Bootstrap Averaging for Model-Based Source Separation in Reverberant Conditions

Swati Chandna and Wenwu Wang, *Senior Member, IEEE*

***Abstract*—Recently proposed model-based methods use time-frequency (T-F) masking for source separation, where the T-F masks are derived from various cues described by a frequency domain Gaussian Mixture Model (GMM). These methods work well for separating mixtures recorded in low-to-medium level of reverberation, however, their performance degrades as the level of reverberation is increased. We note that the relatively poor performance of these methods under reverberant conditions can be attributed to the high variance of the frequency-dependent GMM parameter estimates. To address this limitation, a novel bootstrap-based approach is proposed to improve the accuracy of expectation maximization (EM) estimates of a frequency-dependent GMM based on an *a priori* chosen initialization scheme. It is shown how the proposed technique allows us to construct time-frequency masks which lead to improved model-based source separation for reverberant speech mixtures. Experiments and analysis are performed on speech mixtures formed using real room-recorded impulse responses.**

***Index Terms*—Gaussian mixture model, EM algorithm, bootstrap averaging, model-based source separation, time-frequency masking, reverberant speech mixtures, audio signal processing, spectral histogram**

I. Introduction

Source separation is defined as the problem of separating multiple sources mixed through an unknown mixing system (channel), using only the system outputs (e.g. observed mixtures of speech). Let I denote the number of sources and M denote the number of channels. At discrete time point $n \in \{1, \ldots, N\}$, the system output $x_m(n)$ at the mth channel is a convolutive mixture of the form

$$x_m(n) = \sum_{i=1}^{I} s_i(n) * h_{im}(n) * e_m(n), \tag{1}$$

where $*$ denotes convolution, $s_i(n)$ is the ith source, $h_{im}(n)$ for $m = 1, \ldots, M$, is the room impulse response from source i to channel m, and $e_m(n)$ denotes convolutive noise. The choice of a convolutive noise is made for analytical convenience as it leads to an additive term in the log-magnitude and phase domains, [1]. For each $i = 1, \ldots, I$, let $\mathbf{s}_i = [s_i(1), \ldots, s_i(N)]^T$ denote the source observed at N time points, and similarly, for each $m = 1, \ldots, M$, let $\mathbf{x}_m = [x_m(1), \ldots, x_m(N)]^T$ denote the corresponding mixture vector. Then the problem of source separation deals with the estimation of the source vectors $\mathbf{s}_1, \ldots, \mathbf{s}_I$, given the mixture vectors $\mathbf{x}_1, \ldots, \mathbf{x}_M$. This problem is termed *underdetermined* when the number of observed mixtures, M, is less than the number of sources, I, that comprise the mixture.

In many real-world applications, the population may consist of several sub-populations and a standard distribution is not able to capture the variation over these sub-populations effectively. Finite mixture[1] models, as the name suggests, are extensively used to model such data with a finite mixture of standard distributions. Mixture distributions are extremely popular in areas such as audio-signal processing, image analysis, and geology, where they are used to model spectrograms in the time-frequency domain. Here, the time-frequency analysis of piecewise stationary signals allows the use of GMMs over time frames at each frequency. We shall refer to such frequency-specific GMMs as frequency domain GMMs. Some examples of applications employing GMMs in the frequency domain are [2], [3], [4] in speech signal analysis; [5], [6] in image analysis.

Model-based blind source separation for exactly determined and underdetermined speech mixtures such as [1], [7], [8], [9], are more recent examples of applications in speech analysis involving frequency-specific GMMs. These methods have gained significant popularity due to their simple model-based approach for integration of cues. In these methods, a by-product of the EM algorithm, used to estimate parameters of the frequency domain GMM, is a time-frequency (T-F) mask that allows separation of the target source of interest from the source of interference. These methods perform extremely well for mixtures recorded under low-to-medium levels of reverberation, however, their performance degrades as the reverberation level is increased. The poor performance of such algorithms for reverberant mixtures is attributed to inaccurate EM estimation of the frequency-dependent GMM parameters. More generally, this is due to the absence of an explicit model for reverberation. In addition to this, the frequency domain GMM in these algorithms, [1], [7], [8], relies on the assumption of the cues being independent. As noted in [8, sec. III.A], this assumption of the cues being independent does not hold in practice and is used as a convenient way to make the problem of source separation tractable. Overall, this results in a misspecified mixture model, leading to EM estimates with a very high variance. We show how better EM estimates of the target source parameters can be obtained using

S. Chandna is with the Dept. of Economics, Mathematics and Statistics at Birkbeck, University of London, London WC1E 7HX, UK (email: s.chandna@bbk.ac.uk) and W. Wang is with the Centre for Vision, Speech and Signal Processing (CVSSP), Department of Electrical and Electronic Engineering, University of Surrey, Guildford GU2 7XH, UK (e-mail: w.wang@surrey.ac.uk)

[1]Note that the term 'mixture' here refers to a mixture of distributions and not a mixture of sources.

the proposed bootstrap-based procedure to improve model-based source separation from reverberant speech mixtures. Our method is described for a univariate GMM in the frequency domain. Note that this does not require the observed time domain sample to be univariate, for example, as in the source separation algorithm of [1], where a univariate frequency domain GMM is constructed by transforming a two-dimensional vector observation of speech mixtures.

Bootstrap methods are commonly used to draw inference on statistics of interest when no theoretical results are available, or when inference based on theoretical results is computationally intractable, such as to obtain standard errors and confidence intervals, [10]. Non-traditional applications of bootstrap which show how it can also facilitate a more robust statistical analysis are found, e.g., in machine learning to improve the forecast accuracy of models selected by unstable decision rules, [11], as well as in the area of pattern analysis where it is applied to the problem of fitting ellipse segments to noisy data to eliminate bias in the ellipse estimates [12]. The use of bootstrap for EM estimates of a frequency-dependent GMM, or to improve source separation performance as shown in our work, to the best of our knowledge, has never been mentioned in the literature and is the contribution of this paper. We would like to point out that the proposed idea of bootstrap averaging is very well-suited to the above mentioned source separation problem since the GMM appears in the frequency domain, and hence can be bootstrapped indirectly using the sample in the time domain, details of which are provided in this paper. Results from a set of preliminary experiments using this approach were presented in [13].

More recently, there has been a growing interest in techniques employing non-negative matrix factorization (NMF) and deep neural networks (DNN) to the problem of source separation. The idea with NMF for multichannel source separation, e.g. [14], is to model the power spectrogram of each source in the T-F domain as a product of two non-negative matrices. Then modeling the short-time-Fourier transform (STFT) of each source as the sum of a finite number of latent Gaussian components, an EM implementation and a multiplicative update (MU) approach have been proposed to estimate the matrix factors (determining the variance of the Gaussian components) as well as the unknown convolutive mixing matrix. An NMF and spatial covariance model has also been studied for underdetermined source separation under reverberant conditions [15]. This is based on the EM algorithm for estimation of the model parameters and the authors note the sensitivity of their estimates to parameter initialization as well as degrading performance with increasing reverberation times. Another formulation for multichannel NMF described in [16], clusters the NMF bases according to their spatial properties. DNNs have been used to separate sources from binaural mixtures under reverberation via a binary classification, e.g. [17] and related work by [18], or using a probabilistic time-frequency mask, e.g. [19]. The latter approach integrates binaural cues following the model-based method of [8]. A multichannel source separation method is described in [20].

DNN based approaches are supervised methods that need training data sets, which may not always be available. Model-based approaches such as [1], [9], [8] considered in this paper as well as NMF methods for source separation allow the inclusion of spatial and spectral cues; do not require training data and are easier to deploy in unfamiliar environments. Noting these points, it is of interest to study improvements that can be achieved with such model-based methods and are the subject of this paper. Our approach is illustrated using the method of [8] and can also be adapted, for example, to improve the NMF-based multichannel source separation of [14], as discussed in section VIII.

A. Contributions and organization

Following some background and notation on GMMs and their EM estimation, the contributions of this paper are presented as follows:

1) The proposed idea of bootstrap averaging is described in section III. A simulation example to illustrate the case of sub-optimal EM estimates and the use of averaging to reduce the mean-squared-error is discussed. This experiment shows the benchmark improvement (measured via MSE) that may be achieved via the proposed method in an ideal set-up (without bootstrapping).
2) Model-based source separation focusing on the forms of frequency domain GMMs that appear in such applications is described in section IV.
3) Bootstrap averaging for the source separation algorithm of [8] is presented in section V. Simulation experiments using speech mixtures formed with real room-recorded impulse responses are included in section VI.
4) A further in-depth analysis to understand overall improvements in model-based source separation via the proposed methodology is provided in section VII.

II. Background and Notation

Let $\mathcal{G} \equiv \{g_{Y|j}(y|\boldsymbol{\lambda}_j), j = 1, \ldots, d\}$ denote a set of d probability density functions, each with parameter vector $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_d$, respectively. Let $y_1, \ldots, y_K$ denote a length-K sample from a scalar-valued random process Y, such that each observation of the length-K sample arises from one of the d density functions in $\mathcal{G}$. For $j = 1, \ldots, d$, let $Z_j(y)$ denote an indicator variable which takes the value one if observation y comes from the jth component density $g_{Y|j}(y|\boldsymbol{\lambda}_j)$. Consider the case where $Z_j(y)$ is not observed for $y \equiv y_1, \ldots, y_K$, i.e., the membership of y_i for $i = 1, \ldots, K$ in one of the d components is unknown. Then, the probability density function of Y, denoted as $g_Y(.)$ is obtained by marginalizing the joint density of Y and Z_j over the latent variable Z_j, as

$$g_Y(y|\boldsymbol{\Lambda}, \mathbf{w}) = \sum_{j=1}^{d} g_{Y|Z_j}(y|z_j) g_{Z_j}(z_j) = \sum_{j=1}^{d} g_{Y|j}(y|\boldsymbol{\lambda}_j) w_j, \tag{2}$$

where $g_{Z_j}(z_j) = w_j$ is the probability of the observed y arising from the jth component density with parameter $\boldsymbol{\lambda}_j$. Thus, the weights are non-negative and satisfy the condition $w_1 + \ldots + w_d = 1$. On the left hand side of (2), $\mathbf{w}$ denotes the vector of weights $\mathbf{w} \equiv [w_1, \ldots, w_d]^T$ and $\boldsymbol{\Lambda} \equiv [\boldsymbol{\lambda}_1^T, \ldots, \boldsymbol{\lambda}_d^T]^T$ is the vector of density parameters. So for

a mixture of d Gaussian distributions with mean and variance parameters denoted using μ, σ^2, we have $\boldsymbol{\lambda}_j = [\mu_j, \sigma_j^2]^T$ and $\boldsymbol{\Lambda} = [\mu_1, \sigma_1^2, \ldots, \mu_d, \sigma_d^2]^T$. Equation (2) denotes a weighted mixture of component densities $g_{Y|j}(y, \boldsymbol{\lambda}_j), j = 1, \ldots, d$ with weights $w_1, \ldots, w_d$, respectively [21]. If the component densities $g_Y(y|\boldsymbol{\lambda}_j)$ are Gaussian, then (2) takes the form

$$g_Y(y|\boldsymbol{\Lambda}, \mathbf{w}) = \sum_{j=1}^{d} g_{Y|j}(y|\mu_j, \sigma_j^2) w_j, \tag{3}$$

where $g_{Y|j}(y|\mu_j, \sigma_j^2)$ denotes the Gaussian probability density function with mean μ_j and variance σ_j^2.

A. EM estimation

Let $\boldsymbol{\Psi}$ denote the vector of all unknown parameters of the GMM, i.e. $\boldsymbol{\Psi} = [w_1, \ldots, w_{d-1}, \mu_1, \ldots, \mu_d, \sigma_1^2, \ldots, \sigma_d^2]^T$, and let $\boldsymbol{\Omega}$ denote the parameter space for $\boldsymbol{\Psi}$. The problem of maximum likelihood estimation of the parameters in $\boldsymbol{\Psi}$ is formulated as an incomplete data problem, where the observed vector $\mathbf{y} = [y_1, \ldots, y_K]^T \in \mathbb{R}^K$ is viewed to be incomplete since the corresponding component labels are not available.

