Schwartz and Gannot *EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:43
https://doi.org/10.1186/s13636-021-00228-1

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

A recursive expectation-maximization algorithm for speaker tracking and separation

Ofer Schwartz[1] and Sharon Gannot[2*]

Abstract

The problem of blind and online speaker localization and separation using multiple microphones is addressed based on the recursive expectation-maximization (REM) procedure. A two-stage REM-based algorithm is proposed: (1) multi-speaker direction of arrival (DOA) estimation and (2) multi-speaker relative transfer function (RTF) estimation. The DOA estimation task uses only the time frequency (TF) bins dominated by a single speaker while the entire frequency range is not required to accomplish this task. In contrast, the RTF estimation task requires the entire frequency range in order to estimate the RTF for each frequency bin. Accordingly, a different statistical model is used for the two tasks. The first REM model is applied under the assumption that the speech signal is sparse in the TF domain, and utilizes a mixture of Gaussians (MoG) model to identify the TF bins associated with a single dominant speaker. The corresponding DOAs are estimated using these bins. The second REM model is applied under the assumption that the speakers are concurrently active in all TF bins and consequently applies a multichannel Wiener filter (MCWF) to separate the speakers. As a result of the assumption of the concurrent speakers, a more precise TF map of the speakers' activity is obtained. The RTFs are estimated using the outputs of the MCWF-beamformer (BF), which are constructed using the DOAs obtained in the previous stage. Next, using the linearly constrained minimum variance (LCMV)-BF that utilizes the estimated RTFs, the speech signals are separated. The algorithm is evaluated using real-life scenarios of two speakers. Evaluation of the mean absolute error (MAE) of the estimated DOAs and the separation capabilities, demonstrates significant improvement w.r.t. a baseline DOA estimation and speaker separation algorithm.

Keywords: Array processing, Recursive expectation-maximization algorithm, DOA estimation, LCMV beamforming

1 Introduction

Multi-speaker separation techniques, utilizing microphone arrays, have attracted the attention of the research community and the industry in the last three decades, especially in the context of hands-free communication systems. A comprehensive survey of state-of-the-art multichannel audio separation methods can be found in [1–3].

A commonly used technique for source extraction is the LCMV-BF [4, 5], which is a generalization of the minimum variance distortionless response (MVDR)-BF [6]. In [7], the LCMV-BF was reformulated by substituting the simple steering vectors based on the direct-path with the RTFs encompassing the entire reflection pattern of the acoustic propagation. The authors also presented a method to estimate the RTFs, based on the generalized eigenvalue decomposition (GEVD) of the power spectral density (PSD) matrices of the received signals and the background noise. A multi-speaker LCMV-BF was proposed in [8] to simultaneously extract all individual speaker signals. Moreover, the estimation procedure of the speakers' PSDs was facilitated by the decomposition of the multi-speaker MCWF into two stages, namely multi-speaker LCMV-BF and a subsequent multi-speaker post-filter.

*Correspondence: sharon.gannot@biu.ac.il
[2]Faculty of Engineering, Bar-Ilan University, 5290002 Ramat-Gan, Israel
Full list of author information is available at the end of the article

In [7, 8], the RTFs were estimated using time intervals comprising each of the desired speakers separately assuming a static scenario. Practically, these time intervals need to be detected from data and cannot be assumed to be known.

In [9], time-frames dominated by each of the speakers were identified by estimating the DOA for each frame using clustering of a time-series of steered response power (SRP) estimates. In [10, 11], these frames were identified by exploiting convex geometry tools on the recovered simplex of the speakers' probabilities or the correlation function between frames [12]. In [13, 14], a dynamic, neural-network-based, concurrent speaker detector was presented to detect single speaker frames. A library of these RTFs was collected for constructing an LCMV-BF and for further spatial identification of the speakers. In [15], the speech sparsity in the short-time Fourier transform (STFT) domain was utilized to track the DOAs of multiple speakers using a convolutional neural network (CNN) applied to the instantaneous RTF estimate. Speaker separation was obtained, as a byproduct of the tracking method, by the application of TF masking.

Unfortunately, the existence of single speaker dominant frames is not always guaranteed for simultaneously active speakers. Furthermore, for moving speakers the RTFs estimated by these frames may be irrelevant for subsequent processing. In [16], the sparsity of the speech signal in the STFT domain was utilized to model the frequency bins with complex-Gaussian mixture p.d.f. and the RTFs were offline estimated as part of an expectation-maximization (EM)-MoG procedure. In [17], an offline blind estimation of the acoustic transfer functions was presented using the non-negative Matrix Factorization and the EM algorithm. In [18, 19], an offline estimation of the acoustic transfer functions was done by estimating a latent variable representing the speaker activity pattern. In [20], an *online* estimation of the blocking matrices (required for the generalized sidelobe canceler implementation of the MVDR-BF) associated with each of the speakers was carried out by clustering the DOA estimates from all TF bins. In [21, 22], an online time–frequency masking has been proposed to estimate the RTFs using the EM algorithm and without any prior information on the array geometry or the plane wave assumption.

Common DOA estimators are based on the SRP-phase transform (PHAT) [23], the multiple signal classification (MUSIC) algorithm [24], or Model-based expectation-maximization source separation and localization (MESSL) [25]. In [26–28], the microphone observations were modeled as a mixture of high-dimensional complex-Gaussian with zero-mean, and a spatial covariance matrix that consists of both the speech and the noise power spectral densities (PSDs) was assumed. In [29], a DOA tracking procedure was proposed by applying the Cappé and Moulines recursive EM (CREM) algorithm. Recursive equations for the DOA probabilities and the candidate speakers PSDs were derived, which facilitated online DOA tracking of multiple speakers.

In this paper, an online and blind speaker separation procedure is presented. Multiple RTFs updating is performed using REM model that assumes concurrent activity of speakers. New links are established between the direct-path phase differences and the full RTF of each speaker. The dominant DOAs in each frame are estimated using a dedicated REM procedure. Then, in each frame, the RTFs are initialized by the direct-path phase difference (using the corresponding DOA). Finally, the full RTFs are re-estimated using the LCMV outputs. By examining the LCMV outputs, frames dominated by a single speaker can be detected by comparing the energy of each LCMV output. As a practical improvement, the RTF of a speaker is updated only when the LCMV output corresponding to the relevant speaker is relatively high. Finally, the LCMV-BF is re-employed using the estimated RTFs.

The direct-path phase differences are set using the speakers DOA estimated by an online preliminary stage of multiple concurrent DOA estimation. In this stage, assuming J speakers, J dominant DOAs are estimated in each frame using a novel version of the MoG-REM. Only for the DOA estimation, the sparse nature of the speech is exploited (while it has been proven to be efficient with DOA estimation). The output of many multiple-speaker DOA estimators is actually a probability for an existence of speaker in each DOA, while the final DOA of the speakers is still not clear. In this paper, we design an REM-based concurrent DOA estimation that consists only of J Gaussians. Rather than estimating the probabilities, the DOAs of the speakers are directly estimated using the REM algorithm.

The remainder of this paper is organized as follows. In Section 2, the speaker separation problem is formulated. In Section 3, the proposed dual-stage algorithm is overviewed. In Section 4, the REM procedure for the speaker separation is derived. In Section 5, the REM procedure for the multiple-speaker DOA estimation is derived. In Section 6, the performance of the proposed algorithm is evaluated. Section 7 is dedicated to concluding remarks.

2 Problem formulation

The problem is formulated in the STFT domain, where ℓ and k denote the time-frame index and the frequency-bin index, respectively. The signal observed at the ith microphone is modeled by:

$$Y_i(\ell,k)=\sum_{j=1}^{J} G_{i,j}(\ell,k)S_j(\ell,k)+V_i(\ell,k),\ i=1,\ldots,N \tag{1}$$

where $S_j(\ell,k)$ is the speech signal of the jth speaker, as received by the reference microphone (chosen arbitrary as microphone #1), $G_{i,j}(\ell,k)$ is the RTF from the jth speaker to the ith microphone w.r.t. the reference microphone, and $V_i(\ell,k)$ denotes ambient noise. The number of microphones is N and the number of sources of interest is J.

By concatenating the signals and RTFs in vectors, (1) can be recast as:

$$\mathbf{y}(\ell,k) = \mathbf{G}(\ell,k)\mathbf{s}(\ell,k) + \mathbf{v}(\ell,k), \tag{2}$$

where:

$$\mathbf{y}(\ell,k) = \left[\, Y_1(\ell,k)\ Y_2(\ell,k)\ \ldots\ Y_N(\ell,k) \,\right]^\top \tag{3a}$$

$$\mathbf{G}(\ell,k) = \left[\, \mathbf{g}_1(\ell,k)\ \mathbf{g}_2(\ell,k)\ \ldots\ \mathbf{g}_J(\ell,k) \,\right]^\top, \tag{3b}$$

$$\mathbf{g}_j(\ell,k) = \left[\, G_{1,j}(\ell,k)\ G_{2,j}(\ell,k)\ \ldots\ G_{N,j}(\ell,k) \,\right]^\top \tag{3c}$$

$$\mathbf{s}(\ell,k) = \left[\, S_1(\ell,k)\ S_2(\ell,k)\ \ldots\ S_J(\ell,k) \,\right]^\top, \tag{3d}$$

$$\mathbf{v}(\ell,k) = \left[\, V_1(\ell,k)\ V_2(\ell,k)\ \ldots\ V_N(\ell,k) \,\right]^\top. \tag{3e}$$

The ambient noise is modeled as a zero-mean Gaussian vector with a PSD matrix $\Phi_{\mathbf{v}}(k)$:

$$f\left(\mathbf{v}(\ell,k);\Phi_{\mathbf{v}}(\ell,k)\right) = \mathcal{N}^C(\mathbf{v}(\ell,k);\mathbf{0},\Phi_{\mathbf{v}}(\ell,k)). \tag{4}$$

where:

$$\mathcal{N}^C(\mathbf{z};\mathbf{0},\Phi) = \frac{1}{\pi^N|\Phi|}\exp\left(-\mathrm{Tr}\left[\Phi^{-1}\mathbf{z}\mathbf{z}^{\mathrm{H}}\right]\right), \tag{5}$$

$\mathbf{z}$ denotes a Gaussian vector, Φ is a PSD matrix, Tr [] denotes trace operation and $|\cdot|$ denotes the matrix-determinant operation. The individual speech signals $S_j(\ell,k)$ are also modeled as independent and zero-mean Gaussian processes with variance $\phi_{S_j}(\ell,k)$,

$$f\left(S_j(\ell,k);\phi_{S_j}(\ell,k)\right) = \mathcal{N}^C(S_j(\ell,k);0,\phi_{S_j}(\ell,k)).$$

In the following sections, the frequency index k and time index ℓ are omitted for brevity, whenever no ambiguity arises.

