Hammer *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:16
https://doi.org/10.1186/s13636-021-00203-w

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Dynamically localizing multiple speakers based on the time-frequency domain

Hodaya Hammer, Shlomo E. Chazan, Jacob Goldberger and Sharon Gannot*

Abstract

In this study, we present a deep neural network-based online multi-speaker localization algorithm based on a multi-microphone array. Following the W-disjoint orthogonality principle in the spectral domain, time-frequency (TF) bin is dominated by a single speaker and hence by a single direction of arrival (DOA). A fully convolutional network is trained with instantaneous spatial features to estimate the DOA for each TF bin. The high-resolution classification enables the network to accurately and simultaneously localize and track multiple speakers, both static and dynamic. Elaborated experimental study using simulated and real-life recordings in static and dynamic scenarios demonstrates that the proposed algorithm significantly outperforms both classic and recent deep-learning-based algorithms. Finally, as a byproduct, we further show that the proposed method is also capable of separating moving speakers by the application of the obtained TF masks.

Keywords: DOA, UNET, Tracking

1 Introduction

Localizing multiple sound sources recorded with a microphone array in an acoustic environment is an essential component in various cases such as source separation and scene analysis. The relative location of a sound source with respect to a microphone array is specified in the term of the DOA of the sound wave originating from that location. DOA estimation and tracking are essential building blocks in all modern far-field speech enhancement and recognition for smart home devices as well as robot audition applications. In real-life environments, sound sources are captured by the microphones together with acoustic reverberation. While propagating in an acoustic enclosure, the sound wave undergoes reflections from the room facets and from various objects. These reflections deteriorate speech quality and, in extreme cases, its intelligibility. Furthermore, reverberation increases the time dependency between speech frames, making source DOA estimation a very challenging task.

*Correspondence: sharon.gannot@biu.ac.il
Faculty of Electrical Engineering, Ramat-Gan, Israel

A plethora of classic signal processing-based approaches have been proposed throughout the years for the task of broadband DOA estimation. The multiple signal classification (MUSIC) algorithm [1] applies a subspace method that was later adapted to the challenges of speech processing in [2]. The steered response power with phase transform (SRP-PHAT) algorithm [3] used a generalization of cross-correlation methods for DOA estimation. These methods are still widely in use for both single- and multi-speaker localization tasks. However, in highly reverberant enclosures, their performance rapidly deteriorates [4, 5].

Supervised learning methods can be potentially advantageous for this task since they are data-driven. Deep neural networks can be trained to find the DOA in different acoustic conditions. Moreover, if a network is trained using rooms with different acoustic conditions and multiple noise types, it can be made robust against noise and reverberation even for rooms which were not in the training set. Deep learning methods have recently been proposed for sound source localization. In [6, 7], simple feed-forward deep neural networks (DNNs) were trained using generalized cross correlation (GCC)-based

audio features, demonstrating improved performance as compared with classical approaches. Yet, this method is mainly designed to deal with a single sound source at a time. An extension of the multi-speaker DOA, using a DNN for the estimation task, can be found in [8]. In high reverberation conditions, however, the performance of these algorithms is not satisfactory. In [9] and [10], time-domain features were used demonstrating performance improvement in highly reverberant enclosures. In [11], a CNN-based classification method was applied in the short-time Fourier transform (STFT) domain for broadband DOA estimation, assuming that only a single speaker is active per time frame. The phase component of the STFT coefficients of the input signal were directly provided as input to the CNN. This work was extended by [5] to estimate multiple speakers' DOAs and has shown high DOA classification performance.

The main drawback of most DNN-based approaches, however, is that they only use low-resolution supervision, namely at the time frame level or even utterance-based level, and the network outputs a single localization decision for the entire time frame. For speech signals, however, each time-frequency bin is dominated by a different speaker, a property referred to as W-disjoint orthogonality (WDO) [12]. In the case of multiple speakers, each TF bin can therefore be associated with a different DOA. This high-resolution information can yield an improved DOA estimation also in the entire time frame localization resolution, especially in the case of multiple speakers.

In this study, we present a multi-speaker DOA estimation algorithm that is based on the U-net architecture that infers the DOA of each TF bin. The DOA decisions of all the frequency bands of a single time frame are then aggregated to extract the active speakers at that time frame level. The TF-based classification also facilitates the tracking capabilities of multiple moving speakers. U-Net has been introduced in the medical imaging domain [13] and was recently successfully applied to various audio processing tasks, e.g., for speech dereverberation [14], speaker separation [15], and noise reduction [16], all in the STFT domain, and for speech enhancement in the time-domain [17, 18] also employing self-attention mechanism.

In the current study, we show that U-net architecture is also beneficial in speaker localization and tracking applications. We tested the proposed method on simulated data, using publicly available room impulse responses (RIRs) recorded in a real room [19], as well as real-life experiments recorded at the acoustic lab, Bar-Ilan University. We show that the proposed algorithm significantly outperforms state-of-the-art methods.

