


METHODOLOGY

Open Access



Multi-microphone simultaneous speakers detection and localization of multi-sources for separation and noise reduction

Ayal Schwartz^{1,2}, Ofer Schwartz¹, Shlomo E. Chazan^{1,2} and Sharon Gannot^{1*} 

Abstract

This paper addresses the challenge of online blind speaker separation in a multi-microphone setting. The linearly constrained minimum variance (LCMV) beamformer is selected as the backbone of the separation algorithm due to its distortionless response and capacity to create a null towards interfering sources. A specific instance of the LCMV beamformer that considers acoustic propagation is implemented. In this variant, the relative transfer functions (RTFs) associated with each speaker of interest are utilized as the steering vectors of the beamformer. A control mechanism is devised to ensure robust estimation of the beamformer's building blocks, comprising speaker activity detectors and direction of arrival (DOA) estimation branches. This control mechanism is implemented as a multi-task deep neural network (DNN). The primary task classifies each time frame based on speaker activity: no active speaker, single active speaker, or multiple active speakers. The secondary task is DOA estimation. It is implemented as a classification task, executed only for frames classified as single-speaker frames by the primary branch. The direction of the active speaker is classified into one of the multiple ranges of angles. These frames are also leveraged to estimate the RTFs using subspace estimation methods. A library of RTFs associated with these DOA ranges is then constructed, facilitating rapid acquisition of new speakers and efficient tracking of existing speakers. The proposed scheme is evaluated in both simulated and real-life recordings, encompassing static and dynamic scenarios. The benefits of the multi-task approach are showcased, and significant improvements are evident, even when the control mechanism is trained with simulated data and tested with real-life data. A comparison between the proposed scheme and the independent low-rank matrix analysis (ILRMA) algorithm reveals significant improvements in static scenarios. Furthermore, the tracking capabilities of the proposed scheme are highlighted in dynamic scenarios.

Keywords LCMV beamforming, Relative transfer function estimation, DOA estimation, Speech activity detection, Multi-task deep learning

1 Introduction

In the last two decades, the use of microphone arrays for speech enhancement has surged in popularity. This trend is driven by the potential performance advantages offered by spatial processing. Speech signals frequently face degradation due to ambient noise, reverberation,

and overlapping speakers. Consequently, the significant challenge in speech enhancement is to separate the target speaker from a mixture of speakers, effectively suppressing ambient noise. A comprehensive survey of state-of-the-art multichannel audio separation methods can be found in [1–3].

Many speaker separation and noise reduction methods are based on conventional beamformers, such as minimum variance distortionless response (MVDR)-beamformer (BF) [4–7] and the LCMV-BF [8, 9]. The use of the RTF as steering vector for MVDR beamforming was

*Correspondence:

Sharon Gannot
sharon.gannot@biu.ac.il

¹ Faculty of Engineering, Bar Ilan University, 5290002 Ramat-Gan, Israel

² Origin.AI, Ramat-Gan, Israel



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

introduced in [7] and later extended to LCMV beamforming in the multi-speaker case [10]. The latter work was extended in [11] to simultaneously extract all speakers of interest in the scene. The ability of RTF-based LCMV-BF to extract a set of desired speakers is theoretically analyzed in [12].

The MVDR-BF and the LCMV-BF require the noise spatial correlation matrix that can be estimated using speech absent frames and the RTFs of all speakers of interest. RTF estimation has been an active research field in recent decades. A plethora of estimation methods can be found in the literature, most of them employing least-squares or subspace methods [7, 10, 11, 13–16]. In our work, we employed a method based on the generalized eigenvalue decomposition (GEVD) of the spatial correlation matrix of the received microphone signals and the noise spatial correlation matrix [10, 15].

To accurately estimate the RTFs, prior knowledge of the activity patterns is usually required. Moreover, each of the estimated RTFs should also be consistently associated with the active speakers in the scene.

In [10, 11], the RTFs were estimated in time intervals for which only a single speaker is active while assuming that the activity patterns of the sources are available. Such an assumption cannot be met in realistic scenarios. Therefore, estimating speakers' activity patterns has emerged as an active research topic in recent years.

We first review beamformer implementations employing time frequency (TF) masks in the single-speaker case, i.e., in noise reduction applications. Early works employing speech presence probability (SPP) as a TF mask for estimating RTFs were introduced for controlling an MVDR beamformer, implemented in a generalized sidelobe canceller (GSC) structure [17, 18]. In [19], a complex Gaussian mixture model is employed to estimate the TF mask and the steering vector of an MVDR beamformer. In [20, 21] a DNN is employed to estimate the TF mask. The above model-based and data-driven mask estimation approaches are combined in [22]. In [23], a multi-tap MVDR (a.k.a. convolutive transfer function (CTF)-MVDR beamformer [24]) is implemented using a TF mask estimated from both the audio and visual modalities. In [25], a dynamic scenario is addressed by applying attention-based online estimation of the spatial correlation matrix of the received microphone signals.

The multi-speaker case is now reviewed. In [26], clustering of a series of steered response power (SRP) readings is used for identifying time frames predominantly occupied by single speakers. In [27], the activity of the speakers was estimated by introducing a latent variable with $N + 1$ possible discrete states for a mixture of N speech signals plus additive noise. The activity

is estimated using spatial cues from the observed signals modeled with a Gaussian-mixture-like model. In [28], further improved in [29], it is assumed that the sources do not become simultaneously active. Using this assumption, the RTF of a new speech source is estimated using the estimated RTFs of the already active sources in the environment. The challenge of inferring the activity of the speakers remains open.

In [30, 31], single-speaker dominant frames were identified by utilizing convex geometry tools on the recovered simplex of the speakers' probabilities or the correlation function between frames [32].

In recent years, the capabilities of DNNs have been leveraged for identifying speakers' activities. Both single-microphone [33] and multi-microphone [34] estimators were proposed. These methods classify the activity of each frame into three classes: (1) speech is absent, (2) a single speaker is active, or (3) multiple speakers are concurrently active. To ensure the consistency of the LCMV outputs, it is crucial that each estimated RTF is consistently associated with the same speaker across time frames. In [34], a proposal was made to calculate the cosine distance between the currently estimated RTF and a library of previously acquired RTFs. However, this procedure may face challenges associated with unreliable associations.

Recently, several deep models have been proposed for the task of actual beamforming, going beyond the role of merely controlling conventional beamformers. In [35, 36], beamforming is implemented through a DNN model that provides relevant components. In [35], a DNN-based MVDR beamforming framework is introduced, where the analytical derivation of the steering vector and the inversion of the noise covariance matrix are replaced by two gated recurrent unit (GRU)-Nets. In [36], a novel causal U-net-based multiple-input multiple-output structure is introduced for real-time multichannel speech enhancement. This method maintains the traditional beamforming operation by directly calculating the beamformer weights instead of directly estimating the speakers.

Spectral and spatial features were combined in a deep clustering framework and exploited to improve speech separation in [37]. Inter-microphone phase patterns are provided to the network as additional input features. These additional inputs significantly improve separation performance over the single-microphone case, even with a random microphone-array constellation. While deep spatial processing carries great promise in pushing forward the performance boundaries of conventional beamformers, it still suffers from robustness to mismatches between train and test conditions and from the lack of explainability.

Our work belongs to the class of DNN-controlled conventional beamformers. We focus on the design of DNN-based classifier for estimating both the speakers' activity patterns and the DOAs of the active speakers. The task of DOA estimation using DNN models was widely addressed in the literature [38–44]. In [43], the speech sparsity in the short-time Fourier transform (STFT) domain was utilized to track the DOAs of multiple speakers, using a DNN applied to the instantaneous RTF estimates. In our work, a similar DOA estimation procedure is applied in conjunction with the concurrent speaker detector (CSD) estimation procedure in a deep multi-task architecture. We stress that in our work, the DOA estimates are only utilized for associating the estimated RTFs to speakers. The actual separation is carried out by the RTF-based LCMV-BF [10], which is known to be a low-distortion processor, especially in reverberant environments in which DOA-based steering vectors fail short of providing low-distortion of the desired speaker and strong attenuation of the interfering sources.

