical and binaural models under various conditions which motivated us to combine the models, introducing a new robust algorithm.

In our proposed algorithm, soft T-F masks are estimated to recover the source signals from stereo recordings considering additive and convolutive noise models. This technique also prevents the permutation problem by initializing the T-F masks based on binaural cues. The model parameters and posterior probability of the sources are estimated iteratively using the expectation-maximization (EM) algorithm. The final score of each source being active at each T-F unit is calculated based on a weighted combination of three source models, i.e., ILD, IPD and MV, according to the reliability of each cue. An overview of our method is represented in Fig. 1. Instead of omnidirectional microphone recordings, binaural recordings, are considered in this paper, as often encountered in applications such as hearing aids, robotics and spatial audio.

We examined the performance of our proposed method and the two state-of-the-art algorithms by Mandel *et al.* [15], and Sawada *et al.* [6] in rooms with a typical range of acoustic properties. It is shown that the proposed technique outperforms the two baselines especially under challenging conditions when the sources are close to each other or the reverberation is high.

This paper is organized as follows. Section II discusses the extraction of the cues from the binaural signals in the time and frequency domains with some simplifications. Section III explains the source models and the proposed source separation algorithm in detail. Comparison between the statistical and binaural models is covered in Section IV. Implementation details of the proposed algorithm are discussed in Section V. Experiments are reported in Section VI, followed by Section VII discussing the results. Section VIII summarizes the results and envisages future work.

II. Statistical and Binaural Mixture Models

As mentioned in Section I, we consider both additive and convolutive noise signals associated with statistical and binaural models. Accordingly, the MVs and binaural cues for each source are estimated based on different noise models, as discussed in the following section.

A. Observations in the Time and Frequency Domain

We suppose that there are N sources and M microphones where in the case of binaural recordings $M = 2 \leq N$. It is also assumed that the number of sources, N, is known *a priori*. The recorded signals in a room with reverberation are filtered versions of the source signals added together at each microphone. If $x_k(t)$ is the signal recorded by the kth microphone in the time domain and $s_i(t)$ the ith source, then (1) holds where h_{ik} is the room impulse response (RIR) from source i to microphone k with n_k^a and n_k^c representing the additive and the convolutive noise signals, respectively:

$$x_k(t) = \sum_{i=1}^{N} s_i(t) * h_{ik}(t) * n_k^c(t) + n_k^a(t), \tag{1}$$

where t is the discrete time index, $*$ denotes convolution, and the superscripts a and c represent the additive and convolutive noise terms respectively.

The RIRs become longer with the increase of reverberation, making the process of separation computationally expensive in the time domain [18]. Therefore, using short-time Fourier transform (STFT), the signals are mapped into the T-F domain, where the speech signals are more sparse. Another motivation to work in the T-F domain is that, as suggested by CASA theory, the human auditory system also performs a short-time spectral analysis.

The mixture model (1) can also be represented in T-F domain by replacing the convolution with multiplication based on STFT (assuming a time-invariant mixing system):

$$X_k(m,f) = \sum_{i=1}^{N} S_i(m,f) \cdot H_{ik}(f) \cdot N_k^c(m,f) + N_k^a(m,f), \tag{2}$$

where $X_k(m,f) = \mathcal{F}\{x_k(t)\}$, $S_i(m,f) = \mathcal{F}\{s_i(t)\}$, $H_{ik}(f) = \mathcal{F}\{h_{ik}(t)\}$, $N_k^a(f) = \mathcal{F}\{n_k^a(t)\}$, $N_k^c(f) = \mathcal{F}\{n_k^c(t)\}$, and $\mathcal{F}\{\cdot\}$ denotes the STFT, with $m = 1,\ldots,T$ and $f = 1,\ldots,F$ representing the time frame and frequency channel, respectively. The above model is essentially an integrated model of the two in [6] and [15] used respectively for calculating the mixing vector cue and binaural cues.

For the convenience of analysis in Section IV, we define the contribution of source i to the mixture k at each T-F unit (m,f) using the subscript $k|i$ as follows

$$X_{k|i}(m,f) = S_i(m,f) \cdot H_{ik}(f), \tag{3}$$

For the remaining of the paper, the commonly used assumption that speech signals are sparse in the T-F domain is adopted, as in [5], [6] and [15]. More specifically, we assume that only one source (say, the ith source) is dominant at each T-F unit of the mixture, resulting in a simpler model in the complex T-F domain:

$$X_k(m,f) \approx X_{k|i}(m,f) \cdot N_k^c(m,f) + N_k^a(m,f). \tag{4}$$

B. MV based Classification

With the sparsity assumption in the T-F domain (i.e. that at most one source is active at each T-F point within the mixture), all the columns of the mixing matrix at each T-F point are multiplied by zeros except the one corresponding to the active source. As a result, each observation vector can be considered as a basis vector multiplying the dominant source magnitude. Accordingly, for $k = 1, \ldots, M$, and $M = 2$, a vector representation of equation (4) can be represented as follows (omitting the convolutive noise terms, N_k^c, and also referring to equation (3)):

$$\mathbf{x}(m,f) \approx S_i(m,f)\mathbf{h}_i(f) + \mathbf{n}^a(m,f), \tag{5}$$

where $\mathbf{x}(m,f) = [X_1(m,f), X_2(m,f)]^T$, is the complex 2D observation vector at each T-F unit, $\mathbf{h}_i(f) = [H_{i1}(f), H_{i2}(f)]^T$ the MV for the ith source and $\mathbf{n}^a(m,f) = [N_1^a(m,f), N_2^a(m,f)]^T$ is the additive noise that contains background noise and energy from other sources that are not dominant at that T-F unit. To eliminate the effect of source amplitude variation, the observation vectors are normalized with respect to their magnitudes at each T-F unit, as in [6],

$$\tilde{\mathbf{x}}(m,f) = \frac{\mathbf{x}(m,f)}{\|\mathbf{x}(m,f)\|}, \tag{6}$$

$$\approx \tilde{S}_i(m,f) \cdot \tilde{\mathbf{h}}_i(f) + \tilde{\mathbf{n}}^a(m,f). \tag{7}$$

where $|| \cdot ||$ is Frobenius norm, $\tilde{\mathbf{h}}_i(f) = \frac{\mathbf{h}_i(f)}{\|\mathbf{h}_i(f)\|}$, $\tilde{S}_i(m,f) = \frac{S_i(m,f)}{|S_i(m,f)|}$, $\tilde{\mathbf{n}}^a(m,f) = \frac{\mathbf{n}^a(m,f)}{\|\mathbf{n}^a(m,f)\|}$, and $| \cdot |$ takes the absolute value of its argument. The normalized observation vectors $\tilde{\mathbf{x}}(m,f)$ are then whitened and normalized again as follows:

$$\mathbf{z}(m,f) = \frac{\mathbf{W}(f)\tilde{\mathbf{x}}(m,f)}{\|\mathbf{W}(f)\tilde{\mathbf{x}}(m,f)\|}, \tag{8}$$

where $\mathbf{W}(f)$ is a whitening matrix, with each row being one eigen vector of $E(\tilde{\mathbf{x}}(m,f)\tilde{\mathbf{x}}^H(m,f))$, and $\mathbf{z}(m,f) = [Z_1(m,f), Z_2(m,f)]^T$.

We then apply centroid-based clustering, for each frequency bin, to group $\mathbf{z}(m,f)$ into N clusters, in which each cluster is represented by a centroid, denoted as $\mathbf{a}_i(f)$, where $i = 1, \ldots, N$. The aim is to minimize the Mahalanobis distance between the vectors in each cluster and the centroid of that cluster. In [6] a complex Gaussian density function is employed to do this at each frequency bin with frequency-dependent mean and variance:

$$p(\mathbf{z}(m,f)|\mathbf{a}_i(f), \gamma_i(f)) = \frac{\exp\left(-\frac{\|\mathbf{z}-(\mathbf{a}_i^H\mathbf{z}).\mathbf{a}_i\|^2}{\gamma_i^2}\right)}{(\pi\gamma_i^2)^{M-1}}, \tag{9}$$

where $\mathbf{a}_i(f)$ is the centroid with a unit norm $\|\mathbf{a}_i(f)\|^2 = 1$, and $(\gamma_i(f))^2$ is the variance. For notational convenience, we denote $p(\mathbf{z}(m,f)|\mathbf{a}_i(f), \gamma_i(f))$ as $p_B^i(m,f)$. The distance $\|\mathbf{z}(m,f) - (\mathbf{a}_i(f)^H\mathbf{z}(m,f)).\mathbf{a}_i(f)\|$ is the minimum distance between $\mathbf{z}(m,f)$ and the subspace spanned by $\mathbf{a}_i(f)$ because $(\mathbf{a}_i(f)^H\mathbf{z}(m,f)).\mathbf{a}_i(f)$ is the orthogonal projection of $\mathbf{z}(m,f)$ onto the subspace, where the superscript H is Hermitian (conjugate) transpose. In other words, it shows how probable it is that $\mathbf{z}(m,f)$ belongs to the ith source. Note that, in terms of the above discussions, we can see that the estimated mixing vector $\mathbf{a}_i(f)$, which is obtained from $\mathbf{z}(m,f)$, is related to $\mathbf{h}_i(f)$ by

$$\mathbf{a}_i(f) \approx \frac{\mathbf{W}(f)\tilde{\mathbf{h}}_i(f)}{\|\mathbf{W}(f)\tilde{\mathbf{h}}_i(f)\|}. \tag{10}$$

C. IPD and ILD Based Classification

Considering $X_k(m,f)$ as in (4) for a pair of recordings, $k = 1, 2$, two different ratio cues can be calculated (omitting the additive noise terms, N_k^a):

$$\begin{aligned}\alpha(m,f) &= dB\left(\frac{|X_1(m,f)|}{|X_2(m,f)|}\right)\\ &\approx dB\left(\frac{|H_{i1}(f)|}{|H_{i2}(f)|}\right) + dB\left(\frac{|N_1^c(m,f)|}{|N_2^c(m,f)|}\right),\end{aligned} \tag{11}$$

$$\begin{aligned}\phi(m,f) &= \angle\left(\frac{X_1(m,f)}{X_2(m,f)}\right)\\ &\approx \angle\left(\frac{H_{i1}(f)}{H_{i2}(f)}\right) + \angle\left(\frac{N_1^c(m,f)}{N_2^c(m,f)}\right).\end{aligned} \tag{12}$$

where $dB(\cdot)$ denotes $20\log_{10}(\cdot)$ and $\angle(\cdot)$ finds the phase angle.

Therefore, the level difference related to each source, $ILD = dB(\frac{|H_{i1}|}{|H_{i2}|})$, and the phase difference corresponding to that source, $IPD = \angle(\frac{H_{i1}}{H_{i2}})$, can be estimated as the mean value of the noisy observations, $\alpha(m,f)$, and $\phi(m,f)$, respectively, as long as the T-F units dominated by each source (say ith) are identified.

Assuming that $dB(\frac{|N_1^c(m,f)|}{|N_2^c(m,f)|})$ has a normal distribution with variance of $\eta_i^2(f)$ for ith source, the probability of each T-F unit being dominated by that source based on level differences can be estimated as in [15]:

$$p(\alpha(m,f)|i) = \mathcal{N}(\alpha(m,f)|\mu_i(f), \eta_i^2(f)), \tag{13}$$

where $\mu_i(f) \approx dB(\frac{|H_{i1}|}{|H_{i2}|})$ is the mean value and can be estimated based on maximum likelihood (ML), which is explained in more detail in Section III. Similar to the MV, we denote $p(\alpha(m,f)|i)$ as $p_L^i(m,f)$.