3 Algorithm overview

The proposed algorithm comprises two stages as detailed below and summarized in Fig. 1.

3.1 Speaker extraction

The goal of this paper is to estimate the individual speech signals S_j of the dominant J speakers (while the number of speakers J is assumed fixed and known) using the multi-speaker MCWF or the multi-speaker LCMV beamformer [8].

$$\mathbf{s}_{\mathrm{LCMV}}(\mathbf{G}) \equiv \left(\mathbf{G}^{\mathrm{H}}\Phi_{\mathbf{v}}^{-1}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathrm{H}}\Phi_{\mathbf{v}}^{-1}\mathbf{y}, \tag{6}$$

$$\mathbf{s}_{\mathrm{MCWF}}(\mathbf{G},\Phi_{\mathbf{s}}) \equiv \mathbf{G}^{\mathrm{H}}\Phi_{\mathbf{s}}\left(\mathbf{G}\mathbf{G}^{\mathrm{H}}\Phi_{\mathbf{s}} + \Phi_{\mathbf{v}}\right)^{-1}\mathbf{y} \tag{7}$$

where $\Phi_{\mathbf{s}} = \mathrm{Diag}\left[\phi_{S_1},..,\phi_{S_J}\right]$ is a diagonal matrix (namely, the individual speech signals are assumed mutualy independent). Even though the MCWF usually achieves better noise reduction relative to the LCMV, in many cases the LCMV is preferred due to its distortionless characteristics (especially when a large number of microphones is available). For the main task of this paper, namely speaker separation, the LCMV-BF suffices.

3.2 Parameters estimation

For implementing the LCMV-BF (6), an estimate of the RTF matrix $\mathbf{G}$ is required. The proposed algorithm for blind and online estimation of $\mathbf{G}$ is based on two separated stages:

1. Estimating J dominant DOAs associated with the J dominant active speakers. The DOA of each speaker is chosen from a predefined set of candidate DOAs.
2. Estimating J RTFs $\mathbf{g}_j$ associated with the J dominant DOAs from the first stage. In each frame, the RTFs are initialized by the direct-path TF (based on the DOAs from the previous stage) and then the RTFs are updated using the MCWF outputs.

To concurrently estimate the multiple DOAs and the RTFs of the speakers, the EM [30] formulation is adopted (separately for each task), as described in the following sections. Moreover, to achieve *online* estimation of the RTFs and to maintain smooth estimates over time, a recursive version of the EM algorithm is adopted. A block diagram of the proposed two-stage algorithm is depicted in Fig. 1.

In Section 4, an estimation procedure for the RTFs is proposed while associating an RTF to each speaker using its associated estimated DOA. In Section 5, an estimation procedure of the J-dominant DOAs is proposed.

4 Speaker extraction given the DOAs

To implement the EM algorithm, three datasets should be defined: the *observed data*, the *hidden data*, and the *parameter set*. The observed data in our model is the received microphone signals $\mathbf{y}$. We are proposing to define the individual signals $\mathbf{s}$ as the hidden data. The parameter set Θ is defined as the RTFs $\mathbf{G}$, and the PSD matrix of the speakers $\Phi_{\mathbf{s}}$ such that $\Theta = \{\mathbf{G},\Phi_{\mathbf{s}}\}$.

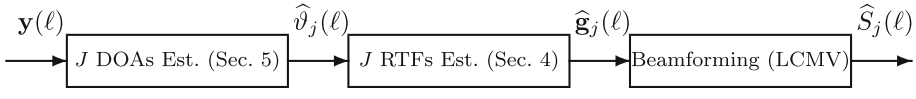


Fig. 1 Block diagram of the proposed two-stage algorithm for online speaker separation

The E-step evaluates the auxiliary function, while the maximization step maximizes the auxiliary function w.r.t. the set of parameters. The batch EM procedure converges to a local maximum of the likelihood function of the observation [30]. To track time-varying RTFs and to satisfy the online requirements, the CREM [31] algorithm is adopted. CREM is based on smoothing of the auxiliary function along the time axis and executing a single maximization per time instance. The smoothing operation is given by [31, Eq. (10)]

$$Q_R\left(\Theta;\hat{\Theta}(\ell-1)\right)=\alpha Q_R\left(\Theta;\hat{\Theta}(\ell-2)\right)+(1-\alpha)Q\left(\Theta;\hat{\Theta}(\ell-1)\right), \tag{8}$$

where $Q_R\left(\Theta;\hat{\Theta}(\ell)\right)$ is the recursive auxiliary function, $\hat{\Theta}(\ell)$ is the estimate of Θ at the ℓth time instance and $0<\alpha<1$ is a smoothing factor.

The term $Q\left(\Theta;\hat{\Theta}(\ell-1)\right)$ is the instantaneous auxiliary function of the ℓth observation, namely the expectation of the log p.d.f. of the complete data (the observed and hidden data) given the observed data and the previous parameter set:

$$Q\left(\Theta;\hat{\Theta}(\ell-1)\right)=E\left\{\log f(\mathbf{y}(\ell),\mathbf{s}(\ell);\Theta)|\mathbf{y}(\ell);\hat{\Theta}(\ell-1)\right\}, \tag{9}$$

The ℓth parameter set estimate $\hat{\Theta}(\ell)$ is obtained by maximizing $Q_R\left(\Theta;\hat{\Theta}(\ell-1)\right)$ w.r.t. Θ.

4.1 Auxiliary function

By applying the Bayes rule, the p.d.f. of the complete instantaneous data is given by:

$$f(\mathbf{y},\mathbf{s};\Theta)=f(\mathbf{y}|\mathbf{s};\Theta)\cdot f(\mathbf{s};\Theta), \tag{10}$$

where the conditional p.d.f. in (10) is given by:

$$f(\mathbf{y}|\mathbf{s};\Theta)=\mathcal{N}^C\left(\mathbf{y},\mathbf{G}\mathbf{s},\Phi_{\mathbf{v}}\right) \tag{11}$$

and the p.d.f. of $\mathbf{s}$ is given by $f(\mathbf{s}|\Theta)=\mathcal{N}^C\left(\mathbf{s},0,\Phi_{\mathbf{s}}\right)$. Finally, the auxiliary function is given by

$$\begin{aligned}Q\left(\Theta;\hat{\Theta}(\ell-1)\right)=E\{&-\pi^N|\Phi_{\mathbf{v}}|-\mathrm{Tr}\left[\Phi_{\mathbf{v}}^{-1}(\mathbf{y}-\mathbf{G}\mathbf{s})(\mathbf{y}-\mathbf{G}\mathbf{s})^{\mathrm{H}}\right]\\&-\pi^J|\Phi_{\mathbf{s}}|-\mathrm{Tr}\left[\Phi_{\mathbf{s}}^{-1}\mathbf{s}\mathbf{s}^{\mathrm{H}}\right];\hat{\Theta}(\ell-1)\}\end{aligned} \tag{12}$$

The EM is notoriously known for converging to local maxima and hence proper initialization is mandatory. In the following section such initialization is discussed.

4.2 Initialization

4.2.1 *Initialization of the individual speaker RTFs*

Since the RTFs of the speakers in $\mathbf{G}$ are time-varying, we propose to reinitialize them in each frame using the estimated DOAs of the speakers. In each new frame, the previous RTFs are discarded and substituted by RTFs which are based on DOA only (as initialization). In the M-step, the RTFs are re-estimated using the smoothed latent-variables. Using the DOAs, the RTFs are initialized by the direct-path transfer function namely the relative phase from the desired speaker to the ith microphone w.r.t. the reference microphone. Accordingly, given the estimates of each speaker DOA $\hat{\vartheta}_j$, the RTFs can be initialized by:

$$\hat{G}_{i,j}=\hat{D}_{i,j}\equiv\exp\left(-\iota\frac{2\pi k}{K}\frac{\tau_i(\hat{\vartheta}_j)}{T_s}\right), \tag{13}$$

where T_s denotes the sampling period and $\tau_i(\hat{\vartheta}_j)$ denotes the time difference of arrival (TDOA) between microphone i and the reference microphone given the jth speaker DOA $\hat{\vartheta}_j$. Note that the DOAs are blindly estimated, as explained in Section 5.

Examining only the horizontal plane, and given the two-dimensional positions of the microphones, the TDOA is given by:

$$\tau_i(\vartheta_j)=\frac{1}{c}\cdot\left[\cos\left(\vartheta_j\right)\quad\sin\left(\vartheta_j\right)\right]\left(\mathbf{x}_i-\mathbf{x}_1\right), \tag{14}$$

where c is the sound velocity and $\mathbf{x}_i$ is the horizontal position of microphone i.

4.2.2 *Initialization of the individual speakers PSD*

Similarly to the RTFs initialization, we propose to reinitialize the PSDs in each frame using the estimated DOAs. The PSDs of the speakers can be initialized by maximizing the p.d.f. of the observations given the relative phase (13):

$$\hat{\Phi}_{\mathbf{s}}=\underset{\Phi_{\mathbf{s}}}{\mathrm{argmax}}\log f(\mathbf{y};\Phi_{\mathbf{s}},\hat{\mathbf{D}}) \tag{15}$$

where

$$f(\mathbf{y};\Phi_{\mathbf{s}},\hat{\mathbf{D}})=\mathcal{N}^C(\mathbf{y};\mathbf{0},\hat{\mathbf{D}}\Phi_{\mathbf{s}}\hat{\mathbf{D}}^{\mathrm{H}}+\Phi_{\mathbf{v}}) \tag{16}$$

and $\hat{\mathbf{D}}$ is an $M\times J$ matrix with elements defined by $\hat{\mathbf{D}}_{i,j}=\hat{D}_{i,j}$. Taking the derivative of the p.d.f. above w.r.t. $\Phi_{\mathbf{s}}$ and equating to zero attains the estimate of the speaker PSDs:

$$\hat{\Phi}_{\mathbf{s}}=\mathbf{s}_{\mathrm{LCMV}}\left(\hat{\mathbf{D}}\right)\mathbf{s}_{\mathrm{LCMV}}^{\mathrm{H}}\left(\hat{\mathbf{D}}\right)-\Phi_{\mathbf{v}_{\mathrm{res}}}\left(\hat{\mathbf{D}}\right), \tag{17}$$

where $\mathbf{s}_{\mathrm{LCMV}}\left(\hat{\mathbf{D}}\right)$ is the multi-speaker LCMV output vector and $\Phi_{\mathbf{v}_{\mathrm{res}}}$ is the residual noise PSD matrix at the output of the multi-speaker LCMV stage,

$$\mathbf{s}_{\mathrm{LCMV}}\left(\hat{\mathbf{D}}\right)\equiv\left(\hat{\mathbf{D}}^{\mathrm{H}}\Phi_{\mathbf{v}}^{-1}\hat{\mathbf{D}}\right)^{-1}\hat{\mathbf{D}}^{\mathrm{H}}\Phi_{\mathbf{v}}^{-1}\mathbf{y}, \tag{18a}$$

$$\Phi_{\mathbf{v}_{\mathrm{res}}}\left(\hat{\mathbf{D}}\right)\equiv\left(\hat{\mathbf{D}}^{\mathrm{H}}\Phi_{\mathbf{v}}^{-1}\hat{\mathbf{D}}\right)^{-1}. \tag{18b}$$

Since $\hat{\Phi}_{\mathbf{s}}$ is defined as diagonal matrix the off-diagonal elements of the estimated matrix in (17) are zeroed-out.

