Modeling the Comb Filter Effect and Interaural Coherence for Binaural Source Separation

Luca Remaggi, Philip J. B. Jackson, Wenwu Wang, *Senior Member IEEE* .

***Abstract*—Typical methods for binaural source separation consider only the direct sound as the target signal in a mixture. However, in most scenarios, this assumption limits the source separation performance. It is well known that the early reflections interact with the direct sound, producing acoustic effects at the listening position, e.g. the so-called comb filter effect. In this article, we propose a novel source separation model, that utilizes both the direct sound and the first early reflection information to model the comb filter effect. This is done by observing the interaural phase difference obtained from the time-frequency representation of binaural mixtures. Furthermore, a method is proposed to model the interaural coherence of the signals. Including information related to the sound multipath propagation, the performance of the proposed separation method is improved with respect to the baselines that did not use such information, as illustrated by using binaural recordings made in four rooms, having different sizes and reverberation times.**

***Index Terms*—Source separation, comb filter effect, RIRs, IPD, ILD, binaural audio, multipath propagation, interaural coherence.**

I. Introduction

Source separation is one of the most investigated fields in the signal processing community. Several application areas can benefit from it. For instance, it can improve target detection performance of passive sonar systems [1]. In biomedical engineering, source separation is often used to analyze electrocardiograms, electroencephalograms, or magnetic resonance images [2]. Work on ancient document restoration has utilized source separation for correcting bleed-through distortion [3]. Source separation has also been used in a large range of speech applications. For instance, it is used for improving speech enhancement [4], crosstalk cancellation [5], and automatic speech recognition systems [6]. It can also be applied to improve hearing aids [7], or improve security systems [8]. Spatial audio can also rely on it, to produce object-based audio [9]. Robust speech processing is another target area [10].

In typical conditions, a sound produced by a source interacts with its environment during propagation, before it reaches a listening position. This multipath propagation is defined by its room impulse response (RIR), i.e. an acoustic signal describing the propagation of sound from source to listening position. RIRs have three parts: direct sound, early reflections, and late reverberation [11]. The direct sound carries information related to the source. Late reverberation provides clues about the size of the environment, without directional information [12]. Instead, early reflections affect the human sound perception, by conveying a directional sense of the geometry of the environment [13]. This generates auditory effects, for instance modifying the source width perception [14]. Moreover, being coherent with the direct sound, strong early reflections modify the perceived sound coloration, by generating a comb filter effect [15]. Hence, acoustic multipath properties should be considered in the design of source separation methods [16].

Many different approaches can be found in the literature to tackle the source separation problem. However, most of them do not explicitly model the acoustic multipath properties. For instance, in the well-known Model-based Expectation Maximization Source Separation and Localization (MESSL) method [17] only the direct sound interaural cues (i.e. the interaural phase difference (IPD) and interaural level difference (ILD)) were modeled, without considering any early reflection effect. Furthermore, although a garbage source was defined to indirectly deal with the late reverberation, there was not any formal attempt to model the reverb.

The aim of this article is to investigate how information related to early reflections can improve source separation methods, in general. Such information can be potentially used in many source separation methods, either unsupervised or supervised. Here, we selected MESSL [17] as a baseline method due to its unsupervised nature, and the convenience in incorporating the early reflections information into its IPD model. We extended MESSL [17], by emulating the comb filter effect produced by the early reflections. To do so, we define parametric functions in the time-frequency (TF) domain, and model the behavior of the IPD, by considering the interaction between the direct sound and the first arriving early reflection. The first reflection is chosen to be included into the model as it is the one that most affects the spatial cues [18]. Similar to MESSL, we also use an ILD model, which considers the direct sound cue, and the garbage source.

In addition to the comb filter effect, we propose a model that separates the reverberation's effect from the rest of the RIR's. This is done by approximating the human capability of separating sounds in reverberant environments. Specifically, we model the interaural coherence (IC) of indivual sources in the mixture, similar to what was introduced in [19]. However, there, the target source was assumed to be in front of the listener. Here, we propose an approach that is not limited by this, but works for any target source position.

The main novelties of this article include:

- a new IPD model, considering both direct sound and first reflection, to approximate the comb filter effect;

Manuscript received ; revised .

The authors are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. W. Wang is also with Qingdao University of Science and Technology, China. Emails: [l.remaggi, p.jackson, w.wang]@surrey.ac.uk.

- an extension of the MESSL IPD model, employing the target signal IC;
- an additional novel source separation method, obtained by combining the two new models above;
- the application of a source and image source localization algorithm to initialize the expectation maximization (EM) algorithm used to estimate the Gaussian mixture model (GMM) parameters, and one deep-learning approach using an MLP architecture with two hidden layers to generate the TF mask.

Since the novel IPD model approximates the early reflection information, the first new pipeline is named as Early Reflection MESSL (ER-MESSL). The second novel pipeline uses the IC of the estimated target signal, hence, its name is IC-MESSL. By combining the new IPD model with the IC based model, we obtain the third proposed method, thus named as ERIC-MESSL. Finally, there is need for the employed EM algorithm to be initialized. Since our proposed methods combine the direct sound and first reflection information, we employ our Image Source Direction and Ranging (ISDAR) [20] to initialize it, by localizing the target source and related image source [21]. A comparative evaluation of early and late models is performed and reported as additional contribution. The challenging two source binaural speech mixture scenario was analyzed, by employing signal and perceptual objective measures. In the experimental section, we also evaluate the improvement given by considering early reflection information in a state-of-the-art deep learning based method, for supervised speech separation. Through this, we further demonstrate that early reflection information improves source separation methods' performance, including deep learning, and that this can be potentially applied to many approaches in the literature.

The overall structure of this article is as follows: in Section II, related source separation methods are discussed; Section III defines the theoretical foundations of the proposed approach. In Sections IV and V, the proposed interaural cue models for the comb filter and IC are presented, respectively. Section VI describes the source separation algorithm. In Section VII, the experiments are described, with related results and discussion. Finally, Section VIII draws the conclusion.

II. Related Work in Speech Source Separation

Many approaches can be found in the literature to tackle the source separation problem. Some of them exploit a-priori information about basis functions representing the signals in the mixture [22]. Others employ the non-negative matrix factorization (NMF) to learn sparse representation of speech sources [23–26]. The independent component analysis (ICA) [27] is also used to decompose the mixture into independent signals, by projecting the mixtures into different domains. Scenarios where multiple microphones are available were also investigated [28–31], e.g. using beamformers [32], [33]. Recently, deep neural networks (DNNs) became widely popular, when large training datasets are available [34–38].

TF masking is a popular approach, which assigns different weights to the mixture, in the TF domain [39]. In [17], the authors presented the MESSL method which uses binaural signals. Two interaural cues were exploited, i.e. the ILD and the IPD, relating the azimuthal sound direction of arrival (DOA) to the head orientation [40]. The method presented in [41] utilized, instead, the so called mixing vector (MV). For each frequency bin, this vector contains the time invariant frequency response component of the room. In both [17] and [41], the probability of each TF point belonging to a specific source in the mixture was determined. From this probability, TF masks were generated. In [42], the two methods proposed in [17] and [41] were combined, constructing a probability distribution that takes into account the three cues ILD, IPD and MV. In [43], a high-dimensional vector, constructed by combining the IPD and ILD cues, was projected onto a 2D space, represented by the sound azimuth and elevation DOA. A regression approach located the sources, and estimated the TF masks. The IC cue was then employed in [44].

In the literature, yet few works can be found that consider both direct sound and early reflections. In [45], the source separation problem was divided into different procedures, by applying deconvolution to each individual reflection. However, the performance degrades with low signal-to-noise ratio (SNR) conditions. In [46], a variation of the ICA method [47] was used to estimate the time-dependent mixing system, considering the multipath propagation. However, with the ICA approach, the effect of its classical permutation problem was exacerbated by the incorrect RIR components' alignment. Deconvolution of the received signals was proposed in [48], by employing simulated RIRs. These RIRs were estimated by matching the temporal support of recorded ones. Nevertheless, binaural effects, such as head shadowing and pinnae influence, were not considered. Multichannel microphone arrays were used in [33], where beamformers were designed to have their directivty patterns characterized by multiple beams, to simultaneously extract direct sound and early reflections. Results show improvement with respect to classical beamforming. However, they were tested only with simulated RIRs. The work in [49] demonstrated the benefit of including reflection information in source separation models, by employing a NMF approach. Nevertheless, only simulated RIRs were employed.

In this article, we consider the first arriving early reflection and related direct sound, to propose a binaural model that increases the robustness in reverberant environments, by estimating TF masks. It is based on [17], nevertheless, the proposed model could be potentially adapted to work with other methods described above, from beamformers to DNNs.

III. Background Definitions

In this section, we provide a general overview of the adopted approach, and discuss the assumptions. The definitions of the general elements of the proposed architecture (e.g. binaural RIRs (BRIRs) and interaural spectrograms) are also given.

A. General Overview of the Proposed Method

Classical source separation methods exploit features related to the direct sound to separate the target sound from a mixture. In [17], the authors presented one of the first models to deal with the reverberation, by proposing the "garbage" source. In this article, we model two perceptual effects: the comb

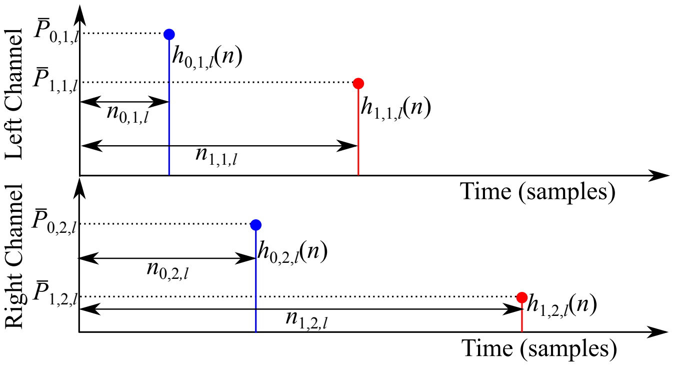

Fig. 1: Example of an ideal BRIR, zoomed into its direct sound (blue) and first reflection (red) components (depicted as Dirac pulses). The top figure shows the RIR related to sensor $i = 1$, whereas the bottom one the RIR at sensor $i = 2$. The amplitudes and delays are defined in Equation (2).