The main contribution of our work is casting the time-domain DOA estimation problem into a time-frequency segmentation problem. The proposed method improves the DOA estimation performance with respect to (w.r.t.) the state-of-the-art (SOTA) approaches, which are frame-based, and facilitates simultaneous tracking of multiple moving speakers.

2 Multiple-speaker localization algorithm

In this section, we describe the proposed algorithm, including the feature extraction, the network architecture, and the training procedure.

2.1 Multi-microphone time-frequency features

Consider an array with M microphones acquiring a mixture of N speech sources in a reverberant environment. The ith speech signal $s^i(t)$ propagates through the acoustic channel before being acquired by the mth microphone:

$$z_m(t) = \sum_{i=1}^{N} \left\{ s^i * h_m^i \right\}(t), \qquad m = 1, \ldots, M, \tag{1}$$

where h_m^i is the RIR relating the ith speaker and the mth microphone. In the STFT domain, (1) can be written as (provided that the frame-length is sufficiently large w.r.t. the filter length):

$$z_m(l,k) = \sum_{i=1}^{N} s^i(l,k) h_m^i(l,k), \tag{2}$$

where l and k are the time frame and the frequency indices, respectively.

The STFT (2) is complex-valued and hence comprises both magnitude and phase information. It is clear that the magnitude information alone is insufficient for DOA estimation. It is therefore a common practice to use the phase of the TF representation of the received microphone signals, or their respective phase-difference, as they are directly related to the DOA in non-reverberant environments. We decided to use an alternative feature, which is generally independent of the speech signal and is mainly determined by the spatial information. For that, we have selected the relative transfer function (RTF) [20] as our feature, since it is known to encapsulate the spatial fingerprint for each sound source. Specifically, we use the instantaneous relative transfer function (iRTF), which is the bin-wise ratio between the mth microphone signal and the reference microphone signal $z_{\text{ref}}(l,k)$:

$$\text{iRTF}(m,l,k) = \frac{z_m(l,k)}{z_{\text{ref}}(l,k)}. \tag{3}$$

Note that the reference microphone is arbitrarily chosen. Reference microphone selection is beyond the scope of this paper (see [21] for a reference microphone selection method). The input feature set extracted from the recorded signal is thus a 3D tensor $\mathcal{R}$:

$$\mathcal{R}(l,k,m) = \left[\Re\mathfrak{e}(\text{iRTF}(m,l,k)), \Im\mathfrak{m}(\text{iRTF}(m,l,k)) \right]. \tag{4}$$

The tensor $\mathcal{R}$ is constructed from $L \times K$ bins, where L is the number of time frames and K is the number of frequencies. Since the iRTFs are normalized by the reference microphone, the latter is excluded from the features. Then, for each TF bin (l,k), there are $P = 2(M-1)$ channels, where the multiplication by 2 is due to the real and imaginary parts of the complex-valued feature. For each TF bin, the spatial features were normalized to have a zero mean and a unit variance. Other feature extraction methods can be considered. In Section 3, we show that the features described above are a suitable choice for the localization task.

2.2 U-Net for DOA estimation

The WDO assumption [12, 22] implies that each TF bin (l,k) is dominated by a single speaker. Consequently, as the speakers are spatially separated, i.e., located at different DOAs, each TF bin is dominated by a single DOA. We first accurately estimate the speaker direction at every TF bin from the given mixed recorded signal. Then, we extract the speakers' locations at each time frame.

We formulated the DOA estimation as a classification task by discretizing the DOA range. The resolution was set to 5°, such that the DOA candidates are in the set $\Theta = \{0°, 5°, 10°, \ldots, 180°\}$. It is natural to view the DOA estimation as a regression problem. The regression output is a Gaussian unimodal distribution. Casting the problem as a classification yields a multi-modal distribution which is more suitable for the case of several speakers. Let $D_{l,k}$ be a random variable (r.v.) representing the active dominant direction, recorded at bin (l,k). Our task boils down to deducing the conditional distribution of the discrete set of DOAs in Θ for each TF bin, given the recorded mixed signal:

$$\mathcal{P}_{l,k}(\theta) = p(D_{l,k} = \theta|\mathcal{R}), \quad \theta \in \Theta. \tag{5}$$

For this task, we use a DNN. The network output is an $L \times K \times |\Theta|$ tensor, where $|\Theta|$ is the cardinality of the set Θ. Under this construction of the feature tensor and output probability tensor, a pixel-to-pixel approach [23] for mapping a 3D input "image," $\mathcal{R}$, and a 3D output "image," $\mathcal{P}$, can be utilized. A U-net is used to compute (5) for each TF bin. The pixel-to-pixel method is beneficial in two ways. First, for each TF bin in our input image, the network estimates the DOA distribution separately. Second, the TF supervision is carried out with the spectrum of the different speakers. The U-Net hence takes advantage of the spectral structure and the continuity of the sound sources in both the time and frequency axes. These structures contribute to the pixel-wise classification task and prevent discontinuity in the DOA decisions over time. In our implementation, we used a U-net architecture, similar to the one described in [24].