4.3 Instantaneous expectation and maximization steps

Examining (12), the E-step in the ℓth time instance boils down to the calculation of $\hat{\mathbf{s}}$ and $\hat{\mathbf{s}\mathbf{s}^{\mathrm{H}}}$, where for

any stochastic variable a, $\hat{a} \equiv E\left\{a|\mathbf{y};\hat{\Theta}\right\}$. Using the multi-speaker MCWF [8], the following expressions are obtained:

$$\hat{\mathbf{s}} = \mathbf{W}_{\text{MCWF}}^{\text{H}}\left(\hat{\mathbf{D}}\right)\mathbf{y} \tag{19a}$$

$$\hat{\mathbf{s}\mathbf{s}}^{\text{H}} = \hat{\mathbf{s}}\hat{\mathbf{s}}^{\text{H}} + \left(I - \mathbf{W}_{\text{MCWF}}^{\text{H}}\left(\hat{\mathbf{D}}\right)\hat{\mathbf{D}}\right)\hat{\Phi}_{\mathbf{s}} \tag{19b}$$

where

$$\mathbf{W}_{\text{MCWF}}\left(\hat{\mathbf{D}}\right) \equiv \left(\hat{\mathbf{D}}\hat{\mathbf{D}}^{\text{H}}\hat{\Phi}_{\mathbf{s}} + \Phi_{\mathbf{v}}\right)^{-1}\hat{\mathbf{D}}\hat{\Phi}_{\mathbf{s}} \tag{20}$$

is the multi-speaker MCWF. Using the expectations above, the instantaneous auxiliary function $Q\left(\Theta;\hat{\Theta}\right)$ is given by:

$$Q\left(\Theta;\hat{\Theta}\right) = -\log(\pi^{N}|\Phi_{\mathbf{v}}|) - \text{Tr}\left[\mathbf{y}\mathbf{y}^{\text{H}} - \mathbf{G}\hat{\mathbf{s}}\mathbf{y}^{\text{H}} - \mathbf{y}\hat{\mathbf{s}}^{\text{H}}\mathbf{G}^{\text{H}} + \mathbf{G}\hat{\mathbf{s}\mathbf{s}}^{\text{H}}\mathbf{G}^{\text{H}}\right] - \log(\pi^{J}|\Phi_{\mathbf{s}}|) - \text{Tr}\left[\Phi_{\mathbf{s}}^{-1}\hat{\mathbf{s}\mathbf{s}}^{\text{H}}\right]. \tag{21}$$

Substituting the auxiliary function (21) in the recursive equation from (8) and following some algebraic simplifications, the implementation of (8) can be summarized according to the following recursive equations:

$$\mathbf{A}(\ell) = \alpha_{a}\mathbf{A}(\ell-1) + (1-\alpha_{a})\mathbf{y}\hat{\mathbf{s}}^{\text{H}} \tag{22a}$$

$$\mathbf{B}(\ell) = \alpha_{B}\mathbf{B}(\ell-1) + (1-\alpha_{B})\hat{\mathbf{s}\mathbf{s}}^{\text{H}}. \tag{22b}$$

Using $\mathbf{A}(\ell)$ and $\mathbf{B}(\ell)$, the recursive auxiliary function can be rewritten as

$$Q_{R}\left(\Theta;\hat{\Theta}\right) = -\log(\pi^{N}|\Phi_{\mathbf{v}}|) - \text{Tr}\left[\mathbf{y}\mathbf{y}^{\text{H}} - \mathbf{G}\mathbf{A}^{\text{H}}(\ell) - \mathbf{A}(\ell)\mathbf{G}^{\text{H}} + \mathbf{G}\mathbf{B}(\ell)\mathbf{G}^{\text{H}}\right] - \log(\pi^{J}|\Phi_{\mathbf{s}}|) - \text{Tr}\left[\Phi_{\mathbf{s}}^{-1}\mathbf{B}(\ell)\right] \tag{23}$$

Similar to the batch EM procedure, the M-step is obtained by maximizing $Q_{R}\left(\Theta;\hat{\Theta}\right)$ w.r.t. the problem parameters. The speaker PSDs and the RTFs estimates are then given by:

$$\hat{\Phi}_{\mathbf{s}}(\ell) = \mathbf{B}(\ell) \tag{24a}$$

$$\hat{\mathbf{G}}(\ell) = \mathbf{A}(\ell)\mathbf{B}^{-1}(\ell). \tag{24b}$$

Since $\Phi_{\mathbf{s}}$ is defined as diagonal matrix the off-diagonal of its estimate should be zeroed-out. Note that the RTFs are discarded in each new frame and reinitialized using the DOA based steering vector (see (13)). Nevertheless, the RTFs are re-estimated by the updated recursive-variables $\mathbf{A}$ and $\mathbf{B}$ (see (24b)). These variables are only slightly updated form frame to frame using the smoothing factor α. Therefore, the final estimate of the RTFs is only slightly updated. The re-initialization of the RTFs in each frame only influences the estimates of $\hat{\mathbf{s}}$ and $\hat{\mathbf{s}\mathbf{s}}^{\text{H}}$ as used in (22).

4.4 Practical considerations

Due to the intermittent nature of the speech signal, a few speakers may be non-active in several frames. This will result a few elements on the main diagonal of $\Phi_{\mathbf{s}}$ that are close to zero. Note that as the number of speakers J is set in advance, it might be larger than the instantaneous number of active speakers in several frames.

As a matter of fact, bins where only a single speaker is dominant should be the preferred for the task of estimating the RTFs, since the other speakers do not bias the estimate. To determine these TF bins, the power ratio between the desired speech and other interfering speech signals, denoted as desired speaker-to- interferes ratio (DSIR), may be examined for each TF bin according to $\text{DSIR}_{j}(\ell) = \frac{\mathbf{B}_{j,j}(\ell)}{\sum_{i=1}^{J}\mathbf{B}_{i,i}(\ell)}$. Using the PSD matrix initialization (17), the RTFs should be estimated only if the DSIR_{j} obtains a high value. In that case, the RTFs are estimated by applying the following simplified formula:

$$\hat{\mathbf{g}}_{j}(\ell) = \begin{cases} \frac{\mathbf{A}_{j}(\ell)}{\mathbf{B}_{j,j}(\ell)} & \text{DSIR}_{j}(\ell) > \eta \\ \hat{\mathbf{g}}_{j}(\ell-1) & \text{otherwise} \end{cases} \tag{25}$$

where η is some predefined threshold.

To summarize this part of the proposed algorithm, the REM procedure for estimating the individual speaker signals given the DOAs is given in Algorithm 1.

Algorithm 1: Online REM-based RTFs estimation.

Inputs: J dominant DOAs $\hat{\vartheta}_{j}(\ell)$, and microphone signals $\mathbf{y}(\ell)$.
while *new frame* ℓ *is obtained* **do**
  for *each frequency* k **do**
    Initialize $\hat{\mathbf{G}}(\ell)$ using $\hat{\mathbf{D}}$ (13) and $\hat{\Phi}_{S}(\ell)$ using (17)
    E-step:
      Estimate $\hat{\mathbf{s}}$ and $\hat{\mathbf{s}\mathbf{s}}^{\text{H}}$ using (19)
      Update $\mathbf{A}(\ell)$ and $\mathbf{B}(\ell)$ using (22)
    M-step:
      Estimate $\hat{\Phi}_{S}(\ell)$ using (24a)
      Calculate DSIR_{j}
      Estimate $\hat{\mathbf{g}}_{j}(\ell)$ using (25)
  end
end
Output: J RTFs $\hat{\mathbf{g}}_{j}(\ell)$.

5 DOA estimation

For the estimation of the speakers' DOA, we take a different statistical model, and assume hereinafter that the W-disjoint orthogonality property of the speech [32, 33] holds. This assumption was shown to be beneficial in handling multi-speaker DOA estimation tasks [25–29].

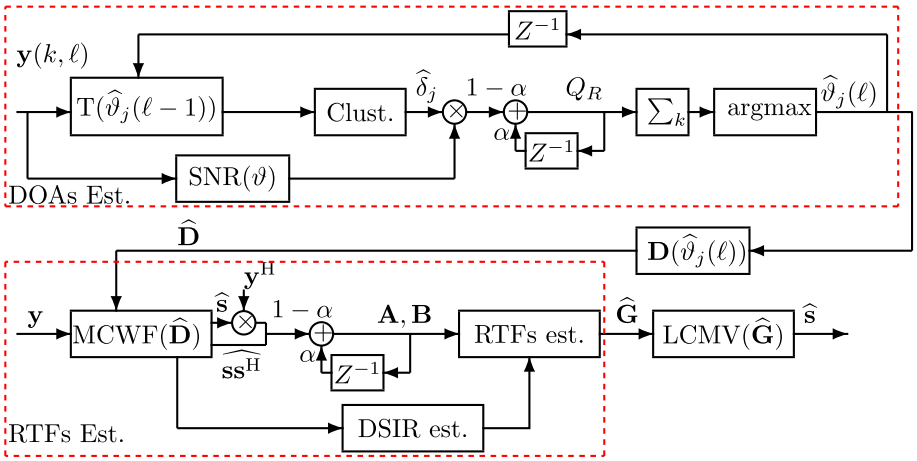


Fig. 2 Diagram of the online two stages REM-based multiple-speaker DOAs estimation and separation

Using this TF sparsity assumption, the signal observed at the ith microphone can remodeled as described in [29]:

$$Y_i(\ell,k) = \sum_{j=1}^{J} \delta_j(\ell,k) D_i(\vartheta_j,k) S_j(\ell,k) + V_i(\ell,k), \quad (26)$$

where the variables $\delta_j(\ell,k)$ are indicators that the jth speaker is active at the (ℓ,k)th TF bin. A disjoint activity of the speakers can be imposed by allowing the J indicators $\delta_j(\ell,k)$ to have only a single non-zero element per each TF bin. The RTF $D_i(\vartheta_j,k)$ is solely defined by the direct-path, as given in (13).