For each $i = 1, \ldots, K$, let $\mathbf{z}_i = [z_1(y_i), \ldots, z_d(y_i)]^T$ denote the length-d vector of indicator variables where the index of its non-zero entry indicates the component to which the ith observation y_i belongs. Let $\mathbf{y}_C = [\mathbf{y}, \mathbf{Z}]$, with $\mathbf{Z} = [\mathbf{z}_1^T, \ldots, \mathbf{z}_K^T]^T \in \{0, 1\}^{K \times d}$, denote the complete-data matrix. The EM algorithm forms the log-likelihood function $L_C(\boldsymbol{\Psi})$ based on the complete-data $\mathbf{y}_C$ as,

$$\log L_C(\boldsymbol{\Psi}) = \sum_{i=1}^{K} \sum_{j=1}^{d} z_{ij} \{\log g(y_i|\mu_j, \sigma_j^2) + \log w_j\}, \tag{4}$$

where $z_{ij} = (\mathbf{z}_i)_j$, and circumvents the problem of unobserved component-labels by working iteratively with the conditional expectation of the complete-data log-likelihood given the observed sample vector $\mathbf{y}$. More specifically, the E-step computes: $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(m)}) = E(L_C(\boldsymbol{\Psi})|\mathbf{y}, \hat{\boldsymbol{\Psi}}^{(m)})$, using the fit $\hat{\boldsymbol{\Psi}}^{(m)}$ at the mth iteration. The M-step on the $(m+1)$th iteration involves computing the global maxima of $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(m)})$ w.r.t $\boldsymbol{\Psi}$ over the parameter space $\boldsymbol{\Omega}$ to get the updated estimate $\hat{\boldsymbol{\Psi}}^{(m+1)}$, [22]. The EM algorithm is initialized with parameter values in $\boldsymbol{\Psi}^{(0)}$ and subsequently the iterative E- and M-steps are alternated repeatedly until the difference between the observed data log-likelihood function $L(\boldsymbol{\Psi})$ computed at $\boldsymbol{\Psi}^{(m+1)}$ and $\boldsymbol{\Psi}^{(m)}$ changes by a small amount, i.e. stop at stage m when

$$\left| \frac{L(\boldsymbol{\Psi}^{(m+1)})}{L(\boldsymbol{\Psi}^{(m)})} - 1 \right| < \epsilon, \tag{5}$$

where ϵ denotes the desired tolerance, [22]. The EM algorithm is sensitive to the choice of starting values or initialization, and therefore it is important to use robust initialization schemes, [23], [22]. For our experiments in the next section, we use the search/run/select (S/R/S) initialization scheme of [23] which is known to perform well in practice. The three step strategy is to first (i) search for p initial positions, for example based on random starts using an EM run; next, (ii) run the EM algorithm at each initial position for a fixed number of times, say L; and finally (iii) select the solution that provides the best likelihood among all the $L \times p$ trials.

III. The Proposed Method

We propose a bootstrap averaging approach where for each GMM parameter, the EM estimates (based on the *a priori* chosen initialization scheme) computed from bootstrap replicates of the observed sample are averaged to reduce the variance, while leaving their bias unchanged.

Let $\mathbf{y} = [y_1, \ldots, y_K]^T$ denote a length-K sample obtained by independently drawing samples from the probability distribution F_Y. Let $\hat{\theta}(\mathbf{y}) \equiv \hat{\theta}$ denote a scalar-valued statistic of interest derived from $\mathbf{y}$. Consider the averaged estimator

$$\hat{\theta}_{\mathrm{A}}(B) = \frac{\hat{\theta}_1 + \ldots + \hat{\theta}_B}{B}, \tag{6}$$

based on B samples $\mathbf{y}_1, \ldots, \mathbf{y}_B$ from $F_{\mathbf{y}}$, with the bth estimate $\hat{\theta}_b$ derived from the bth length-K sample $\mathbf{y}_b = [y_{b,1}, \ldots, y_{b,K}]^T$. Then the bias and variance of $\hat{\theta}_{\mathrm{A}}$, are given by

$$\begin{aligned} \mathrm{BIAS}(\hat{\theta}_{\mathrm{A}}) &= E(\hat{\theta}_{\mathrm{A}}) - \theta \\ &= \frac{1}{B}[\mathrm{BIAS}(\hat{\theta}_1) + \ldots + \mathrm{BIAS}(\hat{\theta}_B)] \\ &= \mathrm{BIAS}(\hat{\theta}); \end{aligned} \tag{7a}$$

$$\mathrm{VAR}(\hat{\theta}_{\mathrm{A}}) = \frac{1}{B^2} \left[\sum_{j=1}^{B} \mathrm{VAR}(\hat{\theta}_j) + \sum_{\substack{j,k=1 \\ j \neq k}}^{B} \mathrm{Correlation}(\hat{\theta}_j, \hat{\theta}_k) \right]. \tag{7b}$$

Then if $\hat{\theta}_1, \ldots, \hat{\theta}_B$ are pairwise uncorrelated, we get

$$\mathrm{VAR}(\hat{\theta}_{\mathrm{A}}) = \frac{\mathrm{VAR}(\hat{\theta})}{B}. \tag{8}$$

Thus, on averaging, under the pairwise uncorrelated assumption, the bias remains unchanged whereas the variance is reduced. Since $\mathrm{MSE}(\hat{\theta}_{\mathrm{A}}) = \{\mathrm{BIAS}(\hat{\theta}_{\mathrm{A}})\}^2 + \mathrm{VAR}(\hat{\theta}_{\mathrm{A}})$, it follows that $\mathrm{MSE}(\hat{\theta}_{\mathrm{A}}) \leq \mathrm{MSE}(\hat{\theta})$. This provides motivation for the averaged estimator $\hat{\theta}_{\mathrm{A}}$, when $\hat{\theta}$ is known to have a small bias but high variance.

In practice, the underlying distribution of the observed sample is unknown. We propose the idea of constructing the averaged estimator by bootstrapping the given sample $\mathbf{y}$ to obtain bootstrap samples $\mathbf{y}_1^*, \ldots, \mathbf{y}_B^*$, from which the corresponding bootstrap estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ can be derived. Then, we define the bootstrap sample version of (6) as

$$\hat{\theta}_{\mathrm{A}}^*(B) = \frac{\hat{\theta}_1^* + \ldots + \hat{\theta}_B^*}{B}. \tag{9}$$

The bootstrap samples are easily generated to be independent of each other, however, the corresponding estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$, may be correlated. Since the correlation term in (7b) is weighted by a factor of $1/B^2$, choosing a bootstrap size B such that the sum of pairwise correlations is negligible in comparison to B, is sufficient for a reduction in the variance of the estimate. This leads to

$$\mathrm{BIAS}(\hat{\theta}_{\mathrm{A}}^*) = \frac{1}{B}[\mathrm{BIAS}(\hat{\theta}_1^*) + \ldots + \mathrm{BIAS}(\hat{\theta}_B^*)], \tag{10a}$$

$$\text{VAR}(\hat{\theta}^*_{\text{A}}) = \frac{1}{B^2}[\text{VAR}(\hat{\theta}^*_1) + \ldots + \text{VAR}(\hat{\theta}^*_B)]. \quad (10b)$$

Note that $\text{BIAS}(\hat{\theta}^*_j) = E(\hat{\theta}^*_j) - \theta$ for any $j = 1, \ldots, B$, approximates $\text{BIAS}(\hat{\theta}) = E(\hat{\theta}) - \theta$, with the expectation taken over the bootstrap distribution of $\hat{\theta}$ rather than its theoretical distribution. Thus, under an appropriately chosen bootstrap method for $\hat{\theta}$ [10], (10a) and (10b) imply that

$$\text{BIAS}(\hat{\theta}^*_{\text{A}}) \approx \text{BIAS}(\hat{\theta}), \quad (11a)$$

and similarly,

$$\text{VAR}(\hat{\theta}^*_{\text{A}}) \approx \frac{1}{B}[\text{VAR}(\hat{\theta})], \quad (11b)$$

so that $\text{MSE}(\hat{\theta}^*_{\text{A}}) \leq \text{MSE}(\hat{\theta})$. This shows that assuming an appropriate bootstrap method based on a sufficiently large number of bootstrap samples B, a smaller mean squared error estimate can be achieved by using the bootstrap averaged estimator $\hat{\theta}^*_{\text{A}}(B)$ given by (9). Note that the main motivation for proposing the bootstrap averaged EM estimator is the sub-optimal nature of EM estimates of GMMs used to approximate structure in the frequency domain, providing scope for further improvement.

Since our interest in this paper lies in frequency domain GMMs, we prescribe a fast circulant embedding based procedure [24, Chapter 7], [25], which has the ability to correctly mimic the underlying dependence structure in the frequency domain. Details of the bootstrap procedure for use with frequency domain GMMs arising in model-based source separation are provided in section VI.

A. A Simulation Example

To get an indication of the scope of improvement via the averaging approach, we consider the averaged estimate $\hat{\theta}_A(B)$ given by (6), computed from B randomly generated realizations, without bootstrapping. We illustrate the case of high variance EM estimates using a misspecified GMM and show how the averaged estimator provides a smaller MSE. For simplicity, we work with EM estimates of realizations generated from a GMM in the *time domain*– a term used in the rest of the paper to refer to any GMM not in the frequency domain.

We consider a mixture of two Gaussian distributed random variables where each component is generated from an autoregressive process of order one, denoted AR(1), i.e.

$$y_t = \phi_j y_{t-1} + \epsilon_{j,t}, \;\; j = 1, 2, \quad (12)$$

where $\epsilon_{j,t} \sim \mathcal{N}(0, \sigma^2_{\epsilon,j})$ is the error term and ϕ_j denotes the AR coefficient, corresponding to the jth component of the Gaussian mixture. Then, clearly for each $j = 1, 2$, the true component means $\mu_j = E(\epsilon_j) = 0$ and the true component variances $\sigma^2_j = \sigma^2_{\epsilon,j}/(1 - {\phi_j}^2)$. Now from eqn. (12), it follows that $(y_t - \phi_j y_{t-1})/\sigma_{\epsilon,j}$ is distributed as a standard normal random variable. Thus, conditional on the history of the process till time $t-1$, denoted as $\mathcal{F}_{t-1}$, the cumulative distribution function (cdf) of y_t is given by

$$F(y_t|\mathcal{F}_{t-1}) = \sum_{j=1}^{d} \Phi\left(\frac{y_t - \phi_j y_{t-1}}{\sigma_{\epsilon,j}}\right) w_j, \quad (13)$$

where $\Phi(.)$ denotes the cdf of the standard normal distribution. Following the notation in section II-A, let $\mathbf{y} = [y_1, \ldots, y_K]^T$ denote a length-K GMM sample obtained using (12) with $\mathbf{Z} = [\mathbf{z}_1^T, \ldots, \mathbf{z}_K^T]^T \in \{0, 1\}^{K \times d}$ as the indicator variable denoting the mixture component to which each y_i belongs, and $\mathbf{\Psi} = [w_1, \phi_1, \phi_2, \sigma_{\epsilon,1}, \sigma_{\epsilon,2}]^T$ as the vector of all unknown parameters. We write the log-likelihood function $L_C(\mathbf{\Psi})$ based on the complete data $y_C = [\mathbf{y}, \mathbf{Z}]$ as

$$\log L_C(\mathbf{\Psi}) = \sum_{i=2}^{K} \Bigg\{ \sum_{j=1}^{d} z_{ij} \log(w_j) - \sum_{j=1}^{d} z_{ij} \log(\sigma_{\epsilon,j}) - \sum_{j=1}^{d} z_{ij} \frac{(y_t - \phi_j y_{t-1})^2}{2\sigma^2_{\epsilon,j}}) \Bigg\}, \quad (14)$$

where $z_{ij} = (\mathbf{z}_i)_j$ is the jth entry of $\mathbf{z}_i$. Again, the EM algorithm can be used to obtain the unknown parameters. Note that the formulation described above appropriately standardizes the AR component to take the dependence across time points into account. This is in contrast to a standard GMM where samples for each mixture component are independently Gaussian distributed. To understand the improvement via averaging, a simple simulation study is performed as described below. The main steps of our simulation study are as follows:

(i) We first simulate (e.g. [22]) a length-K realization $\mathbf{y} = [y_1, \ldots, y_K]^T$ from a $d = 2$-component GMM with $\mathbf{w} = [1/2, 1/2]$, where each component follows an AR(1) model with coefficients $\phi_1 = 0.3$, $\phi_2 = 0.8$, and noise variances $\sigma^2_{\epsilon,1} = 2.25$, $\sigma^2_{\epsilon,2} = 7.84$.