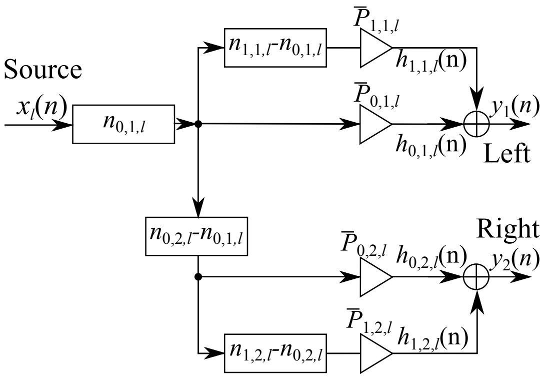

Fig. 2: Schematic representation of the comb filter effect created for the two received sounds ($y_1(n)$ and $y_2(n)$), given the sound produced at the l-th source $x_l(n)$. The direct sounds and reflections, together with the related delays (▭) and attenuation factors (▷) are the same as those defined in Fig. 1.

filter and IC. Through the former we aim to model the first early reflection, in a constructive fashion, to enhance the sound produced by the target speaker. The latter models the reverberation, by aiding the garbage source in suppressing it.

B. Proposed Method Assumptions

In the proposed source separation method, assumptions were made, defining its scientific boundaries as follows:

- The number of sources L is known a-priori;
- Source signals are sparse in the TF domain;
- The mixing system is time invariant;
- The first reflection has a dominant specular component;
- Sources are sufficiently far from the reflectors;
- The first early reflection is coherent with the direct sound.

Although L has to be known a-priori, there is no restriction on it with respect to the number of microphones M, thus, the method can be also applied to underdetermined scenarios. Sparsity over the TF domain corresponds to the assumption of having, for each TF bin, only one of the sources dominating the mixture. Sources and microphones are assumed to be static within a static environment, i.e. the mixing system is time invariant. Where the first reflection has a dominant specular component, it is detected from RIRs to initialize the EM re-estimation. The sources have to be distant enough from the reflectors, in order to have the first reflection arriving between 5 ms and 40 ms later than the direct sound. Finally, the assumption of coherence between the first reflection and direct sound allow them to be modeled as a comb filter. The later reflections, having a more stochastic nature, are assumed to be incoherent and modeled through the IC, with the reverb.

C. Binaural Room Impulse Response

A RIR is a signal that characterizes the acoustics of an environment with respect to source and sensor positions. RIRs that are recorded by microphones in ear canals of a dummy head, are usually known as BRIRs. They are defined as:

$$I_{i,l}(n) = \sum_{e=0}^{T_m} h_{e,i,l}(n - n_{e,i,l}) + w_{i,l}(n), \qquad (1)$$

where $i \in [1,2] \in \mathbb{N}$ and l are the microphone and source indexes, respectively; n is the discrete time index, T_m indicates the last early reflection, and $w_{i,l}(n)$ represents the late reverberation, whereas e is the reflection index ($e = 0$ indicates the direct sound). $h_{e,i,l}$ is a function describing the reflection. $n_{e,i,l}$ represents the reflection times of arrival (TOAs).

Following the assumption of having dominant specular components, the early reflections are approximated by Dirac deltas $\delta(n)$ of different amplitudes $\overline{P}_{e,i,l}$. For source separation purpose, we consider the direct sound and first reflection components (i.e. $e = \{0, 1\}$) (see Fig. 1):

$$\begin{aligned} h_{0,1,l}(n) &= \overline{P}_{0,1,l}\delta(n - n_{0,1,l}); \\ h_{1,1,l}(n) &= \overline{P}_{1,1,l}\delta(n - n_{1,1,l}); \\ h_{0,2,l}(n) &= \overline{P}_{0,2,l}\delta(n - n_{0,2,l}); \\ h_{1,2,l}(n) &= \overline{P}_{1,2,l}\delta(n - n_{1,2,l}). \end{aligned} \qquad (2)$$

D. Comb Filter and Interaural Coherence

In environments where the first reflection is delayed between 5 ms and 40 ms to the direct sound, the coloration of the sound perceived is different from the one produced [14]. In signal processing, the superimposition of a signal with its delayed version is the result of comb filtering the signal, hence, we model this perceptual effect as a comb filter effect (see Fig. 2).

Reverberation is a diffuse component of the RIR that makes source separation more challenging by smearing the target signal, both temporally and spatially. Thus it is useful for robust separation to suppress it. With spaced microphones, reverberation signals are decorrelated above a certain frequency [50]. With binaural microphones, IC measures the two signals correlation, hence we use it to model the reverberation.

E. Interaural Spectrogram

Following the definition of BRIR in Equation (1), the mixtures received at the i-th sensor can be written as:

$$y_i(n) = \sum_{l=1}^{L} x_l(n) * I_{i,l}(n) * w_{i,l}(n), \qquad (3)$$

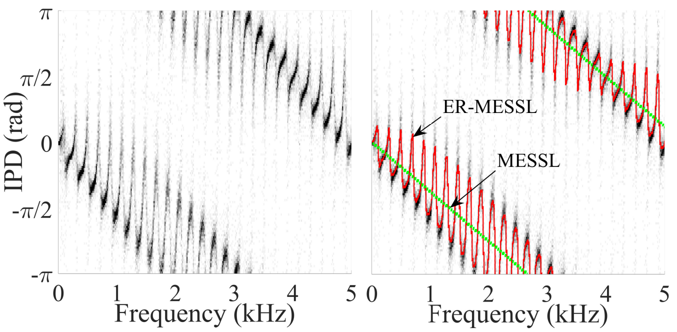

Fig. 3: The figure on the left shows the IPD as a function of frequency for a single source convolved with an ideal BRIR formed by only direct sound and first reflection. On the right, the same IPD function is simultaneously fitted by the MESSL IPD model [17] (the straight green line), and our comb filter based ER-MESSL IPD model (the fluctuating red curve).

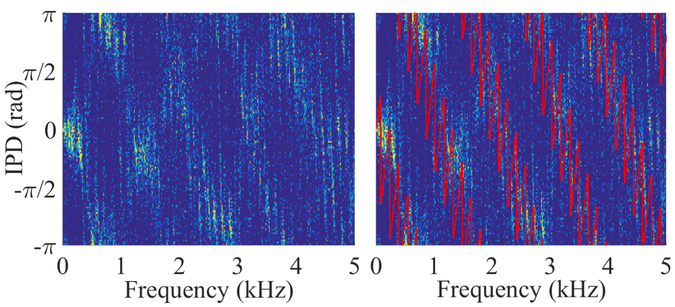

Fig. 4: On the left the IPD function for a mixture of two sources is shown. On the right, our comb filter based ER-MESSL IPD model (the fluctuating red curve) is employed to fit one of the two sources in the same IPD function.

where $x_l(n)$ is the signal generated by the l-th source, $w_{i,l}(n)$ is the convolutive white Gaussian noise, L is the number of sources, and "$*$" is the convolution operator. Since the human auditory system analyzes the received mixtures in the TF domain [51], we use the the short-time Fourier transform (STFT) to calculate the TF representation of $y_i(n)$:

$$y_i(m,\omega) = \sum_{l=1}^{L} x_l(m,\omega) I_{i,l}(\omega) w_i(m,\omega), \tag{4}$$

where m is the discrete time frame index, whereas ω is the angular frequency. $I_{i,l}(\omega)$ is not time dependent, by assuming the mixing system to be time-invariant. Considering binaural systems, the interaural spectrogram is defined as [17]:

$$y^{\mathrm{IS}}(m,\omega) = \frac{y_1(m,\omega)}{y_2(m,\omega)} = 10^{\alpha^{\mathrm{ILD}}(m,\omega)/20} \exp[j\phi^{\mathrm{IPD}}(m,\omega)], \tag{5}$$

where $\alpha^{\mathrm{ILD}}(m,\omega)$ and $\phi^{\mathrm{IPD}}(m,\omega)$ are the ILD and IPD of the observation, respectively, and $j = \sqrt{-1}$.

IV. Modeling the Comb Filter Effect

The IPD and ILD cues can be modeled to generate probability distributions for identifying the dominant source, given each TF bin. The novel IPD model that approximates the comb filter effect is proposed in this section. Furthermore, the ILD model (that was presented in [17]) is described. Finally, these two are combined into a joint probability distribution.

In the proposed model (as in MESSL [17]), sound sources are assumed to be spatially quasi-static: they have to be static within the time interval under investigation. Nonetheless, as a potential extension for future work, one could employ a tracking system, that would provide the model with updated time delays (i.e. $n_{e,i,l}$). Using audio only, beamformers could be used to estimate constantly the DOAs of the direct sound and early reflections. Alternatively, one could track sources by employing a particle filter [52], or a multimodal approach [53].

A. Interaural Level and Phase Differences

The proposed IPD model is defined to match the behavior of the observed IPD and is different from previous work where only the direct sound information was used [17]. By assuming ideal BRIRs as formed by direct sound and first reflection (see Fig. 1), the two channel frequency responses are:

$$\begin{aligned}\hat{I}_{1,l}(\omega) &= \overline{P}_{0,1,l} \exp[-j\omega n_{0,1,l}] + \overline{P}_{1,1,l} \exp[-j\omega n_{1,1,l}]);\\ \hat{I}_{2,l}(\omega) &= \overline{P}_{0,2,l} \exp[-j\omega n_{0,2,l}] + \overline{P}_{1,2,l} \exp[-j\omega n_{1,2,l}]).\end{aligned} \tag{6}$$

Their ratio is the interaural frequency response model:

$$\hat{I}_l(\omega) = \frac{\hat{I}_{1,l}(\omega)}{\hat{I}_{2,l}(\omega)} = \\ \frac{\overline{P}_{0,1,l} + \overline{P}_{1,1,l} \exp[-j\omega(n_{1,1,l} - n_{0,1,l})]}{\overline{P}_{0,2,l} \exp[-j\omega(n_{0,2,l} - n_{0,1,l})] + \overline{P}_{1,2,l} \exp[-j\omega(n_{1,2,l} - n_{0,1,l})]}. \tag{7}$$

The phase of this equation, denoted as $\hat{I}_l^{\mathrm{ang}}(\omega)$, corresponds to the proposed IPD model, and it is one of the main novelties of this article. For the l-th source, the difference between the observed IPD $\phi^{\mathrm{IPD}}(m,\omega)$ and its model is the phase residual:

$$\hat{\phi}_l^{\mathrm{IPD}}(m,\omega;\mathbf{C}_l) = \phi_l^{\mathrm{IPD}}(m,\omega) - \hat{I}_l^{\mathrm{ang}}(\omega;\mathbf{C}_l), \tag{8}$$

that is wrapped into the interval $[-\pi\ \pi)$; and:

$$\mathbf{C}_l = [n_l^{\mathrm{DS}}, n_l^{\mathrm{DF}}, n_l^{\mathrm{ST}}, \overline{P}_{0,1,l}, \overline{P}_{1,1,l}, \overline{P}_{0,2,l}, \overline{P}_{1,2,l}], \tag{9}$$

where $n_l^{\mathrm{DS}} = n_{0,2,l} - n_{0,1,l}$, $n_l^{\mathrm{DF}} = n_{1,1,l} - n_{0,1,l}$, and $n_l^{\mathrm{ST}} = n_{1,2,l} - n_{1,1,l}$. An example of the IPD model fitting an ideal IPD observation is shown in Fig. 3, together with a visual comparison of the MESSL IPD model [17]. The ideal IPD observation was obtained from a synthetic BRIR composed of only direct sound and first reflection. From this figure, it is clear that our proposed ER-MESSL IPD model fits the observed data better than MESSL, by considering the comb filter effect. In Fig. 4, we also report the IPD function related to a mixture of two sources, generated using recorded BRIRs. The two sources' contributions are well visible from the figure on the left, as two linear patterns having opposite gradients. From the figure on the right, it is also visible that our proposed ER-MESSL model fits one of the two sources.