Due to the fact that all the measured phases are wrapped to the range $(-\pi, \pi]$, they cannot be mapped to their corresponding interaural time difference (ITD) uniquely. To avoid this ambiguity, a top-down process is suggested in [15] where the equally spaced ITDs corresponding to azimuths from $-90°$ to $90°$ are mapped to the corresponding IPDs without ambiguity. Then the difference between the observed and the predicted IPDs gives the phase residuals $\hat{\phi}(m,f;\tau) = \angle(e^{j\phi(m,f)}e^{-j2\pi f\tau(f)})$ that can be modelled by a normal distribution for each candidate ITD, τ, as explained in [15]:

$$p(\hat{\phi}(m,f)|i,\tau) = \mathcal{N}(\hat{\phi}(m,f;\tau(f))|\xi_{i\tau}(f), \sigma_{i\tau}^2(f)) \tag{14}$$

where $\xi_{i\tau}$ is the mean and $\sigma_{i\tau}$ the standard deviation. Similar to MV and ILD, we denote $p(\hat{\phi}(m,f)|i,\tau)$ as $p_P^{i,\tau}(m,f)$. The Gaussian distributions are summed over τ with some coefficients in a Gaussian mixture model (GMM) framework to give the marginal distribution for source i at each T-F unit. This formulation has the capability to integrate strong early reflections with the direct sound from the source.

D. Differences between MV and IPD/ILD Cues

Despite the fact that both the MV cue and the IPD/ILD cue are derived from the mixtures (i.e. the same information), we found that there are considerable differences in the behavior of these cues, due to the use of different processing methods (with versus without normalization and whitening) and noise models (additive versus convolutive). Such differences will be analyzed in detail in Section IV, based on both the theoretical models and numerical comparisons, where we can see that these cues present complementary properties in various conditions (e.g. for closely spaced sources and/or for highly reverberant mixtures). This motivates us to integrate the cues for improving the estimation of the T-F masks under the same probabilistic framework as discussed in the above subsections. For enhancing the readability of Section IV and for notational convenience, we first introduce how the cues can be integrated in the source models and how the model parameters are estimated by EM, as discussed in the next section. We then explain in Section IV why these cues are complementary to each other and therefore why it is beneficial to combine them for T-F mask estimation and source separation.

III. Source modeling and separation

A. Model Parameter Estimation from the Mixtures

The parameters of the models described in Sections II-B and II-C can be estimated for each source based on the T-F units dominated by that source. However, the dominant source at each T-F unit is a latent variable, i, which is not directly observed but can be inferred from the observed cues and estimated models. On the other hand, the parameters are also unknown, leading us to apply the EM algorithm which is an iterative method for obtaining ML or maximum a posterior (MAP) estimates of the parameters in statistical models, where the model depends on the expectation of latent variables. We also consider another latent variable which is the time delay, τ, between the left and right recordings corresponding to the dominant source at each T-F unit.

Two probabilities need to be estimated and updated during iteration of the EM algorithm. The first one is $\nu_{i\tau}(m,f)$ which is the occupation likelihood that source i dominates at the (m,f) unit in the mixture. Hence $\sum_{i,\tau}\nu_{i\tau}(m,f) = 1$. The second one is $\psi_{i\tau}$, which is the joint probability of any T-F unit activated by source i at time delay τ, and can be considered as the mixing weights in the GMM, as in [15]. Note that $\psi_{i\tau} = \frac{1}{TF}\sum_{m,f}\nu_{i\tau}(m,f)$, where T and F are the number of time frames and frequency bins, respectively.

The parameters are $\Theta = \{\xi_{i\tau}, \sigma_{i\tau}, \mu_i, \eta_i, \mathbf{a}_i, \gamma_i, \psi_{i\tau}\}$ that maximize the log-likelihood of the observations:

$$\mathcal{L}(\Theta) = \sum_{m,f}\log p(\phi(m,f), \alpha(m,f), \mathbf{z}(m,f)|\Theta) \quad (15)$$

$$= \sum_{m,f}\log\sum_{i,\tau}\Big\{\psi_{i\tau}\cdot p_P^{i,\tau}(m,f) \cdot p_L^i(m,f)\cdot p_B^i(m,f)\Big\}, \quad (16)$$

where $\xi_{i\tau}$, $\sigma_{i\tau}^2$, μ_i, η_i^2, $\mathbf{a}_i$, and γ_i^2 are the mean and variance of the IPDs, the ILDs and the MVs, respectively, for source i and time delay τ. Equation (16) represents a GMM with one Gaussian distribution for each source i and each azimuth (corresponding to each τ). Therefore, there are N (number of sources)×N_τ (number of equally spaced ITDs) Gaussian distributions being mixed by the mixing weight $\psi_{i\tau}$.

We should emphasize here that, in (16), we have followed the original work of Mandel *et al.* in [15] and assumed that the IPD/ILD cues are independent. As a result, the mutual (joint) probability is written as the product of individual probabilities. Such an assumption may not hold in practice, but it provides a convenient way for dealing with the issues related to the optimization of the log-likelihood function, as well as the parameter estimation of the probabilistic model. Due to the independence assumption, when both cues are contaminated by independent noise, they should still be independent. A further study about this assumption can be found in Mandel *et al.* [19].

With the above log-likelihood function, the aim is therefore to estimate the model parameters given the observations of IPD, ILD and MV. This can be achieved by the well-known EM algorithm, based on the units allocated to each source in the mixture spectrograms, and then both the units and the parameters are refined alternately, as discussed next.

B. Expectation-Maximization Algorithm

The EM algorithm is employed to estimate the model parameters and the probability at each T-F point, iteratively. In the Expectation step (E step), it calculates the expected value of the log-likelihood function with respect to the observations ϕ, α and $\mathbf{z}$, under the current estimate of the parameters Θ. In other words, given the estimated parameters, Θ, and the observations, and assuming the statistical independence of the cues [15], the probability of the source i at time delay τ being dominant at T-F unit (m,f) is calculated as:

$$\nu_{i\tau}(m,f) = K\cdot\psi_{i\tau}\cdot p_P^{i,\tau}(m,f)\cdot p_L^i(m,f)\cdot p_B^i(m,f) \quad (17)$$

where $\nu_{i\tau}(m,f)$ is the occupation likelihood of source i with delay τ. Coefficient K can be determined in such a way that $\nu_{i\tau}(m,f)$ adds up to 1 over all sources and time delays at each T-F unit, while the mixing coefficient $\psi_{i\tau}$ is initialized by the PHAT histogram [20]. Other elements of (17), i.e. $p_P^{i,\tau}(m,f)$, $p_L^i(m,f)$, and $p_B^i(m,f)$ can be estimated via (14), (13), and (9), respectively.

The ILD parameters $(\mu_i(f), \eta_i^2(f))$ and the IPD residual parameters $(\xi_{i\tau}(f), \sigma_{i\tau}^2(f))$, are re-estimated for each source and time delay using the estimated occupation likelihood $\nu_{i\tau}(m,f)$ that was calculated in the E-step. The M-step of the algorithm can be defined as follows where the model distributions are Gaussian:

Similar to [15], the ILD parameters are updated as:

$$\mu_i(f) = \frac{\sum_{m,\tau}\alpha(m,f)\nu_{i\tau}(m,f)}{\sum_{m,\tau}\nu_{i\tau}(m,f)}, \quad (18)$$

$$\eta_i^2(f) = \frac{\sum_m(\alpha(m,f)-\mu_i(f))^2\sum_\tau\nu_{i\tau}(m,f)}{\sum_{m,\tau}\nu_{i\tau}(m,f)}. \quad (19)$$

IPD residual parameters are updated:

$$\xi_{i\tau}(f) = \frac{\sum_m\hat{\phi}(m,f;\tau)\nu_{i\tau}(m,f)}{\sum_m\nu_{i\tau}(m,f)}, \quad (20)$$

$$\sigma_{i\tau}^2(f) = \frac{\sum_m\left(\hat{\phi}(m,f;\tau)-\xi_{i\tau}(f)\right)^2\nu_{i\tau}(m,f)}{\sum_m\nu_{i\tau}(m,f)}. \quad (21)$$

Frequency-independent parameters can be estimated by taking the average along the frequency bins. For example, the frequency independent mean of ILD can be calculated as $\mu_i = \frac{1}{F}\sum_f \mu_i(f)$, and likewise for the other ILD/IPD parameters. Such averaging can be used to control the model complexity, as will be discussed in Section VI. However, the MVs are estimated at each frequency independently [6]. Therefore the permutation alignment over frequency bins is still a problem. In addition, it is well-known that the EM algorithm only guarantees a local optimum and, in practice, it is important to set the initial values appropriately to achieve the global optimum. Both issues will be addressed in Section V.

To update the parameters of the mixing vectors, the correlation matrix of weighted samples is required. Since the orientation of a linear subspace (i.e., the basis vector related to each source) can be thought of as its greatest variance [21], the eigenvector corresponding to the maximum eigenvalue, $\max(\lambda)$, of the correlation matrix is assumed as the optimum $\mathbf{a}_i$, as in [6]:

$$\mathbf{R}_i(f) = \sum_{m,\tau} \nu_{i\tau}(m,f)\mathbf{z}(m,f)\mathbf{z}^H(m,f), \tag{22}$$

$$\mathbf{a}_i(f) = \text{eigenvector}(\mathbf{R}_i(f))_{\max(\lambda)}, \tag{23}$$

$$\gamma_i^2(f) = \frac{\sum_{m,\tau} \nu_{i\tau}(m,f)\|\mathbf{z} - (\mathbf{a}_i^H\mathbf{z}).\mathbf{a}_i\|^2}{(M-1)\sum_{m,\tau}\nu_{i\tau}(m,f)}, \tag{24}$$

$$\psi_{i\tau} = \frac{1}{TF}\sum_{m,f}\nu_{i\tau}(m,f), \tag{25}$$

where T and F are the number of time frames and frequency bins, respectively.

Equipped with clear definitions of symbols and descriptions of the proposed algorithm discussed above, we are now able to provide a detailed analysis of the properties of the MV cue in contrast with the IPD/ILD cues, and show that they can complement each other in various acoustic conditions for the improvement of T-F mask estimation, which is our focus in next section.

IV. Comparing the Cues at Different Conditions

A. Complex Mixing Vector Representation

First, we show that the operations of normalization (6) and whitening (8) in the T-F domain have reduced the degrees of freedom of the model (5) when represented by the mixing vector. To see this, we represent the model in (5) with complex $2D$ vectors containing amplitude and phase information as follows (neglecting any effects of noise on the estimates):

$$\begin{bmatrix} |Z_1(m,f)|e^{j\angle Z_1(m,f)} \\ |Z_2(m,f)|e^{j\angle Z_2(m,f)} \end{bmatrix} \approx \tilde{S}_i(m,f)\cdot\begin{bmatrix} a_{i1}(f) \\ a_{i2}(f) \end{bmatrix}, \tag{26}$$

where source i is the dominant source. Therefore at each frequency bin f, we have:

$$|Z_1(m,f)|^2 + |Z_2(m,f)|^2 = 1 \tag{27}$$

$$\angle Z_1(m,f) + \angle Z_2(m,f) \approx \angle a_{i1}(f) + \angle a_{i2}(f) + 2\angle S_i(m,f), \tag{28}$$

$$\angle Z_1(m,f) - \angle Z_2(m,f) \approx \angle a_{i1}(f) - \angle a_{i2}(f), \tag{29}$$

where $\angle Z_1(m,f) + \angle Z_2(m,f)$ and $\angle Z_1(m,f) - \angle Z_2(m,f)$ have uniform and normal distributions, respectively. As shown in (28), $\angle Z_1(m,f) + \angle Z_2(m,f)$ is time-variant since the phase of the source signal changes with respect to time. As a result, it is uninformative and cannot be used to estimate the time-invariant mixing vectors blindly. Instead, the MVs $\mathbf{a}_i(f) = [a_{i1}(f), a_{i2}(f)]^T$ can be evaluated as the main eigenvectors of the covariance matrices $\mathbf{R}_i(f)$ as defined in (22) where we take $\sum_\tau \nu_{i\tau}(m,f) = 1$, according to the aforementioned sparsity assumption that only one source, i.e. $S_i(m,f)$, is active at each (m,f). Consequently, the MVs will have two degrees of freedom: relative amplitude and relative phase, since $\|\mathbf{a}_i(f)\| = 1$ and $\angle a_{i1}(f)$ or $\angle a_{i2}(f) = 0$.