We now summarize the main stages of our contribution. A dual-task DNN-based model is presented. The activity of the speakers (speakers absence, single speaker, or multiple speakers) and the DOAs (only for single speaker frames) are simultaneously estimated. In our experimental study, we demonstrate that employing a multi-task model enhances the performance of each specific task. The network takes inputs in the form of the log-spectrum of the current frame and instantaneous RTF estimate. These inputs provide both spectral and spatial information, contributing to both the CSD and DOA classification tasks. As elaborated, a control mechanism governs the speaker separation process. When the CSD module identifies a single-speaker frame, an update is made to the RTF estimate and to the associated estimated DOA. Consequently, a collection of RTFs is accumulated for each visited DOA, forming a library. This library enhances the robustness of RTF estimation by incorporating an extended history of frames, thus increasing robustness to dynamic scenarios. During multi-speaker frames, these RTFs are utilized to separate the speaker by applying the LCMV beamformer.

The proposed DNN-based model was examined using simulated and actual microphone recordings with static and dynamic speakers. The experimental study is split into three parts: (1) CSD accuracy, (2) DOA estimation accuracy, and (3) speaker separation performance of the entire proposed system. Finally, our separation performance in static scenarios is compared with the ILRMA algorithm [45, 46] and shows both objective and perceptual improvements. The proposed method is further evaluated in dynamic scenarios, demonstrating its ability to adapt rapidly to the changing acoustic scene.

2 Problem formulation and main objectives

This work considers the case of concurrent static or dynamic speakers acquired by a microphone array in a reverberant and noisy environment. The signal captured by the m -th microphone is given in the STFT domain:

$$y_m(n, k) = \sum_{j=1}^{J(n)} g_{m,j}(n, k) s_j(n, k) + v_m(n, k), \quad (1)$$

where n and k represent the frame and frequency indexes, respectively. $y_m(n, k)$ denotes the m -th microphone signal, $s_j(n, k)$ represents the j -th speaker signal as received by the reference microphone, and $g_{m,j}(n, k)$ is its associated RTF relating microphone m to the reference microphone. The ambient noise is denoted by $v_m(n, k)$. The variable $J(n)$ denotes the number of active speakers in frame n , which is unknown in advance. The signal model in (1) can be recast in a vector form,

$$\begin{aligned} \mathbf{y}(n, k) &= \sum_{j=1}^{J(n)} \mathbf{g}_j(n, k) s_j(n, k) + \mathbf{v}(n, k) \\ &= \mathbf{G}(n, k) \mathbf{s}(n, k) + \mathbf{v}(n, k) \end{aligned} \quad (2)$$

where

$$\mathbf{y}(n, k) = [y_1(n, k), \dots, y_M(n, k)]^\top, \quad (3a)$$

$$\mathbf{g}_j(n, k) = [g_{1,j}(n, k), \dots, g_{M,j}(n, k)]^\top, \quad (3b)$$

$$\mathbf{G}(n, k) = [\mathbf{g}_1(n, k), \dots, \mathbf{g}_{J(n)}(n, k)], \quad (3c)$$

$$\mathbf{v}(n, k) = [v_1(n, k), \dots, v_M(n, k)]^\top \quad (3d)$$

$$\mathbf{s}(n, k) = [s_1(n, k), \dots, s_{J(n)}(n, k)]^\top. \quad (3e)$$

For conciseness, the indices n and k will be omitted unless needed.

Define J the total number of speakers in the scene, namely the union of all sets of active speakers per frame. By definition, $J \geq J(n), \forall n$. The primary goal of this work is to extract a noiseless and undistorted version of the individual speaker signals $s_j, j = 1, \dots, J$, as received by the reference microphone.

As elaborated in the introduction, the multi-speaker linearly constrained minimum variance beamformer (LCMV-BF) [11] will serve as the backbone tool for speaker extraction in our work:

$$\hat{\mathbf{s}}(n, k) = \mathbf{W}_{\text{LCMV}}^H(n, k) \mathbf{y}(n, k) \quad (4)$$

where

$$\mathbf{W}_{\text{LCMV}}(n, k) = \Phi_{\mathbf{v}}^{-1} \mathbf{G} \left(\mathbf{G}^H \Phi_{\mathbf{v}}^{-1} \mathbf{G} \right)^{-1} \quad (5)$$

and the matrix $\Phi_{\mathbf{v}}$ denotes the noise spatial power spectral density (PSD) matrix. The above LCMV formulation defines an $M \times J(n)$, time-varying and frequency-dependent, filtering matrix $\mathbf{W}_{\text{LCMV}}(n, k)$, such that each entry of the $J(n) \times 1$ output vector $\hat{\mathbf{s}}(n, k)$ is dominated by one of the desired speakers. An accurate estimate of the noise PSD matrix and the RTFs will guarantee that the desired source component at each output is a distortionless replica of the source as received by the reference microphone with all other interfering sources entirely suppressed and the noise signal attenuated. Note that when $J(n) = 1$, the LCMV-BF degenerates to the MVDR-BF.

The noise PSD matrix and the RTFs associated with the speakers are usually not known in advance, and their blind estimation is, therefore, the main goal of this work. The noise spatial PSD matrix can be estimated using speech-absent frames, while the RTFs can be estimated using single-speaker frames, namely frames in which only a single speaker is active.

The specific objectives of this work involve classifying frames of received signals based on their activity and consistently associating single-speaker frames with each speaker in the scene. This process facilitates the construction of the LCMV-BF (5), subsequently extracting the individual speakers of interest.

3 Dual task CSD and DOA classifier

In this section, we present a dual-task DNN model that determines, per frame, the activity of the speakers and their corresponding DOA.

The concurrent speaker detector (CSD) branch classifies each frame to either (1) speech absence, (2) single-speaker activity, or (3) multi-speaker activity. Simultaneously, the model classifies each single-speaker frame to a DOA range chosen from a predefined set of possible DOAs. Hence, our dual-task classifier has two outputs, the CSD with three classes as explained above, and the DOA estimator with N classes corresponding to N DOA ranges.

Estimating the noise spatial PSD matrix necessitates speech-absent frames (Class #0). The RTF estimation requires single-speaker frames (Class #1). Subsequently, to preserve estimation consistency, each estimated RTF is associated with a specific DOA range, and its corresponding PSD matrix is archived in a library of PSD matrices per DOA to enhance the robustness of future estimates. In frames when two or more speakers are

concurrently active (Class #2), no estimation procedure is applied, and the previous BF weights are frozen. The time-varying LCMV-BF is then implemented using the estimated noise PSD and the relevant RTFs.

3.1 Multi task classification

As elaborated above, the proposed DNN model has two simultaneous outputs for each frame n . The first is the CSD:

$$\text{CSD}(n) = \begin{cases} 0 & \text{Noise only } (J(n) = 0) \\ 1 & \text{Single-speaker } (J(n) = 1) \\ 2 & \text{Multi-speaker } (J(n) > 1) \end{cases} \quad (6)$$

The second output is the DOA range estimate for single-speaker frames, namely when $\text{CSD}(n) = 1$. The DOA output is not considered in the other cases. Let θ be the angle of the source with respect to the microphone array. The permissible DOA values, $\theta \in [0^\circ, 180^\circ]$, are split into N equal ranges. The DNN model classifies each single-speaker frame to a range:

$$\text{DOA}(n) = \begin{cases} 0 & \theta \in [0, \frac{180^\circ}{N}) \\ 1 & \theta \in [\frac{180^\circ}{N}, 2\frac{180^\circ}{N}) \\ \vdots & \vdots \\ N-1 & \theta \in [(N-1)\frac{180^\circ}{N}, 180] \end{cases} \quad (7)$$

3.2 Input features

In this work, we use the frame-based log-spectrum of the reference microphone and the spatial cues (as defined later) as the input features. It was experimentally verified that adding spatial cues to the models' input improves the CSD accuracy.

The log-spectrum values are normalized across the frequency index, obtaining zero mean and unity variance,

$$\mathbf{a}(n) = \text{Normalize} \left([\log |y_1(n, 1)|, \dots, \log |y_1(n, K)|]^\top \right) \quad (8)$$

where for each matrix \mathbf{X} , $\text{Normalize}(\mathbf{X})$ returns \mathbf{X} with each column normalized to zero-mean and standard deviation equals 1.