The ILD cue, $\alpha_l^{\mathrm{ILD}}(m,\omega)$, is modeled, similar to [17], by considering directly the frequency-dependent BRIR, as:

$$a_l^{\mathrm{ILD}}(\omega) = 20\log_{10}\left|\frac{I_{1,l}(\omega)}{I_{2,l}(\omega)}\right|, \tag{10}$$

where "$|\cdot|$" indicates the absolute value.

B. Interaural Cue Probability Distributions

For the ILD cue, the probability of each TF bin being associated to source l can be written as a Gaussian distribution [42]:

$$p(\alpha^{\mathrm{ILD}}(m,\omega)|l) = \mathcal{N}(\alpha^{\mathrm{ILD}}(m,\omega)|\mu_l^{\mathrm{ILD}}(\omega), \sigma_l^{\mathrm{ILD}^2}(\omega)), \tag{11}$$

where $\mu_l^{\mathrm{ILD}}(\omega)$ is the mean, and $\sigma_l^{\mathrm{ILD}^2}(\omega)$ is the variance.

Regarding the IPD cue, a top-down approach is used to wrap the signal phase between $\pm\pi$ [17]. $\hat{\phi}_l^{\mathrm{IPD}}(m,\omega;\mathbf{C}_l)$ is modeled by a Gaussian distribution:

$$\begin{aligned} &p(\hat{\phi}^{\mathrm{IPD}}(m,\omega)|l,\mathbf{C}_l) = \\ &= \mathcal{N}(\hat{\phi}^{\mathrm{IPD}}(m,\omega;\mathbf{C}_l)|\mu_l^{\mathrm{IPD}}(\omega;\mathbf{C}_l), \sigma_l^{\mathrm{IPD}^2}(\omega;\mathbf{C}_l)), \end{aligned} \tag{12}$$

where $\mu_l^{\mathrm{IPD}}(\omega;\mathbf{C}_l)$ and $\sigma_l^{\mathrm{IPD}^2}(\omega;\mathbf{C}_l)$ are the IPD distribution mean and variance, respectively.

To sum up, by assuming the IPD and ILD observations as being conditionally independent given their related parameters, their probability distributions can be combined as:

$$\begin{aligned} &p(\alpha^{\mathrm{ILD}}(m,\omega), \hat{\phi}^{\mathrm{IPD}}(m,\omega)|l,\mathbf{C}_l) = \\ &= \mathcal{N}(\alpha^{\mathrm{ILD}}(m,\omega), \hat{\phi}^{\mathrm{IPD}}(m,\omega;\mathbf{C}_l)|\Xi_l), \end{aligned} \tag{13}$$

where $\Xi_l = \{\mu_l^{\mathrm{ILD}}(\omega), \sigma_l^{\mathrm{ILD}^2}(\omega), \mu_l^{\mathrm{IPD}}(\omega;\mathbf{C}_l), \sigma_l^{\mathrm{IPD}^2}(\omega;\mathbf{C}_l)\}$. This probability distribution identifies the proposed comb filter model, that was conceived to approximate the interaction between the received direct sound and first early reflection, i.e. two strongly coherent signals. This model does not take into account either later reflections or reverberation, which are, in this article, dealt by the IC model.

V. Modeling the Interaural Coherence

To suppress reverberation, the idea is to identify those areas in the TF domain that are dominated by the direct sound, and the strong early reflections. The direct sound and a strong reflection recorded at the two ears are highly correlated and coherent. In contrast, the late reverberation is diffuse, and does not present correlation between the binaural signals, at every frequency. Thus, we use the IC to create a probability mask, based on the coherence level, for every TF bin [19].

A. Interaural Coherence TF Mask

The process we employed to calculate the IC of a signal follows an approach that was originally proposed in [54], for dereverberation. For each TF bin, the auto-power spectral density of the two channels $i = \{1,2\}$ is calculated as:

$$\Phi_i(m,\omega) = \kappa\Phi_i(m-1,\omega) + (1-\kappa)|y_i(m,\omega)|^2, \tag{14}$$

where $0 \leq \kappa \leq 1$ is a smoothing factor determined as $\kappa = 1/(\tau \cdot f_s)$, with $\tau = 10\,\mathrm{ms}$ being a time constant and f_s the sampling frequency [55]. The cross-power spectral density between the two channels is:

$$\Phi_{1,2}(m,\omega) = \kappa\Phi_{1,2}(m-1,\omega) + (1-\kappa)y_1(m,\omega)y_2^*(m,\omega), \tag{15}$$

with $[\cdot]^*$ indicating the complex conjugate operation. From (14) and (15), the magnitude squared coherence is:

$$\Gamma_{1,2}(m,\omega) = \frac{\Phi_{1,2}(m,\omega)}{\Phi_1(m,\omega)\Phi_2(m,\omega)}. \tag{16}$$

The values of $\Gamma_{1,2}(m,\omega)$ are constrained between 0 and 1, thus, $\Gamma_{1,2}(m,\omega)$ is employed as the TF soft mask that models the IC. To do so, it will be used as prior mask during the posterior probability calculation, that will be described in Section VI-B[1]. $\Gamma_{1,2}(m,\omega)$ is computed from the observation by employing the equations defined in [55].

The aim of modeling the IC is to suppress remaining early reflections and late reverberation, i.e. the BRIR parts that are not modeled by the comb filter. A similar approach to calculate an IC based TF mask was employed in [19]. However, there, the target source was assumed to be in front of the listener. Here, we do not make any assumption regarding the position of the target source. Its position is estimated by ISDAR, the algorithm described later, in Section VI-C. Having the target source position, we then calculate $\Gamma_{1,2}(m,\omega)$ by analyzing the BRIR related to the estimated DOA.

B. The Garbage Source

Late reflections and reverberation are problematic components of the acoustics that are undesiderable in the comb-filter model, proposed in Section IV, as their first-order statistics are unreliable. Hence, the IC model described above is used to suppress these components of the BRIRs by consideration of their second-order statistics. In addition to this, we utilize a garbage source, as in [17]. It represents noise dominating the TF bins that are not claimed by any of the other sources.

The parameters Ξ^{G} used to model the garbage source are the same as those used by the other sources to define the distribution in Equation (13). The difference is the initialization, since the garbage source is used to model the noise sources, such as background noise, measurement noise, and reverberation.

VI. Source Separation Model Reestimation

The EM is described here, along with the log-likelihood used to optimize the parameters of the proposed models.

A. Parameter Estimation from Mixtures

The parameters characterizing the interaural cue probability models are $\Omega_l = \{\Xi_l,, \beta_{l,\mathbf{C}_l}\}$, where $\beta_{l,\mathbf{C}_l}$ is the marginal class membership, described as the joint probability of each TF bin being dominated by source l with the IPD model parameters $\mathbf{C}_l$: $\beta_{l,\mathbf{C}_l} = p(l,\mathbf{C}_l)$. These parameters can be estimated for a specific source l. This is a trivial problem upon the availability of the dominant source information for each TF bin. However, whether the source l is dominating a specific TF bin is not directly observable from the mixtures. On the other hand, l can be inferred from the interaural cues and observed models, that are not known a-priori. This missing data problem is solved by the EM algorithm.

The log-likelihood of the observations can be then defined as in [17], however, with the additional IC distribution:

$$\begin{aligned} \mathcal{L}(\Omega) &= \sum_{m,\omega}[\log p(\alpha^{\mathrm{ILD}}(m,\omega), \hat{\phi}^{\mathrm{IPD}}(m,\omega), |\Omega) + \log\Gamma_{1,2}(m,\omega)] \\ &= \sum_{m,\omega}\log\sum_{l,\mathbf{C}_l}\beta_{l,\mathbf{C}_l}p(\alpha^{\mathrm{ILD}}(m,\omega)|l)p(\hat{\phi}^{\mathrm{IPD}}(m,\omega)|l,\mathbf{C}_l)\Gamma_{1,2}(m,\omega). \end{aligned} \tag{17}$$

[1] This has been implemented using the MESSL open source code's option allowing the definition of prior masks: https://github.com/mim/messl.

This definition assumes that the IC, IPD and ILD cues are independent. As a result, the joint probability is written as the product of individual probabilities. In addition, the number of sources must be specified a-priori [17]. Note that the inclusion of the IC into the log-likelihood function is different from previous approaches, such as [19]. There, the IC mask was multiplied by the TF representation of the mixture. Equation (17) represents the proposed ERIC-MESSL.