The indicators $\boldsymbol{\delta} = [\delta_1,\ldots,\delta_J]^\top$ will be used as the hidden data under this formulation. The parameter set is accordingly defined as the DOAs ϑ_j and the speakers PSD ϕ_{S_j} such that $\Theta = \{\vartheta_j, \phi_{S_j}\}_{j=1}^{J}$.

Unlike [29], where the probabilities of each candidate DOA are estimated, in this paper the J dominant DOAs are determined from the DOA candidate set. In [29], a subsequent pick peaking stage is therefore required. In the proposed algorithm, the J DOAs are estimated during the M-step.

5.1 Auxiliary function

Using Bayes rule, the p.d.f. of the complete data is given by:

$$f(\mathbf{y},\boldsymbol{\delta};\Theta) = f(\mathbf{y}|\boldsymbol{\delta};\Theta)\cdot f(\boldsymbol{\delta}), \quad (27)$$

where the conditional p.d.f. in (27) is composed as a weighted sum of J Gaussians:

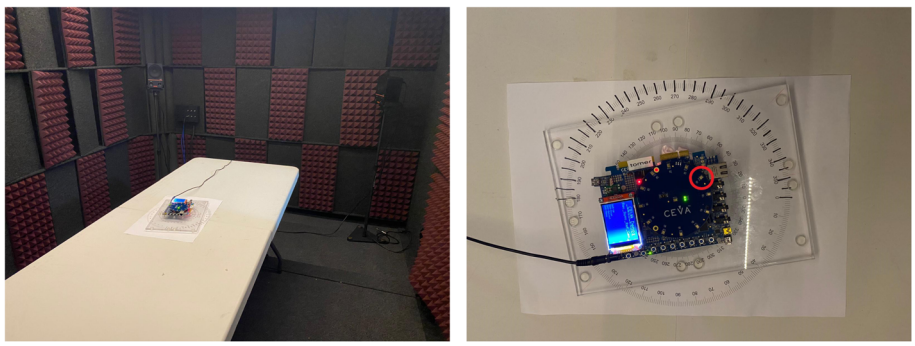
$$f(\mathbf{y}|\boldsymbol{\delta};\Theta) = \sum_{j=1}^{J} \delta_j \mathcal{N}^C\left(\mathbf{y},\mathbf{0},\hat{\mathbf{D}}(\vartheta_j)\hat{\mathbf{D}}^{\mathrm{H}}(\vartheta_j)\phi_{S_j} + \Phi_{\mathbf{v}}\right). \quad (28)$$

The p.d.f. of the indicators is $f(\boldsymbol{\delta}) = \sum_{j=1}^{J} p_j \delta_j$, with p_j the probabilities of activity for each speaker and $\sum_{j=1}^{J} p_j = 1$. These probabilities may be initialized as $1/J$.

5.2 Initial estimation of the speech PSDs

It was shown in [29] that for each DOA ϑ_j, the corresponding speech PSDs are independent of the E-step and thus can be estimated prior to the EM iterations. For each DOA ϑ_j, the corresponding PSD is estimated by maximizing the relevant Gaussian:

$$\hat{\phi}_{S_j} = \underset{\phi_{S_j}}{\operatorname{argmax}} \log \mathcal{N}^C\left(\mathbf{y},\hat{\mathbf{D}}(\vartheta_j)\hat{\mathbf{D}}^{\mathrm{H}}(\vartheta_j)\phi_{S_j} + \Phi_{\mathbf{v}}\right). \quad (29)$$



(a) CEVA acoustic lab (b) CEVA aquisition board

Fig. 3 Recording room and CEVA-DSP platform (one of the microphones is marked by red circle)

Table 1 Trajectories of the two speakers for all experiments

	Speaker	
Experiment	**Male**	**Female**
# 1	Standing at 90°	Standing at 270°
# 2	Standing at 90°	Counterclockwise Surrounding from 270°
# 3	Standing at 90°	Clockwise Surrounding from 270°
# 4	Counterclockwise Surrounding from 90°	Standing at 90°
# 5	Counterclockwise Surrounding from 90°	Counterclockwise Surrounding from 270°
# 6	Counterclockwise Surrounding from 90°	Clockwise Surrounding from 270°
# 7	Clockwise Surrounding from 90°	Standing at 90°
# 8	Clockwise Surrounding from 90°	Counterclockwise Surrounding from 270°
# 9	Clockwise Surrounding from 90°	Clockwise Surrounding from 270°

In [29], it was shown that using the Fisher-Neyman factorization, the log-likelihood above can be expressed as

$$\log \mathcal{N}^{C}\left(\mathbf{y}, \mathbf{0}, \hat{\mathbf{D}}(\vartheta_j)\hat{\mathbf{D}}^{\mathrm{H}}(\vartheta_j)\phi_{S_j} + \Phi_{\mathbf{v}}\right) =$$
$$\log \mathcal{N}\left(\hat{S}_{\mathrm{MVDR}}, 0, \phi_{S_j} + \phi_{\mathbf{v},\mathrm{res}}\right) + \log\left(\frac{\phi_{\mathbf{v},\mathrm{res}}}{|\Phi_{\mathbf{v}}|\,\pi^{N-1}}\right) - \mathbf{y}^{\mathrm{H}}\Phi_{\mathbf{v}}^{-1}\mathbf{y} + \frac{\left|\hat{S}_{\mathrm{MVDR}}\right|^2}{\phi_{\mathbf{v},\mathrm{res}}}, \tag{30}$$

where $\hat{S}_{\mathrm{MVDR}}(\vartheta_j) = \frac{\hat{\mathbf{D}}^{\mathrm{H}}(\vartheta_j)\Phi_{\mathbf{v}}^{-1}\mathbf{y}}{\hat{\mathbf{D}}^{\mathrm{H}}(\vartheta_j)\Phi_{\mathbf{v}}^{-1}\hat{\mathbf{D}}(\vartheta_j)}$ is the output of the MVDR BF steered towards the j-th speaker, and $\phi_{\mathbf{v},\mathrm{res}}(\vartheta_j) = \left(\hat{\mathbf{D}}^{\mathrm{H}}(\vartheta_j)\,\Phi_{\mathbf{v}}^{-1}\hat{\mathbf{D}}(\vartheta_j)\right)^{-1}$ is the residual noise power at the output of the corresponding MVDR BF. Taking the derivative of the log-likelihood in (30) w.r.t. ϕ_{S_j} and equating to zero results in:

$$\hat{\phi}_{S_j} = \left|\hat{S}_{\mathrm{MVDR}}(\vartheta_j)\right|^2 - \phi_{\mathbf{v},\mathrm{res}}(\vartheta_j). \tag{31}$$

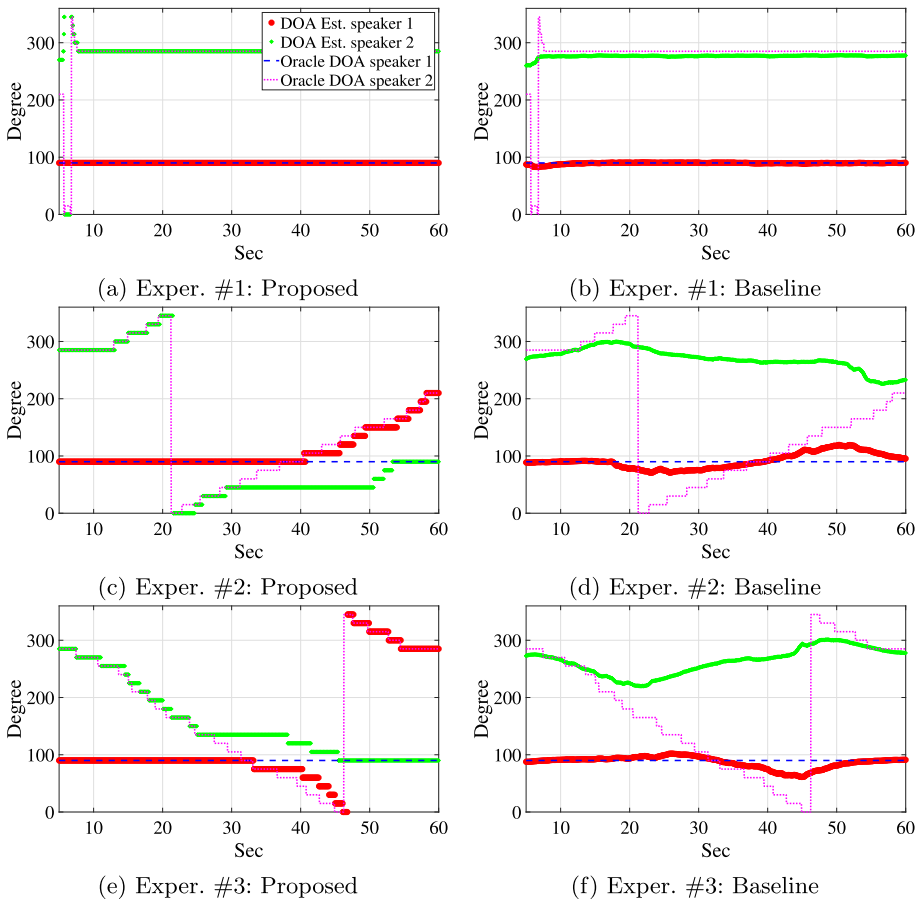


Fig. 4 DOAs estimation for two speakers of the proposed and baseline algorithms for Experiments #1, #2, and #3