In this work, we use an *instantaneous* estimate of the RTFs as the spatial cues, obtained by the GEVD-based method [15], as explained in the sequel.

Let $\mathbf{z}(n)$ be the whitened microphone signals, $\mathbf{z}(n) = \Phi_{\mathbf{v}}^{-H/2} \mathbf{y}(n)$, where $\Phi_{\mathbf{v}}^{1/2}$ is the square root of the noise PSD matrix, namely $\Phi_{\mathbf{v}} = \Phi_{\mathbf{v}}^{H/2} \Phi_{\mathbf{v}}^{1/2}$, obtained using, e.g., Cholesky decomposition. The respective spatial PSD matrix of the whitened microphones is estimated by averaging the context frames $n - m_1, \dots, n + m_2$:

Table 1 Source-array constellation and room configuration

Parameter	Range	Comments
T_{60}	0.3, ..., 0.55 s	
Array orientation	0°, ..., 360°	
Room dimensions	4 – 40 sqm	
Array position	All over the room	At least 0.5 m from the walls
Speakers position	All over the room	At least 0.5 m between speakers
Directional noise position	All over the room	At least 2 m from array
Diffuse noise SNR	10–20 dB	
Speaker to mic distance	1–1.5 m	

$$\hat{\Phi}_{\mathbf{z}}(n) = \sum_{m=n-m_1}^{n+m_2} w_m \mathbf{z}(m) \mathbf{z}^H(m) \quad (9)$$

where w_m is a weighting factor, emphasizing the current frame. Denote the principal eigenvector of $\hat{\Phi}_{\mathbf{z}}(n)$ as $\hat{\psi}(n)$. The *instantaneous* RTF estimate is given by:

$$\hat{\mathbf{g}}(n) = \frac{\Phi_{\mathbf{v}}^{H/2} \hat{\psi}(n)}{\mathbf{e}_1^T \Phi_{\mathbf{v}}^{H/2} \hat{\psi}(n)} \quad (10)$$

where $\mathbf{e}_1^T = [1 \mathbf{0}_{1 \times M-1}]$. Note that the first element of $\hat{\mathbf{g}}(n)$ can be omitted from the input matrix since it always equals 1. The spatial cues are separated into real and imaginary components and normalized across the frequency and microphone indexes:

$$\mathbf{B}(n) = \text{Normalize} \left(\begin{bmatrix} \text{Re}(\hat{\mathbf{g}}(n, 1), \dots, \hat{\mathbf{g}}(n, K)) \\ \text{Im}(\hat{\mathbf{g}}(n, 1), \dots, \hat{\mathbf{g}}(n, K)) \end{bmatrix} \right). \quad (11)$$

The feature matrix $[\mathbf{a}(n) \ \mathbf{B}^T(n)]$ constitutes the input for each frame n .

The algorithm's latency, attributed to the estimation of the spatial cues used for classifying the current frame, is determined by the frame length plus the number of future frames (m_2). Although reducing m_2 can decrease latency, our experiments demonstrated that doing so had a negative impact on the results.

3.3 Database construction

The training comprises various simulated recordings outlined below to capture a diverse range of real-life scenarios. The number of microphones and the constellation of the microphone array remain consistent across all training and test conditions. The configurations for source-array constellations and room parameters are illustrated in Table 1. This table encompasses permissible ranges for various factors, such as room dimensions, microphone array position

and orientation, speaker positions, speaker-to-microphone distance, directional noise position, diffuse noise signal-to-noise ratio (SNR), and reverberation level (T_{60}).

The activity patterns of speakers are randomly determined to imitate realistic scenarios. Each activity class and each DOA range comprise an equal number of utterances. While the training data comprises only static speakers, the algorithm was also tested with dynamic speakers.

3.4 DNN architecture

The proposed DNN model comprises three convolutional layers, succeeded by three fully connected (FC) layers. Subsequently, two distinct branches are established for each classification task: (1) an FC layer with N outputs for the DOA classification and (2) an FC layer with three outputs for the speaker activity classification. The activation function employed in each layer's output is rectified linear unit (ReLU), except the final layer, where Softmax activation is utilized, producing N outputs for DOA range and three for the speaker activity. The final decision for both the CSD and DOA classifiers is made by selecting the class with the highest probability. The model incorporates dropout, batch normalization, and weight constraint operations to mitigate overfitting. The categorical cross-entropy (CCE) loss function was employed during the model training with the adaptive moment estimation (ADAM) optimizer.

To tailor the loss function to the specific goal of enhancing LCMV beamforming, three crucial updates were incorporated for both the CSD and DOA tasks. Let pCSD and aCSD be the predicted and actual CSD, respectively. Similarly, let pDOA and aDOA be the predicted and actual DOA, respectively.

3.4.1 CSD Loss

In our experiments, we observed that our CSD estimator tends to misclassify between single-speaker and multi-speaker classes. Detecting multi-speaker activity when only a single speaker is active does not significantly impact the performance of the LCMV-BF. However, identifying single-speaker activity during multi-speaker scenarios may adversely affect the performance of the RTF estimator. To mitigate the effects of the latter error, we suggest altering the CSD loss as follows:

$$\text{Loss}_{\text{CSD}} = \begin{cases} \alpha \cdot \text{CCE}_{\text{CSD}} & \text{pCSD} = 1 \ \& \ \text{aCSD} = 2 \\ \text{CCE}_{\text{CSD}} & \text{otherwise} \end{cases}, \quad (12)$$

where CCE_{CSD} is the CCE between the actual CSD and the model prediction, and $\alpha > 1$ a scaling parameter.

3.4.1.1 DOA loss While small DOA errors may be tolerable, large DOA errors may severely degrade the consist-

ency of the LCMV outputs. We, therefore, amplify the loss function when the DOA error is larger:

$$\text{Loss}_{\text{DOA}} = \frac{|p\text{DOA} - a\text{DOA}|}{N} \text{CCE}_{\text{DOA}} \quad (13)$$

where CCE_{DOA} is the CCE between the actual DOA and the model prediction.

3.4.2 Weighted loss

Recall that the DOA estimates in speaker-absent or multi-speaker frames are discarded, and the corresponding loss is set to zero:

$$\text{Loss}_{\text{DOA}} = \begin{cases} 0 & a\text{CSD} = 0, 2 \\ \text{Loss}_{\text{DOA}} & a\text{CSD} = 1 \end{cases}. \quad (14)$$

Consequently, approximately 2/3 of the utterances are not used for the DOA classification. We therefore over-stress Loss_{DOA} by a factor $\beta > 1$:

$$\text{Loss} = \beta \text{Loss}_{\text{DOA}} + \text{Loss}_{\text{CSD}}. \quad (15)$$

4 Determining the number of active sources

Denote the set of active DOAs at frame n as $\overline{\text{DOA}}(n)$:

$$\overline{\text{DOA}}(n) = \{\text{DOA}_1(n), \text{DOA}_2(n), \dots, \text{DOA}_{J(n)}(n)\}, \quad (16)$$

where $\text{DOA}_j(n) \in \{0, \dots, N-1\}$, is the class number attributed to the DOA range of an active source $j = 1, 2, \dots, J(n)$. The procedure for determining $\overline{\text{DOA}}(n)$ is outlined below:

- 1 If, in a single-speaker frame, the speaker's angle is classified to the j -th DOA range, then $\overline{\text{DOA}}(n) = j$, is added to the set of active sources $\overline{\text{DOA}}(n)$. If an adjacent DOA range is already active, it is substituted by the currently estimated DOA; otherwise, a new active direction is declared. If $|\overline{\text{DOA}}(n)| = M-1$, the oldest DOA is removed from the set of active sources, and the new DOA takes its place.
- 2 In the multi-speaker case, namely $\text{CSD}=2$, $\overline{\text{DOA}}(n)$ remains intact.
- 3 A DOA is removed from the list of currently active DOAs if it was inactive for Q consecutive frames (further denoted "expiry time").

5 RTFs and noise PSD matrix estimation

A procedure that utilizes the dual-task CSD and DOA classifiers for estimating the noise PSD matrix and the speakers' RTFs is now described.