B. Expectation-Maximization (EM)

The EM algorithm is used to estimate the parameters and probability at each TF bin. $\Gamma_{1,2}(m,\omega|l)$ is considered as a prior, and not updated during the iterations. During the E-step, the occupation likelihood of source l with parameters $\mathbf{C}_l$ is calculated for each TF bin, given $\alpha^{\mathrm{ILD}}(m,\omega)$ and $\hat{\phi}^{\mathrm{IPD}}(m,\omega)$:

$$\begin{aligned}\nu_l(m,\omega|\mathbf{C}_l) = \beta_{l,\mathbf{C}_l} p(\alpha^{\mathrm{ILD}}(m,\omega)|l)\\ \cdot\, p(\hat{\phi}^{\mathrm{IPD}}(m,\omega)|l,\mathbf{C}_l) p(\Gamma_{1,2}(m,\omega)|l).\end{aligned} \tag{18}$$

This expectation is then used in the M-step, to re-estimate the parameters, and maximize the likelihood. The ILD parameters are updated as [42]:

$$\begin{aligned}&\mu_l^{\mathrm{ILD}}(\omega) = \frac{\sum_{m,\mathbf{C}_l} \alpha^{\mathrm{ILD}}(m,\omega)\nu_l(m,\omega|\mathbf{C}_l)}{\sum_{m,\mathbf{C}_l} \nu_l(m,\omega|\mathbf{C}_l)},\\ &\sigma_l^{\mathrm{ILD}^2}(\omega) =\\ &\frac{\sum_m(\alpha^{\mathrm{ILD}}(m,\omega) - \mu_l^{\mathrm{ILD}}(\omega))^2 \sum_{\mathbf{C}_l} \nu_l(m,\omega|\mathbf{C}_l)}{\sum_{m,\mathbf{C}_l} \nu_l(m,\omega|\mathbf{C}_l)},\end{aligned} \tag{19}$$

whereas the IPD residual parameters are updated as:

$$\begin{aligned}&\mu_l^{\mathrm{IPD}}(\omega|\mathbf{C}_l) = \frac{\sum_m \hat{\phi}_l(m,\omega|\mathbf{C}_l)\nu_l(m,\omega|\mathbf{C}_l)}{\sum_m \nu_l(m,\omega|\mathbf{C}_l)},\\ &\sigma_l^{\mathrm{IPD}^2}(\omega|\mathbf{C}_l) =\\ &\frac{\sum_m(\hat{\phi}_l(m,\omega|\mathbf{C}_l) - \mu_l^{\mathrm{IPD}}(\omega|\mathbf{C}_l))^2\nu_l(m,\omega|\mathbf{C}_l)}{\sum_m \nu_l(m,\omega|\mathbf{C}_l)}.\end{aligned} \tag{20}$$

Also the marginal class membership is updated:

$$\beta_{l,\mathbf{C}_l} = \frac{1}{B}\sum_{m,\omega} \nu_l(m,\omega|\mathbf{C}_l), \tag{21}$$

where B is the total number of TF bins.

The model parameters that are found during the last EM iteration are selected as the final estimation. Probabilistic masks are generated by marginalizing over the estimated $\mathbf{C}_l$:

$$M_l(m,\omega) = \sum_{\mathbf{C}_l} \nu_l(m,\omega|\mathbf{C}_l). \tag{22}$$

The separated source signal l can finally be obtained as:

$$\hat{y}_{i,l}(m,\omega) = y_i(m,\omega)M_l(m,\omega), \qquad \forall m,\ \forall\omega. \tag{23}$$

The seven interaural model parameters defined in $\mathbf{C}_l$ are treated in the EM as hidden variables. Specifically, they are modeled as discrete random variables, where the sets of allowed values are specified a-priori, as in [17]. The parameters in $\mathbf{C}_l$ are not internally updated by the EM algorithm. Instead, every allowed value combination is tested [17]. The combination that maximizes the log-likelihood is then chosen. Since the proposed IPD model in ER-MESSL and ERIC-MESSL is composed of seven parameters $\mathbf{C}_l$ (Equation (9)), it involves a seven dimensional space when trying to find the best combination of them, hence it is computationally expensive. Therefore, the amplitudes $\overline{P}_{e,i,l}$ are fixed; only the initialized value is allowed. The time-dependent parameters' allowed ranges were found empirically, as in Table I.

C. Model Initialization

The initialization part plays a crucial role for the EM algorithm performance, since the log-likelihood is not convex. A poor initialization leads to local maxima, thus affecting the source separation results. The estimated source and image source positions are used to initialize the time-dependent parameters n_l^{DF}, n_l^{DS} and n_l^{ST}. Instead, the amplitudes $\overline{P}_{0,1,l}$, $\overline{P}_{1,1,l}$, $\overline{P}_{0,2,l}$, $\overline{P}_{1,2,l}$ are initialized by analyzing the BRIR that is related to the estimated DOA. Therefore, the early reflection information is not pre-estimated, but found and refined by the proposed system at each iteration. The microphone array is only used to initialize the EM algorithm.

In [17], only the direct sound was used to model the source, and the parameters were initialized by using the GCC-PHAT algorithm [56]. In our proposed method, correct localization of the first reflection is also crucial. Source and image source positions are estimated through our ISDAR method [20]. This method relies on RIRs recorded via a multichannel microphone array, placed at the same listener position. We chose this since, to our knowledge, no method in the literature can reliably localize reflections, given binaural recordings. However, other kinds of approaches could be also employed, for instance, audio-visual based methods [57].

ISDAR is based on spherical coordinates. Direct sound and reflection TOAs $\hat{n}_{e,i,l}$ are estimated through the clustered dynamic programming projected phase-slope algorithm (C-DYPSA), that we proposed in [20], whereas azimuth DOAs $\Theta_{e,l}$ are estimated through the delay-and-sum beamformer [20], [58]. Considering the listener at the center of the coordinate system, the radial distances of the source and image source are calculated as $\rho_{e,l} = \frac{1}{M}\sum_{i=1}^{M}(\hat{n}_{e,i,l}c_0)$, where c_0 is the sound speed, and $\hat{n}_{e,i,l}$ is either the estimated direct sound ($e = 0$) or first reflection ($e = 1$) TOA. The source and image source positions in the Cartesian coordinate system are given by $b_{x,e,l} = \rho_{e,l}\cos\Theta_{e,l}$ and $b_{y,e,l} = \rho_{e,l}\sin\Theta_{e,l}$. Knowing the listener position, these values are converted into TDOAs to populate Equation (9). The amplitudes $\overline{P}_{e,i,l}$ are calculated by directly analyzing the BRIRs at the reflection TOA $\hat{n}_{e,i,l}$.

Regarding the ILD distribution, the value of the ILD prior mean is estimated by utilizing a set of synthetic binaural RIRs, as in [17]. The garbage source is initialized to have a uniform distribution across IPD, and a uniform ILD distribution with zero mean for all frequencies.

VII. Experiments and Results

In this section, the results of a set of experiments are described. In these experiments, we consider mixtures of speech signals in four different recorded environments. When only the IC is modeled, and MESSL is used to model only

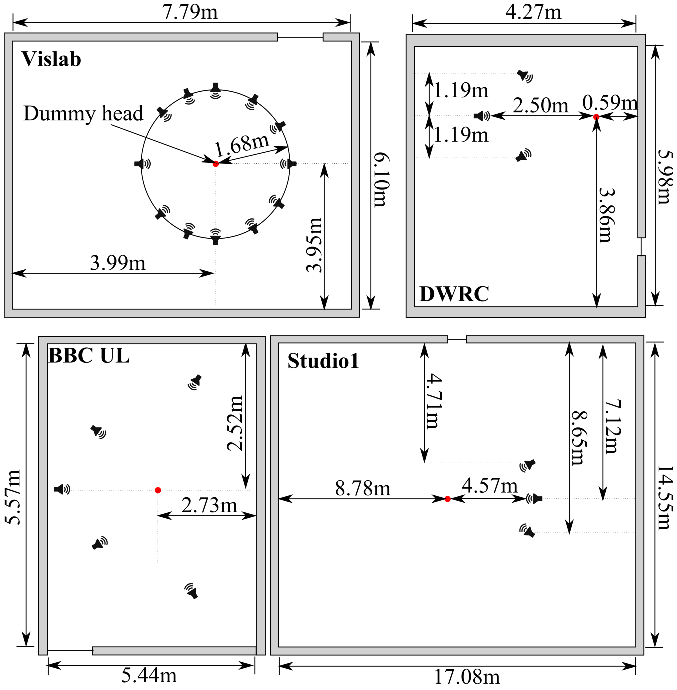

Fig. 5: Plan views of the four recorded rooms. The red circles represent the position of the dummy head, whereas the loudspeakers are depicted using their stylized symbol.

TABLE I: Range sizes for the allowed values around the initialized IPD model parameters.

	Vislab	DWRC	BBC UL	Studio1
n_l^{DF}, n_l^{DS}, n_l^{ST}	$\pm 0.13\,\mathrm{ms}$	$\pm 0.13\,\mathrm{ms}$	$\pm 0.19\,\mathrm{ms}$	$\pm 0.31\,\mathrm{ms}$

TABLE II: Recorded room RT60s, averaged over the $\frac{1}{3}$ octave bands between 500 Hz and 4 kHz, DRRs, and TISAs, averaged over all the tested combinations. L_{TOT} is the number of loudspeakers. The loudspeaker positions are reported as lateral angles with respect to the dummy head orientation.

	Vislab	DWRC	BBC UL	Studio1
RT60 (s)	0.32	0.27	0.28	0.94
DRR (dB)	17.8	3.9	15.7	6.0
AVG TISA (Deg)	75	37	71	32
$\mathbf{L}^{\mathrm{TOT}}$	7	3	5	3
Lateral angles (Deg)	0, ±30, ±60, ±90	0, ±27	0, ±37, ±110	0, ±27

the direct sound, the proposed method is named as IC-MESSL. When the comb filter effect is modeled, extending MESSL in that sense, without considering any prior knowledge regarding the IC, the proposed method is ER-MESSL. Otherwise, if both the comb filter and the IC are modeled, the novel method is named as ERIC-MESSL. The three proposed methods are compared to MESSL [17]. The ranges of allowed parameters for the comb filter model are in Table I, for each dataset.

At the end of this section, we also show that other separation algorithms would benefit from the inclusion of early reflection information. We extend a deep learning based state-of-the-art method. Different from MESSL, which is an unsupervised method, the deep learning approach is used to demonstrate that improvements can be achieved also for supervised methods.

A. Datasets

BRIRs[2] were recorded in four rooms, characterized with different size and reverberation time (RT60). The four rooms are named as "Vislab", "Digital World Research Centre" (DWRC), "BBC Usability Laboratory" (BBC UL), and "Studio1". Their plan views are shown in Fig. 5, whereas the RT60s are in Table II, together with the number of loudspeaker positions L_{TOT} and their lateral angles. Two different dummy heads were employed (i.e. a Cortex Manikin Mk2 Binaural Head and Torso Simulator and a Neumann KU100 dummy head), depending on their availability for the recordings. To obtain data for the initialization, a 48-channel bi-circular array with a typical microphone spacing of 21 mm and an aperture of 212 mm was utilized to record RIRs [20][3]. The dummy head and bi-circular array were recorded separately, to avoid interference effects. All the recordings were made by employing the swept-sine technique [59], with $f_s = 48\,\mathrm{kHz}$.