(ii) Given $\mathbf{y}$ and the number of components $d = 2$, the EM algorithm is used to estimate parameters of the GMM. We used Biernacki's search/run/select initialization strategy, first searching for $p = 5$ initial positions using a short EM run with tolerance in eqn. (5) fixed at $\epsilon = 10^{-2}$, each based on random starts initialized using the sample mean and variance of $\mathbf{y}$, see e.g. [22, p. 55]. Next, starting at each of these p initial positions, we ran short EM runs repeatedly for $L = 20$ times with tolerance $\epsilon = 10^{-5}$. Of all the $p \times L = 100$ solutions, the one corresponding to the highest likelihood was chosen as the starting point for the final long EM run for which we fixed $\epsilon = 10^{-10}$.

(iii) Next, we repeat steps (i) and (ii) above, for a fixed number of times say B, to generate time series samples $\mathbf{y}_1, \ldots, \mathbf{y}_B$ and the corresponding EM estimates $\theta_1, \ldots, \theta_B$, for each GMM parameter $\theta \in \mathbf{\Psi}$. A consequence of the non-identifiability of GMMs is the permutation of component labels of the estimated parameters [22]. Consequently, for each parameter, the set of d component parameter estimates are sorted consistently across all B replications, before averaging. In our example, estimated means of the two components are very close to each other (due to the true means fixed to zero), however there is a large difference between the estimated variances (due to $\sigma^2_1 << \sigma^2_2$), which gives us the permutation that allows us to consistently order the estimated means and weights for each replication, before averaging to construct $\hat{\theta}_{\text{A}}(B)$.

(iv) The MSEs of these estimates are computed over R Monte Carlo iterations, i.e. steps (i)-(iii) are repeated R times, subsequently we compute

$$\text{MSE}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^{(r)}-\theta)^2; \text{MSE}(\hat{\theta}_{\text{A}}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_{\text{A}}^{(r)}-\theta)^2,$$

where the superscript $.^{(r)}$ denotes the estimate at the rth, $r = 1, \ldots, R$, Monte Carlo iteration, and dependence of $\hat{\theta}_{\text{A}}$ on B (fixed) is suppressed.

Our study is based on $R = 500$ Monte Carlo iterations. Before a MSE comparison, we draw attention to Fig. 1 which displays the true component variances σ_j^2 (solid line) and a subset of the corresponding EM estimates $\hat{\sigma}_j^{2(r)}$ (dotted) and averaged estimates $\hat{\sigma}_{\text{A},j}^{2(r)}(B)$ (dashed) for a subset of Monte Carlo iterations with lengths $K = 600$ ((a) and (b)) and $K = 1200$ ((c) and (d)), respectively. The high variation in EM estimates of component variances, particularly for $\hat{\sigma}_2^2$ (note the difference in scale on the y-axis) is clearly visible. This is due to our simulation design where components of each Monte Carlo realization are generated from an AR model, imposing dependence structure in time which is completely ignored when fitting a standard GMM to these realizations. As expected, the averaged estimates $\hat{\sigma}_{\text{A},j}^{2(r)}(B)$ have little or no variation and are very closely aligned with the true values in each case. This is reflected in a comparison of the MSE of

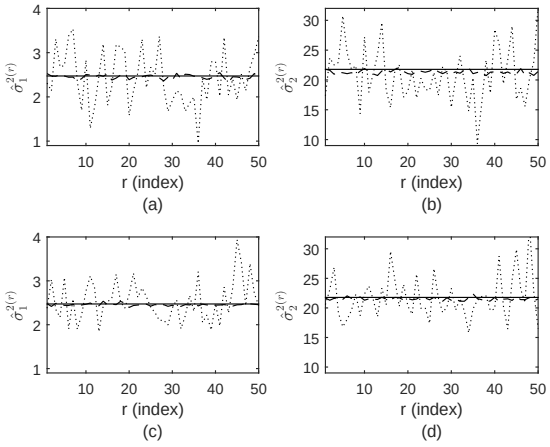

Fig. 1. The true GMM variance parameter values (solid) in comparison with the EM estimates (dotted) and averaged estimates (dashed) using $B = 200$. The subplots display component variance estimates of the form $\hat{\sigma}_j^{2(r)}$ across a subset of Monte Carlo iterations indexed using r (as in the text) for (a) $j = 1$, $K = 600$, (b) $j = 2$, $K = 600$, (c) $j = 1$, $K = 1200$, and (d) $j = 2$, $K = 1200$.

the EM estimates $\hat{\theta}$ with the MSE of the averaged estimates $\hat{\theta}_{\text{A}}(B)$, for example, as shown for the component variances in Table I. This table displays the ratios of MSE, precisely: $\{\text{MSE of } \hat{\sigma}_{\text{A},j}^2\}$ /$\{\text{MSE of } \hat{\sigma}_j^2\}$, for $j = 1, 2$. We see that as the sample size is increased from $K = 600$ to $K = 1200$, the reduction in MSE of the EM estimate of σ_1^2 via averaging approaches 0.0055. This follows from eqns. (7a) and (8) which imply that

$$\frac{\text{MSE}(\hat{\theta}_{\text{A}})}{\text{MSE}(\hat{\theta})} = \frac{\{\text{BIAS}(\hat{\theta}_{\text{A}})\}^2 + \frac{1}{B}\text{VAR}(\hat{\theta})}{\{\text{BIAS}(\hat{\theta})\}^2 + \text{VAR}(\hat{\theta})} \approx 1/B = 0.0050,$$

with $B = 200$, when the bias in $\hat{\theta}$ is close to zero and as $\text{BIAS}(\hat{\theta}_{\text{A}}) = \text{BIAS}(\hat{\theta})$. On the other hand, when the bias is away from zero, the ratio will contain contribution from both the bias and the variance terms and we shall not expect the ratio to be around 0.0050. This is what we observe for σ_1^2 when K is small, and for all K's in the case of σ_2^2. This is because of the small bias in the parameter estimates of the first component in comparison to the second component. This is exactly what we expect as the first component is generated using an AR(1) model with AR coefficient $\phi_1 = 0.3$ imposing significantly weaker dependence in comparison to the relatively stronger dependence due to $\phi_2 = 0.8$ used to generate the second component of the GMM. Of course, in practice when the true model is unknown, bootstrap will be used and relatively larger MSE ratios will be observed. The results in Table I show the benchmark improvement that may be achieved over EM estimates via averaging.

TABLE I
RATIO OF THE MSE OF $\hat{\theta}_{\text{A}}$ TO THE MSE OF $\hat{\theta}$ (EM), FOR $\theta \in \{\sigma_1^2, \sigma_2^2\}$.

K	$\text{MSE}(\hat{\theta}_{\text{A}})/\text{MSE}(\hat{\theta})$	
	σ_1^2	σ_2^2
600	0.0061	0.0181
800	0.0065	0.0160
1000	0.0063	0.0137
1200	0.0055	0.0116

Our experiments show that in the case of a misspecified GMM, EM estimates based on a robust initialization strategy can still be very unstable leading to estimates with high variance. We see that in such cases, the averaging approach can lead to estimates with relatively smaller MSE. For simplicity and convenience, we have simulated this scenario in the time domain. In practice, it shall find applications in the frequency domain, as illustrated via the source separation application in the remaining part of the paper. Our simulation experiments in section VI show that the proposed methodology works for EM estimates of the frequency domain GMM.

IV. SOURCE SEPARATION

Our focus is on exactly determined and underdetermined source separation for speech mixtures using the model-based approaches involving frequency domain GMMs. Model-based methods often achieve underdetermined source separation by relying on the assumption that any two distinct speech sources are disjoint in the T-F domain, formally, known as the W-disjoint orthogonality (WDO) condition [26]. This assumption reduces the source separation problem to identifying the dominant source at each T-F point. Since the WDO condition is only approximately true in a reverberant environment, probabilistic methods using statistical models for a chosen set of cues, [1], [8], [9], are considered more suitable than the binary approach [27].

A common feature of the probabilistic methods is that given an observed speech mixture, the overall likelihood for cue model parameters takes the form of a partially observed fre-

quency domain GMM, which must be estimated. We provide the exact forms of such frequency domain GMMs below.

A. Model-Based EM Separation

Let $\mathbf{x}(n) = [x_1(n), x_2(n)]^T$ denote a two-channel mixture vector, where components $x_1(n)$ and $x_2(n)$ are formed by convolutions of the form (1). Here we focus on the two-channel or the binaural case ($M = 2$) as in [1] and [8], and use their notation with $l(n) \equiv x_1(n)$ and $r(n) \equiv x_2(n)$, so that $\mathbf{x}(n) = [l(n), r(n)]^T$. Suppose that the speech mixture is observed at N time points, then let $\mathbf{l} = [l(1), \ldots, l(N)]^T$, and $\mathbf{r} = [r(1), \ldots, r(N)]^T$ denote components of the observed binaural mixture with $\mathbf{L} = [L(\omega, t)] \in \mathbb{C}^{W \times T}$ and $\mathbf{R} = [R(\omega, t)] \in \mathbb{C}^{W \times T}$ denoting their STFT matrices, respectively. Here ω and t denote the frequency bin and time frame indices, respectively; W denotes the number of frequency bins and T denotes the number of time frames.

The method proposed in [9] (not limited to two-channels) performs classification of the T-F points into one of the I classes (or I sources) based on the mixing vector (cue) in the T-F domain, i.e. $\mathbf{X}(\omega, t) = [L(\omega, t), R(\omega, t)]^T$ for each (ω, t) pair in a frequency bin-wise manner. This is done by employing a complex Gaussian density function for $\mathbf{X}(\omega, t)$ for each ω, i.e. $p(\mathbf{X}(\omega, t)|\mathbf{a}_i(\omega), \sigma_i^2(\omega)) \sim \mathcal{N}^{\mathbb{C}}(\mathbf{a}_i(\omega), \sigma_i^2(\omega))$ where $\mathbf{a}_i(\omega)$ is the mean vector (of the left and right speech mixture components) with $||\mathbf{a}_i(\omega)|| = 1$ and $\sigma_i^2(\omega)$ denotes the common variance. Then the density of $\mathbf{X}(\omega, t) \equiv \mathbf{X}(t)$ (ω fixed) is given by

$$p(\mathbf{X}(t)|\boldsymbol{\theta}) = \sum_{i=1}^{I} \beta_i(\omega) p(\mathbf{X}(\omega, t)|\mathbf{a}_i(\omega), \sigma_i^2(\omega)), \qquad (15)$$

where $\boldsymbol{\theta} \equiv (\mathbf{a}_1(\omega), \sigma_1(\omega), \beta_1(\omega), \ldots, \mathbf{a}_I(\omega), \sigma_I(\omega), \beta_I(\omega))$ is the parameter set and $\beta_i(\omega)$ is the fraction of T-F points that belong to class $i \in \{1, \ldots, I\}$, so that $0 < \beta_i(\omega) < 1$, and $\sum_{i=1}^{I} \beta_i(\omega) = 1$. Clearly, the above equation represents a complex-valued, I-component GMM with weights $\beta_i(\omega)$, $M = 2$ dimensional mean vector $\mathbf{a}_i(\omega)$ and variance $\sigma_i^2(\omega)$. The E-step computes $p(C_i|\mathbf{X}(\omega, t), \boldsymbol{\theta})$ where C_i denotes the ith class, for each $i = 1, \ldots, I$ and ω. Then a binary T-F mask is derived by identifying the dominant source based on a comparison of probabilities $p(C_i|\mathbf{X}(\omega, t), \boldsymbol{\theta})$ for source i with probabilities $p(C_j|\mathbf{X}(\omega, t), \boldsymbol{\theta})$ for source j, $i \neq j$, at each (ω, t).