5.1 Noise PSD estimation

The noise PSD matrix is updated during frames that are classified by the CSD as noise-only (Class #0), using the following adaptation rule (per frequency k):

$$\Phi_{\mathbf{v}}(n) = \begin{cases} \gamma_{\text{nos}} \Phi_{\mathbf{v}}(n-1) + \\ (1 - \gamma_{\text{nos}}) \mathbf{y}(n, k) \mathbf{y}^H(n), & p\text{CSD} = 0 \\ \Phi_{\mathbf{v}}(n-1) & \text{otherwise} \end{cases} \quad (17)$$

where γ_{nos} is the learning rate.

5.2 RTF estimation

The LCMV-BF (5) necessitates estimates of the RTFs associated with the dominant active speakers. The LCMV criterion may support up to $M-1$ speakers.

A note on the steering vector of the BF is in place. In this work, we use an RTF-based BF rather than the simpler DOA-based BF due to its higher capabilities in suppressing the competing speakers [7, 10]. To maintain consistency in the LCMV outputs, each estimated RTF is associated with a specific DOA range. Thus, the DOAs are merely used as attributes of the estimated RTFs and do not directly construct the steering vectors.

The following procedure is applied. First, we construct a library of PSD matrices with each entry attributed to one of the DOA ranges. Namely, a frame classified as $\text{CSD} = 1$ and $\text{DOA} = j$, is used to update the PSD matrix $\Phi_j(n)$ (per frequency k):

$$\Phi_j(n) = \begin{cases} \delta_j \Phi_j(n-1) + \dots \\ (1 - \delta_j) \mathbf{y}(n) \mathbf{y}^H(n), & p\text{CSD}(n) = 1, \\ \Phi_j(n-1) & p\text{DOA} = j \\ & \text{otherwise} \end{cases} \quad (18)$$

where δ_j is the learning rate. The RTFs are assumed to be time-varying since the sources may move. Hence, it is required that they will be continuously updated by applying the GEVD procedure (10). This may impose a large computational burden that may be alleviated using subspace tracking methods [13, 47]. Note that while (9) is an instantaneous estimate of the correlation matrix of the (whitened) received microphone signals, $\Phi_j(n, k)$ in (18) is estimated using a longer averaging and hence provides a more accurate RTF estimate.

6 The construction of the LCMV beamformer

The LCMV can accommodate a maximum of $M-1$ constraints per time frame. The rest of the degrees of freedom are allocated to noise reduction. To achieve higher noise reduction capabilities, it is recommended that the

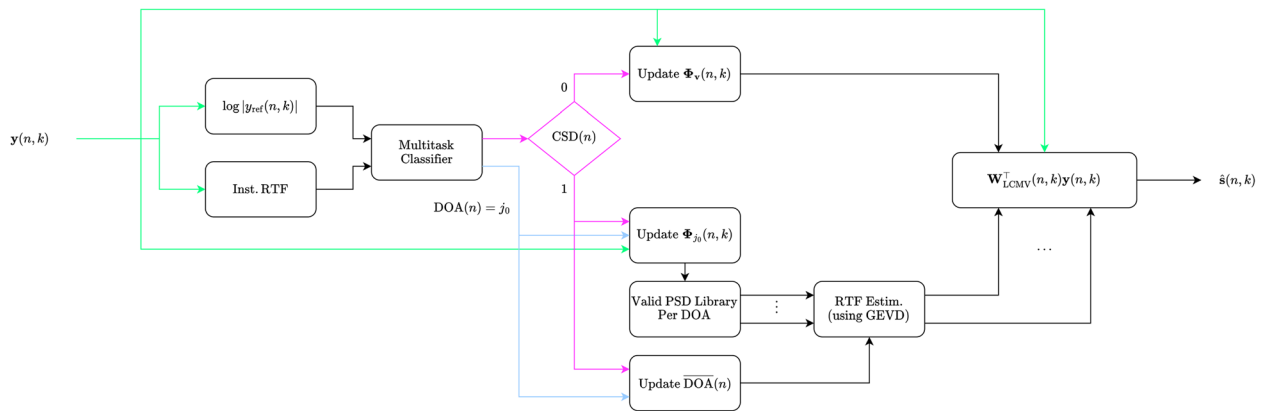


Fig. 1 Processing flow

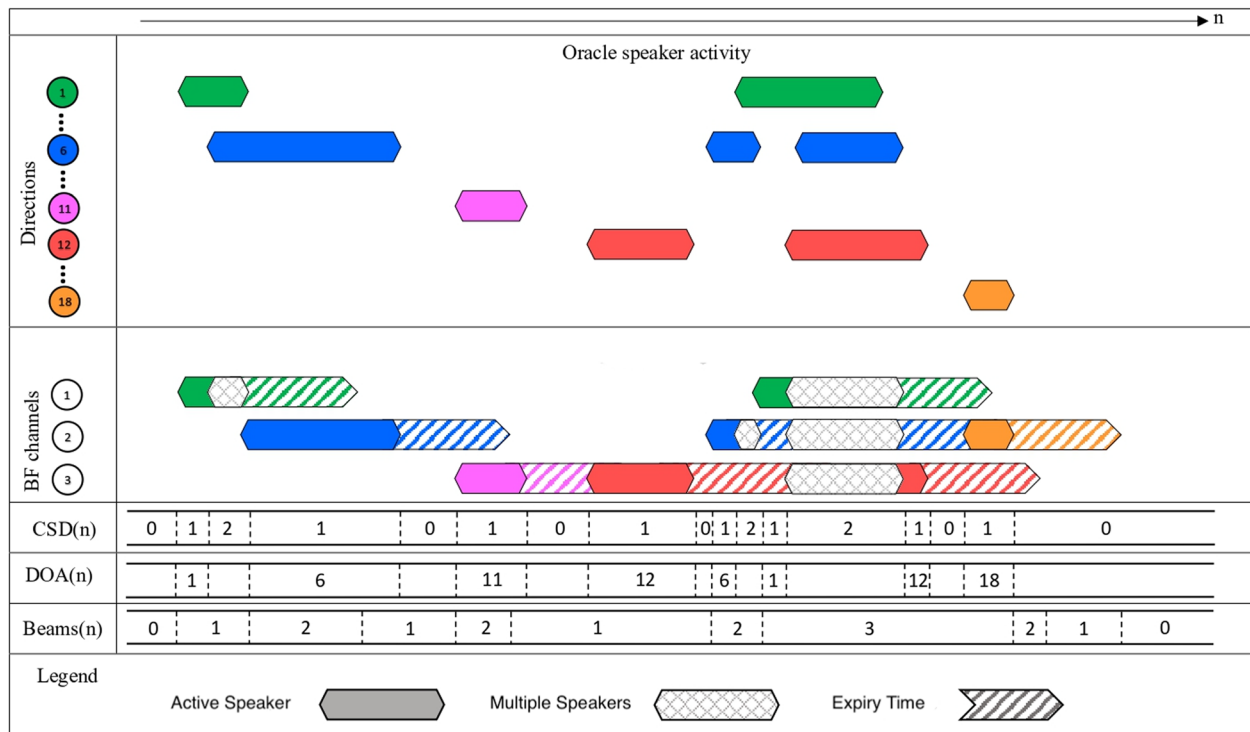


Fig. 2 Schematic activity patterns illustrating the algorithm flow outlined in Sect. 4. For illustration, a resolution of 10° was chosen. The actual activities per DOA are depicted, together with the corresponding CSD(n), and DOA(n), in frames when only a single speaker is active. The content of $\overline{DOA}(n)$ can be deduced from the color code of the constructed beams

number of constraints be kept as low as possible without sacrificing the suppression of the competing speakers. It is, therefore, imperative to construct the LCMV-BF with only the RTFs related to the currently active speakers.

The entire processing flow, from the noisy inputs to the enhanced and separated speakers, is summarized in Algorithm 1 and is illustrated in Fig. 1.

Figure 2 illustrates a sample processing flow. The upper panel of the figure illustrates the actual activities

of all speakers. In this example, five directions are active throughout the utterance, with no more than three sources concurrently active. We represent the various directions using a color code. Due to the dynamic nature of the problem, it is important to note that the same speaker can occupy different directions.