The input to the network is the feature tensor $\mathcal{R}$ (see (4)). In our U-net architecture, the input shape is (L,K,P) where $K = 256$ is the number of frequency bins, $L = 256$ is the number of frames, and $P = 2M - 2$ where M is the number of microphones.

TF bins in which there is no active speech are non-informative. Therefore, the estimation is carried out only on speech-active TF bins. As we assume that the acquired signals are noiseless, we define a TF-based voice activity detector (VAD) as follows:

$$\mathrm{VAD}(l,k) = \begin{cases} 1 & |z_{\mathrm{ref}}(l,k)| \geq \epsilon \\ 0 & \text{o.w.} \end{cases}, \tag{6}$$

where ϵ is a threshold value. In noisy scenarios, we can use a robust speech presence probability (SPP) estimator instead [25].

The DOAs should only be estimated on a time frame basis. Hence, we aggregate over all active frequencies at time frame l to obtain a frame-wise probability:

$$p_l(\theta) = \frac{1}{K'} \sum_{k=1}^{K} \mathcal{P}_{l,k}(\theta)\mathrm{VAD}(l,k), \qquad \theta \in \Theta \tag{7}$$

where K' is the number of frequency bands for which (6) exceed the threshold at the lth time frame. We thus obtain for each time frame a posterior distribution over all possible DOAs. If the number of speakers is known in advance, we can choose the directions corresponding to the highest posterior probabilities. If an estimate of the number of speakers is also required, it can be determined by applying a proper threshold. We dub our algorithm time-frequency direction-of-arrival net (TF-DOAnet). Figure 1 summarizes the TF-DOAnet network architecture. The algorithm is summarized in Table 1.

2.3 Model training

The supervision in the training phase is based on the WDO assumption in which each TF bin is dominated by (at most) a single speaker. The training is based on simulated data generated by a publicly available RIR generator software[1], efficiently implementing the image method [26]. A four microphone linear array was simulated with $(8,8,8)$ cm inter-microphone distances. Similar microphone inter-distances were used in the test phase. For each training sample, the acoustic conditions were randomly drawn from one of the simulated rooms of different sizes and different reverberation levels RT_{60} as described in Table 2. The microphone array was randomly placed in the room in one out of six arbitrary positions.

For each scenario, two clean signals were randomly drawn from the Wall Street Journal 1 (WSJ1) database [27] and then convolved with RIRs corresponding to two

[1] Available online at `github.com/ehabets/RIR-Generator`.

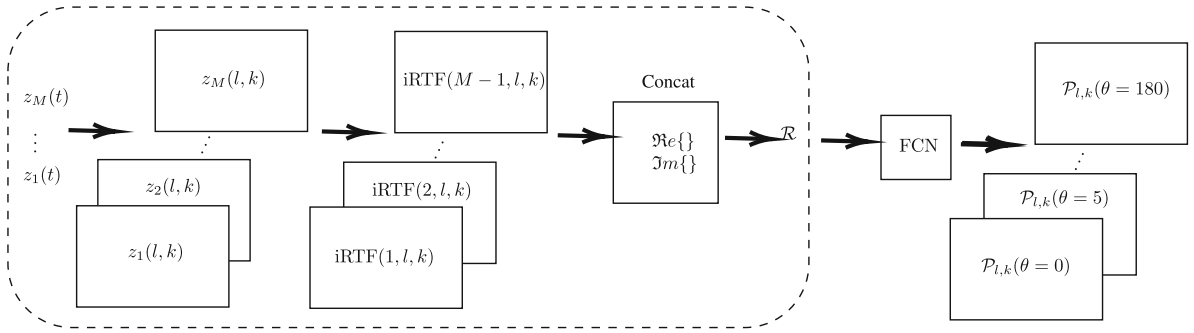


Fig. 1 Block diagram of the TF-DOAnet algorithm. The dashed envelope describes the feature extraction step

different DOAs in the range $\Theta = \{0, 5, \ldots, 180\}$. The sampling rate of all signals and RIRs was set to 16 KHz. The speakers were positioned on a radius of $r = 1.5$ m from the center of the microphone array. To enrich the training diversity, the radius of the speakers was perturbed by a Gaussian noise with a variance of 0.1 m. The DOA of each speaker was calculated w.r.t. the center of the microphone array.

The contributions of the two sources were then summed with a random signal to interference ratio (SIR) selected in the range of SIR $\in [-2, 2]$ to obtain the received microphone signals. Next, we calculated the STFT of both the mixture and the STFT of the separate signals with a frame-length $K = 512$ and an overlap of 75% between two successive frames.