This result is consistent with the fact that the covariance matrices are positive-semidefinite and symmetric [22] and so Hermitian in the complex domain with all the eigenvalues being real and simple [23]. Hence, the eigenvectors (mixing vectors) will be like $[r\ cr]^T$ where $r \in \mathbb{R}$ and $c \in \mathbb{C}$, with relative phase and amplitude containing the whole information.

On the other hand, as illustrated in Section II-B, the MV related to each source at a given frequency f can be considered as the centroid of that source's (whitened and normalized) observation vectors $\mathbf{z}(m,f) = [Z_1(m,f), Z_2(m,f)]^T$ where $m = 1,\ldots,T$ at that frequency. Therefore, the MV of the source at a given frequency, $\mathbf{a}_i(f)$, can be represented by the observation vectors $\mathbf{z}(m,f)$ of that given source (assuming all other sources to be inactive). To show this, similar to the notation used in equation (4), we define the contributions of source i to $Z_k(m,f)$ as $Z_{k|i}(m,f)$, where $Z_{k|i}(m,f) = a_{ik}\tilde{S}_i(m,f)$, $k = 1,2$, and likewise, its contributions to $\mathbf{z}(m,f)$ as $\mathbf{z}_{|i}(f) = [Z_{1|i}(m,f), Z_{2|i}(m,f)]^T$. When only source i is active, according to (26), we have $\mathbf{z}(m,f) = \mathbf{z}_{|i}(m,f)$.

We now present an example to demonstrate the relationship between $\mathbf{z}(m,f)$ and the MVs with a scatter plot. To this end, we generate the observed signals by convolving two random utterances from the TIMIT dataset [24] with binaural RIRs (BRIRs) of room A [25] (as listed in Table II in Section VI) for sources at 0° and 10° azimuths, one at a time. For example, we can allow one utterance to be active (e.g., source s_1 placed at 0° azimuth), by switching off the other (e.g. source s_2 placed at 10°). In this way, the observed signals x_k, $k = 1,2$, would contain only the contributions from source s_1. Then the signals x_k are transformed to T-F domain and concatenated to produce complex observation vectors at each T-F unit, in this case $\mathbf{x}(m,f) = \mathbf{x}_{|1}(m,f)$, which are further processed in terms of (6) and (8) to produce the normalized and whitened observation signals $\mathbf{z}(m,f) = \mathbf{z}_{|1}(m,f)$. The observation vector, $\mathbf{z}_{|1}(m,f)$, with corresponding MV $\mathbf{a}_1(f)$ as its centroid, is represented by $\angle Z_{1|1}(m,f) - \angle Z_{2|1}(m,f)$, and $\tan^{-1}(|Z_{1|1}(m,f)|/|Z_{2|1}(m,f)|)$, which are associated to the phase and level differences of source $S_1(m,f)$, respectively. The observation vectors $\mathbf{z}_{|1}(m,f)$ at the frequency band of 3.85 kHz are the circles plotted in Fig. 2(a). When only source $S_2(m,f)$ is active (by switching off source $S_1(m,f)$), we can similarly visualize $\angle Z_{1|2}(m,f) - \angle Z_{2|2}(m,f)$, and $\tan^{-1}(|Z_{1|2}(m,f)|/|Z_{2|2}(m,f)|)$ as the triangles in Fig. 2(a). It can be seen that all the points are confined to a quadrant of a unit cylinder shell due to the normalization which can be unwrapped to a 2D plane as shown in Fig. 2(b). Now, this seems to suggest that the MV does not provide extra information compared to IPD and ILD. However, as will be clear in the following sections, the scatter plots and probability

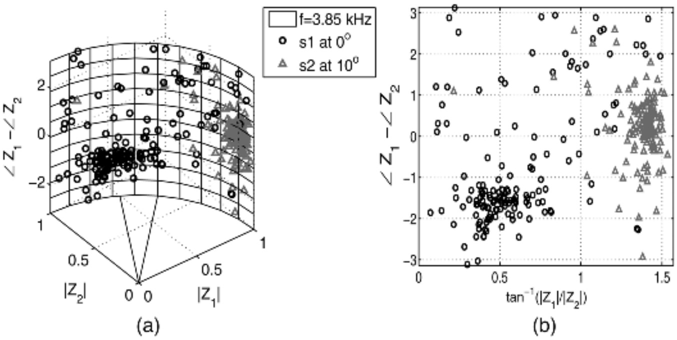


Fig. 2. 2D representation of the observation vectors in frequency channel = 3.85 kHz after normalization and whitening on a (a) unit cylinder wall, and (b) unwrapped 2D plane, for two different sources at 0° and 10° azimuths.

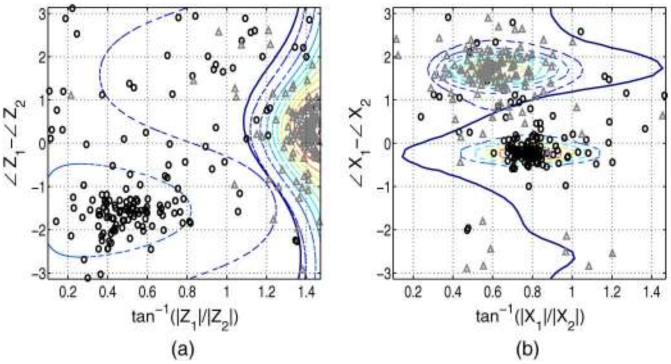


Fig. 3. Scatter plots and probability contours (dashed lines) for sources in room A at 0° in ○ and 10° in Δ with decision boundaries shown by solid lines based on (a) mixing vectors and (b) binaural cues in frequency $f = 3.85$ kHz.

distributions of MV and binaural cues are different and they introduce different behavior under various conditions.

B. Closely Spaced Sources

In this section we compare the behavior of the MV and binaural cue (ILD/IPD) distributions and show that MV distributions are more distinct compared to the joint probability of ILD and IPD, when the sources are close to each other. Equal probability contours are used to illustrate the multivariate distributions in 2D space [26]. We consider the same example as shown in Fig. 2 in Section IV-A, where the observed signals (whitened and normalized) $\mathbf{z}_{|1}$ and $\mathbf{z}_{|2}$ were obtained in the same way by placing respectively source 1 at 0° azimuth and source 2 at 10° azimuth, one at a time.

To calculate the equal probability contours based on the MV, the (mean and variance) model parameters, $\mathbf{a}_i(f)$, and $\gamma_i^2(f)$, were estimated for each source $i = 1, 2$ based on the whitened observation signals $\mathbf{z}_{|i}(m, f) = [Z_{1|i}(m, f), Z_{2|i}(m, f)]^T$ using (22)-(24). For example, for the estimation of $\mathbf{a}_1(f)$, and $\gamma_1^2(f)$, source 1 alone was active when capturing the observation signals $\mathbf{z}_{|i}$. As a result, $\sum_\tau \nu_{1,\tau}(m, f) = 1$, and equation (22) can be simplified as $\mathbf{R}_1(f) = \sum_m \mathbf{z}_{|1}(m, f)\mathbf{z}_{|1}^H(m, f)$. Hence for the frequency band at $f = 3.85$ kHz, the MV parameters $\mathbf{a}_1(f)$ and $\gamma_1^2(f)$ were estimated according to (22)–(24). The same procedure was followed when calculating $\mathbf{a}_2(f)$ and $\gamma_2^2(f)$ assuming that only source 2 was active in that band. We then plotted the equal probability contours of two MV probabilities under the two sets of MV parameters calculated above, as shown by dashed lines in Fig. 3(a), as follows. First, two sets of MV probabilities $p(\mathbf{z}_{|i}|\mathbf{a}_i(f), \gamma_i(f))$ were calculated using (9) for $i = 1, 2$, with $\angle Z_{1|i} - \angle Z_{2|i}$ taking discrete values from $(-\pi, \pi]$ and $\frac{|Z_{1|i}|}{|Z_{2|i}|}$ from -20 dB to 20 dB. This corresponds to changing $\tan^{-1}\frac{|Z_{1|i}|}{|Z_{2|i}|}$ from approximately 0.1 to $\frac{\pi}{2}$. Then, Matlab's *contour* function was employed to draw the equal probability contours in dashed lines based on these two sets of calculated probabilities.

To show how distinguishable the sources are, the variables $\{\mathbf{z} = [Z_1, Z_2]^T\}$ in 2D space should be divided into two (or more) groups, which are associated with the samples from each source. The decision boundaries, or the borders between these regions, are drawn with a solid line where the two sets of MV probabilities are equal. In other words, when $p(\mathbf{z}_{|1}|\mathbf{a}_1(f), \gamma_1(f)) = p(\mathbf{z}_{|2}|\mathbf{a}_2(f), \gamma_2(f))$. We also show the scatter plots based on the whitened observations from the clean source signals, $Z_{1|i}(m, f)$ and $Z_{2|i}(m, f)$ at $f = 3.85$ kHz, i.e. by plotting the quantities $\tan^{-1}\frac{|Z_{1|i}(m,f)|}{|Z_{2|i}(m,f)|}$ versus $\angle Z_{1|i}(m, f) - \angle Z_{2|i}(m, f)$, for both sources $i = 1, 2$. Note that, each scatter point corresponds to a time frame m, as f has been fixed to 3.85 kHz in this plot. Since the model parameters were estimated with the same observation vectors, the equal probability contours and the scattered samples are consistent.