Arrangements. Two further measures characterize the datasets: the direct to reverberant ratio (DRR) [60], and the average target-interferer separation angle (AVG-TISA). These will allow a more comprehensive discussion over the separation performance achieved. DRR is calculated as the ratio between the energy carried by the direct sound and the rest of the BRIR. AVG-TISA is the mean lateral angle separating the target source from the interferer, considering all the possible target-interferer combinations. DRR and AVG-TISA characterizing the four datasets are reported in Table II, together with the related RT60s, and DRRs.

Rooms. Vislab was an acoustically treated room at the University of Surrey, where the "Surrey Sound Sphere", having radius of 1.68 m, was assembled. The loudspeakers were clamped on the sphere equator. The dummy head employed was the Cortex Manikin Mk2 Binaural Head and Torso Simulator. Both dummy head and bi-circular microphone array were placed at the sound sphere center.

DWRC is furnished as a living room-like area. Its acoustics are representative of typical domestic living rooms. A Cortex Manikin Mk2 Binaural Head and Torso Simulator sat on a sofa. The bi-circular array was positioned right behind it.

BBC UL is a room at the BBC R&D center, in Salford, UK. Similar to DWRC, it is furnished to resemble a typical living room environment. A Neumann KU100 dummy head was positioned on an armchair and the bi-circular array of microphones was separately measured at the same position.

Since the RT60s related to the three already introduced rooms were similar, an additional room was chosen: Studio1, a large recording studio at the University of Surrey. A Cortex Manikin Mk2 Binaural Head and Torso Simulator was used as dummy head. The loudspeaker positions were selected to have their height similar to the dummy head's. The microphone array was positioned about 2 m far from the dummy head. Therefore, the image source positions found by this array were first manually modified, according to the dummy head position, before being used to initialize the EM. Depending on the

[2] Available at http://cvssp.org/data/s3a, DOI: 10.15126/surreydata.00844867

[3] DOIs: 10.15126/surreydata.00812228 and 10.15126/surreydata.00808465

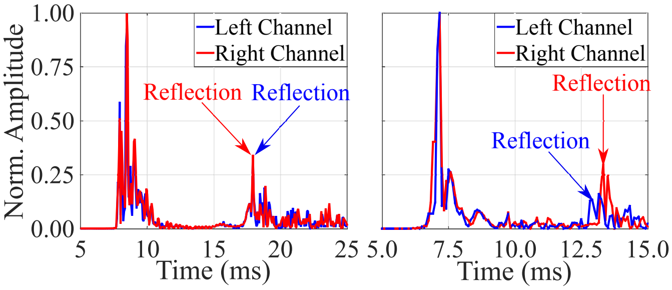

Fig. 6: Two BRIR absolute values, for a frontal source, zoomed into their direct sound and first reflection. On the left, reflection is generated by the floor, thus it arrives at the two ears simultaneously; on the right, reflection arrives from a lateral wall, thus there is a difference in TOAs and amplitudes.

loudspeaker-microphone positions in each room, reflections are generated from either the floor or lateral walls. Examples of RIRs for these two cases are depicted in Fig. 6.

The Utterances. Fifteen utterances, of 3 s length, were randomly selected from the TIMIT acoustic-phonetic continuous speech corpus [61]. For each combination of target source and interferer(s), $U = 15$ random combinations of the fifteen utterances were selected and tested. Therefore, the number of mixtures generated and tested for each dataset is:

$$\Upsilon = \binom{L^{\text{TOT}}}{L} U, \tag{24}$$

where the symbol "()" represents the binomial coefficient, L is the number of sources in the mixture, and L^{TOT} is the total number of loudspeaker positions available in the dataset. The utterances were normalized before applying the convolutions to have the same root mean square energy.

B. Evaluation Metrics

The source to distortion ratio (SDR) metric is based on signal energy ratios, thus, is typically reported in dB. Following Equation (4), the ideal target signal l, that arrives at channel i free from any interference and noise, can be defined as:

$$y^{\text{tar}}_{i,l}(m,\omega) = x_l(m,\omega) I_{i,l}(\omega). \tag{25}$$

Hence, the source $\hat{y}_{i,l}(m,\omega)$, separated by a source separation method as in Equation (23), can be decomposed as [62]:

$$\hat{y}_{i,l}(m,\omega) = y^{\text{tar}}_{i,l}(m,\omega) + E_{\text{interf}} + E_{\text{noise}} + E_{\text{artif}}, \tag{26}$$

where E_{interf} is the interference error term, E_{noise} the noise error term, and E_{artif} errors provided by general artifacts. We chose the SDR, since it emphasizes all the three error terms [62]:

$$\text{SDR} = 10 \log_{10} \frac{||y^{\text{tar}}_{i,l}(m,\omega)||}{||E_{\text{interf}} + E_{\text{noise}} + E_{\text{artif}}||^2}, \tag{27}$$

where $||\cdot||$ represents the Euclidean norm operator. Once the SDR for each of the Υ combinations of sources is obtained, the overall result for the dataset is calculated as their mean $\overline{\text{SDR}} = \frac{1}{\Upsilon}\sum_{\upsilon=1}^{\Upsilon} \text{SDR}_{\upsilon}$,, where υ is the tested mixture index. As clean reference, we employed the target utterance convolved with the related BRIR direct sound. This is also used for the other performance metrics, described below. To extract the direct sound component from the BRIRs, we truncated them by using a Hamming window, centered at the direct sound TOA.

The perceptual evaluation of speech quality (PESQ) has been widely employed to evaluate processed speech quality [63]. This is related to the Mean Opinion Score (MOS) of human subjective assessments, therefore, the PESQ unit of measure is MOS. Before proceeding with the PESQ value calculation, $\hat{y}_{i,l}(m,\omega)$ and $y^{\text{tar}}_{i,l}(m,\omega)$ are aligned in time, in terms of amplitudes and delays, by employing Wiener filters [63]. Through two parameters that model symmetric and asymmetric disturbances, a parametric function is then employed, mapping the differences between the processed version of $\hat{y}_{i,l}(m,\omega)$ and $y^{\text{tar}}_{i,l}(m,\omega)$, to subjective assessment results [63]. The overall PESQ is the mean over the Υ target-interferer combinations, as $\overline{\text{PESQ}} = \frac{1}{\Upsilon}\sum_{\upsilon=1}^{\Upsilon} \text{PESQ}_{\upsilon}$.

Another aspect that has to be evaluated in speech signals separated via source separation algorithms is intelligibility. To do so, we employ the extended short-time objective intelligibility (ESTOI) metric [64]. ESTOI is a function of the separated signal $\hat{y}_{i,l}(m,\omega)$ and the clean reference $y^{\text{tar}}_{i,l}(m,\omega)$. The goal of ESTOI is to produce an index (that we name as ESTOI_{υ}) that is monotonically related to the intelligibility of $\hat{y}_{i,l}(m,\omega)$ [64]. The overall ESTOI is the mean over the Υ target-interferer combinations: $\overline{\text{ESTOI}} = \frac{1}{\Upsilon}\sum_{\upsilon=1}^{\Upsilon} \text{ESTOI}_{\upsilon}$.

C. Control Masks

Performance bounds are needed to perform a fair evaluation of source separation systems [65]. Reference signals are generated from the mixtures, for comparison with the output of the proposed source separation methods. For the lower bound, random TF masks were applied to the mixture. For the upper bound, we chose to calculate the ideal binary mask $M^{\text{IBM}}_l(m,\omega)$, also known as ORACLE mask [66]. It is generated, for each source l, by comparing the l-th signal energy $E^{\text{tar}}_l(n,\omega)$, for each TF bin, with respect to the interferers' $E^{\text{int}}_{l'}(m,\omega)$ in the mixture:

$$M^{\text{IBM}}_l(m,\omega) = \begin{cases} 1, & E^{\text{tar}}_l(n,\omega) > E^{\text{int}}_{l'}(m,\omega), \quad \forall l \neq l' \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

where l' is referred to a source that is other than l. This equation could have also been defined by looking at the source that is louder than the sum of all other sources, instead of the loudest in general. Nevertheless, for our experiments in this article, this would not change the results, since we are focusing on cases where there are only two sources in the mixtures.

D. Source Separation Experiments

The experiments performed were focused on analyzing the source separation performance, employing mixtures composed of two sources ($L = 2$), i.e. target and interferer. These experiments were designed to compare our three novel methods (i.e. IC-MESSL, ER-MESSL and ERIC-MESSL) with the baseline (i.e. MESSL [17]), that models only the direct sound IPD, by

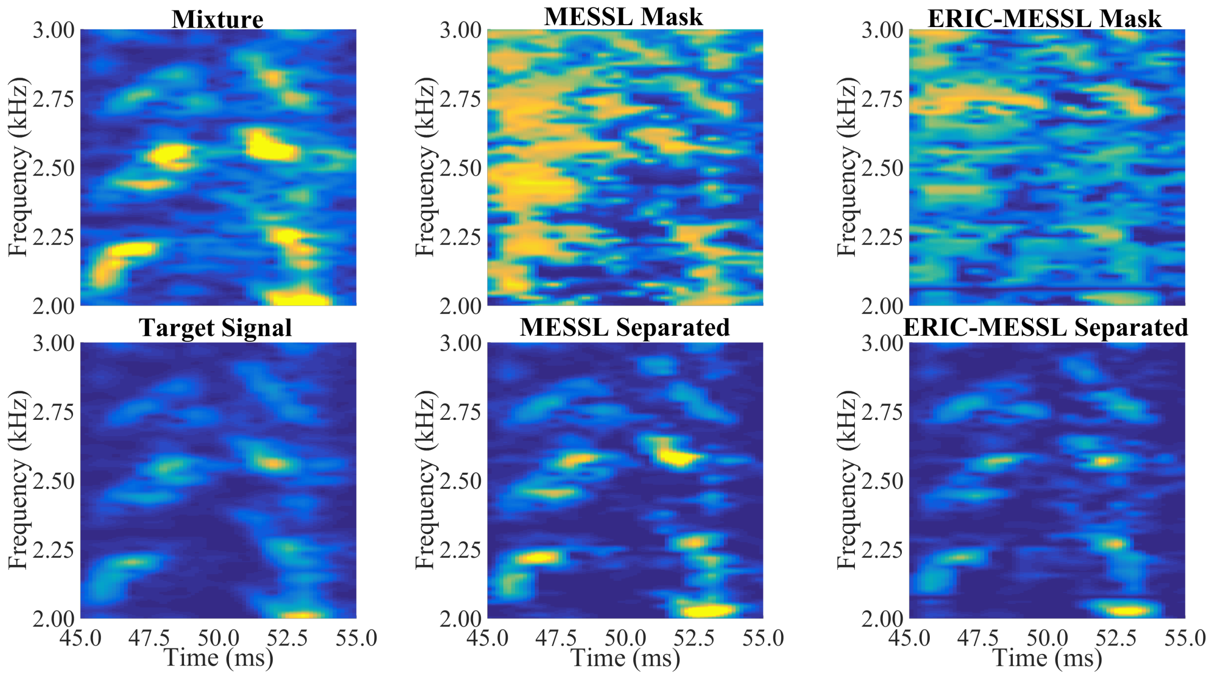

Fig. 7: The top three figures show a zoom into a mixture TF domain absolute value, the related TF masks generated by MESSL, and the TF mask estimated by the proposed ERIC-MESSL. The bottom three figures show the same TF bins of the target signal, the signal separated by MESSL, and ERIC-MESSL, respectively.