We then constructed the audio feature tensor $\mathcal{R}$ as described above. In the training phase, both the location and a clean recording of each speaker were known; hence, they could be used to generate the labels. For each TF bin (l, k), the dominant speaker was determined by:

$$\text{dominant speaker} \leftarrow \underset{i}{\text{argmax}} |s^i(l,k) h^i_{\text{ref}}(l,k)|. \tag{8}$$

The ground-truth label $D_{l,k}$ is the DOA of the dominant speaker. The training set comprised 4 h of recordings with 30,000 different scenarios of mixtures of two speakers. It is worth noting that as the length of each speaker recording was different, the utterances may also include non-speech or single-speaker frames. The network was trained to minimize the cross-entropy between the correct and the estimated DOA. The cross-entropy cost function was summed over all the images in the training set. The network was implemented in Tensorflow with the ADAM optimizer [28]. The number of epochs was set to be 100, and the training stopped after the validation loss increased for 3 successive epochs. The mini-batch size was set to be 64 images.

Table 1 The TF-DOAnet multi-speaker localization algorithm

- Compute the iRTF features from the multi-microphone recordings.
- Apply the U-net network to classify each TF bin to one of the possible DOAs.
- Based on the U-net results, decide the locations of the active speakers at each time frame.

3 Experimental study

3.1 Datasets

We evaluated the TF-DOAnet and compared its performance to both classic and DNN-based algorithms. To objectively evaluate the performance of the TF-DOAnet, we first simulated two rooms that were different from the rooms in the training set. Then, we tested our TF-DOAnet with real RIR recordings in different rooms. Finally, a real-life scenario with fast moving speakers was recorded and tested. For each test scenario, we selected two speakers from the test set of the WSJ1 database [27] and placed them at two different angles between 0 and 180° relative to the microphone array, at a distance of either 1 m or 2 m. The signals were generated by convolving the signals with RIRs corresponding to the source positions and with either simulated or recorded acoustic scenarios. The SIR was tested in accordance with the DOA literature.

3.2 Performance measures

Two different measures to objectively evaluate the results were used: the mean absolute error (MAE) and the localization accuracy (Acc.). The MAE, computed between the true and estimated DOAs for each evaluated acoustic condition, is given by

$$\text{MAE}(^\circ) = \frac{1}{N \cdot C} \sum_{c=1}^{C} \min_{\pi \in S_N} \sum_{n=1}^{N} |\theta^c_n - \hat{\theta}^c_{\pi(n)}|, \tag{9}$$

where N is the number of simultaneously active speakers and C is the total number of speech mixture segments considered for evaluation for a specific acoustic condition. The term π is the permutation and S_N represents the permutation possibilities. The true and estimated DOAs for the nth speaker in the cth mixture are denoted by θ^c_n and $\hat{\theta}^c_n$, respectively.

Table 2 Configuration of training data generation. All rooms are 2.7 m in height

Simulated training data					
	Room 1	Room 2	Room 3	Room 4	Room 5
Room size	(6 × 6) m	(5 × 4) m	(10 × 6) m	(8 × 3) m	(8 × 5) m
RT_{60}	0.3 s	0.2 s	0.8 s	0.4 s	0.6 s
Signal	Noiseless signals from WSJ1 **training** database				
Array position in room	6 arbitrary positions in each room				
Source-array distance	1.5 m with added noise with 0.1 variance				

The localization accuracy is given by

$$\text{Acc.}(\%) = \frac{\hat{C}_{\text{acc.}}}{C} \times 100 \tag{10}$$

where $\hat{C}_{\text{acc.}}$ denotes the number of speech mixtures for which the localization of the speakers is accurate. We considered the localization of speakers for a speech frame to be accurate if the angular distance between the true and the estimated DOA for all the speakers was less than or equal to 5°.

3.3 Compared algorithms

We compared the performance of the TF-DOAnet with two frequently used baseline methods, namely the MUSIC and SRP-PHAT algorithms. In addition, we compared its performance with the CNN multi-speaker DOA (CMS-DOA) estimator [5][2]. To facilitate the comparison, the MUSIC pseudo-spectrum was computed for each frequency subband and for each STFT time frame, with an angular resolution of 5, over the entire DOA domain. Then, it was averaged over all frequency subbands to obtain a broadband pseudo-spectrum followed by averaging over all the time frames L. Next, the two DOAs with the highest values were selected as the final DOA estimates. Similar post-processing was applied to the computed SRP-PHAT pseudo-likelihood for each time frame.

3.4 Speaker localization results

3.4.1 Static simulated scenario

We first generated a test dataset with simulated RIRs. Two different rooms were used, as described in Table 3. For each scenario, two speakers (male or female) were randomly drawn from the WSJ1 test database and placed at two different DOAs within the range $\{0, 5, \ldots, 180\}$ relative to the microphone array. Since the length of each speaker recording is different, the test dataset also includes non-speech or single-speaker frames. We assume the minimum angle between 2 speakers to be 20°, which, for the radius of $\approx$ 1.5 m from the microphone array, implies that the speakers are practically standing shoulder to shoulder. Each speaker has a different signal length in the mixture. The microphone array was similar to the one used in the training phase. The assumption that we are familiar with the microphone array is fairly common and realistic. For instance, the microphone array in a conference room, in smart devices, or even in phones, is known in advance. Using the RIR generator, we generated the RIR for the given scenario and convolved it with the speakers' signals.