For binaural cues, the level and phase differences of each source i with no interference at the same frequency ($f = 3.85$ kHz) were calculated based directly on $X_{1|i}(m, f)$ and $X_{2|i}(m, f)$ using (11) and (12), by replacing $X_k(m, f)$ in these equations with $X_{k|i}(m, f)$ for $k = 1, 2$ (again due to the assumption that only source i is active). The scattered samples in Fig. 3(b) were obtained based on the observations $X_{1|i}(m, f)$ and $X_{2|i}(m, f)$ of each source i. Then the model parameters, $\mu_i(f), \eta_i(f), \xi_{i,\tau}(f)$ and $\sigma_{i,\tau}(f)$, were estimated according to (18)–(21) using the observed $\alpha(m, f)$ and $\hat{\phi}(m, f)$ values for each source. In this case $\psi_{i,\tau}$ was set to a normal distribution over τ, whose mean was estimated via the PHAT-histogram [20] and variance is fixed to 1. Then the probability, given that source i is active, of any phase difference $\angle X_1(m, f) - \angle X_2(m, f)$ from $-\pi$ to π was calculated as $p_P^i(m, f) = \sum_\tau p_P^{i,\tau}(m, f)$ based on the GMM of source i using (14), with $\angle X_1(m, f) - \angle X_2(m, f)$ as observations. The relative amplitude, $\tan^{-1}\frac{|X_1(m,f)|}{|X_2(m,f)|}$, was also varied from 0.1 to $\frac{\pi}{2}$ and the probability of any relative amplitude belonging to each source, i, i.e. $p_L^i(m, f)$, was computed based on (13), with $\tan^{-1}\frac{|X_1(m,f)|}{|X_2(m,f)|}$ as observation. The equal probability contours in Fig. 3(b) were estimated from a set of variables $\mathbf{x}(m, f) = [X_1(m, f), X_2(m, f)]$ similarly to that used in Fig. 3(a). Also, $\angle X_1(m, f) - \angle X_2(m, f)$ ranges from $(-\pi, \pi]$, and $\frac{|X_1(m,f)|}{|X_2(m,f)|}$ ranges from -20 dB to 20 dB. However, in contrast to the MV probability defined in (9), the binaural probability is calculated as the product of (14) and (13). Note that, in the above analysis, we have used the frequency-independent mode in the EM algorithm as discussed in Section VI.

Now we can see that the two MV based clusters in Fig. 3(a) are more distinct compared to the binaural based clusters and probability contours in Fig. 3(b) when the sources are close to each other. This suggests that MVs with the statistical model perform better than the binaural cues for closely spaced sources.

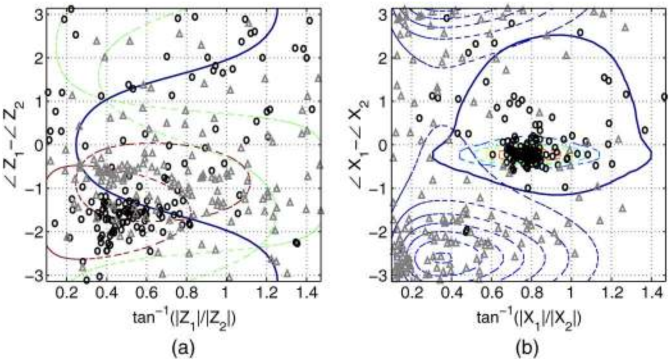


Fig. 4. Scatter plots and probability contours (dashed lines) for sources in room A at 0° in ∘ and 80° in Δ with decision boundaries shown by solid lines based on (a) mixing vectors and (b) binaural cues in frequency $f = 3.85$ kHz.

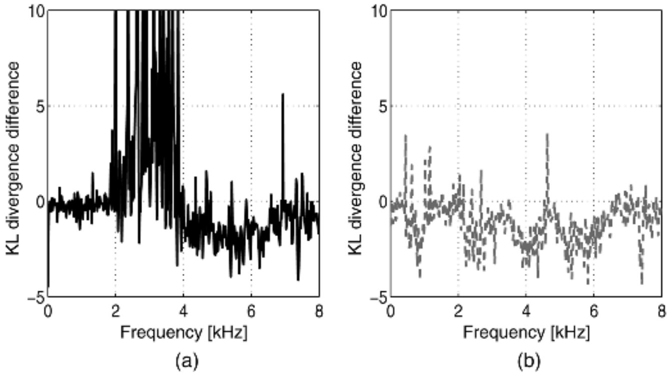


Fig. 5. The difference between the KL divergences obtained respectively from the MV and binaural models is shown here. The KL divergence between the two source models is calculated based on binaural cues and mixing vectors in room A with $T_{60} = 0.32$ s where one source is at 0° and the second source is at (a) 10° and (b) 80°.

Fig. 4 displays the scatter plots, the equal probability contours and decision boundaries for sources at 0° and 80° in room A obtained similarly. It can be seen that when the sources are well away from each other with 80° difference in azimuth, binaural cue source models are quite distinct whereas the observation vectors have more overlap, which is opposite to what has been observed for closely spaced sources.

We have also examined how distinct the source models are over frequency based on the Kullback–Leibler (KL) divergence [27] between the source models for two sources at 0° and 10° or 80° azimuths. The KL divergence (κ) for MV and binaural models are defined respectively as follows[1] :

$$\kappa^{\mathrm{MV}}(f) = \sum_m \Bigg\{ p(\mathbf{z}(m,f)|\mathbf{a}_1(f),\gamma_1(f)) \cdot \log \frac{p(\mathbf{z}(m,f)|\mathbf{a}_1(f),\gamma_1(f))}{p(\mathbf{z}(m,f)|\mathbf{a}_2(f),\gamma_2(f))} \Bigg\}, \quad (30)$$

where the probability density function $p(\mathbf{z}(m,f)|\mathbf{a}_i(f),\gamma_i(f))$, $i = 1,2$ has already been defined in (9), and

$$\kappa^{\mathrm{Binaural}}(f) = \sum_m p(\mathbf{x}(m,f)|1) \log \frac{p(\mathbf{x}(m,f)|1)}{p(\mathbf{x}(m,f)|2)}, \quad (31)$$

TABLE I
KL-DIVERGENCE BETWEEN THE CLEAN AND NOISY SIGNAL MODELS FOR THREE DIFFERENT CUES AND TWO TYPES OF NOISE AVERAGED OVER ALL FREQUENCIES

-	additive noise	convolutive noise
mixing vector (MV)	**2.10**	2.31
IPD	2.70	**2.01**
ILD	3.39	**3.29**

where $p(\mathbf{x}(m,f)|i) = p_P^i(m,f) \cdot p_L^i(m,f)$, and the calculation of $p_L^i(m,f)$ and $p_P^i(m,f)$ based on (13) and (14), respectively, has been described earlier in this section.

We evaluate the difference between the KL divergences obtained from MV and binaural models[2], i.e. $\Delta\kappa(f) = \kappa^{\mathrm{MV}}(f) - \kappa^{\mathrm{Binaural}}(f)$. When $\Delta\kappa(f) > 0$, the MV cue is more discriminative as compared with the binaural cues, and vice versa. As shown in Fig. 5(a), MV based source models are well separated even when the sources are close to each other (10° azimuth) especially in the frequency range $2-4$ kHz where ILD and IPD are not very reliable [9]. On the other hand, when the sources are positioned away from each other (80° azimuthal displacement) the IPD/ILD source models become more distinct compared to those based on MVs (see Fig. 5(b)). This suggests that MV and binaural models play complementary roles for different source positioning which motivated us to combine the statistical and binaural models and introduce a new algorithm that, as we will show in Section VI, works better than the methods using the individual cues for various source configurations and conditions.

C. High Reverberation

Next, we examined the effect of two types of noise on the cues. First, speech shaped noise was generated by averaging the spectra of the anechoic recordings of 15 utterances used in the experiments (see Section VI-A). Then the generated noise was added to a clean signal to produce a corrupted signal, similar to [28]. The clean signal was one of the utterances convolved with anechoic BRIR (see Section VI-A). The same utterance was also convolved with the BRIR of the reverberant room D (as in Table II) to introduce convolutive noise. To measure the relative level of this convolutive noise we divided room D's BRIR at 32 ms, which is also half of the window length (64 ms), and zero-padded each remaining part to have two RIRs representing the direct sound with desired early reflections and late reverberation noise. The two parts were then convolved with the original utterance and the relative energy of the signals was measured to be approximately 5 dB for room D. Accordingly, the level of speech shaped noise was set to yield an SNR of 5 dB in the anechoic room.

The model parameters of the source, $\Theta = \{\xi,\sigma,\mu,\eta,\mathbf{a},\gamma\}$, were estimated under three different conditions: 1− anechoic room, 2− anechoic room with additive noise, and 3− reverberant room, to investigate the effect of additive and convolutive noise. The degradation from the original models is measured based on the KL divergence [27] between the pdfs of the noisy observations and those corresponding to the clean anechoic signal.

[1] Note that, the discrete model probability is normalized over m such that $\sum_m p(\mathbf{z}(m,f)|\mathbf{a}_i(f),\gamma_i(f)) = 1, i = 1,2$. This rule also applies to the calculation of other KL divergence in this paper.

[2] Note that, κ was calculated over the following range of values for the parameters: $\angle X_1 - \angle X_2 \in (-\pi,\pi]$, $\frac{|X_1|}{|X_2|} \in [-20,20]$ dB, $\angle Z_1 - \angle Z_2 \in (-\pi,\pi]$, and $\frac{|Z_1|}{|Z_2|} \in [-20,20]$ dB.

Unlike the KL divergence defined earlier that measures the distance of the pdfs obtained from different sources at different positions in the same environment in Section IV.B, the KL divergence here measures the distance of the pdfs obtained from the same source under different conditions, either between the noiseless anechoic environment and the additive-noise-corrupted anechoic environment, or between the noiseless anechoic environment and the convolutive-noise-corrupted reverberant environment. Take the KL divergence based on MV cues, for example. Suppose the parameter set $\{\mathbf{a}(f), \gamma(f)\}$ is obtained from a source in the noiseless anechoic situation, and $\{\mathbf{a}^{\mathrm{n}}(f), \gamma^{\mathrm{n}}(f)\}$ is estimated from the same source at the same input angle in a noisy environment, then the KL divergence here is calculated as:

$$\kappa_B^{\mathrm{n}}(f) = \sum_m \left\{ p(\mathbf{z}(m,f)|\mathbf{a}(f), \gamma(f)) \cdot \log \frac{p(\mathbf{z}(m,f)|\mathbf{a}(f), \gamma(f))}{p(\mathbf{z}(m,f)|\mathbf{a}^{\mathrm{n}}(f), \gamma^{\mathrm{n}}(f))} \right\}. \tag{32}$$

In a similar way, the KL divergences based on IPD cues and ILD cues, i.e. κ_P^{n} and κ_L^{n}, can also be obtained.

The results are given in Table I, which demonstrate that the MV model is more affected by high reverberation with higher KL divergence ($\kappa_B^{\mathrm{n}} = 2.31$) compared to the same level of additive noise ($\kappa_B^{\mathrm{n}} = 2.10$). On the other hand, binaural cues are more robust to reverberation especially IPD with $\kappa_P^{\mathrm{n}} = 2.01$, but more sensitive to additive noise with $\kappa_P^{\mathrm{n}} = 2.70$, which confirms that MV and binaural cues play complementary roles for dealing with different types of noise. This provides further evidence that combing the cues can lead to a method that is more robust to both additive and convolutive noise. Moreover, we can see that MV and IPD are more reliable as compared to ILD with less deviation from the original models, exhibiting a smaller KL divergence. This observation motivated us to assign different weights for each cue (as explained in Section V-C).

V. Practical Implementation Issues of the Proposed Algorithm

A. Dealing with the Permutation Problem and Initialization

Since the EM algorithm can be initialized either from the E-step or the M-step and also there is commonly no prior information about the mixing vectors, we propose to initialize the probabilistic mask first and then estimate the initial values of $\mathbf{a}_i(f)$ and $\gamma_i(f)$ based on the masked spectrogram. More specifically, we initialize the mask based on the IPD and ILD cues derived from the binaural model, and let the program run for two iterations without any MV contribution.

For the first iteration, we set $p_B^i(m,f) = 1$ for all time frames m and frequencies f in (17) to remove the effect of the MV contribution. Once the mask $M_i(m,f) \equiv \sum_\tau \nu_{i\tau}(m,f)$ is obtained after two iterations based on only the information in the binaural cues, the parameters of the MV distributions, $(\mathbf{a}_i(f), \gamma_i^2(f))$, are estimated from the next M-step to prevent the permutation problem, as explained in [29].