The method proposed in [1] works with the interaural spectrogram which is given by the ratio of $L(\omega, t)$ to $R(\omega, t)$, and can be expressed as

$$\frac{L(\omega, t)}{R(\omega, t)} = 10^{\alpha(\omega,t)/20} \mathrm{e}^{\mathrm{i}\phi(\omega,t)}, \qquad (16)$$

in terms of the interaural level difference (ILD) denoted by $\alpha(\omega, t)$, and the interaural phase difference (IPD) $\phi(\omega, t)$ and where i denotes the unit imaginary number. Gaussian distributions are found appropriate for both $\alpha(\omega, t)$ and $\phi(\omega, t)$ and the corresponding densities are chosen to be of the form $p(\alpha(\omega, t)|\mu_i(\omega), \eta_i^2(\omega)) \sim \mathcal{N}(\alpha(\omega, t)|\mu_i(\omega), \eta_i^2(\omega))$ and $p(\phi(\omega, t)|\xi_{i\tau}(\omega), \gamma_{i\tau}^2(\omega)) \sim \mathcal{N}(\phi(\omega, t)|\xi_{i\tau}(\omega), \gamma_{i\tau}^2(\omega))$. Then, assuming that T-F points from the same source and at the same delay τ are independently distributed, the joint density function of $\alpha(\omega, t)$ and $\phi(\omega, t)$ is expressed as

$$\begin{aligned} p(\phi(\omega, t), \alpha(\omega, t)|\Theta_{i\tau}) &= p(\phi(\omega, t)|\xi_{i\tau}(\omega), \gamma_{i\tau}^2(\omega)) \\ &\quad .p(\alpha(\omega, t)|\mu_i(\omega), \eta_i^2(\omega)) \\ &\quad .p(i, \tau), \end{aligned} \qquad (17)$$

where $p(i, \tau) \equiv \psi_{i\tau}$ is the joint probability of any T-F point being in source i at delay $\tau \in \mathcal{T}$, where $\mathcal{T}$ denotes the set of admissible values for delay τ. Let $\Theta = [\Theta_{i\tau}; i = 1, \ldots, I; \tau \in \mathcal{T}]$ where $\Theta_{i\tau} = \{\xi_{i\tau}(\omega), \gamma_{i\tau}(\omega), \mu_i(\omega), \eta_i(\omega), \psi_{i\tau}\}$ denote the complete parameter set. Then, the total probability density is given by

$$\begin{aligned} p(\phi(\omega, t), \alpha(\omega, t)|\Theta) &= \sum_{i,\tau} p(\phi(\omega, t), \alpha(\omega, t)|\Theta_{i\tau}) \\ &= \sum_{i,\tau} \psi_{i\tau} \{p(\alpha(\omega, t)|\mu_i(\omega), \eta_i^2(\omega)) \\ &\quad .p(\phi(\omega, t)|\xi_{i\tau}(\omega), \gamma_{i\tau}^2(\omega))\}, \end{aligned} \qquad (18)$$

which clearly represents a real-valued GMM with one Gaussian per (i, τ) combination and mixing weights $\psi_{i\tau}$, given the assumed Gaussian distributions for the ILD and IPD. Again, the EM algorithm is implemented to estimate the unknown parameters in Θ. Here initializations for parameter estimation via the EM algorithm are chosen informatively as discussed in [1], with the main objective of achieving the best possible local maximizer and in order to avoid spurious estimates.

The iterative E-step computes the conditional probability of the spectrogram point (ω, t) coming from source i and delay τ, given the observed interaural cues $\phi(\omega, t)$ and $\alpha(\omega, t)$ and the current Θ, i.e.,

$$p((\omega, t) \in (i, \tau)|\phi(\omega, t), \alpha(\omega, t), \Theta) \equiv \nu_{i\tau}(\omega, t), \qquad (19)$$

using which MLEs of the unknown parameters are calculated in the M-step, [1, eqn. (18)]. Repeated iterations of the E- and M-steps are performed to obtain final estimates of the parameters, and subsequently $\nu_{i\tau}(\omega, t)$ in the final E-step is computed using (19). Clearly, summing $\nu_{i\tau}(\omega, t)$ over all possible delays τ gives the probability of the ith source being dominant at the time-frequency point (ω, t). Therefore, for each source i, a probabilistic T-F mask denoted as $\mathbf{M}_i = [M_i(\omega, t)] \in [0, 1]^{W \times T}$ is computed as,

$$M_i(\omega, t) = \sum_{\tau} \nu_{i\tau}(\omega, t), \qquad (20)$$

which allows estimation of the I source vectors of interest from the observed binaural mixtures.

Recently, [7], [8] combined the mixing vector model of [9] with the ILD and IPD models of [1] to perform source separation based on the combined set of parameters denoted as $\Gamma = [\Gamma_{i\tau}, i = 1, \ldots, I; \tau \in \mathcal{T}]$ with $\Gamma_{i\tau} = \{\mathbf{a}_i(\omega), \sigma_i(\omega), \xi_{i\tau}(\omega), \gamma_{i\tau}(\omega), \mu_i(\omega), \eta_i(\omega), \psi_{i\tau}\}$. Here the total probability density for a given (ω, t) i.e. $\sum_{i,\tau} p(\phi(\omega, t), \alpha(\omega, t), \mathbf{X}(\omega, t)|\Gamma_{i\tau})$ is a GMM of the form

$$\begin{aligned} &\sum_{i,\tau} \psi_{i\tau} \{p(\alpha(\omega, t)|\mu_i(\omega), \eta_i^2(\omega)).p(\phi(\omega, t)|\xi_{i\tau}(\omega), \gamma_{i\tau}^2(\omega)) \\ &\qquad .p(\mathbf{X}(\omega, t)|\mathbf{a}_i(\omega), \sigma_i^2(\omega))\}. \end{aligned} \qquad (21)$$

The initialization strategy from [1] is easily adapted to deal with the additional mixing vector cue as discussed in [8, sec. V]. Subsequently, the EM algorithm is used to derive a probabilistic T-F mask. It is shown that the probabilistic mask obtained as a result of this joint model leads to improvements in separation performance measured by SDR over the methods of [1] and [9]. We use this joint model of [8] to study improvement via the proposed bootstrap averaging approach. Here, our main objective is to show how the proposed bootstrap averaging technique can be implemented to improve the EM estimates of frequency domain GMM appearing in [8], and to improve the source separation performance for reverberant mixtures.

V. Bootstrap Averaging for Source Separation

To immediately illustrate the need for improvement in EM estimates of the frequency domain GMM arising in the source separation algorithms described above, a comparison of the EM estimates of the ILD mean parameter $\hat{\mu}(\omega)$ computed using the algorithm of [8] with the ground-truth values $\mu(\omega)$ is provided in Fig. 2. It displays the ground-truth ILD mean (black) and its EM estimate (grey) for two-source binaural mixtures formed by convolving two randomly chosen speech signals from the TIMIT data set with impulse responses measured by Hummersone [28] under anechoic (*Room A*) and reverberant conditions (*Room D* with $RT_{60} = 0.89$s), with the two sources placed at 0° and 30° (further details are provided in section VI). The top row of the plot corresponds to (a) $\mu_1(\omega)$ – the ILD mean for s_1, and (b) $\mu_2(\omega)$ – the ILD mean for s_2, for the anechoic mixture; subplots (c) and (d) similarly correspond to the ILD mean for s_1 and s_2 in the reverberant room. Clearly, the estimated ILD mean follows the ground truth ILD very closely for the anechoic case, especially for source 2 (Fig. 2(b)). On the other hand, we see large variation in the EM estimates of the ILD mean parameter, [8] in the reverberant case. In this section, we describe the bootstrap averaging algorithm to yield improved estimates of the frequency domain GMM parameters (in Γ) employed in the framework of [8].

A convenient way to bootstrap the frequency-dependent GMM parameter estimates is to bootstrap the observed mixture vector

$$\underline{\mathbf{x}} = \begin{bmatrix}\mathbf{x}(1) & \dots & \mathbf{x}(N)\end{bmatrix} = \begin{bmatrix} l(1) & \dots & l(N) \\ r(1) & \dots & r(N)\end{bmatrix}$$

to obtain time domain bootstrap samples, denoted as $\underline{\mathbf{x}}_1^*, \dots, \underline{\mathbf{x}}_B^*$ from which bootstrap estimates of the cue model parameters can be obtained directly using the algorithm of [8]. Since model-based source separation relies on the interaural spectrogram derived from the speech mixture vector, it is important to use a bootstrap procedure that appropriately mimics the frequency domain dependence in the given sample. Bootstrap samples of the mixture vector are obtained using the circulant embedding based approach of [25]. The basic idea in [25] is to generate portions of realizations with spectral density given by an estimated spectral density derived from the observed vector-valued time series via circulant embedding. The procedure of [25] is easily implemented using a FFT which makes it computationally efficient and hence very attractive for our application where the length-N of the observed speech mixtures is usually very large.

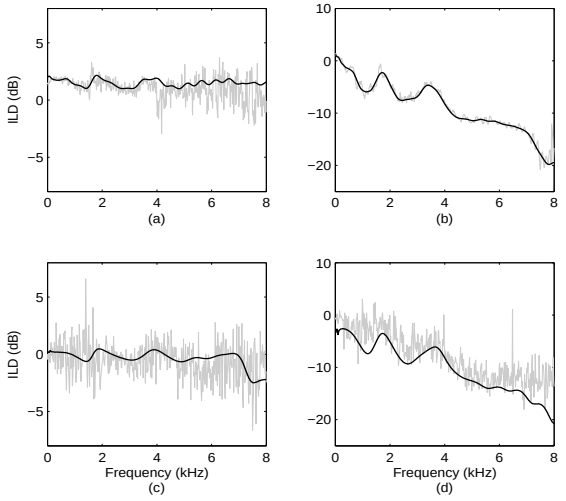

Fig. 2. The mean ILD estimates $\hat{\mu}_i(\omega)$ (solid grey) from the joint model of [8] vs. frequency (kHz) in an anechoic environment for (a) $i = 1$ and (b) $i = 2$; and a reverberant (Room D, $RT_{60} = 0.89$s – see sec. VI for further details) room environment for (c) $i = 1$ and (d) $i = 2$. The solid black line in each subplot shows the ground-truth ILD mean (dB) $\mu_i(\omega)$. Sources s_1 and s_2 chosen from the TIMIT data set were placed at $\varphi = 0^\circ$, and $\varphi = 30^\circ$, respectively.

Consider a bivariate discrete time second-order stationary process $\mathbf{V}_t = [X_t, Y_t]^T$, where $t \in \mathbb{Z}$ denotes the time index. Without loss of generality, assume that each component process has a mean of zero. Given a length-N realization $\mathbf{V}_1, \dots, \mathbf{V}_N$ from the vector process $\mathbf{V}_t$, the following bootstrap algorithm [25], allows us to generate bootstrap time series samples.