The results for the TF-DOAnet compared with the competing methods are depicted in Table 4. The tables demonstrate that the deep-learning approaches outperform the classic approaches. The TF-DOAnet achieved very high scores and outperforms the DNN-based CMS-DOA algorithm in terms of both MAE and accuracy. Note that the results in Table 4 are reported at a frame-based resolution, where each frame may consist one or two speakers.

3.4.2 Static real recordings scenario

The best way to evaluate the capabilities of the TF-DOAnet is testing it with real-life scenarios. For this purpose, we first carried out experiments with real measured RIRs from a multi-channel impulse response database [19], recorded in our lab. The database comprises RIRs measured in an acoustics lab for three different reverberation times of $RT_{60} = 0.160, 0.360$, and 0.610 s. The lab dimensions are $6 \times 6 \times 2.4$ m.

The recordings were carried out with different DOA positions in the range of $[0°, 180°]$, in steps of 15°. The sources were positioned at distances of 1 m and 2 m from the center of the microphone array. The recordings were

Table 3 Configuration of test data generation. All rooms are 3 m in height

Simulated test data		
	Room 1	Room 2
Room size	(5 × 7) m	(9 × 4) m
RT_{60}	0.38 s	0.7 s
Source-array distance	1.3 m	1.7 m
Signal	Noiseless signals from WSJ1 **test** database	
Array position in room	4 arbitrary positions in each room	

[2]The trained model is available here https://github.com/Soumitro-Chakrabarty/Single-speaker-localization.

Table 4 Results for two different test rooms with simulated RIRs

Test room	**Room 1**		**Room 2**	
Measure	MAE	Acc.	MAE	Acc.
MUSIC [2]	27.95	28.34	31.62	18.38
SRP-PHAT [3]	28.25	27.23	36.61	36.28
CMS-DOA [5]	12.87	73.09	24.0	39.25
TF-DOAnet (our algorithm)	**1.58**	**97.45**	**2.76**	**93.0**

carried out with a linear microphone array consisting of 8 microphones with three different microphone spacings. For our experiment, we chose the [8, 8, 8, 8, 8, 8, 8] cm setup. In order to construct an array setup identical to the one in the training phase, we selected a sub-array of the four center microphones out of the total 8 microphones in the original setup. Consequently, we used a uniform linear array (ULA) $M = 4$ elements with an inter-microphone distance of 8 cm.

The results for the TF-DOAnet compared with the competing methods are depicted in Table 5. Again, the TF-DOAnet outperforms all competing methods, including the CMS-DOA algorithm. Note that the results are reported per time frame and not per utterance, and hence, the inferior results may be expected. Interestingly, for the 1 m case, the best results for the TF-DOAnet were obtained for the highest reverberation level, namely $RT_{60} = 610$ ms, and for the 2 m case, for $RT_{60} = 360$ ms. While surprising at the first glance, this can be explained using the following arguments. There is an accumulated evidence that reverberation, if properly addressed, can be beneficial in speech processing, specifically for multi-microphone speech enhancement and source extraction [20, 29, 30] and for speaker localization [31, 32]. In reverberant environments, the intricate acoustic propagation pattern constitutes a specific "fingerprint" characterizing the location of the speaker(s). When reverberation level increases, this fingerprint becomes more pronounced and is actually more informative than its an-echoic counterpart. An inference methodology that is capable of extracting the essential driving parameters of the RIR will therefore improve when the reverberation is higher. If the acoustic propagation becomes even more complex, as is the case of high reverberation and a remote speaker, a slight performance degradation may occur, but as evident from the localization results, for sources located 2 m from the array, the performance for $RT_{60} = 610$ ms is still better than the performance for $RT_{60} = 160$ ms.

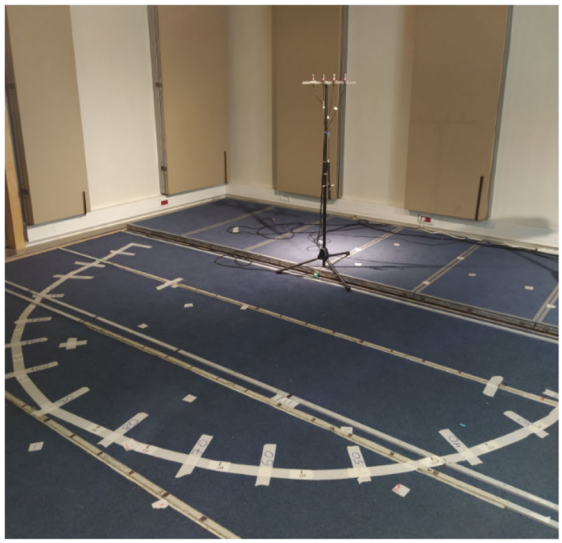
It is worth noting that the test samples were not part of the training phase. The network was not fine-tuned for these test conditions. Yet, since we trained the network with the same RIR generator (with different conditions), it is likely that the results on the simulated test set will be high. The RIR generator cannot capture the accurate sound propagation in real acoustic environments. Therefore, with real recordings, the network performance is likely to be inferior.