The active DOAs along the frame axis and the corresponding number of the BF outputs are depicted in the central panel. Note that in this illustration, $M = 4$,

and hence, the maximum number of outputs (i.e., BF beams) is $M - 1 = 3$. The colors of the BF outputs correspond to the active DOA colors. We can easily distinguish between frames with a single speaker and those with multiple speakers. The depiction also highlights the expiration time of each beam, the tracking of angularly adjacent DOAs (see the gradual change from “pink” to “red”), and the substitution of the oldest DOA with the newest one (see the transition from “blue” to “orange”) when the number of allowed constraints is exhausted. As per our design paradigm, the number of LCMV constraints is upper limited by $M - 1$. Consequently, we discard the beam associated with the oldest active speaker. For instance, in the right-hand section of the figure illustration, the “orange” beam replaces the “blue” beam. If the number of instantaneously active speakers exceeds the permissible number of constraints, namely if $J(n) > M - 1$, the LCMV beamformer may fail to extract the desired speaker and/or direct a null towards the interference sources. Note that in the illustration, $M - 1 = 3$; hence, no more than three beams can be simultaneously formed. In this example, the total number of sources in the scene is $J = 5$, but this does not hinder the ability of the beamformer to extract the desired source while attenuating the interfering sources.

Finally, in the bottom panel, the CSD class, DOA class, and the number of active BF outputs are also numerically designated.

Algorithm 1 Processing flow

```

while observing new frame  $n$  do
  Input: Noisy signal in the STFT domain  $\mathbf{y}(n, k)$ 
  DNN-based classifier:
  Construct the input feature matrix
   $[\mathbf{a}(n) \mathbf{B}^T(n)]$  (8),(9),(11)
  Obtain  $\text{CSD}(n)$  and  $\text{DOA}(n)$ .
  Update PSD matrices:
  if  $\text{CSD}(n) = 0$  then
    Update the noise PSD  $\Phi_v$  (17)
  else if  $\text{CSD}(n) = 1$  & DOA =  $j$  then
    Update the PSD  $\Phi_j(n)$  associated with
    DOA $_j(n)$  (18)
  Determine the active speakers:
  Update  $\overline{\text{DOA}}(n)$  as outlined in Sec. 4
  RTF estimation stage:
  Estimate the RTFs using the PSD matrices attributed
  to the DOAs in the current  $\overline{\text{DOA}}(n)$  using GEVD (10)
  Beamforming stage:
  Construct the LCMV-BF  $\mathbf{W}_{\text{LCMV}}(n, k)$  (5) using the
  estimated RTFs and the estimated noise PSD matrix
  Output: The LCMV outputs:
   $\hat{\mathbf{s}}(n, k) = \mathbf{W}_{\text{LCMV}}^H(n, k)\mathbf{y}(n, k)$ 

```

7 Experimental study

In this section, we assess the performance of the proposed scheme. We conduct separate evaluations of the building blocks and the overall system, comparing the results with different baseline techniques: (1) CSD results are compared with those from [44], (2) DOA results are compared with the steered response power with phase transform (SRP-PHAT) algorithm [48], and (3) speaker separation capabilities (in static scenarios) are compared with the ILRMA algorithm [45, 46]. For dynamic scenarios, we provide a qualitative demonstration of the performance of the proposed method.

7.1 Database generation

Train and test data generation is now discussed.

7.1.1 Training data

The proposed classifier was trained using simulated data. Signals from the TIMIT database [49] were convolved with synthetic room impulse responses (RIRs), generated using open-source software package¹ for RIR generation, which effectively implements the image method [50]. The reverberation level was randomly drawn in the range $T_{60} = 0.3 - 0.55$ s. The same array, comprising $M = 4$ microphones, was used throughout the training and test simulations. The microphones were organized in a semi-circle with a 10-cm radius and equal inter-microphone distance. We also used a similar structure in the real-life experiments. The DOA resolution of our model was set to 10° ($N = 18$ classes). The distance between the speakers and the microphone array center was set in the range $1 \div 1.5$ m.

Mixtures with a maximum of two concurrent speakers were simulated. Each signal is a summation of two partially overlapping single-speaker signals, thus constructing frames with 0,1,2 concurrent speakers, such that all values of the CSD are covered uniformly. The speakers’ signals were added with signal-to-interference ratio (SIR) in the $-5 \div 5$ dB range. To diversify the training data, the DOAs of the speakers were randomized for each utterance to cover the permissible range of directions uniformly. In the training stage, each speaker was static, while for the test database, speakers were free to move.

Three background noise types were added: (1) directional noise with SNR equal to 20 dB played from arbitrary positions in the room, (2) diffuse noise with SNR randomly drawn in the range $10 \div 20$ dB, and (3) spatially white sensor noise with SNR set to 30 dB. Throughout the experiments, the signal-to-noise ratio is measured with respect to the stronger speaker. Overall, the training

¹ <https://github.com/ehabets/RIR-Generator>

set comprises 500 simulated signals, each approximately 40-s long (5.55 h). We work in 16K Hz sampling rate, and the window size of the STFT was 2048 samples with an overlap of 1024 samples. The various simulation parameters are summarized in Table 1.

7.1.2 Test data: static speakers

Ten 40-s-long signals and static speakers were generated. Each signal starts with a 3-s-long noise-only segment², then each speaker speaks alone for 10 s, and finally, the two speakers are concurrently speaking for approximately 10 s. The speech signals are randomly drawn from the TIMIT test set. The DOAs of the speakers is randomized for each utterance.

7.1.3 Test data: simulated dynamic speakers

Ten 40-s-long signals and dynamic speakers were generated, which we have about 6500 frames. The speakers move at a speed of one meter every 3 s. In the simulated dynamic scenario, the first speaker starts at 0°, walks towards 140°, and then returns to 0°. The second speaker alternately moves between 160° and 180°. The speakers' movement is simulated using a signal generator³. Each speaker was static for a sufficiently long time before the concurrent speakers period to facilitate the RTF estimation. Overall, we used ten sentences in this test scenario.

7.1.4 Test data: real dynamic speakers

We conducted experiments in real dynamic scenarios to further evaluate the proposed method. The recordings took place at the acoustic lab at Bar-Ilan University, allowing for a wide range of reverberation levels. In this specific case, we examined a reverberation level of $T_{60} = 390$ ms. The setup involved four AKG CK32 microphones arranged on a semi-circle array constellation, similar as much as possible to the simulated array, assembled on a plastic construction. During the experiments, speakers walked naturally along an arc, keeping a distance of approximately 2.2 m from the center of the microphone array. We recorded three signals in this test scenario. To facilitate the evaluation of the separation algorithm using real recordings, we individually recorded each speaker and the noise signal. Then, we combined these recordings to create the mixture.

² Note that it is not mandatory to start the recording with noise-only segments. In cases where a speaker is active from the beginning of the utterance, the noise spatial PSD matrix will be initialized with an identity matrix. This should not have a major impact on the source extraction capabilities and may only degrade the noise reduction capabilities.

³ <https://github.com/ehabets/Signal-Generator>

7.2 Performance measures

Each building block of the proposed system was separately evaluated, and then the entire system's performance was assessed. The following performance measures were used.

The performance of the CSD was evaluated by analyzing confusion matrices.

The performance of the DOA classifier was evaluated by analyzing the histogram of the estimation errors, as detailed below.

When assessing the separation and noise reduction capabilities of the algorithms, three measures were employed, namely, short-term objective intelligibility (STOI) [51], signal-to-interference ratio (SIR) [52], and scale-invariant signal-to-distortion ratio (SI-SDR) [53]. Only double-talk segments, namely if $p_{\text{CSD}} = 2$, were used for this evaluation.

7.3 CSD performance

In this section, the performance of the proposed CSD is compared to the multi-channel concurrent speakers detector (MCCSD) [44]. There are two main differences between the proposed model and the MCCSD. First, the input to MCCSD is the individual log-spectrum $\log |y(n, k)|$ of all microphone signals (with the past and future context frames); hence, no phase information is considered. In the proposed model, the log-spectrum of the reference microphone is used together with an estimate of all RTFs, which considers the acoustic propagation between the microphones. Second, in the proposed model, another output is defined, namely a classification of the DOA range. Overall, the proposed model is only 10% computationally more intensive than the MCCSD.