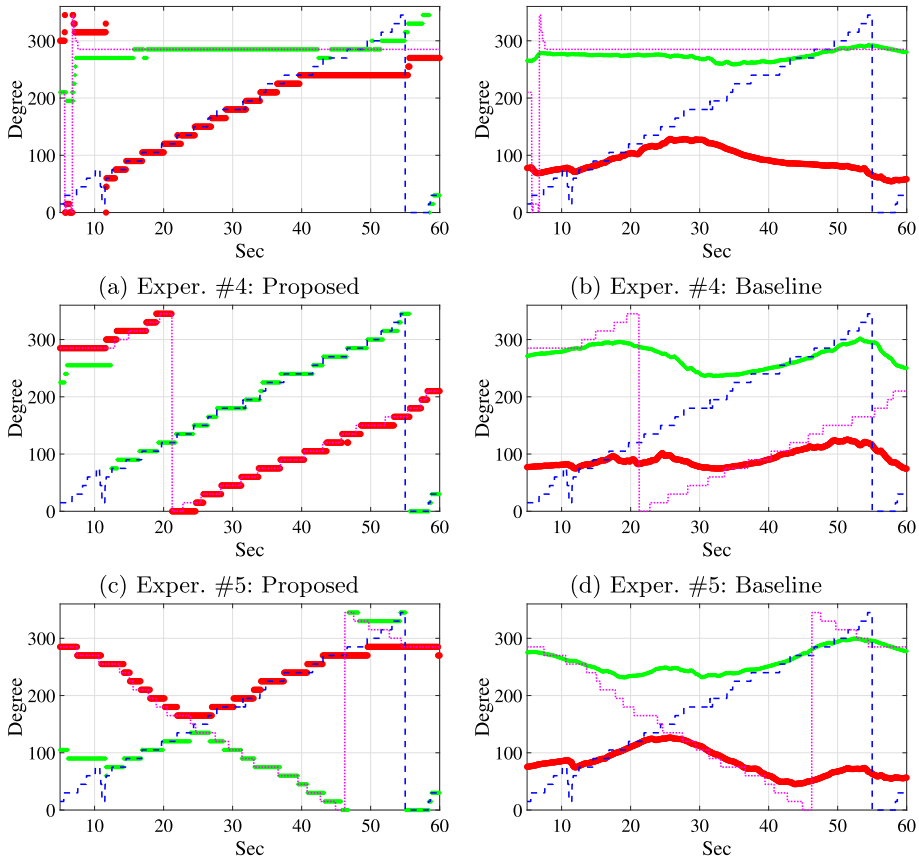


Fig. 5 DOAs estimation for two speakers of the proposed and baseline algorithms for Experiments #4, #5, and #6

Substituting the estimate of the speech PSDs in the log p.d.f. in (30), yields

$$\log \mathcal{N}^{C}\left(\mathbf{y}, \mathbf{0}, \hat{\mathbf{D}}(\vartheta_j)\hat{\mathbf{D}}^{\mathrm{H}}(\vartheta_j)\phi_{S_j} + \Phi_{\mathbf{v}}\right) \overset{C}{=} \mathrm{SNR}\left(\vartheta_j\right) - \log \mathrm{SNR}\left(\vartheta_j\right), \tag{32}$$

where $\mathrm{SNR}\left(\vartheta\right) \equiv \frac{\left|\hat{S}_{\mathrm{MVDR}}(\vartheta)\right|^2}{\phi_{\mathbf{v},\mathrm{res}}(\vartheta)}$ is the posterior SNR of each speaker, and $\overset{C}{=}$ stands for equality up to a constant independent of the relevant parameters.

5.3 The EM iterations

Using the log p.d.f. from (32) and the definitions (27)–(28), the auxiliary function is given by:

$$\begin{aligned} Q\left(\vartheta_j; \hat{\vartheta}_j(\ell-1)\right) &\overset{C}{=} E\left\{\log f(\mathbf{y}, \boldsymbol{\delta}; \Theta) | \mathbf{y}; \hat{\vartheta}_j(\ell-1)\right\} \\ &= \sum_{j=1}^{J} \hat{\delta}_j \Big(\mathrm{SNR}\left(\vartheta_j\right) - \log \mathrm{SNR}\left(\vartheta_j\right)\Big), \end{aligned} \tag{33}$$

where $\hat{\delta}_j$ is the expected indicator $\hat{\delta}_j \equiv E\left\{\delta_j | \mathbf{y}; \hat{\Theta}(\ell-1)\right\}$. According to [29], the expressions for the indicators can be simplified to:

$$\hat{\delta}_j = \frac{p_j T\left(\hat{\vartheta}_j(\ell-1)\right)}{\sum_{j=1}^{J} p_j T\left(\hat{\vartheta}_j(\ell-1)\right)}, \tag{34}$$

where $T\left(\vartheta_j\right) \equiv \frac{1}{\mathrm{SNR}(\vartheta_j)} \exp\left(\mathrm{SNR}\left(\vartheta_j\right)\right)$ is the sufficient statistics.

Using the expected indicators $\hat{\delta}_j$ and the auxiliary function in (32), the smoothing stage in (8) is summarized according to the following recursive equation:

$$\begin{aligned} Q_R\left(\vartheta_j; \hat{\vartheta}_j(\ell-1)\right) = \alpha_R Q_R\left(\vartheta_j; \hat{\vartheta}_j(\ell-2)\right) + \\ (1-\alpha_R)\hat{\delta}_j\left[\mathrm{SNR}\left(\vartheta_j\right) - \log \mathrm{SNR}\left(\vartheta_j\right)\right]. \end{aligned} \tag{35}$$

Note that the term $-\log \mathrm{SNR}\left(\vartheta_j\right)$ in (35) can be omitted because it probably does not influence the maximization[1]. The M-step is obtained by maximizing $Q_R\left(\vartheta_j; \hat{\vartheta}_j(\ell-1)\right)$ w.r.t. ϑ_j and p_j for $j = 1, \ldots, J$:

[1] The function $f(x) = x - \log x$ is a monotonically increasing function when $x > 1$ and indeed we can usually assume that the a posteriori signal-to-noise ratio (SNR) to be larger than 1.

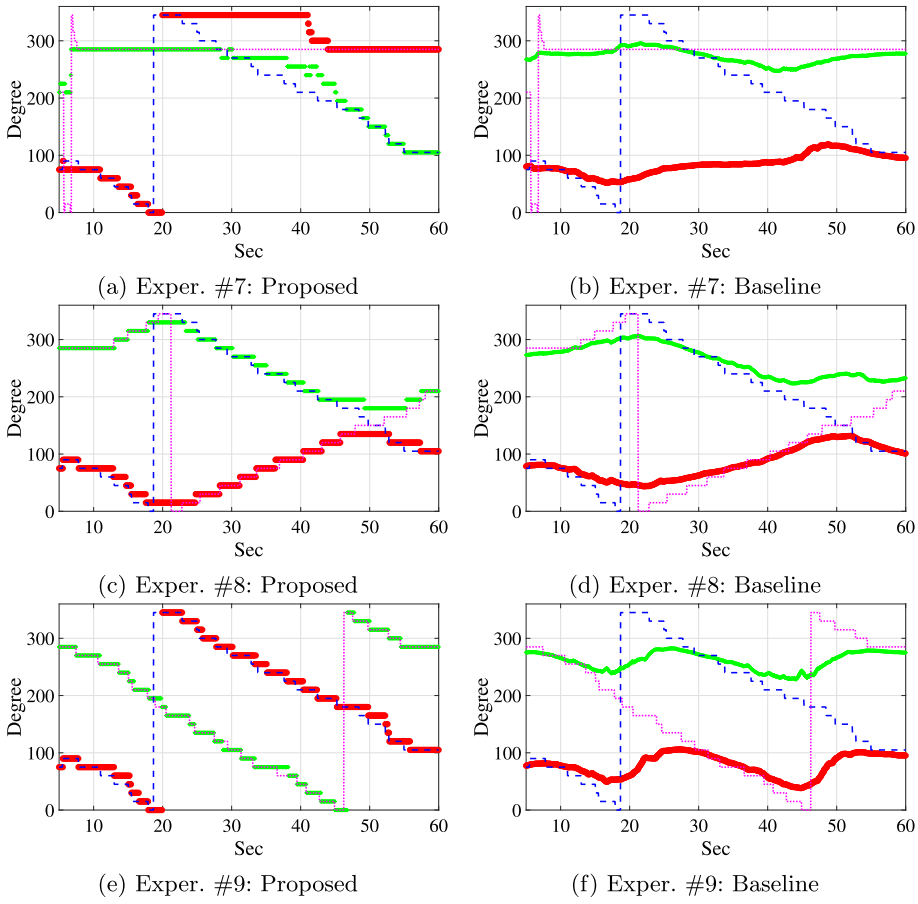


Fig. 6 DOAs estimation for two speakers of the proposed and baseline algorithms for Experiments #7, #8, and #9

$$\hat{\vartheta}_j(\ell) = \underset{\vartheta_j}{\operatorname{argmax}} \sum_k Q_R\left(\vartheta_j; \hat{\vartheta}_j(\ell-1)\right), \tag{36a}$$

$$p_j(\ell) = \frac{1}{K}\sum_k \hat{\delta}_j(k). \tag{36b}$$

Note that the estimation of the DOA is obtained using all frequency bins. Since there is no closed-form expression for $\hat{\vartheta}_j(\ell)$, the term $\sum_k Q_R\left(\vartheta_j; \hat{\vartheta}_j(\ell-1)\right)$ should be calculated for each possible ϑ_j. Practically, a set of DOA candidates can be predefined (for example $0°, \ldots, 345°$ with a $1°$ resolution) and $\sum_k Q_R\left(\vartheta_j; \hat{\vartheta}_j(\ell-1)\right)$ can be calculated only for these candidates. Then, separately to each source j, the DOA which maximizes $\sum_k Q_R\left(\vartheta_j; \hat{\vartheta}_j(\ell-1)\right)$ is selected as the jth speaker DOA.

The estimated probability of each speaker p_j (Eq. (36a)) may be utilized to discard the redundant speaker in the beamforming stage (see the third block in Fig. 1). When p_j is lower than a predefined threshold, it implies that the jth speaker is inactive and the final beamforming may include only the other active speakers.

The REM algorithm for estimating the desired speaker DOA is summarized in Algorithm 2 and depicted in the block diagram in Fig. 2.

Algorithm 2: Online REM-based multiple DOA Estimation.

Inputs: Microphone signals $\mathbf{y}(\ell)$.
while *new frame* ℓ *is obtained* **do**
  E-step:
  for *each frequency* k **do**
    Calculate SNR $\left(\vartheta_j\right)$ for each ϑ_j candidate.
    Calculate $\hat{\delta}_j$ using $\hat{\vartheta}_j(\ell-1)$ by (34).
    Update $Q_R\left(\vartheta_j; \hat{\vartheta}_j(\ell-1)\right)$ for each ϑ_j candidate using (35)
  end
  M-step:
    Find $\hat{\vartheta}_j(\ell)$ using (36a) and $p_j(\ell)$ using (36b)
end
Output: J DOAs $\hat{\vartheta}_j(\ell)$

Table 2 MAE results in degrees

	Experiment								
Algorithm	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**
Baseline	7.5	61.0	44.3	55.9	34.0	43.3	59.1	25.4	28.3
Proposed	0.3	10.8	10.2	32.5	18.1	9.7	21.8	11.7	6.6

6 Performance evaluation

The performance of the proposed algorithm is evaluated using recorded signals of two concurrent speakers on the two presented tasks: (1) online DOA estimation and (2) time-varying source separation. Correspondingly, we used two quality measures: (1) the mean absolute error (MAE) between the estimated and oracle DOAs and (2) the power level between the speakers at the output, as a measure of the separation capabilities.