3.4.3 Real-life dynamic scenario

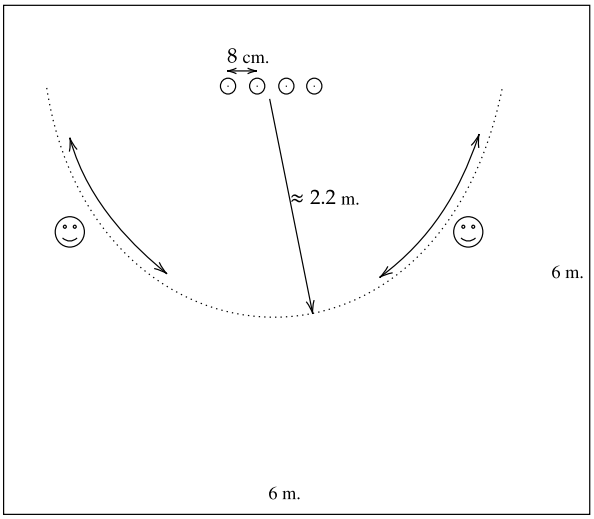
To further assess the capabilities of the TF-DOAnet, we also carried out experiments in real dynamic scenarios. The recordings took place at the acoustic lab, Bar-Ilan University, for which the reverberation level can be set in a wide range. We examined two reverberation levels, namely $RT_{60} = 390$ ms and $RT_{60} = 720$ ms. The microphone array consisted of 4 microphones with an inter-microphone spacing of 8 cm. The speakers walked naturally on an arc at a distance of about 2.2 m from the center of the microphone array. For each RT_{60}, two experiments were recorded. The two speakers started at the angles 20° and 160° and walked until they reached 70° and 100°, respectively, turned back and walked to their starting point. This was done several times throughout the recording. The input SIR values of the first and second speakers are SIR $= -0.12, 0.12$ dB, respectively; hence, both speakers have almost identical power. In the first room setup ($RT_{60} = 390$ ms), the speed of the two moving speakers was 0.34 and 0.35 m/s, respectively. For the second setup ($RT_{60} = 720$ ms), the speakers' speed was 0.28 and 0.31 m/s, respectively. Figure 2a depicts the real-life experiment setup and Fig. 2b depicts a schematic diagram of the setup of this experiment. The ground truth labels

Table 5 Results for three different rooms at distances of 1 m and 2 m with measured RIRs

Distance	**1 m**						**2 m**					
RT_{60}	0.160 s		0.360 s		0.610 s		0.160 s		0.360 s		0.610 s	
Measure	MAE	Acc.	MAE	Acc.	MAE	Acc.	MAE	Acc.	MAE	Acc.	MAE	Acc.
MUSIC	18.7	57.6	19.2	53.2	21.9	42.9	18.4	54.1	26.1	35.8	25.4	32.2
SRP-PHAT	9.0	39.0	13.9	39.4	18.6	29.9	9.7	36.0	16.5	24.7	27.7	21.3
CMS-DOA	1.6	76.3	7.3	75.2	8.4	71.9	5.1	79.5	9.7	60.1	17.5	40.0
TF-DOAnet	**1.3**	**97.5**	**3.5**	**83.5**	**0.9**	**98.3**	**5.0**	**89.5**	**1.7**	**95.7**	**4.8**	**84.2**

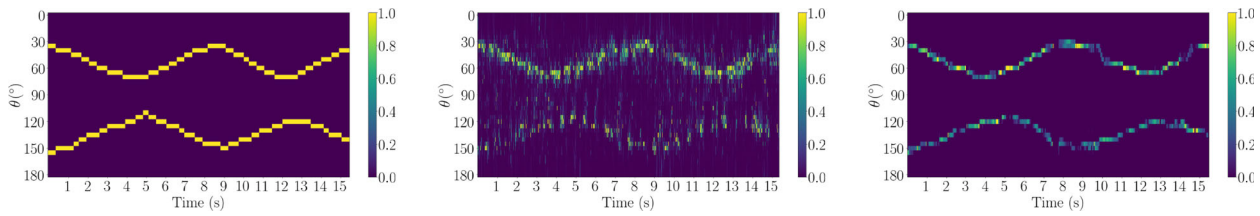


(a) Room view.



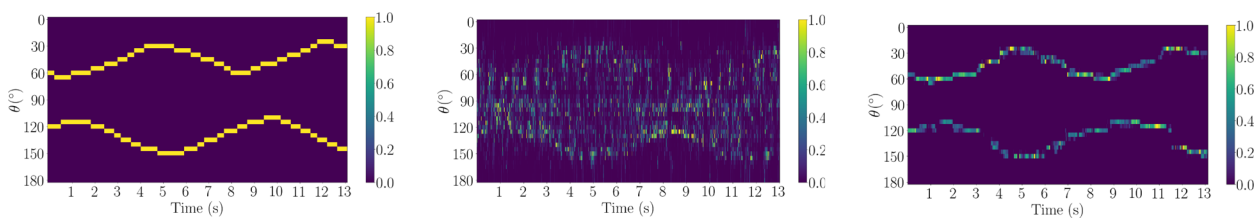
(b) Speakers' trajectory.