TABLE III: $\overline{\text{SDRs}}$ (left) and $\overline{\text{PESQs}}$ (right) obtained by separating the target speech from a two-talker mixture.

$\overline{\text{SDR}}$(dB)	Vislab	DWRC	BBC UL	Studio1	AVG	$\overline{\text{PESQ}}$(MOS)	Vislab	DWRC	BBC UL	Studio1	AVG
Random	−0.43	−0.61	−0.96	0.06	−0.49	**Random**	1.36	1.45	1.45	1.37	1.38
MESSL [17]	4.53	2.54	5.47	0.58	3.28	**MESSL [17]**	1.96	1.93	2.06	1.82	1.94
IC-MESSL	4.80	**2.73**	5.79	0.65	3.49	**IC-MESSL**	1.98	**1.95**	**2.07**	**1.87**	1.97
ER-MESSL	4.98	2.68	5.67	0.67	3.50	**ER-MESSL**	2.00	1.93	2.06	1.83	1.96
ERIC-MESSL	**5.14**	2.70	**5.89**	**0.75**	**3.62**	**ERIC-MESSL**	**2.01**	**1.95**	**2.07**	**1.87**	**1.98**
ORACLE	6.21	5.04	6.82	0.88	4.66	**ORACLE**	2.34	2.45	2.45	1.96	2.30

TABLE IV: $\overline{\text{ESTOIs}}$ obtained by separating the target speech from a two-talker mixture.

$\overline{\text{ESTOI}}$	Vislab	DWRC	BBC UL	Studio1	AVG
Random	0.19	0.17	0.19	0.05	0.15
MESSL [17]	0.28	0.22	0.30	0.07	0.22
IC-MESSL	0.29	0.23	0.31	0.07	0.23
ER-MESSL	0.29	0.23	0.30	0.08	0.23
ERIC-MESSL	**0.29**	**0.24**	**0.31**	**0.10**	**0.24**
ORACLE	0.34	0.29	0.36	0.10	0.27

TABLE V: P-values obtained from a paired t-test that compared the SDRs using MESSL, with the SDRs using each of the three proposed methods.

	Vislab	DWRC	BBC UL	Studio1	AVG
IC-MESSL	**0.0 %**	**0.0 %**	**0.0 %**	7.9 %	**0.0 %**
ER-MESSL	**0.0 %**	8.6 %	**0.0 %**	12.0 %	**0.0 %**
ERIC-MESSL	**0.0 %**	68.9 %	**0.0 %**	**4.1 %**	**0.0 %**

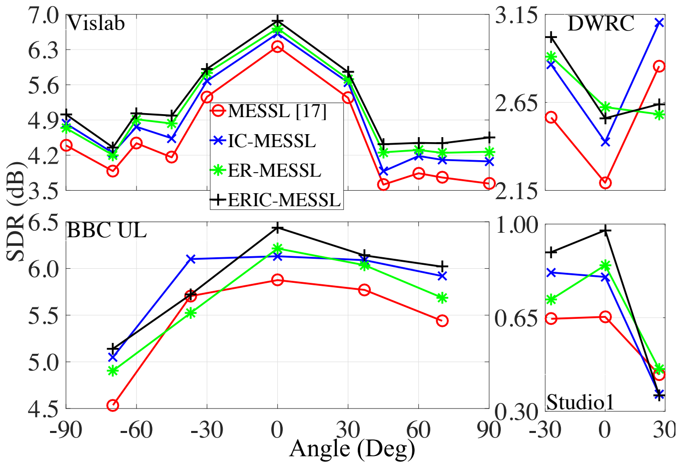

Fig. 8: SDRs obtained by separating a target speech from a two-talker mixture. These results refer to different target source positions, averaged over every interferer position.

calculating the $\overline{\text{SDR}}$ and $\overline{\text{PESQ}}$ scores. Results obtained by applying the ideal masks are also reported as reference.

The number of maximum iterations for the EM algorithm was set, for all the experiments, to be 16. The smoothing factor to calculate the IC was set to be $\kappa = 0.5$. The BRIRs and the utterances introduced in Section VII-A were utilized to create the reverberant mixtures described in Equation (3). Since the BRIRs were recorded having, within the same dataset, constant distance between loudspeakers and listening position, the target-to-interferer ratio (TIR) in the mixture was equal to 0 dB. This choice was made to focus the evaluation on the source separation methods' performance, by avoiding their dependency on the variation in utterance energy and source distance. Furthermore, TIR equal to 0 dB represents a challenging case, where no distinction can be made between target and interferer by looking at their energy levels.

Examples of masks generated by MESSL and the proposed ER-MESSL are depicted in Fig. 7. We can observe that differences between the two masks are pronounced. These differences lead to the TF representation of the signal separated through ERIC-MESSL to be more similar to the groundtruth target signal, when compared to MESSL's separated signal.

For our experiments we used the open-source code of MESSL, where we set to the frequency-dependent parameter modeling option. The tested MESSL model, hence, includes a non-parametric modeling of the "impurities" around the direct sound component. Nevertheless, in MESSL, the early reflection model was not directly defined through parameters. Instead, we drive our system to extract the information related to both direct sound and early reflection. We also use the frequency-dependent parameter modelling (pre-implemented in MESSL) to model the impurities around the estimation.

E. Source Separation Results

The SDR side of Table III shows that ERIC-MESSL, the proposed source separation method that models both the comb filter and IC, outperforms the baseline (i.e. the MESSL method [17]), when applied to any of the four datasets. Furthermore, it provides better performance if compared to the other proposed methods. However, for the DWRC dataset, the other proposed method IC-MESSL produces the highest SDR. This is due to strong reflections arriving from different directions with respect to the direct sound, which corresponds to a lower impact of the comb filter effect [15]. Observing PESQ in Table III, in general, the two proposed methods that model the IC (i.e. IC-MESSL and ERIC-MESSL) have comparable results, and are both better than the other methods. However, in acoustically controlled environments, such as Vislab, the first reflection direction is initialized more accurately by ISDAR, and the comb filter model performs better, with ERIC-MESSL having a higher PESQ. This shows the importance of an accurate initialization of the GMM parameters. Similar trends are reported in Table IV, where the ESTOIs related to the proposed methods are greater than the baseline. ESTOI results show ERIC-MESSL to be the best proposed method, providing a greater intelligibility for every dataset.

In general, DWRC and Studio1 are more challenging datasets, producing low SDR, PESQ and ESTOI values for every tested method. The reason can be found in Table II: they have low DRRs and narrow AVG-TISAs. Low DRR entails difficulties for each of the algorithms, since the IPD curve, that was described in Fig. 3, is highly distorted by the strong reverberation. At the same time, narrow AVG-TISA affects the overall results, since small angles between target and interferer correspond to small variations between the IPD and ILD cues related to the two signals in the mixture.

Assuming the Υ SDR results of each dataset as being normally distributed, the paired t-test was performed to determine whether the results, generated through the three proposed methods, are significantly different from the ones obtained by MESSL. In Table V, the p-values are reported. They represent the probability of rejecting the hypothesis that the two sets under investigation are statistically different (i.e. a low p-value means that the two sets are statistically different). By looking at the results averaged over all the datasets by comparing every tested sample, with a significance level of 5 %, we can state that the results of IC-MESSL, ER-MESSL, and ERIC-MESSL are statistically different from those of MESSL. Moreover, by looking at each dataset singularly, results show that the three proposed methods are statistically different from MESSL in Vislab and BBC UL. However, in DWRC and Studio1 this is valid only for IC-MESSL and ERIC-MESSL, respectively. These results confirm what was already shown in Table III, where the improvement given by IC-MESSL, ER-MESSL, and ERIC-MESSL is, in general, higher in BBC UL and Vislab than in DWRC and Studio1. The statistical significance of the results demonstrates the key point of the manuscript, which is about the importance of considering early reflection information when constructing a source separation model.

For the four datasets, the SDR results can also be reported as a function of the target source location, as shown in Fig. 8. For each target source position, within the dataset, the SDR is calculated by considering each of the correspondent interferer locations. Then, the obtained SDRs are averaged over these interferer positions, leading to one result for each target source location. Due to the cone of confusion, which is well-known for IPD based localization methods [67], it is not possible to discriminate between the IPD of two sources lying at the same lateral angle. Therefore, results are reported in terms of lateral angle, rather than azimuth. Apart from DWRC, the general trend of the results suggests that source separation performs better in situations where the target is frontal to the listener. This situation was, in fact, one of the classical assumptions made to evaluate source separation methods [17]. By reporting results as in Fig. 8, we overcome this assumption. The proposed ERIC-MESSL performs better than the others for almost every position of the target source. For the few positions where it is not the best, either the proposed IC-MESSL or ER-MESSL has higher SDRs. In DWRC, the loudspeaker positioned at 27° stood next to a chest of drawers, that produces scattering. This conflicts with the overall assumption of having reflections with a dominant specular component. Therefore, the localization of the first reflection, for modeling the comb filter, is affected by estimation errors. Similar to 0° in DWRC and 27° in Studio1, for $-37°$ in BBC UL, strong lateral reflections arrive before those from the direct sound direction, making the IC dominate the comb filtering effect [15]. Similar results can be observed in Fig. 9, where the PESQ results are reported as a function of the target source location. It is evident how the proposed ERIC-MESSL, which combines the two proposed models, outperforms, in general the baseline MESSL [17]. Furthermore, these PESQ results also show what was already observed in Fig. 8 for the SDRs (and discussed above), ERIC-MESSL mainly suffers when early reflections are not completely specular.

The majority of the setups that we tested, had a certain

TABLE VI: $\overline{\text{SDRs}}$ (left) and $\overline{\text{PESQs}}$ (right) obtained by separating the target speech from a two-talker mixture. These results are calculated by considering only recording setups where direct sound and first reflection have same DOA.