Bootstrap algorithm:

1) Choose the embedding size $m_1 > 2(N-1)$ such that $m_1 = 2^g$ for some $g \in \mathbb{Z}^+$. Estimate the spectral matrix $\hat{\mathbf{S}}_{\mathbf{V}}(f_l)$, $f_l = l/(m_1\Delta)$, $l = 0, 1, \dots, m_1 - 1$, with Δ denoting the sampling interval, using one of the recommended spectral estimation methods [25], such as multitaper, Welch's Overlapped Segment Averaging (WOSA) etc.
2) Set $\boldsymbol{\lambda}_l = \hat{\mathbf{S}}_{\mathbf{V}}^T(f_l)/\Delta, l = 0, \dots, m_1 - 1$. For each $l = 0, \dots, m_1 - 1$, determine the 2×2 unitary matrix $\mathbf{U}_l$ and the diagonal matrix $\mathbf{D}_l$ such that $\boldsymbol{\lambda}_l = \mathbf{U}_l \mathbf{D}_l \mathbf{U}_l^H$, where $.^H$ denotes conjugate transpose.
3) Simulate two real bivariate independent standard normal vectors $\mathcal{Z}_l^{(\alpha)} \sim N(\mathbf{0}, \mathbf{I}_2); \alpha = 1, 2$, and set $\mathbf{C}_l = \mathbf{U}_l \mathbf{D}_l^{1/2} (\mathcal{Z}_l^{(1)} + \mathrm{i}\mathcal{Z}_l^{(2)})$.
4) Define $\tilde{\mathbf{V}}_j = m_1^{-1/2} \sum_{l=0}^{m_1-1} \mathbf{C}_l \mathrm{e}^{-\mathrm{i}2\pi lj/m_1}, j = 0, \dots, m_1 - 1$, which can be computed easily via an FFT. Then for each j, $\tilde{\mathbf{V}}_j$ is a complex-valued bivariate vector. $\mathrm{Re}\{\tilde{\mathbf{V}}_n\}, n = 0, 1, \dots, N-1$ and $\mathrm{Im}\{\tilde{\mathbf{V}}_n\}, n = 0, 1, \dots, N-1$ are two independent length-N bootstrap replications.

Remark: The consistency of the spectral estimators allowed for this algorithm [29, p. 785] guarantees that asymptotically these

bootstrap samples have the specified second-order structure. This has also been verified empirically, [24].

Note that most of the existing bootstrap procedures are only applicable to second-order stationary data. Since short segments of speech ($30-50$ms) are considered to be second-order stationary [30], we shall apply the bootstrap procedure to short segments of the observed speech mixture. The algorithm for bootstrap-based source separation under the model-based framework of [8] is outlined below.

(i) Divide $\underline{\mathbf{x}} \equiv [\underline{\mathbf{x}}(1), \ldots, \underline{\mathbf{x}}(N)]$ into adjacent pseudo second-order stationary blocks of length-$\tilde{N}$. The jth block is given by

$$\underline{\mathbf{z}}_j = \underline{\mathbf{x}}(2-j+(j-1)\tilde{N} : 1+j(\tilde{N}-1));$$

$j = 1, \ldots, N_b$, where N_b denotes the number of blocks required to cover the full length-N of $\underline{\mathbf{x}}$. For each j, $\underline{\mathbf{z}}_j$ is a matrix of size $2 \times \tilde{N}$.

(ii) For each block $\underline{\mathbf{z}}_j$, implement the bootstrap algorithm of [25] (given above), to generate B bootstrap samples, each of length $\tilde{N}-1$ (to avoid generating the end point twice due to adjacent blocks). So for the jth block, $j = 1, \ldots, N_b$, obtain $\underline{\mathbf{z}}^*_{j,b}, b = 1, \ldots, B$. For each block, the algorithm of [25] is implemented with the multitaper spectral estimation technique, which for a given length-$\tilde{N}$ bivariate time series $\underline{\mathbf{z}}$ is a 2×2 matrix given by:

$$\hat{\mathbf{S}}_{\underline{\mathbf{z}}}(\omega) = \frac{\Delta}{P} \sum_{p=1}^{P} \left\{ \left| \sum_{n=1}^{\tilde{N}} u_{n,p} \underline{\mathbf{z}}(n) \mathrm{e}^{-\mathrm{i}2\pi\omega n\Delta} \right|^2 \right\}, \quad (22)$$

where $|\omega| \leq 1/(2\Delta)$, and $\{u_{n,p}\}_{n=1}^{\tilde{N}}$ is the pth data taper. Tapering prevents the spectra from the problem of leakage and the application of P orthogonal data tapers as in the multitaper approach leads to a consistent spectral estimate [31]. Then the bth full length-N bootstrap sample for $\underline{\mathbf{x}}$ is given by

$$\underline{\mathbf{x}}^*_b = \left[\underline{\mathbf{z}}^*_{1,b} \quad \cdots \quad \underline{\mathbf{z}}^*_{N_b,b} \right]_{2 \times N},$$

and $\underline{\mathbf{x}}^*_1, \ldots, \underline{\mathbf{x}}^*_B$ are the B bootstrap samples.

(iii) The source separation algorithm of [8] is applied to each of the B bootstrap samples $\underline{\mathbf{x}}^*_1, \ldots, \underline{\mathbf{x}}^*_B$, individually. This leads to B bootstrap estimates for each parameter in $\Gamma_{i\tau}$. For each i and τ, let $\Gamma^*_{i\tau,1}, \ldots, \Gamma^*_{i\tau,B}$ denote the bootstrap parameter set where $\Gamma^*_{i\tau,b} = \{\mathbf{a}^*_{i,b}(\omega), \sigma^*_{i,b}(\omega), \xi^*_{i\tau,b}(\omega), \gamma^*_{i\tau,b}(\omega), \mu^*_{i,b}(\omega), \eta^*_{i,b}(\omega), \psi^*_{i\tau,b}\}$ contains model parameter estimates derived from the bth bootstrap sample $\underline{\mathbf{x}}^*_b$. This is used to construct the bootstrap averaged estimates for each frequency dependent parameter.

(iv) The algorithm of [8] allows us to compute bootstrap T-F masks $\mathbf{M}^*_{i,b} = [M^*_{i,b}(\omega,t)] \in [0,1]^{W \times T}$ corresponding to the bootstrap estimates of GMM parameters in $\Gamma^*_{i\tau,b}$ using each of the $b = 1, \ldots, B$ bootstrap replications. The averaged bootstrap T-F mask given by $\mathbf{M}^*_{i,\mathrm{A}} = [M^*_{i,\mathrm{A}}(\omega,t)]$ where

$$M^*_{i,\mathrm{A}}(\omega,t) = \frac{M^*_{i,1}(\omega,t) + \ldots + M^*_{i,B}(\omega,t)}{B}, \quad (23)$$

is used for recovering the source vectors $\mathbf{s}_1, \ldots, \mathbf{s}_I$ from the observed speech mixture vectors in $\underline{\mathbf{x}}$.

Note that the time-frequency mask given by (20) is a by-product of the EM algorithm, computed using the output of the E-step (19) which gives the probability of a spectrogram point (ω,t) coming from source i and delay τ, conditional on the interaural cues $\alpha(\omega,t)$ and $\phi(\omega,t)$ (estimated using the spectrogram of the observed speech mixture), in addition to the parameter estimates from the final M-step of the EM algorithm. For each bootstrap replication, the E-step allows us to compute this probability or T-F masks, conditional on the interaural cues $\alpha^*(\omega,t)$ and $\phi^*(\omega,t)$, estimated from the spectrogram of the bootstrap speech mixture. From [24, Chapter 7], we know that the bootstrap procedure leads to samples which replicate the true frequency domain statistics, i.e. the spectra of bootstrap samples, on average, mimics the theoretical spectra of the process generating the observed sample. Thus, if the bootstrap averaged GMM parameter estimates lead to a smaller MSE in comparison with the original EM estimates, each bootstrap T-F mask is a reasonable estimate (of the true T-F mask). The average of the bootstrap T-F masks provides a simple way to construct an overall T-F mask based on the bootstrap data. We verify the performance of the bootstrap averaged estimates and the bootstrap averaged T-F mask for the task of source separation in our experiments below.

VI. Experiments and Results

A. Set-up

We present the experimental set-up used to test the proposed bootstrap averaging technique for source separation as described above using speech mixtures formed with real room-recorded impulse responses. We use the TIMIT data set from which 15 utterances are randomly selected to form convolutive mixtures using binaural room impulse responses (BRIRs), [8]. The BRIRs were captured by Hummersone [28] using a Head and Torso Simulator (HATS). The HATS and the sources were placed at a height of 2.8m in the room and were separated by a distance of 1.5m. The target source is placed exactly in front of HATS, i.e. at zero degree relative to HATS and the two interfering sources are positioned symmetrically on the left and right hand side of the target source in an arc at azimuth denoted by φ (in degrees). The BRIRs were measured in 5 different rooms corresponding to five different reverberation levels given by RT_{60} and azimuths ranging from -90° to 90° at 5° intervals. The sampling frequency denoted as f_s is 16 kHz and the sampling interval is $\Delta = 1$s. From the chosen set of 15 utterances, we combined two (three) speech signals (about 3s, shortened to 2.5s for consistency) with BRIRs from *Room D* which corresponds to the highest reverberation time of 0.89s within the recorded data set to construct two sets of mixtures: (i) 15 two-source mixtures, and (ii) 15 three-source mixtures, for each azimuth $\varphi = 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ$.

To implement the algorithm described in section V, we divide $\underline{\mathbf{x}}$ into pseudo stationary blocks of length 30ms which correspond to $\tilde{N} = f_s \times 0.003 = 16000 \times 0.003 = 480$ samples and $N_b = 84$ adjacent blocks. Multitaper spectral estimates for

the bivariate time series in each block are computed as given in (22). We employ $P = 8$ sine tapers $\{u_{n,p}\}$, where

$$u_{n,p} = \frac{2}{(\tilde{N}+1)^{1/2}} \sin \frac{(p+1)\pi n}{\tilde{N}+1}, n = 1, \ldots, \tilde{N}, \quad (24)$$

leading to a bandwidth of $W_{\tilde{N}} = (P+1)/2(\tilde{N}+1) = 0.0094$ [31]. Then following the steps in section V, the simulation procedure of [25] is applied to the time series in each block with $B = 500$ to obtain B bootstrap samples $\underline{\mathbf{x}}_1^*, \ldots, \underline{\mathbf{x}}_B^*$ of the full length-N speech mixture. Next, parameter sets $\Gamma_{i\tau,1}^*, \ldots, \Gamma_{i\tau,B}^*$ containing bootstrap estimates and time-frequency masks $\mathbf{M}_{i,1}^*, \ldots, \mathbf{M}_{i,B}^*$ for each source index $i = 1, \ldots, I$, and admissible τ are derived using the source separation algorithm of [8]. We conduct all our experiments with the additional *garbage source* which aims to account for spectrogram points where reverberation (rather than one of the I sources) dominates, [1].

B. Comparison of Cue Parameter Estimates

Here we focus on estimates of the frequency domain GMM parameters, i.e. elements of Γ. We compare estimates obtained using the algorithm of [8], and the bootstrap averaged estimates, with their ground-truth values. Here ground-truth refers to the parameter values that would be obtained if each source was observed in isolation. We focus on the ILD, IPD, and the mixing vector cue mean parameters. The ground-truth for these parameters is obtained as described below. The ground-truth ILD mean is computed from the isolated one-source direct-path mixture, i.e. μ_i is obtained by convolving only the ith source with the direct-path impulse response denoted as $[\tilde{h}_{il}(n), \tilde{h}_{ir}(n)]^T$ to form the convolutive mixture vector $\tilde{\mathbf{x}}^{(i)}(n) = [\tilde{l}^i(n), \tilde{r}^i(n)]^T$, where $\tilde{.}$ denotes direct-path and superscript $.^{(i)}$ indicates that the mixture only involves the ith source. Then from the definition of ILD, it follows that

$$\mu_i(\omega) = \frac{1}{T} \sum_{t=1}^{T} 20 \log_{10} \frac{|\tilde{L}^{(i)}(\omega, t)|}{|\tilde{R}^{(i)}(\omega, t)|}, \quad (25)$$

where $[\tilde{L}^{(i)}(\omega, t)]_{W \times T}$, and $[\tilde{R}^{(i)}(\omega, t)]_{W \times T}$ denote the STFTs of direct path mixture vectors $[\tilde{l}^i(1), \ldots, \tilde{l}^i(N)]$ and $[\tilde{r}^i(1), \ldots, \tilde{r}^i(N)]$, respectively. Alternatively, from [1], the ground-truth ILD mean may be computed directly from the direct-path impulse response as

$$\mu_i(\omega) = 20 \log_{10} \left(\frac{|\tilde{H}_{il}(\omega)|}{|\tilde{H}_{ir}(\omega)|} \right), \quad (26)$$

where $\tilde{H}_{il} = \mathcal{F}\{\tilde{h}_{il}\}$, and similarly, $\tilde{H}_{ir} = \mathcal{F}\{\tilde{h}_{ir}\}$, $\mathcal{F}\{.\}$ denoting the Fourier transform.