The results of the CSD were obtained using the dynamic test data, as detailed in Sect. 7.1.3. Tables 2, 3, and 4 depict the confusion matrices of the MCCSD, the proposed CSD without the DOA classification branch, and the proposed CSD with the DOA classification branch, respectively.

It can be observed that the proposed model is more accurate than the MCCSD, even without the additional DOA classification branch. When the DOA classification branch is incorporated, the accuracy of Class #1 classification is higher. This may be attributed to the additional DOA loss function during single-speaker frames, improving its detection accuracy. Moreover, we note from Table 4 that misclassifying Class #2 frames as Class #1 occurs less frequently compared to other schemes. This has a beneficial effect on the overall performance of the separation algorithm, as such an error could lead to incorrect estimation of the steering vector, potentially directing the beam away from the sources of interest (either desired or interfering).

Table 2 Confusion matrix of MCCSD [%]

Estimated \ True	Class 0	Class 1	Class 2
Class 0	88.3	12.5	0.4
Class 1	9.9	75.4	15.8
Class 2	1.8	12.1	83.8

Table 3 Confusion matrix of CSD without DOA classifier [%]

Estimated \ True	Class 0	Class 1	Class 2
Class 0	90.4	3.7	0.4
Class 1	8.6	78.3	9.4
Class 2	1.0	18.0	90.2

Table 4 Confusion matrix of CSD with DOA classifier [%]

Estimated \ True	Class 0	Class 1	Class 2
Class 0	91.1	1.9	0.0
Class 1	8.3	85.9	4.7
Class 2	0.6	12.2	95.3

7.4 Performance of the DOA classifier

In this section, the performance of the proposed DOA-range classifier is compared with the classical SRP-PHAT [48] algorithm. In our model, the DOA output of the DNN model is only valid for frames classified by the CSD as single speaker frames.

We evaluated the DOA classification accuracy using speech utterances with a single speaker moving along an arc from 0° to 180° in reverberant environments for several reverberation levels and SNR levels. The speaker's movement speed was set to 0.33 m/s. As a baseline algorithm, we chose the SRP-PHAT [48]. Let

$$\hat{\Phi}_{\mathbf{y}}(n, k) = \sum_{m=n-m_1}^{n+m_2} w_m \mathbf{y}(m) \mathbf{y}^H(m) \quad (19)$$

be an instantaneous estimate of the spatial PSD matrix of the received microphone signals, obtained by averaging the context frames $n - m_1, \dots, n + m_2$.

Assuming far-field propagation (an assumption that is violated in a reverberant environment and/or when the source is close to the microphone array), the propagation between j -th speaker and microphone q is determined by the propagation delay:

$$G_{j,q} = \exp\left(-\iota \frac{2\pi k}{K} \frac{\tau_{j,q}}{T_s}\right), \quad (20)$$

where $\tau_{j,q}$ is the time delay between the j -th speaker and microphone q .

The SRP-PHAT DOA estimator is obtained by maximizing:

$$\hat{j}_{\text{SRP}}(n) = \operatorname{argmax}_j \sum_{q_1=1}^M \sum_{q_2=q_1+1}^M \sum_k \frac{\hat{\Phi}_{\mathbf{y},q_1q_2}(n, k)}{\left| \hat{\Phi}_{\mathbf{y},q_1q_2}(n, k) \right|} \frac{G_{j,q_1}^*}{G_{j,q_2}^*}, \quad (21)$$

where $\hat{\Phi}_{\mathbf{y},q_1q_2}$ are the elements of the spatial PSD matrix, with q_1, q_2 representing the indexes of a pair of microphones.

For a fair comparison with the proposed DOA estimator, we scan over all ranges defined in (7).

The DOA estimation plays a crucial role in supporting the speaker separation task. When evaluating the estimation error of the DOA, we categorize the errors into three levels. Recall that in our analysis, the resolution of the DOA estimates is 10° .

- 1 *Successful estimation*: The estimated DOA class is correct, namely $\text{pDOA} = \text{aDOA}$.
- 2 *Low estimation error*: If $|\text{pDOA} - \text{aDOA}| \leq 20^\circ$, the estimation error is categorized as low. Such an error may either occur if a speaker is mistakenly associated with another adjacent speaker or when the same speaker is mistakenly classified into two DOAs. The latter error will increase the number of constraints in the LCMV design and could potentially degrade the noise reduction capabilities of the beamformer.
- 3 *High estimation error*: If $|\text{pDOA} - \text{aDOA}| > 20^\circ$, the estimation error is categorized as high. In such cases, the speech frames may be miscategorized into the wrong DOA range, leading to inaccurate estimates of the PSD matrices. Consequently, this will result in erroneous estimates of the RTE.

The proposed DOA estimator and the SRP-PHAT algorithm are compared in Fig. 3 for different reverberation times and $\text{SNR} = 20$ dB. We analyze the errors according to the categories explained above. It is evident that the performance of the proposed DOA estimator is almost independent of the reverberation levels and that it consistently outperforms the SRP-PHAT algorithm across all examined values of T_{60} . Notably, the analysis of the bar plots reveals that the proposed method yields mostly successful estimates, with a very low occurrence of high estimation errors.

7.5 Overall speaker separation capabilities

In this section, we analyze the speaker separation capabilities of the proposed algorithm, including its CSD,

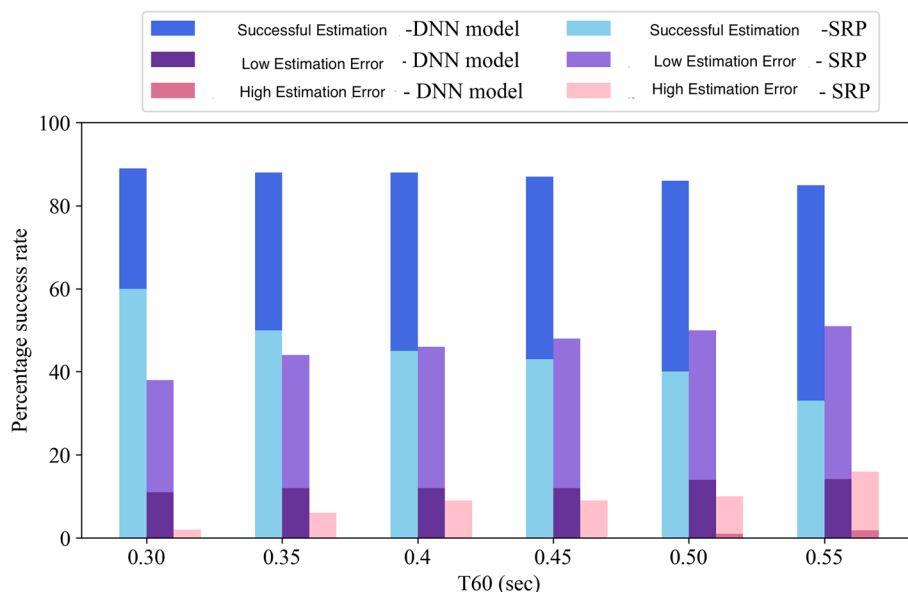


Fig. 3 DOA estimation: performance comparison between the proposed DNN-based the SRP-PHAT algorithms for different reverberation times

DOA blocks, and the online LCMV-BF. We compare the proposed algorithm in the static case with the ILRMA algorithm [45, 46]. ILRMA is an offline algorithm based on independent vector analysis (IVA) and non-negative matrix factorization (NMF) that extracts a predefined number of sources from a mixture, typically equal to the number of microphones used. Due to the inherent permutation ambiguity, we report the SIR results for the correct permutation.

In Table 5, two simulated scenarios are considered: static and dynamic. It is essential to highlight that, in our analysis, speakers are permitted to move only during single-speaker segments. This restriction is imposed because the weights of the BF cannot update if more than one speaker is active. In the dynamic scenario, the velocity of the speakers is set to approximately 0.3 m/s. The ILRMA algorithm is only examined in the static scenario, as it is not designed to operate in dynamic cases.