Similar to [15], we initialize $\psi_{i\tau}$ with one Gaussian distribution for each source, say i, over τ with mean values τ_i corresponding to the direct sound estimated by PHAT-histogram [20]. It is important to set an appropriate window length for the PHAT-histogram approach. As the window length increases, the number of segments available to generate the histogram of time delays decreases, making the estimated pdfs unreliable [20]. We examined various windows and achieved the best result with $L_{PHAT} = 512$.

Initial ILD parameters are set to zero mean and 10 dB standard deviation with phase residuals' means and variances being set to zero and one, respectively. After two iterations the probabilistic mask is applied to initialize the MV parameters. Thereafter, the occupation likelihoods are re-estimated and used to update all model parameters in subsequent iterations.

To deal with the T-F units dominated by reverberation which do not fit into the source models and degrade the parameter estimation, [15] considers a garbage source. Assuming a diffuse sound field due to reverberation, the ILD and IPD of the garbage source should have broad distributions as the energy comes from all directions with equal probability. Here the garbage source is treated as another sound source with large initial variance.

B. Weighted Cue Likelihoods

In the first stage of combining the binaural and statistical cues, we assumed that each cue is as influential as the others, so we simply added their log-likelihoods to estimate the joint probability of each source being active at each T-F unit. However, as explained in IV-C, the cues are not equally reliable especially in the presence of reverberation. For example, the IPD cue seems to be more robust in reverberant conditions compared to the ILD cue. Therefore, it is more appropriate to adjust the contribution of the cues by giving a different weight to each of them before combining them.

The idea of cue weighting is related to that of [30] in which different distributions are weighted and combined to achieve a model that fits the real data better. In the absence of compelling statistical counter evidence, a natural choice of the pdf for modelling the cues is the normal distribution for which no further assumption is needed. The Gaussian (normal) distribution was employed here for consistency with Mandel *et al.* [15] and Sawada *et al.* [6]. It is also simple, with minimized entropy, and fast efficient parameter re-estimation via a straightforward EM algorithm. Moreover, the possibility of extension to GMMs provides potential for greater flexibility and precision in modeling the underlying statistics of sample data.

Another motivation for cue weighting is to make the algorithm more comprehensive compared to that of Mandel *et al.* [15] where the cues are weighted equally with different modes introducing various degrees of freedom for parameters. We decided to make the modes more general by substituting the coefficients with adjustable weights:

$$\log(\nu_{i\tau}) \propto W_P \cdot \log \psi_{i\tau} p_P^{i\tau} + W_L \cdot \log p_L^i + W_B \cdot \log p_B^i \tag{33}$$

where W_P, W_L and W_B control the influence of IPD, ILD and MV cues, respectively, at each T-F point (m,f).

Here, we investigated weights that are fixed over time and frequency. However, based on Duplex theory [31], human perception treats ILD as more reliable at high frequencies, as opposed to IPD which is favoured at low frequencies. Therefore, further investigations are justified to assign weights for each cue accordingly. In our work, the weights are found empirically based on a brute force grid search approach as detailed in Section VII-D.

C. Pseudocode of the Proposed Algorithm

The whole algorithm is summarized in Algorithm 1.

Algorithm 1 Soft mask generation to recover speech sources

```
Task: Binaural speech source separation
Input: $l(t) = x_1(t), r(t) = x_2(t), W_P, W_L, W_B$
Output: $\hat{s}_i(t)$ the estimate of the $i$th source
Initialization: $\xi_{i\tau} = 0, \sigma_{i,\tau} = 1, \mu_i = 0, \eta_i = 10$ dB,
    $\psi_{i\tau} = \mathcal{N}(\tau|\tau_i, 1)$.
$L(m,f) = \mathrm{STFT}(l(t)), R(m,f) = \mathrm{STFT}(r(t))$
    $\alpha(m,f) = |(L(m,f)/R(m,f))|$(11)
    $\phi(m,f) = \angle(L(m,f)/R(m,f))$(12)
    $\mathbf{x}(m,f) = [L(m,f), R(m,f)]^T$
    $\tilde{\mathbf{x}}(m,f) = \mathbf{x}(m,f)/\|\mathbf{x}(m,f)\|$ {Normalization} (6)
    $\mathbf{z}(m,f) =$ Pre - whitening and normalizing ($\tilde{\mathbf{x}}(m,f)$) (8)
for $rep = 1 : 16$ do
    $p_P^{i,\tau}(m,f) = \mathcal{N}(\hat{\phi}(m,f;\tau)|\xi_{i,\tau}(f), \sigma^2_{i,\tau}(f))$ (14)
    $p_L^i(m,f) = \mathcal{N}(\alpha(m,f)|\mu_i(f), \eta_i^2(f))$ (13)
    if $rep < 2$ then
        $p_B^i(m,f) = 1$
    else
        $p_B^i(m,f) = \mathcal{N}(\mathbf{x}(m,f)|\mathbf{a}_i(f), \gamma_i^2(f))$ (9)
        {after 2 iterations the BSS parameters are initialized}
    end if
    $\mathcal{L}_{i,\tau}(m,f) = W_P \log \psi_{i,\tau} p_P^{i,\tau}(m,f) +$
        $W_L \log p_L^i(m,f) + W_B \log p_B^i(m,f)$ (33)
    $\mathcal{L}(rep) = \sum_{m,f} \log \sum_{i,\tau} \exp(\mathcal{L}_{i,\tau}(m,f))$ (16)
    $\nu_{i,\tau}(m,f) =$
        $\{\exp(\mathcal{L}_{i,\tau}(m,f))\}/\{\sum_{i,\tau} \exp(\mathcal{L}_{i,\tau}(m,f))\}$ (17)
    Update $\mu_i(f), \eta_i^2(f), \xi_{i,\tau}(f), \sigma^2_{i,\tau}(f)$ (18)–(21)
    {For frequency independent mode, the average of the
    parameters along $f$ is used, e.g.,
    $\mu_i(f) = \mu_i = \frac{1}{F}\sum_f \mu_i(f)$, and likewise for $\eta_i^2(f)$,
    $\xi_{i,\tau}(f), \sigma^2_{i,\tau}(f)$.}
    if $rep \geq 2$ then
        Update $\mathbf{a}_i(f), \gamma_i^2(f)$ (22)–(24)
    end if
    Update $\psi_{i,\tau}$ (25)
end for
$M_i(m,f) = \sum_\tau \nu_{i,\tau}(m,f)$
$\hat{s}_i(t) = \mathrm{ISTFT}(L \cdot M_i)/2 + \mathrm{ISTFT}(R \cdot M_i)/2$
```

VI. Experiments

This section explains how we selected utterances and convolved them with BRIRs with various acoustic properties to generate the virtual microphone signals including realistic room effects. Mixtures of 2 and 3 speakers with different relative positions were created to examine the effect of source configuration on the performance of the algorithms. These provide tests for determined (2-source) and underdetermined (3-source) cases. The Mandel *et al.* [15], Sawada *et al.* [6] and our proposed algorithms were then applied to the mixtures to recover the source signals. The quality of the recovered signals was evaluated both in terms of signal distortion and perceptual speech quality.

A. Data Source Selection

Similar to [15], we chose the TIMIT data set which is a continuous speech corpus containing 6300 utterances spoken by 630 native American English speakers [24]. 15 utterances, spoken by both male and female speakers, with approximately the same length (about 3 s), were selected randomly and then shortened to 2.5 s for consistency. The two common sentences spoken by all speakers (sa1 and sa2) were removed from the selection set to avoid mixtures containing identical word sequences, which would violate the assumption of sparsity and be unlikely from a practical perspective. All the utterances were also normalized to have equal root mean square amplitude.

Several RIR data sets were investigated to find the most appropriate one for our aim which was evaluating the effects of source configuration and room reverberation on the performance of the algorithms. The BRIRs measured by Hummersone [32] were selected. These were recorded using a dummy head and torso in 5 different types of room, named as X, A, B, C and D at the University of Surrey. One advantage of this database over other datasets, such as [33], is its higher angular resolution which enabled us to evaluate the performance of the algorithms over different configurations with finer resolution. The other positive aspect of this dataset is that the BRIRs were measured in rooms with different acoustical properties, which facilitates comparison of the algorithms over a range of conditions. Table II shows the acoustical properties of the rooms in which the signals were recorded. For the anechoic condition, X, the impulse responses were recorded in a very large room and the reflections were then truncated. The head related transfer function (HRTF) is incorporated in the BRIR which makes the signals similar to what a person would hear in that position.

For each T_{60} and angular configuration, 15 pairs from those 15 selected utterances were chosen in such a way that no signal would be mixed with itself. The mixtures were then generated by simply adding the reverberant target and interferer signals which is equivalent to assuming superposition of their respective sound fields. Even though the time-frequency masks to recover all the sources at different azimuths are calculated in our proposed algorithm, the algorithms' performance is reported based on the quality of the recovered target source located at the 0° azimuth, while the interferer's azimuth varied from 10° to 90° with steps of 5°. All sources were 1.5 m away from the head (this defines 17 different configurations). This is an ecologically valid approach to investigating the effect of target-interferer angular displacement on the system performance, given that we typically turn to face the target [34]. In the case of 2-source mixtures, the interferer was located on the right of the target, whereas for 3-source mixtures, the two interferers were located symmetrically on the right and left of the target source, as in [15].

Since there were 5 different rooms and 17 different configurations, 85 sets of mixtures were created each of which contained 15 different mixtures (1275 mixtures in total). Mandel's algorithm (based on only binaural cues), Sawada's algorithm (based on the MVs) and our proposed algorithm were used to separate the source signals.

For our proposed algorithm we examined various window lengths and found the optimum 1024 sample Hann window (64 ms with $f_s = 16$ kHz) with 75% overlap. To recover the target signal, as shown in line 25 in Algorithm 1, the average of the separated signals at the left and right microphones is calculated. Although this summation favors frontal sources, we have applied the same routine as in the two baselines [6], [15] to facilitate a fair comparison. Each recovered signal was then compared to the original utterance to measure the performance of the algorithm. This evaluation will be explained in more detail below.

B. Evaluation

We considered two measures of the speech separation accuracy: signal-to-distortion ratio (SDR) [35] and Perceptual Evaluation of Speech Quality (PESQ) [36]. The performance of the algorithms was primarily evaluated based on the SDR. The SDR is the ratio of the energy in the recovered signal resembling the original signal to the remaining energy related to interference from other signals and unexplained artefacts. Since sound coloration by room reflections is acceptable to some extent by human listeners, it can be counted towards the target energy. Accordingly, the recovered signals were not compared with the original signals but a filtered version of them which was also normalized for any delays or scaling. Therefore, we applied an FIR Wiener filter (up to 32 ms) to the original signal with the recovered target signal as the reference signal, as in [15]. Thus, any energy in the estimated signal corresponding to a filtered version of the original utterance was considered as an acceptable representation of the target signal. Any remaining energy was assumed as distortion [15].

Although SDR is an objective evaluation method based on physical signal characteristics and is widely used, it may not always correlate well with perceived sound quality. Consequently, we also applied PESQ to evaluate the algorithms for human applications. PESQ is highly correlated with the mean opinion score (MOS) of human listeners, and provides an objective measure of the most perceptually relevant signal characteristics. PESQ provides a score in the range of 1 to 5 where 1 is bad and 5 is great.