6.1 Recording setup

Overall, nine experiments were conducted. In each experiment, two 60-s long speech signals of male and female speakers were separately recorded in the acoustic lab at CEVA Inc. premises, as shown in Fig. 3a. CEVA Inc. DSP platform was used as the acquisition device. The circular array with 5 cm diameter comprises six microphones at the perimeter. The device is depicted in Fig. 3b.

The two speakers were either standing or moving with various trajectories around the array, approximately 1 m from the array center. The speed of the moving speaker was approximately 1 m/s. The various source trajectories are described in Table 1.

The reverberation time was approximately adjusted to $T_{60} = 0.2$ sec using the room panels and additional furniture. The utterances were generated by adding the two separately recorded speech signals together with both spectrally and spatially white noise, with power of 40 dB below the power of the overall speech signals.

The sampling frequency was 16 kHz, and the frame length of the STFT was set to 128 ms with 32 ms overlap. The resolution of the candidates DOA was set to 15° in the range $[0^\circ : 345^\circ]$. The frequency band $500 - 3500$ Hz was used for the DOA estimation.

6.2 Baseline method

The proposed algorithm was compared with a baseline localization and separation algorithm conceptually based on [20]. The steps of the baseline algorithm are:

Table 3 MAE results in degrees for various T_{60}

T_{60}	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**
Baseline	64.38	70.40	75.30	77.70	80.04
Proposed	11.12	20.13	32.23	26.85	31.08

Table 4 MAE results in degrees for various SNRs

SNR	**40**	**30**	**20**	**10**	**0**
Baseline	64.32	67.81	70.99	73.22	77.34
Proposed	11.21	10.97	12.47	11.75	14.49

1. Calculate the SRP-PHAT [23] outputs for each TF bin and DOA candidate in the range $[0^\circ : 15^\circ : 345^\circ]$, and then find the DOA with the maximum SRP value for each TF bin.
2. Cluster the DOAs from all bins to two clusters (assuming two speakers) using the REM-MoG algorithm[2] [31] and determining the two dominant DOAs by taking the centroid of each cluster.
3. For each cluster, estimate the associated RTF using the classical cross-spectral method [34, Eq. (9)] using a recursive version as implemented in [20, Eq.(19)].
4. Implement the LCMV-BF using the two estimated RTFs.

6.3 Tracking results

The tracking algorithms estimate the two dominant DOAs for each frame. Let $\hat{\vartheta}_1(\ell)$ and $\hat{\vartheta}_2(\ell)$ be an estimate of the DOAs of the two speakers at frame ℓ, as obtained by either the proposed and baseline algorithms, and $\vartheta_1(\ell)$ and $\vartheta_2(\ell)$, be the oracle DOAs, respectively. Define the MAE as:

$$\mathrm{MAE} = \frac{1}{2L}\sum_{\ell} \min \ldots \left(\left|\hat{\vartheta}_1(\ell) - \vartheta_1(\ell)\right| + \left|\hat{\vartheta}_2(\ell) - \vartheta_2(\ell)\right|, \left|\hat{\vartheta}_1(\ell) - \vartheta_2(\ell)\right| + \left|\hat{\vartheta}_2(\ell) - \vartheta_1(\ell)\right| \right), \tag{37}$$

where L is the number of frames in the utterance. The oracle DOAs were obtained by apply the proposed algorithm to the separated inputs $\mathbf{x}_{\mathrm{M}}$ and $\mathbf{x}_{\mathrm{F}}$ while assumed a single speaker.

The trajectories of the estimated DOAs for both proposed and the baseline algorithms for all nine experiments are depicted in Figs. 4, 5, and 6, together with the oracle trajectories. The MAEs for all cases are presented in Table 2.

Looking at the tracking curves and the MAEs, the proposed algorithm clings well to the oracle speakers DOA contours, and significantly outperforms the baseline algorithm.

Note that when the speakers' trajectories intersect, the estimates may suffer from unavoidable permutation ambiguity. Consequently, while both trajectories are accurately estimated, the association between them and the speakers may switch after the intersection point (see

[2] Adopted from Section 2.4 in [31], where only the means of the Gaussians were estimated, and the variances and probabilities were set as constants (the probabilities are set to $\frac{1}{2}$ and the variances to 2).

Table 5 SIR results for all cases

		Algorithm							
		Proposed				**Baseline**			
Experiment$\hat{\text{O}}$	**SIR(Input)**	$\text{SIR}_1(\hat{\mathbf{D}})$	$\text{SIR}_1(\hat{\mathbf{G}})$	$\text{SIR}_2(\hat{\mathbf{D}})$	$\text{SIR}_2(\hat{\mathbf{G}})$	$\text{SIR}_1(\hat{\mathbf{D}})$	$\text{SIR}_1(\hat{\mathbf{G}})$	$\text{SIR}_2(\hat{\mathbf{D}})$	$\text{SIR}_2(\hat{\mathbf{G}})$
#1	2.85	7.83	11.58	3.22	7.69	7.83	11.51	3.13	7.84
#2	2.12	4.92	5.83	3.24	5.17	4.17	4.41	1.53	3.29
#3	2.29	4.16	5.75	4.16	6.68	5.14	5.82	2.88	5.04
#4	5.02	6.99	8.54	3.60	4.50	6.38	8.42	2.97	3.49
#5	3.54	2.36	5.99	9.37	12.32	2.75	4.31	7.19	8.54
#6	4.03	6.74	8.79	3.42	5.43	3.18	4.33	6.13	6.80
#7	3.70	3.47	5.27	5.57	7.03	4.66	6.92	2.62	3.74
#8	2.86	2.85	4.47	6.41	8.14	2.56	4.03	5.92	7.85
#9	2.66	7.69	11.36	3.53	7.83	3.26	6.04	5.92	7.74
Mean	2.66	7.69	11.36	3.53	7.83	3.26	6.04	5.92	7.74

Experiments #2, #3, #4, #6, #7, and #8). Note, that by definition (37), the MAE is agnostic to such permutations.

The intersecting trajectories may also result in significant errors when the DOAs of the speakers become closer (see Experiments #2, #3, #4, #7, and #8). This is also reflected in the relatively high MAE values (see Experiments #4, #5, and #7). Higher MAE values are also encountered at the initial convergence period (see Experiments #4, #5, and #6).

The performance improvement of the proposed algorithm may be attributed to the MVDR-BF front-end, which is capable of suppressing the interference sources, as opposed to the SRP-PHAT front-end, which is adopted by the baseline method.

The proposed DOA estimator is also evaluated in comparison with the baseline algorithm in multiple reverberation times (T_{60}) and SNR values. Two moving speakers were simulated by convolving randomly selected male or female utterances with the room response, simulated using an open source signal generator[3]. The microphone signals are then contaminated by a directional, spectrally pink, noise source in several SNR levels. The trajectories of the two speakers were set as clockwise and counter-clockwise.

The MAEs for different values of T_{60} and for SNR=40 dB are presented in Table 3. The MAEs for different SNR levels and for $T_{60} = 0.2$ are presented in Table 4.

The performance of both the baseline and the proposed algorithms degrades as the reverberation level increases. However, the accuracy of the proposed algorithm is significantly higher and is limited by $\sim 30^\circ$. Similar trends can be observed in Table 4, with a significant advantage of the proposed algorithm, with errors kept in the range of $11^\circ - 14.5^\circ$.

[3] https://www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator

6.4 Separation results

The separation capabilities of the proposed and baseline algorithms were assessed by evaluating the speaker-to-interference ratio (SIR) improvement. For convenience, all examined scenarios comprised one male and one female speakers. The microphone signals are thus given by:

$$\mathbf{y} = \mathbf{x}_{\text{M}} + \mathbf{x}_{\text{F}} + \mathbf{v}, \tag{38}$$

where $\mathbf{x}_{\text{M}}$ and $\mathbf{x}_{\text{F}}$ denote the reverberant male and female signals, respectively, as captured by the microphones, and $\mathbf{v}$ denotes a spatially and spectrally white noise signal.

Both the DOAs $\hat{\mathbf{D}}$ and the RTFs $\hat{\mathbf{G}}$ matrices were estimated from the mixed signals $\mathbf{y}$ and the corresponding LCMV-BFs were constructed. The beamformers were then independently applied to the male and female components of the received microphone signals:

$$\mathbf{s}^{\text{M}}_{\text{LCMV}}\left(\hat{\mathbf{G}}\right) \equiv \left(\hat{\mathbf{G}}^{\text{H}}\Phi_{\mathbf{v}}^{-1}\hat{\mathbf{G}}\right)^{-1}\hat{\mathbf{G}}^{\text{H}}\Phi_{\mathbf{v}}^{-1}\mathbf{x}_{\text{M}}, \tag{39a}$$

$$\mathbf{s}^{\text{M}}_{\text{LCMV}}\left(\hat{\mathbf{D}}\right) \equiv \left(\hat{\mathbf{D}}^{\text{H}}\Phi_{\mathbf{v}}^{-1}\hat{\mathbf{D}}\right)^{-1}\hat{\mathbf{D}}^{\text{H}}\Phi_{\mathbf{v}}^{-1}\mathbf{x}_{\text{M}}, \tag{39b}$$

$$\mathbf{s}^{\text{F}}_{\text{LCMV}}\left(\hat{\mathbf{G}}\right) \equiv \left(\hat{\mathbf{G}}^{\text{H}}\Phi_{\mathbf{v}}^{-1}\hat{\mathbf{G}}\right)^{-1}\hat{\mathbf{G}}^{\text{H}}\Phi_{\mathbf{v}}^{-1}\mathbf{x}_{\text{F}} \tag{39c}$$

$$\mathbf{s}^{\text{F}}_{\text{LCMV}}\left(\hat{\mathbf{D}}\right) \equiv \left(\hat{\mathbf{D}}^{\text{H}}\Phi_{\mathbf{v}}^{-1}\hat{\mathbf{D}}\right)^{-1}\hat{\mathbf{D}}^{\text{H}}\Phi_{\mathbf{v}}^{-1}\mathbf{x}_{\text{F}}. \tag{39d}$$