Table 5 Performance measures for the proposed LCMV-BF with control based CSD-DOA classifiers. The STOI measure is presented as an improvement from noisy to enhanced signal. The SI-SDR and the SIR measures indicate the *improvement* compared to the mixture signals

	Algorithm	STOI	SI-SDR	SIR	STOI	SI-SDR	SIR	STOI	SI-SDR	SIR
		SNR				15dB				
		20dB				10dB				
Static scenario	ILRMA [45, 46]	81→90	6.2	8.1	79→90	5.5	7.6	75→86	5.8	7
	Proposed	81→99	15.5	16.5	79→98	10.5	10.9	75→95	9.8	7
Dynamic scenario	Proposed	72→99	8.8	14.5	68→97	8.5	14.3	63→92	7.8	12.1
		RT60				0.4sec				
		0.3sec				0.5sec				
Static scenario	ILRMA [45, 46]	80→91	6.2	8.1	75→82	5.4	6.6	70→77	5.1	4.3
	Proposed	80→98	10.9	16.5	75→91	10.2	15.3	70→90	9.7	14.6
Dynamic scenario	Proposed	70→97	8.8	14.5	64→95	7.2	11.2	60→87	6.5	9.4
		SIR in				-5dB				
		0dB				-10dB				
Static scenario	ILRMA [45, 46]	80→91	6.2	8.1	40→63	6.7	9.5	25→42	6.8	9.3
	Proposed	80→98	10.6	16.5	40→82	10.9	15.6	25→78	10.1	15.8
Dynamic scenario	Proposed	70→98	8.8	14.5	25→88	7.5	14.2	10→51	7.2	13.8

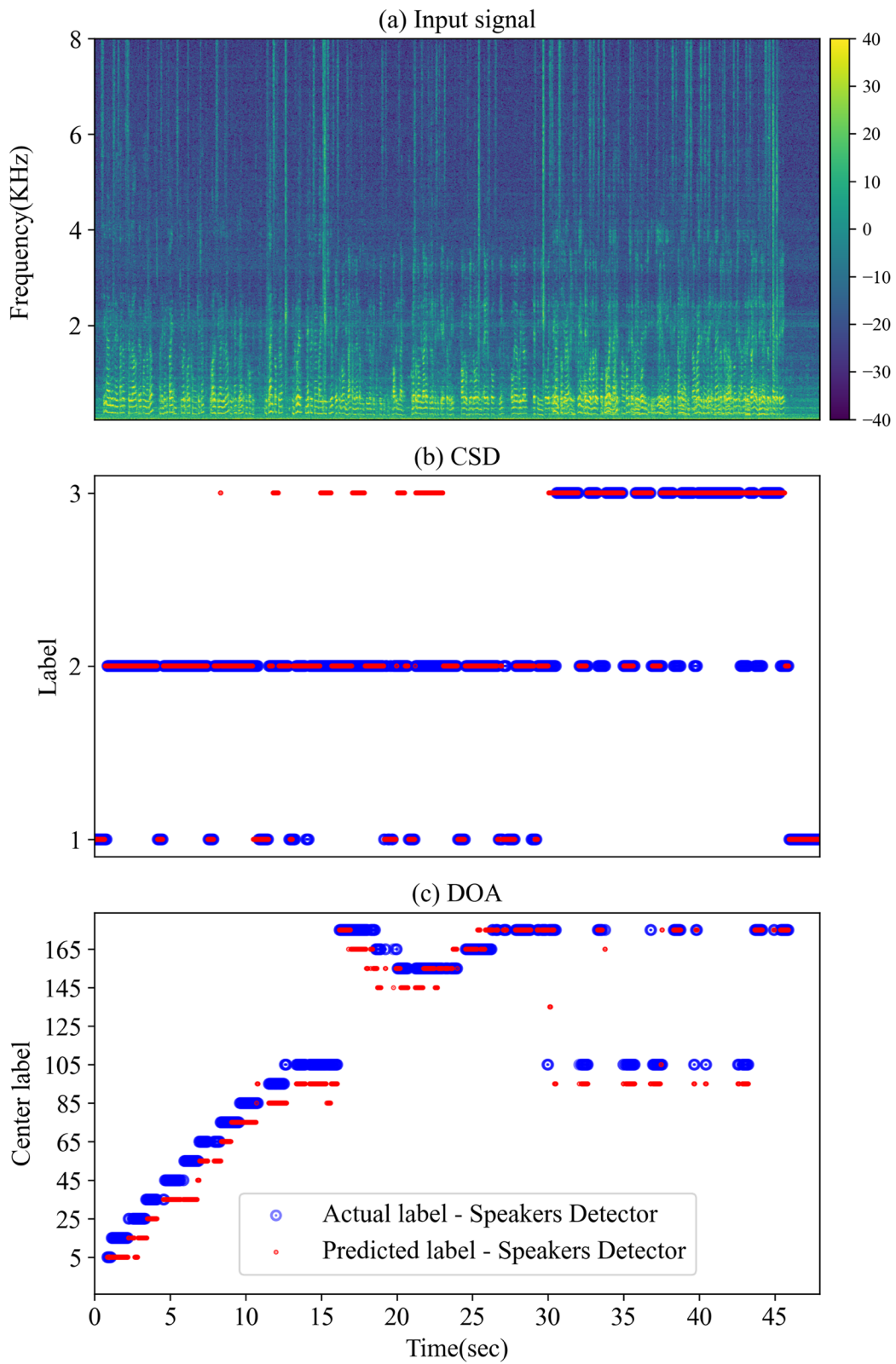


Fig. 4 CSD and DOA classification results for a sample real recording

The proposed algorithm consistently outperforms ILRMA across all tested cases. Furthermore, it is evident that even in dynamic scenarios, the proposed algorithm effectively enhances the input measures, yielding results comparable to those in the static scenario despite the dynamic behavior of the speakers.

Figure 4 depicts an instance of a real recording with an input SNR of 10 dB and SIR of 0 dB. The successful classification results of both the CSD and DOA classifiers are clearly evident, even for this challenging real-life dynamic scenario. After averaging the outcomes from three actual recordings, it was found that the proposed CSD and DOA classifiers correctly classified 86.1% and 88.4% of the frames, respectively. Moreover, from inspecting the estimated DOA trace of the demonstrated utterance, we see that the error usually does not exceed 10° , namely, most of the errors are low estimation errors. The average signal-to-distortion ratio (SDR) and SIR improvements (of the three real recordings) were 8.6 dB and 12.1 dB, respectively.

Several factors may limit the performance of the proposed scheme. First and foremost, the algorithm is “blind” during Class #2 frames, namely, the beamformer’s building blocks are not updated. Hence, if the source(s) move when concurrent speaker activity is detected, the “frozen” beamformer weights will become outdated. Consequently, the beams will not be properly directed, resulting in a performance drop. Another notable limitation of the proposed scheme is actually a limitation of the LCMV beamformer. The LCMV criterion can only support $M - 1$ constraints. Hence, if the number of sources of interest (both desired and interference) is larger than this threshold, the constraints cannot be satisfied, hindering the ability to extract the desired source. Besides, the CSD and DOA detectors learn from training data like all neural network (NN)-based methods. In cases of mismatch between the training and test data, e.g., higher reverberation level, lower SNR values, and different noise types, the performance of the detector may degrade. Furthermore, as the beamformer building blocks are utilizing an estimate of spatial PSD matrices, having a reliable estimate of them is important. If the speed of movement of the sources is very high, these estimates may be inaccurate due to the low number of available frames.

8 Conclusions

We introduced an online algorithm for separating static and moving sources. The backbone of the noise reduction and the separation algorithm is an LCMV-BF. In this paper, we proposed a novel control mechanism for

estimating the building blocks of the beamformer. We used a multi-task CSD and DOA classifiers to jointly infer the activities of the speakers and their DOA. The LCMV-BF is constructed with RTF-based steering vectors, and the consistency of the separated outputs is preserved using the associated estimated DOA.