VII. Results

In this section we first examine the performance of Mandel's algorithm and the proposed algorithm with different source model complexities (known as modes in [15]) to choose the one that gives the best results. Once the model complexity is set, the Sawada, Mandel and proposed algorithms are employed to separate the mixtures under various acoustic conditions for both the determined and underdetermined cases, i.e., for 2 sources and 3 sources with just 2 microphones. The detailed results for diverse configurations are reported to study and compare the methods thoroughly. All the results in the following sections were obtained for equal weight $W_P = W_L = W_B = 1$, except those in Section VII-D where different weights are applied to the cues in order to assess the potential performance improvement by cue weighting. Finally, we present separation results for the mixtures corrupted by spatially diffuse noise.

A. Model Complexity

As explained in Section III, model parameters can be frequency-dependent or wrapped up over all frequency bins to be frequency-independent. There are different modes representing different types of source model from having frequency-dependent parameters, where the mean and variance for ILD and IPD distributions are different for each frequency bin, $\Theta^G_{\Omega\Omega}$, to being frequency-independent where the parameters of each source model are the same for all frequency bins, Θ^G_{11}. The superscript G stands for using the garbage source. For the simplest mode, Θ^G_{00}, the means for residual IPD and ILD are set to zero and do not get updated, with ILD variance also set to ∞. In mode Θ^G_{01} the degree of freedom is increased and so the IPD parameters get updated but remain constant across frequency. Mode Θ^G_{10} represents updating ILD cues and fixed IPD parameters. In summary, the indexes 0,1 and Ω stand for 'fixed,' 'frequency-independent' and 'frequency-dependent' parameters for ILD and IPD cues, respectively (see Table I in [15]).

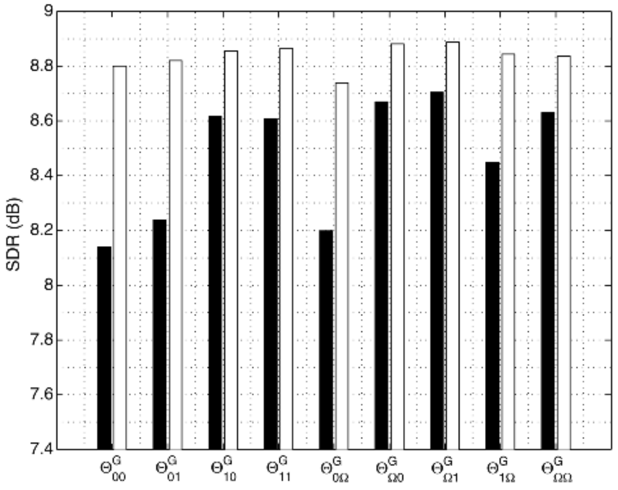


Fig. 6. Performance of the Mandel method [15] (solid bar), and proposed algorithm (white bar) with all possible model complexities averaged over 15 different mixtures in 4 rooms (except anechoic) and 6 different configurations $(15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ)$, 360 mixtures, for the determined (2-source) case.

A pilot study with simulated data [37] and no HRTF showed that the moderate mode of Θ^G_{11} with frequency-independent IPD and ILD cues that incorporated a garbage source gave the best performance for our proposed algorithm, in which the MV-based technique is combined with the binaural cues [38]. Although [15] showed that the most complex model with the garbage source gave the best performance (both ILD and IPD cues being frequency-dependent), they did not examine all of the modes with frequency-independent parameters. In addition, we incorporated the garbage source for all possible modes to have a comprehensive comparison.

Fig. 6 shows that the Mandel algorithm with fixed ILD parameters (Θ^G_{00}, Θ^G_{01}, $\Theta^G_{0\Omega}$) results in lower SDRs. It is due to the fact that the PHAT-based initialization provides some information about ITD of the sources for all the modes whereas ILD information is only incorporated to the modes when updating ILD parameters. Another interesting observation is that by exploiting the garbage source, not only the most complex mode, $\Theta^G_{\Omega\Omega}$, but also simpler modes such as Θ^G_{11} give high SDRs. Thus, it is more efficient to apply Θ^G_{11} with less computational expense for very similar results.

In the case of our proposed method, all the modes gave comparable results with $\Theta^G_{\Omega1}$ (i.e., frequency dependent ILD and frequency-independent IPD) having slightly better performance.

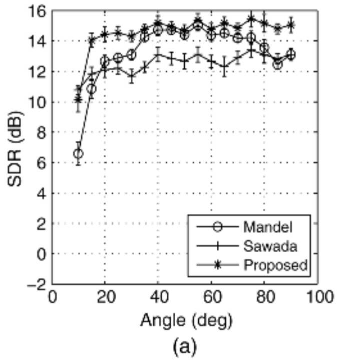


Fig. 7. SDR of the recovered target source averaged over 15 mixtures with mode Θ_{11}, $W_P = W_L = W_B = 1$ at each angular displacement in anechoic conditions, (a) 2-source and, (b) 3-source case.

However, as shown in Section IV-C, ILD models are sensitive to noise and so it is safer to set this cue as frequency independent as in mode Θ_{11}^{G} which gives similar results. In this way the binaural model parameters will be fixed for all frequency bins, preventing the permutation problem introduced by bin-wise clustering and reducing the model complexity.

B. Anechoic Conditions

This section investigates the performance of the algorithms under anechoic conditions, to examine their behavior without the effects of room reflections and reverberation. For these anechoic experiments, pilot tests confirmed our expectation that the garbage source was unnecessary, as there was no reverberation for it to model. Indeed, use of the garbage model produced a slight degradation, so this feature was disabled for these tests, whose results are plotted in Fig. 7 with error bars.

For the determined case in Fig. 7(a) with 2-source mixtures, the proposed method gave an average 3.5 dB SDR improvement over Mandel's approach when the sources were close (15° or less). The advantage reduced to 1.0 dB when the interferer was positioned at 45° or more. The average enhancement over Sawada's approach was approximately 2.0 dB for all target-interferer configurations.

For the underdetermined case, a considerable difference of almost 5.0 dB is seen in Fig. 7(b) between the proposed method over Sawada's but, compared with Mandel's, this large difference only occurs at 10° and is otherwise much more modest. The overall average separation performance for the 3-source case was 9.58 dB for the proposed method, 8.29 dB for Mandel's and 4.67 dB for Sawada's, which is consistent with the anechoic results reported in [15].

C. Reverberation Effect

For any practical system, it is vital to test its performance in typical acoustical conditions including room reflections and reverberation. To study the effect of reverberation on the performance of the algorithms, all the configurations were tested across a range of environmental conditions, as in Table II in Section VI.

Fig. 8 presents the SDRs of the recovered signals with the interfering source positioned at different azimuths. It can be seen that with different T_{60} s and DRRs in all 4 reverberant environments the proposed algorithm shows the best performance. It is also evident that the proposed method outperforms the two baselines especially when the angle between the target and the other source is less than 45°. For example, the average improvement for room A with angles less than 35° over Mandel's is about 1.5 dB, which decreases for larger angular displacements. The Mandel algorithm works well when the sources are well away from each other. Therefore, the average results over all rooms (A, B, C, D) and configurations show a smaller but statistically significant improvement of 0.37 dB with critical p-value of 1.03×10^{-22} (number of mixtures=1020). In case of PESQ, an improvement of 0.026 is shown to be significant with p-value of 3.28×10^{-30}.

TABLE II

Room acoustical properties in initial time delay gap (ITDG), direct-to-reverberant ratio in terms of (DRR) and reverberation time T_{60}[32]

Room	Type	ITDG [ms]	DRR [dB]	T_{60} [s]
A	a medium office	8.72	6.09	0.32
B	a small class room	9.66	5.31	0.47
C	a large lecture theatre	11.9	8.82	0.68
D	a large seminar room	21.6	6.12	0.89

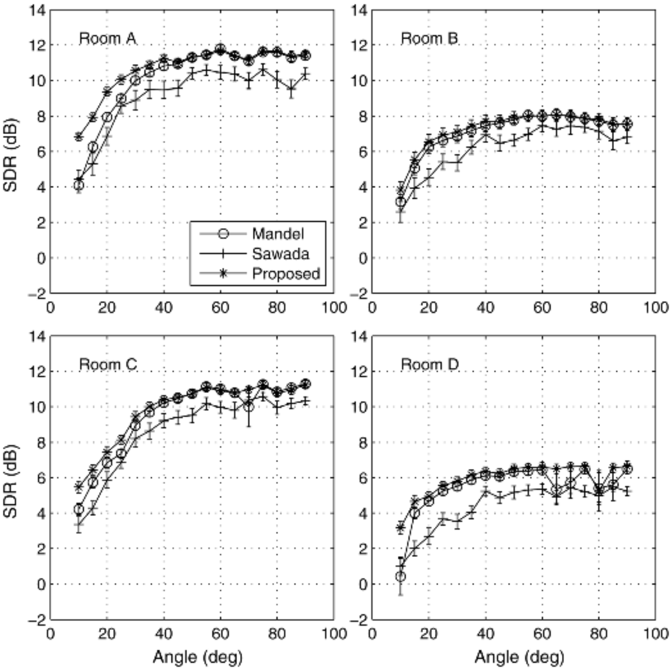


Fig. 8. SDR of the recovered target source averaged over 15 mixtures with mode Θ_{11}^{G}, $W_P = W_L = W_B = 1$ at each angular displacement in 2-source case under different rooms: room A with $T_{60} = 0.32$ s, room B with $T_{60} = 0.47$ s, room C with $T_{60} = 0.68$ s and room D with $T_{60} = 0.89$ s.

The improvement over Sawada is consistent for all the various interferer positions, but varies with environmental conditions. For example, it is especially high in room D with $T_{60} = 0.89$ s. A summary of the results is represented in Tables III and IV.

Fig. 9 presents the results for the underdetermined case with two interfering sources on the right and left hand sides of the target, respectively. It is clear that the proposed method generally outperforms the two baselines. However, there are some weak results at larger azimuths due to poor initialization in room D with its high reverberation. Overall, an average improvement of 0.33 dB over 4 reverberant rooms is achieved, which is significant with p-value $= 1.27 \times 10^{-4}$ (number of mixtures=1020).