$\overline{\text{SDR}}$(dB)	DWRC	BBC UL	Studio1	AVG	$\overline{\text{PESQ}}$(MOS)	DWRC	BBC UL	Studio1	AVG
MESSL [17]	2.00	5.22	0.55	2.59	**MESSL [17]**	1.86	2.04	1.87	1.92
IC-MESSL	2.26	5.57	0.68	2.84	**IC-MESSL**	**1.88**	2.05	2.92	1.95
ER-MESSL	2.43	5.60	0.80	2.94	**ER-MESSL**	1.86	2.06	1.92	1.95
ERIC-MESSL	**2.70**	**5.80**	**0.87**	**3.12**	**ERIC-MESSL**	**1.88**	**2.07**	**1.95**	**1.97**

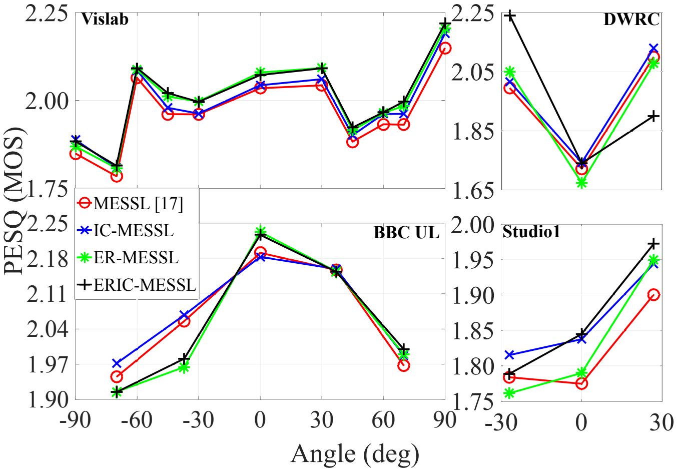

Fig. 9: PESQs obtained by separating a target speech from a two-talker mixture. These results refer to different target source positions, averaged over every interferer position.

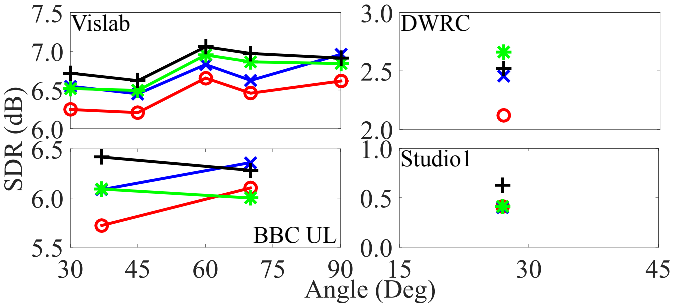

Fig. 10: SDRs for different interferer positions, fixing target at 0°. The black vertical crossed lines refer to ERIC-MESSL, the red circled lines to MESSL [17], the green starred lines to ER-MESSL, and the blue crossed lines to IC-MESSL.

TABLE VII: Evaluation results for the deep learning based methods over Vislab, in terms of SDR, PESQ and ESTOI.

	SDR	PESQ	ESTOI
Direct sound information	8.33	2.51	0.70
Direct sound and early reflection info	**8.80**	**2.59**	**0.73**

configuration that produced, as the first reflection, the one corresponding to the floor (i.e. having same azimuth as the direct sound). Nevertheless, in BBC UL, DWRC, and Studio1, there are cases where the first arriving reflection has a different direction of arrival (DOA) than the direct sound (i.e. coming from a lateral wall). The proposed model does not make any assumption regarding the direction of the reflections, however, the condition that better matches the idea behind it (i.e. a strong comb filter effect) is given by the case of direct sound and early reflection coming from the same direction. To better show the strength of the proposed models, in Table VI, we show the results of the experiments by considering only those situations where direct sound and first reflection have the same DOA. These results show that our methods outperform MESSL with a much wider difference than the overall results in Table III, and ERIC-MESSL is the best.

To analyze the effect of separation angle, the source separation performance was calculated with the frontal loudspeaker (0° azimuth) as the target source, and varying the interferer. The results are reported in Fig. 10, as is typical in the literature for source separation [17], [41], [42]. This kind of visualization allows a better understanding of the source separation performance by varying TISA. By observing the results of Vislab and BBC UL (datasets having loudspeaker positions around the listener), the proposed ERIC-MESSL consistently provides the highest performance. However, for the extreme cases of TISA (i.e. 90° in Vislab and 70° in BBC UL), the proposed IC-MESSL performs better. This behavior is best seen in the proposed ER-MESSL results. As for ERIC-MESSL, ER-MESSL is better than IC-MESSL for almost every TISA, apart from the extreme cases (i.e. 90° in Vislab and 70° in BBC UL). Therefore, we can conclude that the comb filter is, on average, more effective than the IC, apart from large TISAs. For both DWRC and Studio1, all the methods show degradation at low TISA. This is a common source separation problem [17]. Studio1 is also confirmed to be problematic, with SDR lower than 1 dB, for every method.

Regarding the overall computational complexity, the average run time, for a code run in MATLAB R2014b on Intel(R) Core(TM)i7-2600 CPU @ 3.40GHz, 16GB RAM PC is 55 s for ERIC-MESSL and 8 s for MESSL [17]. The parameters are searched within a 7-D space in ERIC-MESSL, making it less efficient than MESSL, where the space was one dimensional.

Early Reflections and Deep Learning. We now evaluate a DNN-based method that is representative of state-of-the-art approaches in speech separation. We modified this reference method to test the key point behind our main work: that the inclusion of early reflection information into source separation methods improves the performance. This test is intended to examine the potential for exploiting this information using a DNN approach, and give a preliminary validation. Further experiments are needed to explore the best way to incorporate early reflection information within DNN architectures for source separation, beyond the present preliminary integration.

The selected pipeline is based on the classic multilayer perceptron (MLP) architecture, as presented in [68]. A similar architecture can be also found in [69]. In our implementation, the MLP has two hidden layers, containing 1024 leaky rectified linear units (ReLU) each. We employed batch normalisation (BN) layers [70] to accelerate convergence, and Adam optimizer [71] with He initialization [72]. The binary crossentropy was used as loss function. The mini-batch size was set to 1000.

Recordings from two male speakers and two female speakers in the TIMIT dataset [61] were used for our experiment. For each of these speakers, ten sentences were randomly selected. The binaural mixtures were generated by convolving the randomly chosen utterances with BRIRs recorded in Vislab. The BRIRs used were the ones recorded for the angles at 0°, $\pm 30^\circ$ and $\pm 60^\circ$. To create the mixtures, each of the 4 speakers was combined to the other 3. For each of these 12 combinations, we associated the 10 sentences. In terms of the product rule for counting, this makes a total of 1200 utterance combinations. Regarding the BRIRs, each of the 5 DOAs was combined to the other 4, making a total of 20 combinations. Convolving utterances with BRIRs, we obtain 24000 mixtures: 19200 were randomly selected for training, the rest for testing.

These 24000 samples comprising the dataset represent all combinations of the BRIR directions convolved with the individual utterances. A distinct set of direction-utterance samples was used for testing and training, although all directions and some utterances did overlap (but not any specific combination). The performance of the methods tested here would likely decrease when generalizing to new unseen utterances and BRIRs, which is however beyond the scope of the present tests. In fact, as mentioned above, this DNN experiment is to demonstrate that, by adding information about early reflections, supervised deep learning based source separation method can also be improved, over the case where only the direct sound is considered, as we observed in the main novelty of this article, i.e. the GMM based unsupervised method.

The training was performed by providing the features related to the IPD as input to the network, and matching with the ORACLE masks in output. In both models, the IPD features were calculated through the approach in Sections III and IV. To evaluate the improvement given by the early reflection information, we have trained one model that considers only the direct sound information [68], and a novel one which we propose to also incorporate the early reflections. The ORACLE masks in output to the training stage were generated from Equation (28), by considering $E_l^{\mathrm{tar}}(n,\omega)$ and $E_l^{\mathrm{int}}(n,\omega)$ related to the direct sound for the model used as in [68], and direct sound plus early reflections for our model. This was done by segmenting the related BRIRs through a Hamming window (5 ms, and 30 ms, respectively).

During the test, the masks predicted by the networks are used to separate the sounds, by employing Equation (23). Results are reported in Table VII. There, it is shown how the model containing information about the early reflections offers better performance with respect to the pipeline which considers only direct sound, for every metric (i.e. SDR, PESQ and ESTOI). This has demonstrated the key idea of the manuscript: early reflections carry important information that is helpful for improving the performance of speech separation models, including both unsupervised (e.g. MESSL) and supervised techniques (e.g. DNNs). However, it is important to stress that MESSL [17] and the methods proposed in Section VI are unsupervised techniques, hence do not need any labeling. Therefore, it is inappropriate to directly compare the results in Table VII with those in Tables III and IV.

VIII. Conclusion

Two room properties (i.e. early reflections and late reverberation) have been modeled for source separation. Depending on whether they are modeled individually or together, three novel source separation methods have been proposed: ER-MESSL, that models the comb filter effect; IC-MESSL, that models the IC; ERIC-MESSL, that combines the two models together.

Experiments were performed by recording four reverberant environments, and comparing the source separation performance of the proposed methods with MESSL's [17]. In general, the proposed ERIC-MESSL outperforms all the other methods. With respect to MESSL, the improvement given by ERIC-MESSL, averaged over the four tested datasets, is about 10 % for SDR and 2 % for PESQ. It was also shown, by running t-tests, that the ERIC-MESSL results are statistically different from MESSL's. Moreover, this experimental analysis revealed that low DRRs and narrow AVG-TISAs led to a degradation of the results. In addition, results were also observed by varying both the target source and interferer positions. Also in this case, it was consistently observed that ERIC-MESSL is, in general, the better model. We conclude that modeling together the comb filter effect and IC is helpful for improving the performance of classical source separation methods. Furthermore, we have also reported an experiment undertaken by including early reflection information into a DNN based state-of-the-art source separation method. Results showed a great improvement, thus confirming the importance of incorporating the early reflection information into both unsupervised and supervised source separation methods.

Future work may be conducted on extending the methods to multichannel arrays of microphones. Furthermore, a combination of audio-visual sensing may be explored, to tackle problematic scenarios where the interferer has a higher level than the target. The proposed models could also be applied to other popular approaches, such as NMF.

Acknowledgments

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and BBC as part of the BBC Audio Research Partnership. The authors would like to thank the reviewers and the associate editor for their helpful comments to improve the article.

References

[1] A. Sutin, B. Bunin, N. Sedunov, L. Fillinger, M. Tsionskiv, and M. Bruno, "Stevens passive acoustic system for underwater surveillance," in *Proc. of the International WaterSide Security Conference*, Carrara, Italy, 2010.