The proposed method was thoroughly evaluated using both simulated and recorded data in both static and dynamic scenarios and was shown to outperform the state-of-the-art ILRMA algorithm. Remarkably, the algorithm, which is trained using only simulated and static data, was able to separate moving sources even in a real-life acoustic environment.

Authors’ contributions

Model development: AS, SC, OS, and SG. Experimental testing: AS. Writing paper: AS, OS, and SG.

Funding

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245, from the Israeli Ministry of Science & Technology, and from Meta.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Consent for publication

All authors agree to the publication in this journal.

Competing interests

The authors declare that they have no competing interests.

Received: 23 February 2024 Accepted: 5 August 2024

Published online: 04 October 2024

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. *Audio Source Separation*, ed. by S. Makino. Signals and communication technology (Springer, Cham, 2018)
4. J. Capon, High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* **57**(8), 1408–1418 (1969)
5. S. Gazor, S. Affes, Y. Grenier, Robust adaptive beamforming via target tracking. *IEEE Trans. Signal Proc.* **44**(6), 1589–1593 (1996)
6. H.L. Van Trees, *Optimum array processing: part IV of detection, estimation, and modulation theory* (John Wiley & Sons, New-York, 2004)
7. S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans. Signal Proc.* **49**(8), 1614–1626 (2001)
8. 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)

9. M.H. Er, A. Cantoni, Derivative constraints for broad-band element space antenna array processors. *IEEE Trans. Acoust. Speech Sig. Process.* **31**(6), 1378–1393 (1983)
10. 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)
11. 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)
12. E.A. Habets, J. Benesty, S. Gannot, P.A. Naylor, I. Cohen, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, On the application of the LCMV beamformer to speech enhancement (IEEE, 2009), pp. 141–144
13. S. Markovich-Golan, S. Gannot, I. Cohen, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Subspace tracking of multiple sources and its application to speakers extraction (IEEE, 2010), pp. 201–204
14. R. Varzandeh, M. Taseska, E.A.P. Habets, in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation (IEEE, 2017), pp. 11–15
15. S. Markovich-Golan, S. Gannot, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method (IEEE, 2015), pp. 544–548
16. C. Li, J. Martinez, R.C. Hendriks, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Low complex accurate multi-source RTF estimation (IEEE, 2022), pp. 4953–4957
17. I. Cohen, S. Gannot, B. Berdugo, An integrated real-time beamforming and postfiltering system for nonstationary noise environments. *EURASIP J. Appl. Signal Process.* **2003**, 1064–1073 (2003)
18. S. Gannot, I. Cohen, Speech enhancement based on the general transfer function GSC and postfiltering. *IEEE Trans. Speech Audio Process.* **12**(6), 561–571 (2004)
19. T. Higuchi, N. Ito, T. Yoshioka, T. Nakatani, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise (IEEE, 2016), pp. 5210–5214
20. J. Heymann, L. Drude, R. Haeb-Umbach, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Neural network based spectral mask estimation for acoustic beamforming (IEEE, 2016), pp. 196–200
21. H. Erdogan, J.R. Hershey, S. Watanabe, M.I. Mandel, J. Le Roux, in *Inter-speech*, Improved MVDR beamforming using single-channel mask prediction networks (ISCA, 2016), pp. 1981–1985
22. T. Nakatani, N. Ito, T. Higuchi, S. Araki, K. Kinoshita, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming (IEEE, 2017), pp. 286–290
23. Y. Xu, M. Yu, S.X. Zhang, L. Chen, C. Weng, J. Liu, D. Yu, in *Proc. Inter-speech 2020*, Neural spatio-temporal beamformer for target speech separation (2020), pp. 56–60. <https://doi.org/10.21437/Interspeech.2020-1458>
24. R. Talmon, I. Cohen, S. Gannot, Convolutional transfer function generalized sidelobe canceler. *IEEE Trans. Audio Speech Lang. Process.* **17**(7), 1420–1434 (2009)
25. T. Ochiai, M. Delcroix, T. Nakatani, S. Araki, Mask-based neural beamforming for moving speakers with self-attention-based tracking. *IEEE/ACM Trans. Audio Speech Lang. Process.* **31**, 835–848 (2023)
26. 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 (IEEE, 2008), pp. 93–96
27. 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)
28. D. Cherkassky, S. Gannot, Successive relative transfer function identification using blind oblique projection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 474–486 (2019)
29. H. Gode, S. Doclo, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Covariance blocking and whitening method for successive relative transfer function vector estimation in multi-speaker scenarios (IEEE, 2023)
30. B. Laufer-Goldshtein, R. Talmon, S. Gannot, Source counting and separation based on simplex analysis. *IEEE Trans. Signal Process.* **66**(24), 6458–6473 (2018)
31. 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)
32. B. Laufer-Goldshtein, R. Talmon, S. Gannot, Audio source separation by activity probability detection with maximum correlation and simplex geometry. *EURASIP J. Audio Speech Music* (2021). <https://rdcu.be/ch29B>
33. 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 (IEEE, 2018), pp. 6712–6716
34. 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)
35. Z. Zhang, Y. Xu, M. Yu, S.X. Zhang, L. Chen, D. Yu, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ADL-MVDR: All deep learning MVDR beamformer for target speech separation (IEEE, 2021), pp. 6089–6093
36. X. Ren, X. Zhang, L. Chen, X. Zheng, C. Zhang, L. Guo, B. Yu, in *Proc. Inter-speech 2021*, A causal U-Net based neural beamforming network for real-time multi-channel speech enhancement (ISCA, 2021), pp. 1832–1836
37. Z.Q. Wang, J. Le Roux, J.R. Hershey, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Multi-channel deep clustering: discriminative spectral and spatial embeddings for speaker-independent speech separation (IEEE, 2018)
38. P.A. Grumiaux, S. Kitić, L. Girin, A. Guérin, A survey of sound source localization with deep learning methods. *J. Acoust. Soc. Am.* **152**(1), 107–151 (2022)
39. S. Adavanne, A. Politis, J. Nikunen, T. Virtanen, Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE J. Sel. Top. Signal Process.* **13**(1), 34–48 (2018)
40. A. Bohlender, A. Spriet, W. Tirry, N. Madhu, Exploiting temporal context in CNN based multisource DOA estimation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 1594–1608 (2021)
41. D. Diaz-Guerra, A. Miguel, J.R. Beltran, Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 300–311 (2020)
42. B. Yang, H. Liu, X. Li, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, SRP-DNN: learning direct-path phase difference for multiple moving sound source localization (IEEE, 2022), pp. 721–725
43. H. Hammer, S.E. Chazan, J. Goldberger, S. Gannot, Dynamically localizing multiple speakers based on the time-frequency domain. *EURASIP J. Audio Speech Music* (2021). <https://rdcu.be/cilAr>
44. S.E. Chazan, H. Hammer, G. Hazan, J. Goldberger, S. Gannot, in *27th European Signal Processing Conference (EUSIPCO)*, Multi-microphone speaker separation based on deep DOA estimation (EURASIP, 2019)
45. D. Kitamura, N. Ono, H. Sawada, H. Kameoka, H. Saruwatari, Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(9), 1626–1641 (2016)
46. D. Kitamura, N. Ono, H. Sawada, H. Kameoka, H. Saruwatari, Determined blind source separation with independent low-rank matrix analysis. *Audio source separation* (Springer International Publishing, Cham, 2018), pp. 125–155. *Signals and Communication Technology*
47. B. Yang, Projection approximation subspace tracking. *IEEE Trans. Signal Process.* **43**(1), 95–107 (1995)
48. J. DiBiase, H. Silverman, M. Brandstein, in *Microphone arrays: Signal processing techniques and applications*, ed. by M. Brandstein, D. Ward. Robust localization in reverberant rooms (Springer-Verlag, Berlin, Heidelberg, 2001), pp. 157–180
49. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. Nist speech disc 1-1.1. NASA STI/Recon Technical Report N. **93**, 27,403 (1993)
50. J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* **65**(4), 943–950 (1979)

51. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* **19**(7), 2125–2136 (2011)
52. E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)
53. J. Le Roux, S. Wisdom, H. Erdogan, J.R. Hershey, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, SDR–half-baked or well done? (IEEE, 2019), pp. 626–630

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.