Now, if the estimated RTFs are approximately equal to the true ones, we expect the algorithm to produce the following two-channel outputs:

$$\mathbf{s}^{\text{M}}_{\text{LCMV}}\left(\hat{\mathbf{G}}\right) \simeq [S_M, 0]^\top \text{ or } [0, S_M]^\top, \tag{40a}$$

$$\mathbf{s}^{\text{F}}_{\text{LCMV}}\left(\hat{\mathbf{G}}\right) \simeq [0, S_F]^\top \text{ or } [S_F, 0]^\top, \tag{40b}$$

where S_F and S_M are the male and female speech signals as observed at the reference microphone. The two alternative outputs result in from the permutation ambiguity problem that was discussed above. This problem my be arbitrarily encountered for each time-frame. If

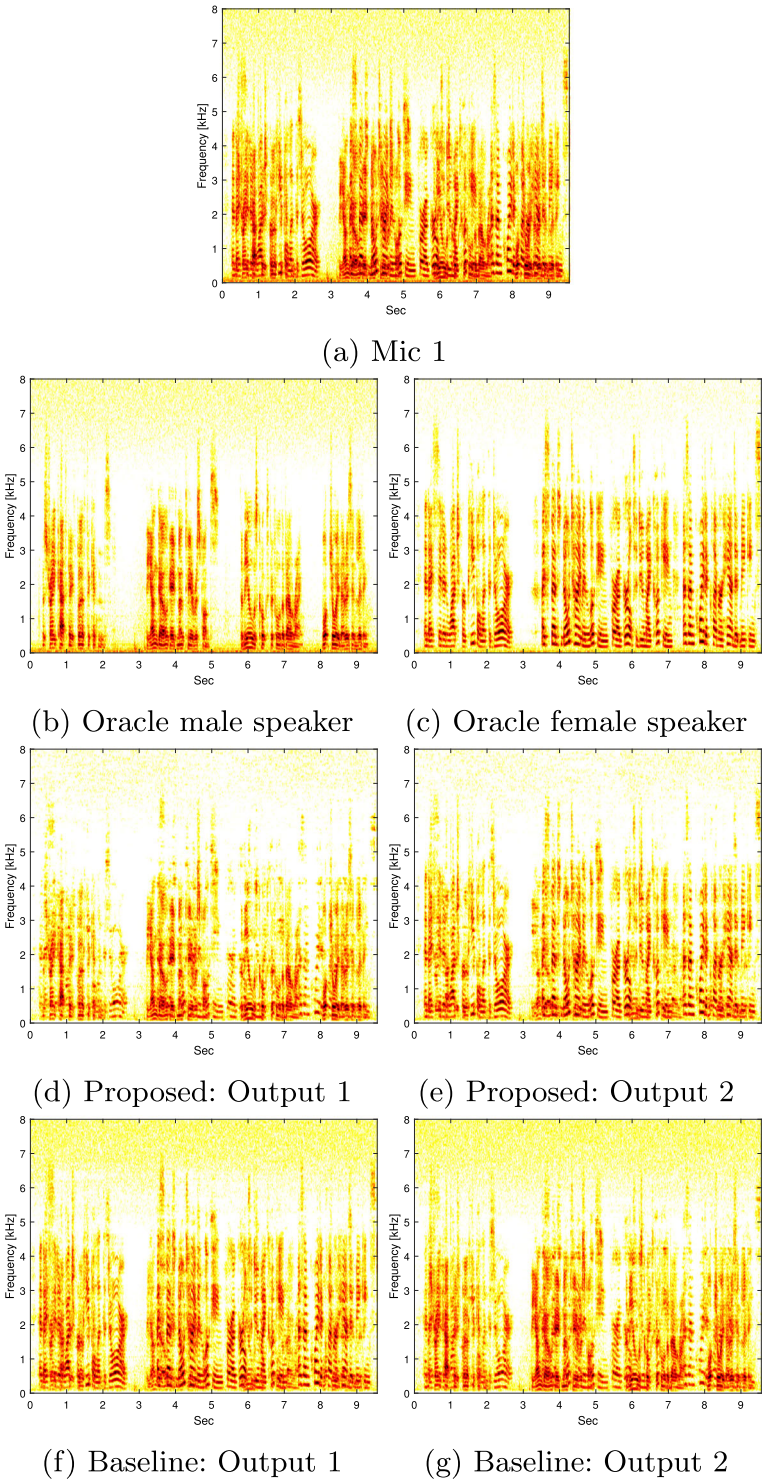


Fig. 7 Example sonograms

Table 6 SIR results along various T_{60}

		Algorithm							
		Proposed				**Baseline**			
T_{60}	SIR(**Input**)	$\text{SIR}_1(\hat{\mathbf{D}})$	$\text{SIR}_1(\hat{\mathbf{G}})$	$\text{SIR}_2(\hat{\mathbf{D}})$	$\text{SIR}_2(\hat{\mathbf{G}})$	$\text{SIR}_1(\hat{\mathbf{D}})$	$\text{SIR}_1(\hat{\mathbf{G}})$	$\text{SIR}_2(\hat{\mathbf{D}})$	$\text{SIR}_2(\hat{\mathbf{G}})$
0.2	2.96	9.53	10.79	4.72	9.57	3.39	5.24	6.25	6.67
0.3	3.19	4.74	6.10	5.17	7.78	3.17	3.16	4.76	4.69
0.4	3.18	3.98	5.15	4.76	7.22	3.22	2.82	4.17	4.23
0.5	3.36	3.94	4.94	4.41	6.23	3.38	2.84	4.08	4.04
0.6	3.48	3.78	4.66	4.16	5.72	3.33	2.69	4.07	4.04

the full RTFs are substituted by the simpler DOAs, the beamformer outputs can be defined with the necessary modifications.

Define, $\mathbf{s}^{\text{M}}_{\text{LCMV},1}\left(\hat{\mathbf{G}}\right)$ and $\mathbf{s}^{\text{F}}_{\text{LCMV},1}\left(\hat{\mathbf{G}}\right)$ as the first output of the beamformers and, correspondingly, $\mathbf{s}^{\text{M}}_{\text{LCMV},2}\left(\hat{\mathbf{G}}\right)$ and $\mathbf{s}^{\text{F}}_{\text{LCMV},2}\left(\hat{\mathbf{G}}\right)$, as the second output. Similar definition apply for the beamformers using the DOAs matrix.

We can now define SIR measures:

$$\text{SIR}_j\left(\hat{\mathbf{G}}\right) = \frac{1}{L}\sum_{\ell}\left|10\log_{10}\frac{\sum_k|\mathbf{s}^{\text{M}}_{\text{LCMV},j}\left(\hat{\mathbf{G}}\right)|^2}{\sum_k|\mathbf{s}^{\text{F}}_{\text{LCMV},j}\left(\hat{\mathbf{G}}\right)|^2}\right|, \quad (41)$$

for $j = 1, 2$, and similarly

$$\text{SIR}_j\left(\hat{\mathbf{D}}\right) = \frac{1}{L}\sum_{\ell}\left|10\log_{10}\frac{\sum_k|\mathbf{s}^{\text{M}}_{\text{LCMV},j}\left(\hat{\mathbf{D}}\right)|^2}{\sum_k|\mathbf{s}^{\text{F}}_{\text{LCMV},j}\left(\hat{\mathbf{D}}\right)|^2}\right|. \quad (42)$$

Since an absolute value of the ratio (in dB) is calculated for each time-frame, these measures are indifferent to the permutation problem.

To evaluate the SIR improvement of all algorithms, we also calculate the input SIR:

$$\text{SIR (Input)} = \frac{1}{L}\sum_{\ell}\left|10\log_{10}\frac{\sum_k||\mathbf{x}_{\text{F}}||^2}{\sum_k||\mathbf{x}_{\text{M}}||^2}\right|. \quad (43)$$

The output SIR results for all experiments are presented in Table 5. It can be verified that $\text{SIR}_j(\hat{\mathbf{D}})$ results are generally higher for the proposed algorithm than for the baseline algorithm, due to the better estimation accuracy of DOAs.

The ratios $\text{SIR}_j(\hat{\mathbf{G}})$ are generally better for the proposed algorithm. The improvement is caused apparently by the MCWF usage for the RTF estimation in (19) which supplies better separation between the speakers within the RTF estimation procedure.

Finally, the algorithms are evaluated by assessing the sonograms of the various outputs for Experiment #9 as depicted in Fig. 7.

Careful examination of the sonograms, demonstrates the improved separation capabilities of the proposed algorithm in comparison with the baseline algorithm. For example, examining the signals in the time periods 2–3 Sec and 5–6 Sec, it can be verified that the proposed algorithm, as compared with the baseline method, better suppresses the female speech at Output 1, while maintaining low distortion for the male speaker.

The proposed speaker separation procedure is also evaluated versus the baseline algorithm for different reverberation levels (T_{60}) and SNR levels. The SIRs for different T_{60} and for SNR=40 dB are presented in Table 6. The SIRs for various SNR values and for $T_{60} = 0.2$ are presented in Table 7.

It is evident from Table 6 that the performance of both the baseline and proposed algorithms degrades with increasing reverberation level and that the proposed algorithm outperforms the baseline algorithm. Analyzing the results in Table 4, it is clearly demonstrated

Table 7 SIR results along various SNRs

		Algorithm							
		Proposed				**Baseline**			
SNR	SIR(**Input**)	$\text{SIR}_1(\hat{\mathbf{D}})$	$\text{SIR}_1(\hat{\mathbf{G}})$	$\text{SIR}_2(\hat{\mathbf{D}})$	$\text{SIR}_2(\hat{\mathbf{G}})$	$\text{SIR}_1(\hat{\mathbf{D}})$	$\text{SIR}_1(\hat{\mathbf{G}})$	$\text{SIR}_2(\hat{\mathbf{D}})$	$\text{SIR}_2(\hat{\mathbf{G}})$
40	2.96	9.53	10.79	4.72	9.57	3.39	5.24	6.25	6.67
30	2.96	9.55	10.84	4.72	9.52	3.33	5.12	6.14	6.27
20	2.96	9.46	10.40	4.57	9.45	3.37	4.79	5.72	5.92
10	2.96	9.55	11.04	4.70	9.01	3.42	4.17	5.47	5.08
0	2.96	9.17	10.19	4.57	7.87	3.52	2.19	5.45	3.19

Table 8 MAE results in degrees for various η values

η	0.1	0.2	0.2	0.4	0.5	0.6	0.7	0.8	0.9
$\text{SIR}_1(\hat{\mathbf{G}})$	6.00	7.80	9.08	10.39	10.79	11.25	11.08	10.29	8.81
$\text{SIR}_2(\hat{\mathbf{G}})$	4.67	6.48	8.01	8.74	9.57	9.64	9.48	8.47	6.19

that the performance of the proposed algorithm does not degrade with decreasing SNR in the range 0–40 dB. Generally, for both the proposed and the baseline algorithms, the utilization of the RTFs enhances the separation capabilities as compared with direct-path only systems.