[2] M. Ungureanu, C. Bigan, R. Strungaru, and V. Lazarescu, "Independent component analysis applied in biomedical signal processing," *Measurement Science Review*, vol. 4, no. 2, pp. 1–8, 2004.

[3] A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique," *International Journal of Document Analysis*, vol. 10, no. 1, pp. 17–25, 2007.

[4] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[5] M. A. Akeroyd, J. Chambers, D. Bullock, Palmer A. R., and A. Q. Summerfield, "The binaural performance of a cross-talk cancellation system with matched or mismatched setup and playback acoustics," *J. Acoustical Society of America*, vol. 121, no. 2, pp. 1056–1069, 2007.

[6] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[7] E. W. Healy, S. E. Yoho, Y. Wang, and Wang D., "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoustical Society of America*, vol. 134, no. 4, pp. 3029–3038, 2013.

[8] C. Crocco, M. Cristiani, A. Trucco, and V. Murino, "Audio surveillance: a systematic review," *ACM Computing Surveys*, vol. 48, no. 4, pp. 52:1–52:46, 2016.

[9] Q. Liu, W. Wang, P. J. B. Jackson, and T. J. Cox, "A source separation evaluation method in object-based spatial audio," in *Proc. of the 23rd European Signal Processing Conference (EUSIPCO)*, Nice, France, 2015.

[10] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. on Advances in Signal Processing*, vol. 2016, no. 1, pp. 7:1–7:19, 2016.

[11] H. Kuttruff, *Room Acoustics - Fifth edition*, Spon press, 2009.

[12] B. Blesser, "An interdisciplinary synthesis of reverberation viewpoints," *J. Audio Engineering Society*, vol. 49, no. 10, pp. 867–903, 2001.

[13] V. Välimäki, J. A. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1421–1448, 2012.

[14] M. Barron, "The subjective effects of first reflections in concert halls - the need for lateral reflections," *J. of Sound and Vibration*, vol. 15, no. 4, pp. 475–494, 1971.

[15] T. Lokki, J. Pätynen, T. Sakar, S. Siltanen, and L. Savioja, "Engaging concert hall acoustics is made up of temporal envelope preserving reflections," *J. Acoustical Society of America Express Letters*, vol. 129, no. 6, pp. EL223–EL228, 2011.

[16] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[17] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[18] S. Bech, "Spatial aspects of reproduced sound in small rooms," *J. Acoustical Society of America*, vol. 103, no. 1, pp. 434–445, 1998.

[19] A. Alinaghi, W. Wang, and P. J. B. Jackson, "Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.

[20] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang, "Acoustic reflector localization: novel image source reversion and direct localization methods," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 2, pp. 296–309, 2017.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Society of America*, vol. 4, no. 65, pp. 943–950, 1979.

[22] G-J. Jang and T-W. Lee, "A maximum likelihood approach to single-channel source separation," *J. of Machine Learning Research*, vol. 23, pp. 1365–1392, 2003.

[23] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. of Interspeech*, Pittsburgh, USA, 2006.

[24] S. Arberet, A. Ozerov, N. Q. K. 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 *Proc. of the 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, Kuala Lumpur, Malaysia, 2010.

[25] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Latent Variable Analysis and Signal Separation: 10th International Conference (LVA/ICA)*. Tel Aviv, Israel, 2012, pp. 322–329, Springer Berlin Heidelberg.

[26] P. Smaragdis, C. Févotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[27] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ICA and time-frequency masking," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2165–2173, 2006.

[28] A. Ozerov and C. Févotte, "Multichannel nonegative 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.

[29] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013.

[30] L. Wang, J. D. Reiss, and A. Cavallaro, "Over-determined source separation and localization using distributed microphones," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1573–1588, 2016.

[31] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[32] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2165–2173, 2006.

[33] I. Dokmanić, R. Scheibler, and M. Vetterli, "Raking the cocktail party," *IEEE J. of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 825–836, 2015.

[34] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[35] A. 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.

[36] X.-L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 5, pp. 967–977, 2016.

[37] J. Du, Y. Tu, L-R. Dai, and C.-H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 8, pp. 1424–1437, 2016.

[38] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 7, pp. 1535–1546, 2017.

[39] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[40] P. M. Hofman and J. Van Opstal, "Spectro-temporal factors in two-dimensional human sound localization," *J. Acoustical Society of America*, vol. 103, no. 5, pp. 2634–2648, 1998.

[41] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

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

[43] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International Journal of Neural Systems*, vol. 25, no. 1, 2015.

[44] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1867–1871, 2010.

[45] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 13, no. 5, pp. 882–895, 2005.

[46] F. Nesta and M. Omologo, "Convolutive underdetermined source separation through weighted interleaved ICA and spatio-temporal source correlation," in *Latent Variable Analysis and Signal Separation: 10th International Conference (LVA/ICA)*. Tel Aviv, Israel, 2012, pp. 222–230, Springer Berlin Heidelberg.

[47] S. Makino, H. Sawada, and T. W. Lee, *Blind Speech Separation*, Springer, 2007.

[48] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 30, pp. 620–633, 2014.

[49] R. Scheibler, D. Di Carlo, A. Deleforge, and I. Dokmanic, "Separake: Source separation with a little help from echoes," *arXiv: CoRR*, vol. abs/1711.06805, 2017.

[50] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*, John Wiley & Sons, Ltd, 2018.

[51] G. J. Brown and M. Cook, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[52] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 1, pp. 216–228, 2007.

[53] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 895–910, 2010.

[54] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, 2004.

[55] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1732–1745, 2010.

[56] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 32, no. 4, pp. 474–484, 2002.

[57] H. Kim, L. Remaggi, P. J. B. Jackson, F. M. Fazi, and A. Hilton, "3D room geometry reconstruction using audio-visual sensors," in *Proc. of the Conference on 3D Vision (3DV)*, Qingdao, China, 2017.

[58] B. D. VanVeen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Acoustic, Speech and Signal Processing Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[59] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. of the 108th Audio Engineering Society Convention (AES)*, Paris, France, 2000.

[60] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J. Acoustical Society of America*, vol. 112, no. 5, Pt. 1, pp. 2110–2117, 2002.

[61] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," Tech. Rep., NIST Interagency, 1993.

[62] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[63] P. C. Loizou, *Speech Enhancement: Theory and Practice - Second Edition*, CRC Press, 2013.

[64] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[65] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010): achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[66] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., chapter 12, pp. 181–197. Kluwer Academic, 2005.

[67] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[68] Q. Liu, Y. Xu, P. J. B. Jackson, W. Wang, and P. Coleman, "Iterative deep neural networks for speaker-independent binaural blind speech separation," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Canada, 2018.

[69] Z.-Q. Wang and D. Wang, "On spatial features for supervised speech separation and its application to beamforming and robust ASR," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Canada, 2018.

[70] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of the International Conference on Machine Learning*, Lille, France, 2015.

[71] D. P. Kingma and J. L. Ba, "ADAM: A method for stochastic optimization," in *Proc. of the International Conference on Learning Representations (ICLR)*, San Diego, USA, 2015.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *Proc. of the International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015.

Luca Remaggi is Audio Research Engineer at Creative Labs, UK, working on cutting edge spatial audio products. Between 2017 and 2019, he was Research Fellow at the Centre for Vision, Speech and Signal Processing, University of Surrey, UK, where he also pursued his PhD, in 2017. His research interest was to investigate the multipath sound propagation combining acoustic and visual data, for applications in spatial audio and source separation. He received the B.Sc. and M.E. degrees in Electronic Engineering from Università Politecnica delle Marche, Italy, in 2009 and 2012, respectively. During his M.E., he has been an intern at the Department of Signal Processing and Acoustics, Aalto University, Finland, where he focused on the sound synthesis of musical instruments.

Philip Jackson is Reader in Machine Audition at the Centre for Vision, Speech & Signal Processing (CVSSP, University of Surrey, UK) with MA in Engineering (Cambridge University, UK) and PhD in Electronic Engineering (University of Southampton, UK). His broad interests in acoustical signals have led to research contributions in sound field control, modeling speech articulation, acoustics and recognition, in audio-visual perception, blind source separation, and spatial audio reverberation, capture, reproduction and quality evaluation [h-index 22; Google Scholar: bit.ly/2oTRw1C]. He led one of four research streams on object-based spatial audio in the S3A programme grant funded in the UK by EPSRC, and enjoys listening.

Wenwu Wang (M02SM11) was born in Anhui, China. He received the B.Sc. degree in 1997, the M.E. degree in 2000, and the Ph.D. degree in 2002, all from Harbin Engineering University, China. He then worked in Kings College London (2002-2003), Cardiff University (2004-2005), Tao Group Ltd. (now Antix Labs Ltd.) (2005-2006), and Creative Labs (2006-2007), before joining University of Surrey, UK, in May 2007, where he is currently a Professor in Signal Processing and Machine Learning, and a Co-Director of the Machine Audition Lab within the Centre for Vision Speech and Signal Processing. He was a Visiting Scholar at Ohio State University, USA, in 2008. He has been a Guest Professor on Machine Perception at Qingdao University of Science and Technology, China, since 2018. His current research interests include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He has (co)-authored over 250 publications in these areas. He and his team have won the Best Paper Award on LVA/ICA 2018, the Best Oral Presentation on FSDM 2016, the Top Paper Award in IEEE ICME 2015, Best Student Paper Award shortlists on IEEE ICASSP 2019 and LVA/ICA 2010. His papers are among the Most Downloaded Papers in IEEE/ACM Transactions on Audio Speech and Language Processing in 2018 and 2019, and Featured Articles in IEEE Transactions on Signal Processing 2013. As a team member, he achieved the 2nd place (among 23 teams) in the DCASE 2019 Challenge Sound event localization and detection, the 3rd place (among 558 submitted systems) in the 2018 Kaggle Challenge "Free-sound general purpose audio tagging", the 1st place (among 35 submitted systems) in the 2017 DCASE Challenge on "Large-scale weakly supervised sound event detection for smart cars", the TVB Europe Award for Best Achievement in Sound in 2016 and the finalist for GooglePlay Best VR Experience in 2017, and the Best Solution Award on the Dstl Challenge "Under-sampled signal signal recognition" in 2012. He is a Senior Area Editor (2019-) for IEEE Transactions on Signal Processing and an Associate Editor (2019-) for EURASIP Journal on Audio Speech and Music Processing. He was an Associate Editor (2014-2018) for IEEE Transactions on Signal Processing. He was a Publication Co-Chair for ICASSP 2019, Brighton, UK.