Similarly, the ground-truth IPD residual mean for the ith source is given by, [1]:

$$\xi_{i\tau}(\omega) = \arg \left(\mathrm{e}^{-\mathrm{i}\tilde{\phi}(\omega,t)} \mathrm{e}^{-\mathrm{i}\omega\tau(\omega)} \right), \quad (27)$$

where $\tilde{\phi}(\omega, t) = \arg(\tilde{H}_{il}(\omega)/\tilde{H}_{ir}(\omega))$, and $\tau(\omega) = \tau_l - \tau_r + \arg(\bar{H}_{il}(\omega)/\bar{H}_{ir}(\omega))$, with $\bar{H}_{il}(\omega) = \mathcal{F}\{\bar{h}_{il}(n)\}$, $\bar{H}_{ir}(\omega) = \mathcal{F}\{\bar{h}_{ir}(n)\}$ and $\bar{h}_{il}(n)$, $\bar{h}_{ir}(n)$ denoting the impulse response truncated at the length of the analysis window. Also $\arg(.)$ denotes the argument, taking values in the interval $(-\pi, \pi]$.

Since $\mathbf{a}_i(\omega)$ denotes the mean of the reverberant mixture vector in the T-F domain, the ground-truth mixing vector mean $\mathbf{a}_i(\omega)$ is computed as described in [9] from the isolated mixture $\check{\mathbf{x}}^{(i)}(n) = [\check{l}^i(n), \check{r}^i(n)]^T$, where $\check{.}$ indicates that the ith source is convolved with the full impulse response. The corresponding EM estimates and bootstrap averaged estimates for each GMM parameter are computed using the algorithm of [8] and bootstrap as described above in section V and VI-A.

Due to the non-identifiability of GMMs, it is important to learn the permutation that allows consistent averaging across bootstrap estimates for each parameter. We described in section III-A, how this may be achieved using the fact that the component variances are well-separated. Different ways to solve the permutation problem due to EM estimation have been studied in the blind source separation literature. In the case of a frequency specific GMM, as in the algorithms of [9], [1], and [8], dealing with the permutation problem is crucial to be able to group together components corresponding to the same source estimated at each frequency. Traditionally, correlation coefficients of *amplitude envelopes* which represent sound source activity, are used to identify the permutation, for example, [32], [33], however, more recently efficient approaches as in [9], have been discussed and are commonly employed. Other techniques solving the permutation problem for frequency domain source separation are discussed in [34] and [35]. Thus, applications employing the EM algorithm for frequency domain source separation commonly have a built-in strategy to deal with the permutation problem, allowing us to average the component parameter estimates consistently across bootstrap replications.

Consider the set of two-source mixtures with the target source s_1 at $\varphi = 0^\circ$ and the interference source s_2 at $\varphi = 30^\circ$. For convenience, we label the 15 two-source mixtures as $k' = 1, \ldots, 15$. Fig. 3 shows a comparison of the ILD mean estimates $\hat{\mu}_i(\omega)$ obtained from the model-based method of [8] (solid grey), and the bootstrap-based estimates $\hat{\mu}_{i,\mathrm{A}}(\omega)$ (solid black), with the ground-truth estimates $\mu_i(\omega)$ (dashed black) over the frequency range (kHz) $[0, f_s/2\Delta] = [0, 8]$. The dotted black lines show the ground-truth ILD mean of each source convolved with impulse response truncated to the length of our analysis window. Note that it follows the direct path ILD mean (dashed black) very closely but has a relatively higher variation. This is due to the early echoes in the impulse response truncated at the window length. The subplots in the left column of Fig. 3 correspond to $\mu_1(\omega)$, i.e. the ILD mean parameter for the target source with subplots in consecutive rows corresponding to the first four mixtures, i.e. $k' = 1, 2, 3, 4$, respectively; similarly, the subplots in the right column correspond to $\mu_2(\omega)$, the ILD mean parameter for the interference source for $k' = 1, 2, 3, 4$, respectively. From the figure we see that the bootstrap averaged ILD mean estimates $\hat{\mu}_{i,\mathrm{A}}^*(\omega)$ (solid black) follow the ground-truth $\mu_i(\omega)$ (dashed black) very closely; ILD mean estimates $\hat{\mu}_i(\omega)$ (solid grey) obtained from the joint method of [8] evidently show large deviations from the ground-truth at each frequency. For a clearer comparison, we compare the absolute error in $\hat{\mu}_i(\omega)$ i.e. $|\hat{\mu}_i(\omega) - \mu_i(\omega)|$ with the absolute error in $\hat{\mu}_{i,\mathrm{A}}^*(\omega)$, i.e. $|\hat{\mu}_{i,\mathrm{A}}^*(\omega) - \mu_i(\omega)|$. Fig. 4 displays absolute error in $\hat{\mu}_i(\omega)$

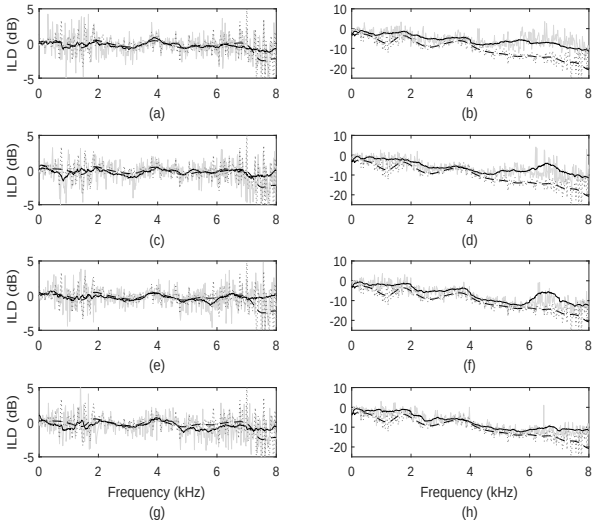

Fig. 3. A comparison of the ground-truth ILD mean (dB) $\mu_i(\omega)$ (dashed black) with the estimate $\hat{\mu}_i(\omega)$ (solid grey) obtained from [8], and the bootstrap averaged estimate $\hat{\mu}^*_{i,\mathrm{A}}(\omega)$ (solid black) vs. frequency (kHz) for (a) $i = 1, k' = 1$, (b) $i = 2, k' = 1$, (c) $i = 1, k' = 2$, (d) $i = 2, k' = 2$, (e) $i = 1, k' = 3$, (f) $i = 2, k' = 3$, (g) $i = 1, k' = 4$, and (h) $i = 2, k' = 4$ with s_1 placed at $\varphi = 0^\circ$, and s_2 placed at $\varphi = 30^\circ$. The ground-truth ILD mean of each source convolved with impulse response truncated to the window length are shown in dotted black.

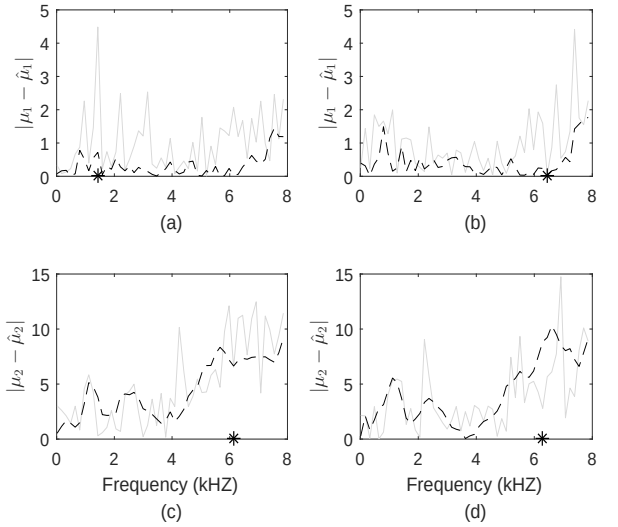

Fig. 4. A comparison of the absolute errors in the ILD mean (dB) estimate $\hat{\mu}_i(\omega)$ from [8] (solid grey), and the bootstrap averaged estimate $\hat{\mu}^*_{i,\mathrm{A}}(\omega)$ (dashed black) vs. frequency (kHz), for (a) $i = 1, k' = 1$, (b) $i = 1, k' = 2$, (c) $i = 2, k' = 1$, and (d) $i = 2, k' = 2$. The asterisk $*$ denotes the frequency of interest in each case.

(solid grey) and $\hat{\mu}^*_{i,\mathrm{A}}(\omega)$ (dashed black) for source and mixture combinations corresponding to the two sources and the first two mixtures, i.e. (a) $i = 1$, $k' = 1$, (b) $i = 1$, $k' = 2$, (c) $i = 2$, $k' = 1$ and (d) $i = 2$, $k' = 2$. For clarity, we have only plotted absolute errors for a set of equally spaced frequencies. From Fig. 4(a)-(b) it is clear that $\hat{\mu}^*_{1,\mathrm{A}}(\omega)$ outperforms $\hat{\mu}_1(\omega)$ at all frequencies, however, for $i = 2$, we observe frequencies where $\hat{\mu}_2(\omega)$ has a smaller error as compared to $\hat{\mu}^*_{2,\mathrm{A}}(\omega)$. The

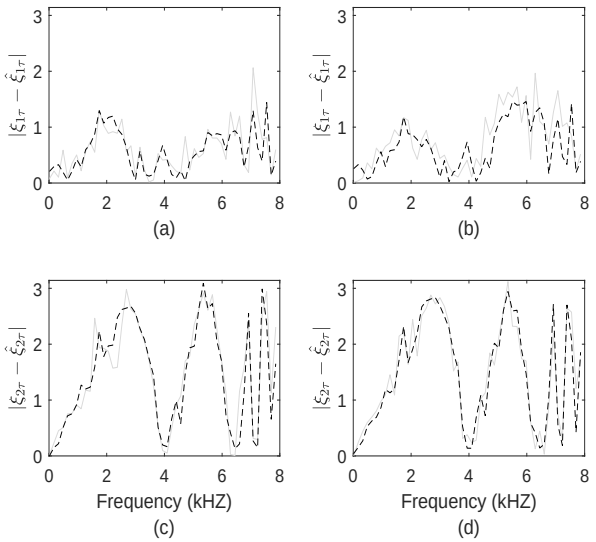

Fig. 5. A comparison of the absolute errors in the IPD mean estimate $\hat{\xi}_{i\tau}(\omega)$ from [8] (solid grey), and the bootstrap averaged estimate $\hat{\xi}^*_{\mathrm{A},i\tau}(\omega)$ (dashed black) vs. frequency (kHz) for (a) $i = 1, k' = 1$, (b) $i = 1, k' = 3$, (c) $i = 2, k' = 1$,, (d) $i = 2, k' = 3$ at the ground truth delay $\tau = 4$.

frequencies marked with $*$ depict four different scenarios and are discussed in the next subsection. Similarly, Fig. 5 displays