TABLE III

RESULTS OF THE BASELINE METHODS AND PROPOSED METHOD WITHOUT ($W_P = W_L = W_B = 1$) AND WITH WEIGHTING ($W_P = 0.8, W_L = 0.1, W_B = 0.5$) FOR ANECHOIC, X, AND REVERBERANT MIXTURES WITH THE AVERAGE OVER ROOMS A, B, C AND D, IN SDR [dB]

Case	Methods	X	A	B	C	D	Mean
2-Src	Sawada	11.83	9.11	6.19	8.63	4.36	7.07
	Mandel	12.53	10.14	7.10	9.51	5.42	8.04
	Unweighted	**14.57**	10.65	7.27	9.79	5.93	8.41
	Weighted	14.03	**10.80**	**7.61**	**10.05**	**6.31**	**8.69**
3-Src	Sawada	4.67	6.43	4.13	6.03	3.30	4.97
	Mandel	8.29	7.81	4.93	7.40	3.97	6.03
	Unweighted	9.58	8.31	5.21	7.69	4.20	6.35
	Weighted	**9.61**	**8.49**	**5.52**	**8.03**	**4.73**	**6.69**

TABLE IV

RESULTS OF THE BASELINE METHODS AND PROPOSED METHOD WITHOUT ($W_P = W_L = W_B = 1$) AND WITH WEIGHTING ($W_P = 0.8, W_L = 0.1, W_B = 0.5$) FOR ANECHOIC, X, AND REVERBERANT MIXTURES WITH THE AVERAGE OVER A, B, C AND D IN PESQ

Case	Methods	X	A	B	C	D	Mean
2-Src	Mixtures	1.93	2.18	1.94	2.18	1.81	2.03
	Sawada	2.86	2.33	2.03	2.31	1.87	2.14
	Mandel	2.57	2.34	2.07	2.34	1.96	2.18
	Unweighted	**2.96**	2.37	2.09	2.36	1.99	2.21
	Weighted	2.93	**2.39**	**2.11**	**2.38**	**2.01**	**2.22**
3-Src	Mixtures	1.62	1.94	1.75	1.98	1.67	1.84
	Sawada	1.92	2.01	1.80	2.02	1.73	1.89
	Mandel	2.13	2.10	1.85	2.14	1.81	1.98
	Unweighted	2.22	2.13	1.85	2.16	1.82	1.99
	Weighted	**2.26**	**2.15**	**1.87**	**2.18**	**1.84**	**2.01**

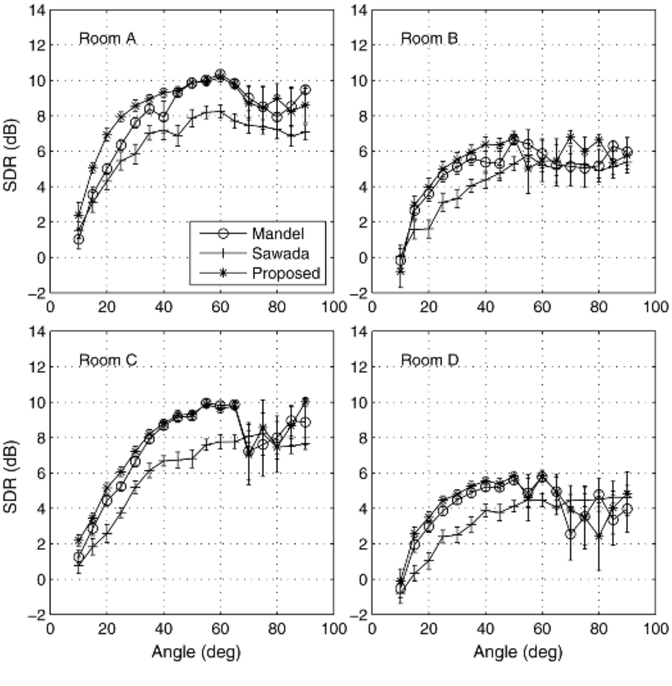


Fig. 9. SDR of the recovered target source averaged over 15 mixtures with mode $\Theta_{11}^{G}, W_P = W_L = W_B = 1$ at each angular displacement in 3-source case under different rooms: room A with $T_{60} = 0.32$ s, room B with $T_{60} = 0.47$ s, room C with $T_{60} = 0.68$ s and room D with $T_{60} = 0.89$ s.

In case of PESQ, an improvement of 0.014 is shown to be significant with p-value of 4.00×10^{-2}.

Furthermore, from Figs. 8 and 9 we see that the performance not only depends on the T_{60} but also the DRR. For example, although the T_{60} of room C is higher than that of room B, the SDRs of the recovered signals are higher in room C due to the higher direct-to-reverberant ratio ($\mathrm{DRR} = 8.8$ dB) compared to that of room B ($\mathrm{DRR} = 5.3$ dB). Therefore, it is important to consider other acoustical factors such as DRR of the rooms to examine and report the performance of an algorithm. The reverberation time (T_{60}) is not the only acoustic parameter that affects the source separation.

D. Cue Weighting

Up to this point, the cues were applied with equal weighting in our experiments (equal to 1), which is not necessarily the best way to model the data and estimate the parameters most reliably, as discussed in Section V-C. Therefore, we decided to adjust the weights of each cue to try to improve the performance with our proposed algorithm.

We first started by adjusting just one cue at a time and keeping the other weights at 1 to discover the general effect of weighting on each cue. As a pilot experiment, mixtures were selected with room A for the sources that were close to each other ($\theta = 10°, 15°, 20°$). We found that values of W_P greater than one increased the SDR of the recovered signals, suggesting that the IPD cue is more reliable than the other two cues which is consistent with our observation in IV-C. Then, we varied W_L ($W_B = W_P = 1$) and observed that giving less weight to ILD increased the quality of the results. This finding also supports the results in IV-C where ILD is degraded due to reverberation. Finally, we examined W_B and discovered that it did not affect the result considerably. Moreover, we observed that the variation of the results over W_P, W_B and W_L was smooth, enabling us to reduce the search resolution to identify the optimum combination. Overall, weighting the MV cue did not change the performance of the algorithm significantly. Weighting the IPD improved the results slightly while ILD weighting had the most influence on the outcome.

Although $W_P = 1.5$ and $W_L = 0.1$ gave the optimum values while the other two cues were fixed at 1, the combination of $W_P = 1.5, W_L = 0.1$ and $W_B = 1$ was not optimal. A coarse search (testing many combinations on all four rooms and various source positions) led us to the optimum set of $W_P = 0.8$, $W_L = 0.1$ and $W_B = 0.5$. It confirms that the relative weights of the cues are more important than the actual coefficients.

We compared the proposed algorithm with no weighting and this optimum weighting of $[0.8, 0.1, 0.5]$ with a t-test which showed that the averaged improvement of 0.32 dB over 240 mixtures was highly significant ($p = 2.55 \times 10^{-27}$).

Although this set of weightings gives the optimum results for binaural mixtures, it should be adjusted for mixtures recorded by alternative configurations, e.g., spaced omnidirectional microphones. Comparing the results in [38] with those represented in Section VII-C, one can see that the improvement (between Mandel's and the unweighted proposed method) based on mixtures without HRTF is higher than that based on binaural recordings. This suggests that the MV contribution is more effective for mixtures without HRTF. Therefore, a different set of weights with higher W_B and lower W_P would improve the performance of the algorithm under those conditions.

In reverberant and anechoic conditions with two and three speakers, the proposed algorithm with weighted cues produced SDRs 0.69 dB and 1.96 dB higher than Mandel's and Sawada's algorithm, respectively. Overall, the proposed method is more robust as compared to the baselines whose performance depends on the type of recording. For example, Mandel's method works better for binaural recordings as it is mainly based on binaural cues, whereas Sawada's method performs better for microphone recordings without HRTF.

E. Spatially Diffuse Noise

We have also evaluated the performance of the proposed algorithm, in comparison with the two baseline algorithms, for separating the mixtures corrupted by spatially diffuse noise. Diffuse noise has the property of sound energy arriving at a sensor from every direction with equal probability. For two sensors sufficiently separated in space (as in our case), we approximately simulate these conditions by adding two independent white noise sequences to the left-channel and right-channel mixture respectively. We have performed two sets of experiments. In the first set of experiments, we repeat the experiments performed in Section VII-B by adding spatially diffuse noise to each of the mixtures used. All the other set-ups (including the parameters set-up, the mode for the IPD/ILD model and the weights for integrating the cues) were exactly the same as those in Section VII-B. In the second set of experiments, we repeat the experiments for reverberant rooms as performed in Section VII-C, where we followed the same set-ups except that we added spatially diffuse noise to each of the mixtures in these new tests. Due to the space constraint, we only report results for room C here (similar performance trends are observed for other rooms). In both sets of experiments, two different levels of noise in terms of signal-to-noise-ratios (SNRs), were tested, 10 dB and 20 dB, respectively. We will only show the results for $\mathrm{SNR} = 10$ dB (again, due to space constraint).

For the anechoic mixtures, the average SDR results are shown in Fig. 10 for $\mathrm{SNR} = 10$ dB. From Fig. 10, it can be observed in the anechoic case that, the proposed algorithm performs better than the MV algorithm (i.e. Sawada's algorithm), especially for the angles between 20° and 60° in diffuse noise. It also outperforms the binaural cue based algorithm (i.e. Mandel's algorithm) for nearly all the angles. By comparing Fig. 10(a) and Fig. 10(b), we can further observe that the performance advantage of the MV cue over the binaural cues in diffuse noise tends to drop considerably with the increase of the number of sources. In the three-source case, our proposed algorithm also performs better than Sawada's algorithm for angles between 20° and 60°, and gives comparable results to Sawada's algorithm for the other angles. We observed in our experiments that when the noise level was not very high, e.g. 20 dB SNR, the binaural cues performed well (results not shown), similar to the case of noise-free conditions (shown in Fig. 7). Yet our proposed algorithm gave consistently better performance as compared with both baseline algorithms, for both two source and three source situations.

For the reverberant case (i.e. room C), the average SDR results are shown in Fig. 11 for $\mathrm{SNR} = 10$ dB. From Fig. 11 with 10 dB noise corruption, it can be observed that, similar to the anechoic case, Mandel's method is greatly affected by the diffuse noise, while Sawada's method is less affected. In this case, Sawada's method exhibits advantages over our proposed algorithm as well as Mandel's method. The reason that the proposed algorithm does not show benefit over Sawada's algorithm in diffuse noise is related to the combination of these cues. However, the results further confirm that the MV cue can be complementary to the IPD/ILD cues since the proposed algorithm improves Mandel's algorithm in diffuse noise. We also observed in our experiments that, when the noise level is not as high, e.g. with 20 dB diffuse noise (results omitted), our proposed algorithm outperforms the two baseline methods, for both two source and three source conditions. Overall, the results are very consistent with the SDR evaluations in the noise-free conditions for room C, as shown previously in Figs. 8 and 9. We would like to note that incorporating a precedence model would be expected to improve the performance of binaural method in reverberation as suggested by our preliminary work in [39].

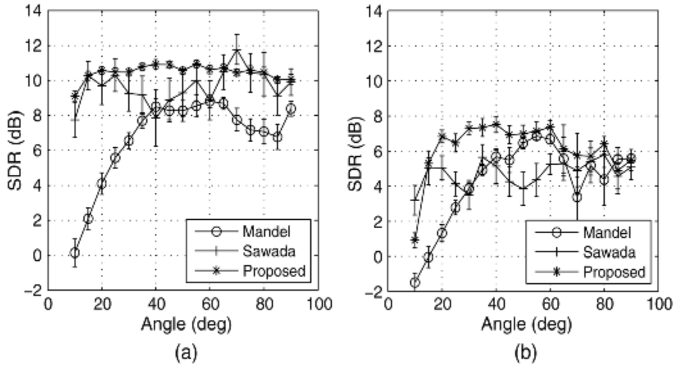


Fig. 10. SDR of the recovered target source averaged over 15 mixtures with mode Θ_{11}, $W_P = W_L = W_B = 1$ at each angular displacement in anechoic conditions, with 10 dB spatially diffuse noise corruption, (a) 2-source and, (b) 3-source case.

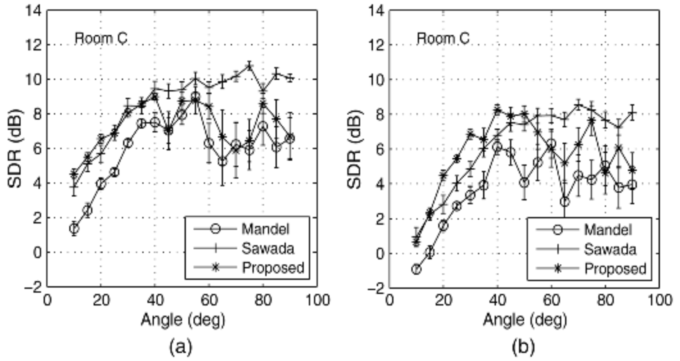


Fig. 11. SDR of the recovered target source averaged over 15 mixtures with mode Θ_{11}, $W_P = W_L = W_B = 1$ at each angular displacement in room C, with 10 dB spatially diffuse noise corruption, (a) 2-source and, (b) 3-source case.