Fig. 2 Real-life experiment setup



(a) Ground truth. (b) CMS-DOA. (c) TF-DOAnet.

Fig. 3 Real-life recording of two moving speakers in a 6 × 6 × 2.4 room with $RT_{60} = 390$ ms



(a) Ground truth. (b) CMS-DOA. (c) TF-DOAnet.

Fig. 4 Real-life recording of two moving speakers in a 6 × 6 × 2.4 room with $RT_{60} = 720$ ms

of this experiment were measured with the Marvelmind indoor 3D tracking set[3].

Figures 3 and 4 depict the results of the two experiments. It is clear that the TF-DOAnet outperformed the CMS-DOA algorithm, especially for the high RT_{60} conditions. Whereas the CMS-DOA fluctuated rapidly, the TF-DOAnet output trajectory was smooth and noiseless.

Table 6 depicts the computational cost of the proposed algorithm in comparison to the CMS-DOA algorithm. It is evident that the number of parameters used by the network of the proposed model is less than half of the respective number of parameters of the CMS-DOA model. Moreover, the processing time of the proposed method is also slightly shorter. Note that the processing of 1-s-long utterance takes 70 ms on NVIDIA DGX V100 (single GPU) machine.

3.5 Blind source separation of dynamical speakers

We next evaluate the applicability of the proposed method to the challenging task of speaker separation. Single microphone approaches, as they only utilize spectral information, have the potential of being robust to the source movement. However, their performance is rapidly deteriorating in reverberant environments [33]. Multichannel speaker separation algorithms can remarkably separate overlapping speakers in static scenarios [34]. In dynamic scenarios, the acoustic propagation from the sources to the microphones are rapidly changing over time. Tracking these acoustic paths is a cumbersome task, and failing to do so may result in significant performance degradation.

We propose here a new blind source separation approach, which can be implemented as a byproduct of the proposed tracking scheme. First, the estimated number of speakers, N, is inferred by selecting directions θ for which $p_l(\theta) > 0.15$ (see Eq. 7). For each speaker, the tracking path, $\hat{\theta}^i(l)$, is found as explained in the previous section. TF masks, $\bar{M}^i(l,k)$, $i = 1,\ldots,N$ are obtained for each tracking path, as explained below.

We first aggregate probabilities from adjacent DOAs:

$$M^i(l,k) = \begin{cases} \mathcal{P}_{l,k}\left(\hat{\theta}^i(l)\right) + \mathcal{P}_{l,k}\left(\hat{\theta}^i(l)+5^\circ\right) & \hat{\theta}^i(l) = 0^\circ \\ \mathcal{P}_{l,k}\left(\hat{\theta}^i(l)-5^\circ\right) + \mathcal{P}_{l,k}\left(\hat{\theta}^i(l)\right) & \\ +\mathcal{P}_{l,k}\left(\hat{\theta}^i(l)+5^\circ\right) & 0^\circ < \hat{\theta}^i(l) < 180^\circ \\ \mathcal{P}_{l,k}\left(\hat{\theta}^i(l)-5^\circ\right) + \mathcal{P}_{l,k}\left(\hat{\theta}^i(l)\right) & \hat{\theta}^i(l) = 180^\circ \end{cases}. \tag{11}$$

Then, we apply a threshold to this mask to mitigate the musical noise phenomenon:

$$\bar{M}^i(l,k) = \begin{cases} M^i(l,k) & M^i(l,k) \geq 0.05 \\ 0.05 & M^i(l,k) < 0.05 \end{cases}. \tag{12}$$

Table 6 Computational cost comparison

	# of parameters [million]	Average inference time of 1-s signal [seconds]
CMS-DOA	8.7	0.09
TF-DOAnet	2.1	0.07

To circumvent source permutation issues, we maintain track smoothness by associating DOA estimates with a specific source, only if the current estimate is within 10° of the estimate at the previous frame. Other, more sophisticated, tracking schemes can be applied, but the heuristic approach proposed here provided satisfactory results for the examined scenarios. More involved scenarios, such as intersecting trajectories that necessitate sophisticated tracking schemes, e.g., Bayesian methods [35, 36], are left for a future study.