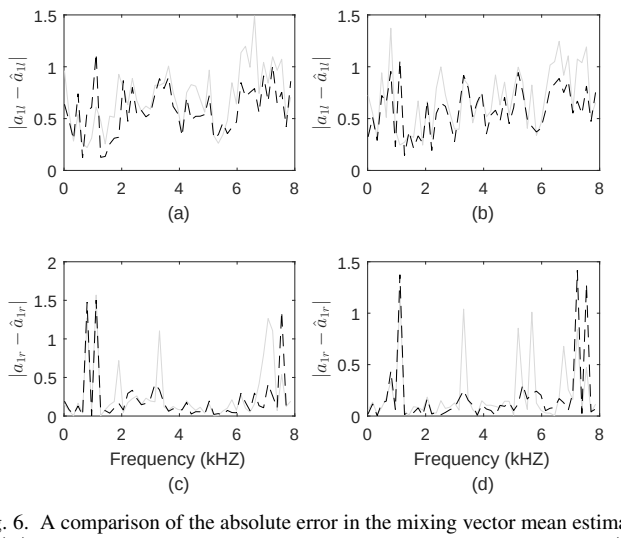

Fig. 6. A comparison of the absolute error in the mixing vector mean estimate $\hat{\mathbf{a}}_1(\omega)$ (solid grey) from [8], and the bootstrap averaged estimate $\hat{\mathbf{a}}^*_{1,\mathrm{A}}(\omega)$ (dashed black) vs. frequency (kHz) for the left component $a_{l1}(\omega)$ of mixtures (a) $k' = 2$, (b) $k' = 4$, and for the right component $a_{r1}(\omega)$ of mixtures (c) $k' = 2$, and (d) $k' = 4$.

absolute error in the IPD mean estimates $\hat{\xi}_{i\tau}(\omega)$ (solid grey) and $\hat{\xi}^*_{i\tau,\mathrm{A}}(\omega)$ (dashed black) for (a) $i = 1$, $k' = 1$, (b) $i = 1$, $k' = 3$, and (c) $i = 2$, $k' = 1$, and (d) $i = 2$, $k' = 3$, at the ground truth delay for a set of equally spaced frequencies. For the target source, the bootstrap averaged IPD mean estimates, seem to have a relatively smaller error. On the other hand, absolute errors for the interference source corresponding to the two set of estimates follow each other very closely. Recall that the mixing vector means are complex-valued. Absolute errors in the components of the mixing vector mean estimates

$\hat{\mathbf{a}}_i(\omega)$ and $\hat{\mathbf{a}}^*_{1,\mathrm{A}}(\omega)$ for the left components of mixtures (a) $k'=2$, (b) $k'=4$; and the right components of mixtures (c) $k'=2$, and (d) $k'=4$ are shown in Fig. 6. Overall, we see relatively smaller errors using the bootstrap based approach. A further analysis is provided in section VII.

Remark: Note that firstly, (i) the bootstrap procedure of [25] allows us to replicate the frequency domain structure of the underlying process generating the observed speech mixture, and secondly, (ii) EM estimates of model parameters from these bootstrap samples are based on informative initializations (using the ILD prior for the ILD parameters and the PHAT histogram for the component weights, [8]), hence, ensuring that EM estimates of parameters using the bootstrapped mixtures are consistent with the true parameter values. This is also evident from Fig. 3 displaying the bootstrap averaged ILD mean estimate which is clearly consistent with the ground truth ILD mean over the entire frequency range.

VII. Further Analysis and Results

To see what exactly differentiates a substantial improvement to none, we consider four possible scenarios using the ILD mean parameter $\mu_i(\omega)$: (i) when the bootstrap averaged estimate $\hat{\mu}^*_{i,\mathrm{A}}(\omega)$ significantly outperforms the estimate $\hat{\mu}_i(\omega)$ from [8], (ii) when both the bootstrap averaged estimate as well as the estimate from [8] have zero absolute error, (iii) when both the estimates have a significant error, and (iv) when the bootstrap averaged estimate has a larger absolute error. The frequencies marked with an asterisk $*$ in Fig. 4 (a)-(d) exactly depict these four scenarios (in order). Let $\omega_a, \omega_b, \omega_c,$ and ω_d denote the frequencies marked in Fig. 4(a)-(d), respectively. We study bootstrap histograms for the ILD mean estimate for each of the four chosen frequencies, for example, bootstrap ILD mean estimates $\hat{\mu}^*_{1,1}(\omega_a), \ldots, \hat{\mu}^*_{1,B}(\omega_a)$ (with $i=1$) derived from mixture $k'=1$ are used to obtain the bootstrap histogram corresponding to frequency ω_a marked in Fig. 4(a), and so on. These histograms are shown in Fig. 7. From Fig. 7(a) we observe that the EM estimate from [8] (square) is at the right extreme of the distribution, with the ground-truth (diamond) very close to the mean of the histogram or the bootstrap averaged estimate (circle). Fig. 7(b) shows that for ω_b, the original estimate as well as the bootstrap-averaged estimate coincide with the ground-truth. Obviously, in this case the estimate from [8] has zero error and bootstrap averaging is not required. However, the fact that the bootstrap estimate also coincides with the ground-truth indicates that the bootstrap methodology has performed impressively well for this frequency. Clearly, the bootstrap histograms of the ILD mean estimate for ω_c and ω_d in Fig. 7(c) and (d), unlike histograms in (a) and (b) do not appear to be normal with the ground-truth located at an extreme of the histogram in each case. Now from (11a) we know that under a suitable bootstrap technique (for $\hat{\theta}$), if the EM estimator $\hat{\theta}$ is unbiased, then the mean of the bootstrap estimates (circle) should also coincide with the true value of θ (diamond). Therefore, for frequencies ω_c and ω_d either the EM estimator $\hat{\mu}_2(\omega)$ has a significant non-zero bias or the bootstrap for $\underline{\mathbf{x}}$ does not lead to ILD mean estimates which accurately represent their underlying second-order statistics. The significant bias in the ILD mean estimates towards 0dB at some frequencies has been independently noted in the literature, e.g., [36] and [37], suggesting that our methodology does not lead to any surprising results.

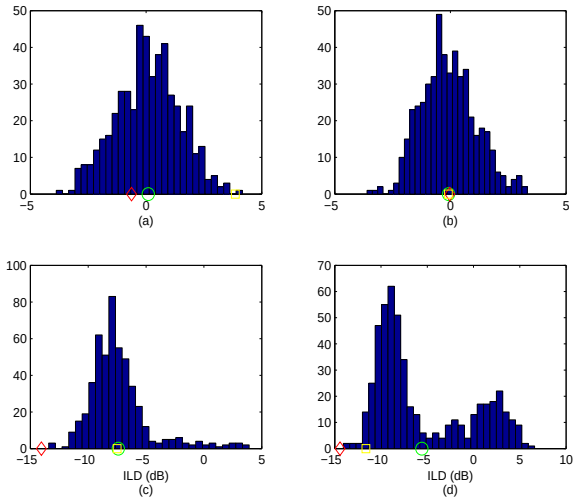

Fig. 7. Bootstrap histograms for the ILD mean estimate (dB) $\hat{\mu}_i(\omega)$ for (a) ω_a, $i=1, k'=1$, (b) ω_b, $i=1, k'=2$, (c) ω_c, $i=2, k'=1$, and (d) ω_d, $i=2, k'=2$. The corresponding ground-truth ILD mean $\mu_i(\omega)$ (red diamond), and estimates $\hat{\mu}_i(\omega)$ from [8] (blue square) and the bootstrap averaged estimate $\hat{\mu}^*_{i,\mathrm{A}}(\omega)$ (green circle) are marked to indicate their position relative to the histogram.

From a practical point of view, since each parameter is frequency dependent, one would consider using bootstrap averaged estimate for a given parameter if it leads to a smaller error over the frequency range. Thus, for a given frequency dependent parameter $\theta(\omega)$ with estimate $\hat{\theta}(\omega)$, it is natural to define a frequency averaged squared error (FASE), as

$$\mathrm{FASE}(\hat{\theta}) = \frac{1}{W}\sum_{\omega=1}^{W} |\hat{\theta}(\omega) - \theta(\omega)|^2, \tag{28}$$

which quantifies the overall error in a parameter estimate by averaging the squared errors for each frequency over the set of chosen frequencies. Here $|.|$ denotes the absolute value of a real or a complex-valued quantity. A comparison of the FASE of EM estimate of each parameter with FASE of the bootstrap averaged estimate shall allow us to see if an overall improvement is achieved. The ratios of FASE for estimates of the ILD, IPD and the mixing vector mean by the bootstrap averaged technique to the corresponding EM estimates from [8] for two-source mixtures with indices $k'=1,2,3,4$ are reported in Table II. Now, the ILD parameters are only frequency dependent, however, IPD parameters are frequency and lag dependent; mean of the mixing vector, on the other hand, is a frequency dependent bivariate complex-valued quantity. For the IPD mean parameter $\xi_{i\tau}$, we compare FASE at the ground truth delay. With the target source at 0° and the interference source placed at 30°, and a sampling frequency of 16000Hz., this is calculated to be 0.00027s, equivalent to 4 samples. For the mixing vector mean $\mathbf{a}_i$, the error in each component (l and r) is compared separately. From the ratio comparison for

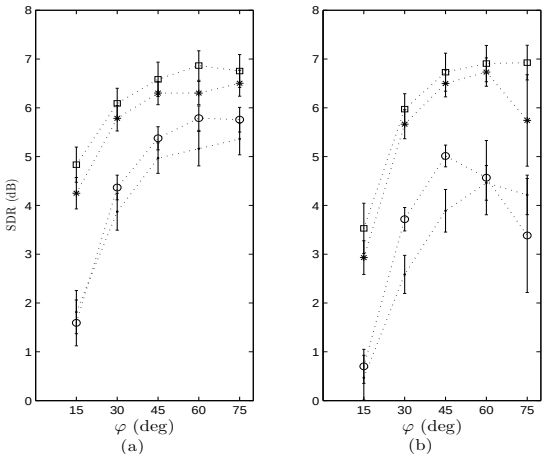

Fig. 8. A comparison of the average SDR over a set of (a) 15 two-source mixtures, and (b) 15 three-source mixtures, for separation performed using the proposed technique (square), the integrated method of [8] (asterisk), the interaural cue-based technique of [1] (circle), and the mixing vector model based method of [9] (dot) for $\varphi = 15^\circ$, 30°, 45°, 60°, and 75°. Error bars show standard error.

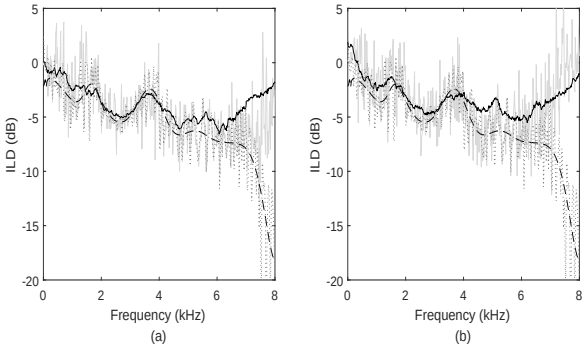

Fig. 9. A comparison of the ground-truth ILD mean (dB) $\mu_i(\omega)$ (dashed black) with the estimate $\hat{\mu}_i(\omega)$ (solid grey) obtained from [8], and the bootstrap averaged estimate $\hat{\mu}^*_{i,\mathrm{A}}(\omega)$ (solid black) vs. frequency (kHz) for (a) $i = 1, k' = 1$ and (b) $i = 1, k' = 2$, with s_1 placed at $\varphi = 15^\circ$, and s_2 placed at $\varphi = 30^\circ$. The ground-truth ILD mean of each source convolved with impulse response truncated to the window length are shown in dotted black.

the target source parameters (s_1) in Table II, we immediately see that the ratios for the ILD, IPD and mixing vector mean parameters are all less than 1 (with the exception of a_{1r} where it is ≈ 1), indicating that the bootstrap averaged estimates lead to an overall smaller error in comparison to the directly obtained EM estimates. We observe a significant improvement via our method for the ILD mean parameters as indicated by the extremely small FASE values for μ_1 across the four mixtures. Relatively higher FASE ratios are observed for the IPD and mixing vector mean parameter estimates. Due to the complex source and lag-dependent structure as well as the number of parameters, it is practically not feasible to perform a FASE comparison for the entire parameter set for all mixtures. The results in Table II, however, assure us of the improvement in the EM estimates via the proposed method. We would like to point out that in general, following [8], unequal weights may be assigned to the three cues in order to optimize the gain from the improved bootstrap averaged estimates.