Finally, the proposed separation technique is evaluated for different values of η. Recall that η is used in (25) to limit the RTF estimation only to a single dominant speaker TF bins. The SIRs for different values of η and for SNR $=$ 40 dB and $T_{60} = 0.2$ are presented in Table 8. It can be verified that the choice of η significantly influences performance and that setting $\eta = 0.6$ yields the best results.

6.5 Comparison with open embedded audition system (ODAS)

In this section, the proposed algorithm is further evaluated versus a state-of-the-art algorithm, namely ODAS[4]. ODAS is an open-source library dedicated to a combined sound source localization, tracking and separation. Two static speakers were simulated using open source signal generator[5]. The DOAs of the speakers were set to 45° and 135° w.r.t. the array center, and their distance from the array was set to 1 m. Clean speech utterances were randomly drawn from a set of male and female speakers. The reverberation time was set to $T_{60} = 0.3$. The performance of the proposed algorithm and of ODAS algorithm were evaluated as a function of the overlap percentage between the speakers. Two widely used speech quality and intelligibility measures, namely perceptual evaluation of speech quality (PESQ) [35] and short-time objective inteligibility measure (STOI) [36], were used to evaluate the performance of the algorithms. The comparison between the algorithms is reported in Table 9. It is clearly demonstrated that the proposed algorithm outperforms the ODAS algorithm in both measures.

Table 9 PESQ (off-brackets) and STOI (in-brackets) results along various time-overlapping between the speakers

Overlapping	Speaker	Input	Proposed	ODAS
20%	1	3.44 (0.94)	3.90 (0.97)	3.45 (0.83)
20%	2	2.90 (0.94)	3.73 (0.98)	2.89 (0.83)
40%	1	2.77 (0.88)	3.63 (0.96)	2.70 (0.79)
40%	2	2.38 (0.89)	2.52 (0.83)	2.41 (0.80)
60%	1	2.33 (0.83)	3.37 (0.95)	2.42 (0.76)
60%	2	2.01 (0.84)	3.20 (0.95)	2.05 (0.75)
80%	1	1.87 (0.79)	3.10 (0.94)	1.97 (0.74)
80%	2	1.84 (0.79)	3.02 (0.94)	1.90 (0.73)

7 Conclusions

We have presented an online algorithm for separating moving sources. The proposed algorithm comprises two stages: (1) online DOA tracking and (2) online RTF estimation. The two stages employ different statistical models. The estimated RTFs are used as building blocks of a continuously-adapted LCMV-BF. The proposed algorithm is compared with a baseline method using real recordings in the challenging task of separating concurrently active and moving sources.

[4] https://github.com/ehabets/Signal-Generator
[5] https://github.com/ehabets/RIR-Generator

Abbreviations

RTF: Relative transfer function; MVDR: Minimum variance distortionless response; DSIR: Desired speaker-to-interferes ratio; STFT: Short-time Fourier transform; PSD: Power spectral density; MAE: Mean absolute error; SNR: Signal-to-noise ratio; MCWF: Multichannel Wiener filter; BF: Beamformer; LCMV: Linearly constrained minimum variance; EM: Expectation-maximization; CREM: Cappé and Moulines recursive EM; SRP: Steered response power; DOA: Direction of arrival; REM: Recursive expectation-maximization; TDOA: Time difference of arrival; RTF: Relative transfer function; MVDR: Minimum variance distortionless response; STFT: Short-time Fourier transform; PSD: Power spectral density; SNR: Signal-to-noise ratio; MCWF: Multichannel Wiener filter; BF: Beamformer; LCMV: Linearly constrained minimum variance; CREM: Cappé and Moulines recursive EM; TF: Time frequency; SIR: Speaker-to-interference ratio; GEVD: Generalized eigenvalue decomposition; MUSIC: Multiple signal classification; MoG: Mixture of Gaussians; PHAT: Phase transform; MESSL: Model-based expectation-maximization source separation and localization; CNN: Convolutional neural network; ODAS: Open embeddeD Audition System; PESQ: Perceptual evaluation of speech quality; STOI: Short-time objective inteligibility measure

Authors' contributions

Model development: OS and SG, Experimental testing: OS, Writing paper: OS and SG. The authors read and approved the final manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

Availability of data and materials

N/A.

Declarations

Consent for publication

All authors agree to the publication in this journal.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] CEVA-DSP Sound Department, Herzelia, Israel. [2] Faculty of Engineering, Bar-Ilan University, 5290002 Ramat-Gan, Israel.

Received: 14 May 2021 Accepted: 18 October 2021
Published online: 04 December 2021

References

1. S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(4), 692–730 (2017)
2. E. Vincent, T. Virtanen, S. Gannot, *Audio Source Separation and Speech Enhancement*. (John Wiley & Sons, New-Jersey, 2018)
3. (S. Makino, ed.), *Audio Source Separation. Signals and Communication Technology*. (Springer, Cham, 2018)
4. B. D. Van Veen, K. M. Buckley, Beamforming: A versatile approach to spatial filtering. IEEE Acoust. Speech Signal Proc. Mag. **5**(2), 4–24 (1988)
5. M. H. Er, A. Cantoni, Derivative constraints for broad-band element space antenna array processors. IEEE Trans. Acoust. Speech Signal Proc. **31**(6), 1378–1393 (1983)
6. H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. (John Wiley & Sons, New-York, 2004)
7. S. Markovich, S. Gannot, I. Cohen, Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1071–1086 (2009)
8. O. Schwartz, S. Gannot, E. A. Habets, Multispeaker LCMV beamformer and postfilter for source separation and noise reduction. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(5), 940–951 (2017)
9. S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, S. Makino, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Speaker indexing and speech enhancement in real meetings/conversations, (2008), pp. 93–96
10. B. Laufer-Goldshtein, R. Talmon, S. Gannot, Source counting and separation based on simplex analysis. IEEE Trans. Signal Process. **66**(24), 6458–6473 (2018)
11. B. Laufer-Goldshtein, R. Talmon, S. Gannot, Global and local simplex representations for multichannel source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 914–928 (2020)
12. B. Laufer Goldshtein, R. Talmon, S. Gannot, Audio source separation by activity probability detection with maximum correlation and simplex geometry. J. Audio Speech Music Proc. **2021**, 5 (2021)
13. S. E. Chazan, J. Goldberger, S. Gannot, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming, (2018), pp. 6712–6716
14. S. E. Chazan, J. Goldberger, S. Gannot, in *The 26th European Signal Processing Conference (EUSIPCO)*. LCMV beamformer with DNN-based multichannel concurrent speakers detector, (Rome, 2018), pp. 1562–1566
15. H. Hammer, S. E. Chazan, J. Goldberger, et al., Dynamically localizing multiple speakers based on the time-frequency domain. J. Audio Speech Music Proc. **2021**, 16 (2021)
16. N. Ito, C. Schymura, S. Araki, T. Nakatani, in *2018 26th European Signal Processing Conference (EUSIPCO)*. Noisy cGMM: Complex Gaussian mixture model with non-sparse noise model for joint source separation and denoising, (2018), pp. 1662–1666
17. A. Ozerov, C. Févotte, Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans. Audio Speech Lang. Process. **18**(3), 550–563 (2009)
18. D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, R. Horaud, S. Gannot, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Exploiting the intermittency of speech for joint separation and diarization, (2017), pp. 41–45
19. D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, R. Horaud, A variational EM algorithm for the separation of time-varying convolutive audio mixtures. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(8), 1408–1423 (2016)
20. N. Madhu, R. Martin, A versatile framework for speaker separation using a model-based speaker localization approach. IEEE Trans. Audio Speech Lang. Process. **19**(7), 1900–1912 (2010)
21. M. Souden, S. Araki, K. Kinoshita, T. Nakatani, H. Sawada, A multichannel MMSE-based framework for speech source separation and noise reduction. IEEE Trans. Audio Speech Lang. Process. **21**(9), 1913–1928 (2013)
22. T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, T. Nakatani, Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust asr. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(4), 780–793 (2017)
23. J. H. DiBiase, H. F. Silverman, M. S. Brandstein, ed. by M. Brandstein, D. Ward. Microphone arrays : Signal processing techniques and applications (Springer, Berlin, Heidelberg, 2001), pp. 157–180
24. R. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
25. M. I. Mandel, R. J. Weiss, D. P. Ellis, Model-based expectation-maximization source separation and localization. IEEE Trans. Audio Speech Lang. Process. **18**(2), 382–394 (2010)
26. O. Schwartz, Y. Dorfan, E. A. Habets, S. Gannot, in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. Multi-speaker DOA estimation in reverberation conditions using expectation-maximization, (2016), pp. 1–5
27. Y. Dorfan, O. Schwartz, B. Schwartz, E. A. Habets, S. Gannot, in *IEEE International Conference on the Science of Electrical Engineering (ICSEE)*. Multiple DOA estimation and blind source separation using estimation-maximization, (2016), pp. 1–5
28. O. Schwartz, Y. Dorfan, M. Taseska, E. A. Habets, S. Gannot, in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*. DOA estimation in noisy environment with unknown noise power using the EM algorithm, (2017), pp. 86–90
29. K. Weisberg, S. Gannot, O. Schwartz, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. An online multiple-speaker DOA tracking using the Cappé-Moulines recursive expectation-maximization algorithm, (2019), pp. 656–660
30. A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol., 1–38 (1977)
31. O. Cappé, E. Moulines, On-line expectation–maximization algorithm for latent data models. J. R. Stat. Soc. Ser. B Stat Methodol.) **71**(3), 593–613 (2009)
32. S. Rickard, O. Yilmaz, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1*. On the approximate w-disjoint orthogonality of speech, (2002), pp. 529–532
33. O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004)
34. O. Shalvi, E. Weinstein, System identification using nonstationary signals. IEEE Trans. Signal Process. **44**(8), 2055–2063 (1996)
35. ITU-T, Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs Rec. ITU-T P. 862 (2021)
36. C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, in *IEEE Transactions on Audio, Speech, and Language Processing, vol. 19*. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech, (2011), pp. 2125–2136

Publisher's Note