VIII. Conclusion

We have studied stereo speech mixtures and analyzed the difference between the MV and the binaural cues. We have shown that the MV cue tends to be more distinct when the sources are close to each other, while the binaural cues, and specially IPD, are more robust to high reverberation for which the MV models degrade. This has led us to combine the cues to compensate for their limitations. We have presented a new algorithm for separating speech mixtures under challenging conditions by considering both additive and convolutive noise models in parallel. It has been shown that this approach improves the quality of the recovered signals in comparison with the two baseline state-of-the-art algorithms named as Mandel [15] and Sawada [6]. We have shown the potential benefits by weighting each cue to adjust their contributions for the T-F unit classification.

Another interesting point is the difference in the performance of the algorithms in four different rooms. We observed that T_{60} is not the only important factor affecting the performance of the algorithms. Other acoustic properties of the recording environment such as DRR also have a great influence on the results. Tests on mixtures corrupted by spatially-diffuse noise also confirmed these findings.

Here the cue weights are fixed over all frequencies whereas frequency-dependent coefficients may yield additional gains in performance. We observed that, the initialization fails at high reverberation, which should be addressed in further work. Finally, as we have concentrated on SDR enhancement, the PESQ results have not changed considerably. This could be achieved by cepstral smoothing to improve the perceptual quality of the signals.

Acknowledgment

We would like to give special thanks to Michael Mandel for giving us his code and assistance generously, and to Hiroshi Sawada for providing help with our implementation of his algorithm. We would also like to thank the Associate Editor and the anonymous reviewers for their contributions to improving the quality of the paper.

References

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, ser. Independent Component Analysis and Applications Series. Boston, MA, USA: Elsevier Science, 2010 [Online]. Available: http://books.google.co.uk/books?id=PTbj03bYH6kC

[2] J. V. Stone, *Independent Component Analysis, A tutorial introduction*. Cambridge, MA, USA: Mass. Inst. of Technol., 2004.

[3] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4–5, pp. 411–430, Jun. 2000.

[4] B. Van Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[5] O. Ylmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2003.

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

[7] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. 8th Int. Conf. Ind. Compon. Anal. Signal Separat.*, Berlin/Heidelberg, Germany, 2009, pp. 734–741 [Online]. Available: http://dx.doi.org/10.1007/978-3-642-00599-2_92, ser. ICA '09, Springer-Verlag

[8] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, "Speech intelligibility and localization in a multi-source environment," *J. Acoust. Soc. Amer.*, vol. 105, no. 6, pp. 3436–3448, Jun. 1999.

[9] W. M. Hartmann, "How we localize sound," *Phys. Today*, pp. 24–29, Nov. 1999.

[10] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: Mass. Inst. of Technol., 1994.

[11] D. L. Wang and G. J. Brown, *Computational Aditory Scene Analysis: Principles, Algorithms and Applications*, D. L. Wang and G. J. Brown, Eds. New York, NY, USA: Wiley Interscience and IEEE Press, 2006.

[12] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," in *Proc. Int. Joint Conf. Neural Netw.*, 2001, vol. 4, pp. 2861–2866.

[13] J. Han and B. Pardo, "Improving separation of harmonic sources with iterative estimation of spatial cues," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, Oct. 2009, pp. 77–80.

[14] J. Woodruff and B. Pardo, "Using pitch, amplitude modulation, and spatial cues for separation of harmonic instruments from stereo music recordings," in *EURASIP J. Adv. Signal Process.*, Jan. 2007, vol. 2007, no. 1 [Online]. Available: http://dx.doi.org/10.1155/2007/86369

[15] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[16] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, Oct. 2007, pp. 139–142.

[17] J. Woodruff, R. Prabhavalkar, E. Fosler-Lussier, and D. Wang, "Combining monaural and binaural evidence for reverberant speech segregation," in *Proc. INTERSPEECH*, 2010, pp. 406–409.

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

[19] M. I. Mandel and D. P. W. Ellis, "The ideal interaural parameter mask: A bound on binaural separation systems," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New York, NY, USA, 2009.

[20] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Trans. Syst., Man, Cybern. C*, vol. 32, no. 4, pp. 474–484, Nov. 2002.

[21] P. O'Grady and B. Pearlmutter, "The lost algorithm: Finding lines and separating speech mixtures," in *EURASIP J. Adv. Signal Process.*, 2008 [Online]. Available: http://asp.eurasipjournals.com/content/2008/1/784296

[22] W. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*. Berlin/Heidelberg, Germany: Springer , 2012 [Online]. Available: http://books.google.co.uk/books?id=3Wz205ve5ioC

[23] D. Serre, *Matrices: Theory and Applications*, ser. ser. Graduate Texts in Mathematics. New York, NY, USA: Springer, 2002 [Online]. Available: http://books.google.co.uk/books?id=3gPezLJQqucC

[24] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus cdrom," . Philadelphia, PA, USA, Linguistic Data Consortium, 1993 [Online]. Available: http://www.idc.upenn.edu/Catalog/LDC93S1.html

[25] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, Sep. 2010.

[26] A. Spanos, *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge, U.K.: Cambridge Univ. Press, 1999 [Online]. Available: http://books.google.co.uk/books?id=G0_HxBubGAwC

[27] J. Hershey and P. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, vol. 4, pp. 317–320.

[28] M. I. Mandel and D. P. W. Ellis, "A probability model for interaural phase difference," in *Proc. ISCA Workshop Statist. Percept. Audio Process. (SAPA)*, 2006, pp. 1–6.

[29] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. IEEE Workshop Applicat. Signal Process. Audio and Acoust.*, Oct. 2007, pp. 139–142.

[30] T. Petsatodis, C. Boukis, F. Talantzis, Z. H. Tan, and R. Prasad, "Convex combination of multiple statistical models with application to VAD," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2314–2327, Nov. 2011.

[31] L. Rayleigh, "On our perception of sound direction," *Philos. Mag.*, vol. 13, no. 74, pp. 214–232, 1907.

[32] C. Hummersone, "A psychoacoustic engineering approach to machine sound source separation in reverberant environments," Ph.D. dissertation, Music and Sound Recording, Univ. of Surrey, Guildford, U.K., 2011.

[33] B. Shinn-Cunningham, N. Kopco, and T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Amer.*, vol. 117, pp. 3100–3115, May 2005.

[34] H.-D. Kim, J. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Target speech detection and separation for humanoid robots in sparse dialogue with noisy home environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Rob. Syst.*, Sep. 2008, pp. 1705–1711.

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

[36] L. D. Persia, D. Milone, H. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Process.*, vol. 88, no. 10, pp. 2578–2583, Oct. 2008.

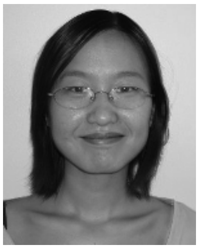
[37] N. D. Gaubitch, "Allen and Berkley image model for room impulse response," Imperial College London, [Online]. Available: http://www.commsp.ee.ic.ac.uk/%7Endg/downloadfiles/mcsroom.m, 1979

[38] A. Alinaghi, W. Wang, and P. J. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 209–212.

[39] A. Alinaghi, W. Wang, and P. Jackson, "Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 684–688.

Atiyeh Alinaghi received the B.Sc. degree in electrical engineering from Sharif University of Technology, Iran, in 2006, the M.Sc. degree in applied digital signal processing with distinction from University of Southampton, U.K., in 2009. She is now pursuing her PhD degree in the area of blind source separation at the Centre for Vision, Speech and Signal Processing (CVSSP) in University of Surrey. Her main interest includes audio signal processing, blind source separation, computational auditory scene analysis and statistical signal processing.

Philip Jackson is Senior Lecturer in Machine Audition at the Centre for Vision, Speech and Signal Processing, University of Surrey, U.K. He joined CVSSP in 2002 after a postdoctoral fellowship at University of Birmingham (U.K.), with PhD from University of Southampton (UK) and MA from Cambridge University (U.K.). Through projects Columbo, BALTHASAR, DANSA, SAVEE, DynamicFaces, QESTRAL, UDRC, POSZ and S3A, he has contributed to active noise control for aircraft, speech aero-acoustics, source separation and articulatory models for automatic speech recognition, audio-visual emotion classification and visual speech synthesis, plus techniques for spatial audio. He has over 100 journal, patent, conference and book publications. He serves as reviewer for the Journal of the Acoustical Society of America, IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, IEEE SIGNAL PROCESSING LETTERS, InterSpeech and ICASSP, and as associate editor for Computer Speech and Language (Elsevier).



Qingju Liu received the B.Sc. degree in electronic information engineering from Shandong University, Jinan, China in 2008, and the Ph.D. degree in signal processing in 2013 under the supervision of Dr. Wenwu Wang within the Machine Audition Group at the Centre for Vision, Speech and Signal Processing (CVSSP) in University of Surrey, Guildford, U.K. Since October 2013, she has been working as a research fellow in CVSSP. Her current research interests include audio-visual signal processing, spatial audio reproduction and machine learning.

Wenwu Wang (M02–SM11) was born in Anhui, China. He received the B.Sc. degree in automatic control in 1997, the M.E. degree in control science and control engineering in 2000, and the Ph.D. degree in navigation guidance and control in 2002, all from Harbin Engineering University, Harbin, China. He then joined Kings College, London, U.K., in May 2002, as a postdoctoral research associate and transferred to Cardiff University, Cardiff, U.K., in January 2004, where he worked in the area of blind signal processing. In May 2005, he joined the Tao Group Ltd. (now Antix Labs Ltd.), Reading, U.K., as a DSP engineer working on algorithm design and implementation for real-time and embedded audio and visual systems. In September 2006, he joined Creative Labs, Ltd., Egham, U.K., as an R&D engineer, working on 3D spatial audio for mobile devices. Since May 2007, he has been with the Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, U.K., where he is currently a Senior Lecturer, and a Co-Director of the Machine Audition Lab. He is a member of the Ministry of Defence (MoD) University Defence Research Collaboration (UDRC) in Signal Processing (since 2009), a member of the BBC Audio Research Partnership (since 2011), and an associate member of Surrey Centre for Cyber Security (since 2014). During spring 2008, he has been a visiting scholar at the Perception and Neurodynamics Lab and the Center for Cognitive Science, The Ohio State University.

His current research interests include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, machine audition (listening), and statistical anomaly detection. He has been a Principal Investigator or Co-Investigator on a number of research grants (total award approximately £10M) funded by the U.K. governmental bodies such as the Engineering and Physical Sciences Research Council (EPSRC), Ministry of Defence (MoD), Defence Science and Technology Laboratory (Dstl), Home Office (HO), and the Royal Academy of Engineering (RAEng), as well as the U.K. industry such as BBC and Samsung (U.K.). He has (co-)authored over 130 publications in these areas, including two books, namely, Machine Audition: Principles, Algorithms and Systems by IGI Global published in 2010 and Blind Source Separation by Springer in 2014. He has been a regular reviewer for many IEEE journals including IEEE TRANSACTIONS ON SIGNAL PROCESSING and IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING, and an associate editor of The Scientific World Journal: Signal Processing. He has also been a chair, session chair or a technical/program committee member on a number of international conferences including Local Arrangement Co-Chair of MLSP 2013, Session Chair of ICASSP 2012, Area and Session Chair of EUSIPCO 2012, Track Chair and Publicity Co-Chair of SSP 2009. He was a Tutorial Co-Speaker on ICASSP 2013.