Subsequently, we proceed with the final step of performing source separation. With the target source placed at 0 degree and the interference source(s) placed at azimuth φ (and $-\varphi$), we perform source separation for a set of 15 two-source and 15 three-source binaural mixtures, i.e. for a total of $15 \times 5 = 75$ two-source mixtures and 75 three-source mixtures. For each mixture, the bootstrap averaged T-F mask $\mathbf{M}^*_{i,\mathrm{A}}$ is computed as in (23), using which the target and the interference source(s) are separated. A comparison of the average SDR for the estimated target sources (over 15 mixtures) with the average SDR obtained from [8], [1], and [9] for each azimuth $\varphi = 15^\circ$, 30°, 45°, 60°, and 75° with (a) two-source and (b) three-source mixtures is displayed in Fig. 8. The improvement in the SDR of the sources separated using our bootstrap averaged T-F mask (squares) and [8] (asterisks) is clearly visible. Comparing the SDR (average) levels obtained using the joint method of [8] in Fig. 8(b), we note that relatively smaller SDRs are obtained either when the two interference sources s_2 and s_3 are placed too close with $\varphi = 15^\circ$ or too far with $\varphi = 75^\circ$ from the target source s_1. It is interesting to note that the bootstrap averaging approach leads to a significantly greater improvement for $\varphi = 75^\circ$ in the case of the three-source mixtures. The difference between the average SDR from the bootstrap average approach and the approach of [8] is calculated to be (in dB): $0.59, 0.31, 0.28, 0.57, 0.25$ for $\varphi = 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ$, respectively. A t-test (5% significance level) confirms that the average gain (in SDR) of 0.4 dB over 75 mixtures (15 mixtures for each φ) is significant. Similarly, the gain in the three-source case is calculated to be $0.60, 0.31, 0.23, 0.18, 1.19$ for $\varphi = 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ$, respectively, and a t-test (5% significance level) confirms that the average gain (in SDR) of 0.5 dB is significant.

The existing model-based source separation methods [8], [1], [9], have focused on the separation of the target source placed at 0° (exactly in front of the head). It is of interest to understand how these methods perform when the target source is positioned laterally at an angle greater than 0°. With the target source placed at 15°, and the interference source placed at azimuth $\varphi = 30^\circ$, we constructed a set of 15 two source mixtures using the binuaral room impulse responses exactly as described above. To understand the gain in performance via the bootstrap approach, a comparison of the ILD mean parameter estimates using the joint model-based method of [8] with the bootstrap averaged estimates is shown in Fig. 9. From the subplots in Fig. 9, we observe a notable improvement in the bootstrap averaged estimate (solid black) at higher frequencies in comparison to the estimate of [8] (solid grey), but only for the second mixture. This is confirmed by the corresponding FASE ratios, with an FASE of 1.19 for $k' = 1$ in comparison to an FASE of 0.70 for $k' = 2$. The relatively larger bias in the bootstrap averaged estimates for the first mixture is a consequence of the inherent bias in the estimates obtained directly from the speech mixture via [8] (solid grey) in this case. A comparison of the directly obtained estimates of [8] (solid grey) when the target source is placed at 0° (as shown

TABLE II
RATIOS OF THE FASE OF $\hat{\theta}_{\mathrm{A}}^{*}$ TO THE FASE OF $\hat{\theta}$ (EM), FOR $\theta \in \{\mu_i, \xi_{i\tau}, \mathbf{a}_i; i = 1, 2, \tau = 4\}$

	FASE($\hat{\theta}_{\mathrm{A}}^{*}$)/FASE($\hat{\theta}$)							
k'	μ_1	μ_2	$\xi_{1,\tau}$	$\xi_{2,\tau}$	$\mathbf{a}_1$		$\mathbf{a}_2$	
					l	r	l	r
1	0.1121	0.6786	0.8893	0.9826	0.6162	1.0481	0.7277	0.7844
2	0.1723	0.9879	0.9811	0.9940	0.6936	0.7232	0.7117	0.8742
3	0.2244	1.1773	0.6664	0.9943	0.7155	0.7809	0.8544	0.9079
4	0.1630	0.9733	0.9901	0.9903	0.7269	0.7577	0.8801	0.8610

in Fig. 3) with the estimates when the target source is placed laterally at 15° (shown in Fig. 9), makes this bias apparent. The SDR for the target source at 15° using the method of [8] is calculated to be 2.22 dB for $k' = 1$ and -1.60 dB for $k' = 2$, in comparison to 2.94 dB for $k' = 1$ and 1.57 dB for $k' = 2$ via bootstrap averaging. The average SDR over 15 mixtures calculated to be 1.49 dB from [8] in comparison to the average SDR of 2.60 dB via the bootstrap averaging approach, confirms improvement via the proposed procedure.

VIII. CONCLUSION

We draw attention to the problem of sub-optimal EM estimates of frequency-dependent GMMs and propose a bootstrap-based method to obtain estimates with smaller MSE. We identify the problem of model misspecification in the area of source separation where the absence of a precise model for reverberation leads to poor separation performance for reverberant speech mixtures. Our simulation experiments with speech mixtures show a clear improvement in estimates of the frequency domain GMM parameters via the proposed bootstrap averaging algorithm. An overall improvement is indicated by the FASE comparison. The averaged T-F mask leads to a higher SDR implying improved source separation. Further improvements in separation performance can be achieved by assigning frequency-specific weights to cues in order to maximize the gain from the bootstrap averaged estimates of cue model parameters. This will be investigated in our future work. Following the recent work by [38], another possible direction is to suitably combine T-F masks estimated from different model-based methods using appropriately bootstrapped speech mixtures to maximize improvement in source separation for reverberant mixtures.

Our bootstrap averaging approach applies to any frequency-specific GMM and hence its use is not limited to model-based source separation. In the multichannel NMF-based method [14], the authors note the poor performance of the EM estimates when assumptions on their model are not satisfied, e.g. due to nonlinear sound effects, longer reverberation times, and non-point sources. This corresponds to a misspecified mixture model in the T-F domain and hence the proposed bootstrap-based method finds application. The alternative MU algorithm for estimation, also discussed in [14], which does not exploit these assumptions is seen to be more robust to such model discrepancies. This issue with the EM implementation (e.g. [14], [15]) is also noted in [16], where only a MU algorithm is studied. An interesting direction for future work would be to understand if the EM implementation of [14] combined with our bootstrap approach can outperform the MU algorithm when assumptions on the mixing model are known to be violated, for example as in the case of professionally produced music recordings, [14].

ACKNOWLEDGEMENT

The authors are grateful to the referees for their very helpful comments leading to a much improved paper. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K014307 and the MOD University Defence Research Collaboration in Signal Processing.

REFERENCES

[1] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 382–394, 2010.

[2] M. N. Stuttle, "A gaussian mixture model spectral representation for speech recognition," Ph.D. dissertation, University of Cambridge, 2003.

[3] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of gaussians," in *Proceedings of the Fourth International Conference on Spoken Language, 1996. ICSLP 96.*, vol. 2. IEEE, 1996, pp. 1229–1232.

[4] M. Nilsson, H. Gustaftson, S. V. Andersen, and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002*, vol. 1. IEEE, 2002, pp. I–525.

[5] Z. K. Huang and K. W. Chau, "A new image thresholding method based on gaussian mixture model," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 899–907, 2008.

[6] A. Deleforge, F. Forbes, S. Ba, and R. Horaud, "Hyper-spectral image analysis with partially latent regression and spatial markov dependencies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 1037–1048, 2015.

[7] A. Alinaghi, W. Wang, and P. J. B. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 209–212.

[8] A. Alinaghi, P. Jackson, Q. Liu, and W. Wang, "Joint mixing vector and binaural model based stereo source separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 22, no. 9, pp. 1434–1448, 2014.

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

[10] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application*. Cambridge University Press, 2007.

[11] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[12] J. Cabrera and P. Meer, "Unbiased estimation of ellipses by bootstrapping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 752–756, 1996.

[13] S. Chandna and W. Wang, "Improving model-based convolutive blind source separation techniques via bootstrap," in *Proceedings of the IEEE Workshop on Statistical Signal Processing (SSP)*, 2014, pp. 424–427.
[14] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.
[15] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proceedings of the 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*. IEEE, 2010, pp. 1–4.
[16] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.
[17] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 2112–2121, 2014.
[18] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1075–1084, 2017.
[19] Y. Yu, W. Wang, and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2016, no. 1, p. 7, 2016.
[20] A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.
[21] D. Böhning and W. Seidel, "Editorial: recent developments in mixture models," *Computational Statistics & Data Analysis*, vol. 41, no. 3, pp. 349–357, 2003.
[22] G. McLachlan and D. Peel, *Finite Mixture Models*. John Wiley & Sons, 2004.
[23] C. Biernacki, G. Celeux, and G. Govaert, "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models," *Computational Statistics & Data Analysis*, vol. 41, no. 3, pp. 561–575, 2003.
[24] S. Chandna, "Frequency domain analysis and simulation of multichannel complex-valued time series," Ph.D. dissertation, Imperial College London, 2013.
[25] S. Chandna and A. Walden, "Simulation methodology for inference on physical parameters of complex vector-valued signals," *IEEE Transactions on Signal Processing*, vol. 61, pp. 5260–5269, 2013.
[26] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.
[27] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE transactions on Signal Processing*, vol. 52, pp. 1830–1847, 2004.
[28] C. Hummersone, "A psychoacoustic engineering approach to machine sound source separation in reverberant environments," Ph.D. dissertation, University of Surrey, UK, 2011.
[29] A. T. Walden, "A unified view of multitaper multivariate spectral estimation," *Biometrika*, vol. 87, pp. 767–788, 2000.
[30] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Prentice-hall Englewood Cliffs, 1978.
[31] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications*. Cambridge University Press, 1993.
[32] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. ICA*, 2000, pp. 215–220.
[33] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l 1-norm minimization," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 81–81, 2007.
[34] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.
[35] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2006, pp. 601–608.
[36] A. Ihlefeld and B. G. Shinn-Cunningham, "Effect of source spectrum on sound localization in an everyday reverberant room," *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 324–333, 2011.
[37] B. Shinn-Cunningham, N. Kopco, and T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *The Journal of the Acoustical Society of America*, vol. 117, p. 3100, 2005.
[38] X. Jaureguiberry, E. Vincent, and R. Gaël, "Fusion methods for speech enhancement and audio source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 7, pp. 1266–1279, 2016.

Swati Chandna received the M.S. degree in Mathematics from the University of Houston, Texas, U.S.A. in 2009, and the Ph.D. degree in statistics from Imperial College London, U.K., in 2013. She worked as a Research Fellow at the University of Surrey, U.K., from 2013 to 2014 and as a Research Associate at the Department of Statistical Science, University College London, U.K. from 2014 to 2017. She is currently a Lecturer in Statistics in the Department of Economics, Mathematics and Statistics at Birkbeck, University of London. Her research interests include nonparametric methods for networks, audio signal processing, frequency domain analysis of multichannel complex-valued time series and its applications in signal processing.

Wenwu Wang (M'02-SM'11) 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 Reader in Signal Processing, 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), an associate member of Surrey Centre for Cyber Security (since 2014), a member of the MRC/EPSRC Microphone Network (since 2015), and a member of the BBC Data Science Research Partnership (since 2017). During spring 2008, he has been a visiting scholar at the Perception and Neurodynamics Lab and the Centre 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 (co)-authored over 200 publications in these areas, including two books Machine Audition: Principles, Algorithms and Systems (IGI Global, 2010) and Blind Source Separation: Advances in Theory, Algorithms and Applications (Springer, 2014). He is currently an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING. He is also Publication Co-Chair of ICASSP 2019 (to be held in Brighton, UK). He was a Tutorial Speaker on ICASSP 2013, UDRC Summer School 2014-2017, SpaRTan/MacSeNet Spring School 2016, and London Intelligent Sensing Summer School 2017.