Once the TF masks are obtained, the separation is implemented by applying the masks to z_{ref}, the mixed signal in the reference microphone.

$$\hat{s}^i(l,k) = z_{\text{ref}}(l,k) \cdot \bar{M}^i(l,k). \tag{13}$$

Figure 5 depicts the mixed signal, described at the previous section, the estimated TF masks and the separated signals. To estimate the masks, we used the tracking path from Fig. 3c. The separation capabilities are clearly demonstrated from these figures. After the application of the proposed algorithm, the output SIR values of the first and second speakers are, respectively, SIR = 6.08 dB and SIR = 7.51 dB, i.e., approximately 7 dB improvement. It is worth noting that separating overlapping dynamic speakers in a highly reverberant room is a challenging task and the obtained results are promising. The reader is also referred to the corresponding audio samples in our website[4].

A note on the validity of the WDO assumption [22] is in place. This widely used assumption underlies many blind audio separation algorithms that apply binary masking. Strictly speaking, this assumption may not hold in reverberant environments for multiple time-frequency bins, due to the "smearing" effect of the reverberation phenomenon. While this may only marginally degrade localization performance in static environments, it can significantly deteriorate speaker separation capabilities, especially in dynamic scenarios. In our experiments, we have shown that even a naïve application of time-frequency masking (see Eq. (12)) can yield satisfactory separation performance. Other, more sophisticated separation schemes that utilize these masks may be applied. Such schemes are left for a future study.

[3] https://marvelmind.com/product/starter-set-ia-02-3d/

[4] www.eng.biu.ac.il/gannot/speech-enhancement/

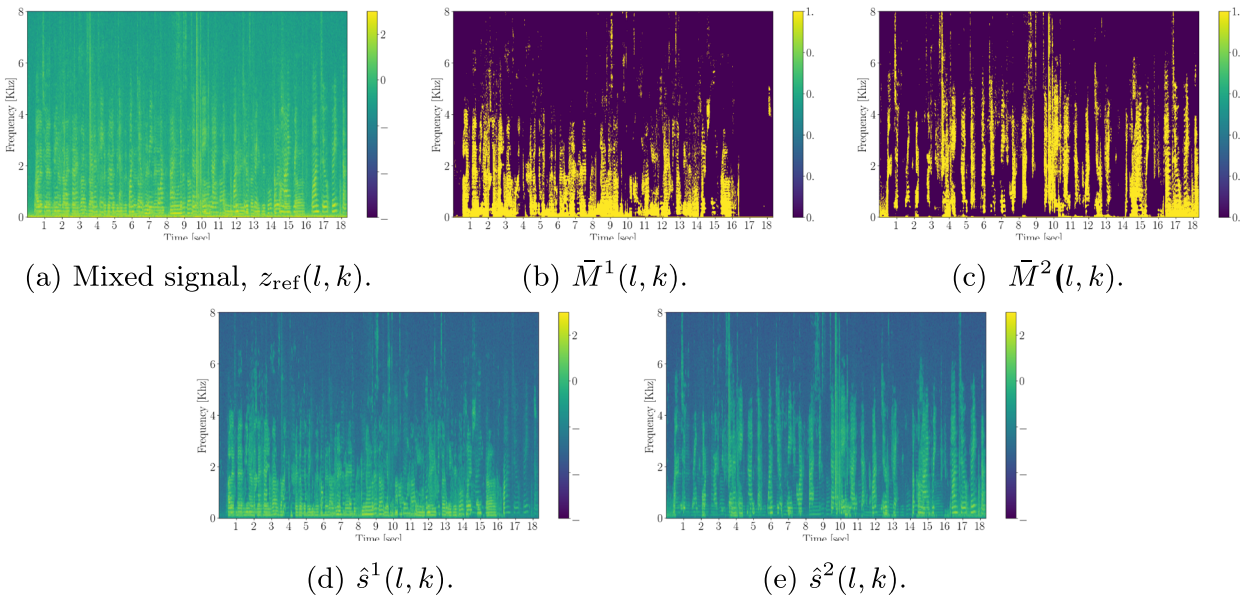


Fig. 5 Real-life separation results of two moving speakers in a 6 × 6 × 2.4 room with $RT_{60} = 390$ ms

4 Conclusions

A joint time-frequency approach was presented in this paper for the DOA estimation task. Instantaneous RTF features were used to train the model. The high TF resolution facilitated the simultaneous tracking of multiple moving speakers. A comprehensive experimental study was carried out with both simulated and real-life recordings. The proposed approach outperformed both the classic and CNN-based SOTA algorithms in all experiments. As a byproduct of the DOA tracking algorithm, we also presented a separation scheme, based on TF masking, which can be applied to moving speakers in a reverberant environment. We believe that the proposed method can be also applicable for localization audio signals other than speech [37].

Abbreviations

TF-DOAnet: Time-frequency direction-of-arrival net; WDO: W-disjoint orthogonality; SOTA: State-of-the-art; FCN: Fully convolutional network; DDESS: Deep direction estimation for speech separation ; NMF: Non-negative matrix factorization; SDR: Signal to distortion ratio; TF: Time-frequency; BF: Beamformer; BSS: Blind source separation; DOA: Direction of arrival; SPI: Speaker position identifier; GSC: General sidelobe canceller; DSE: Deep single expert; MVDR: Minimum variance distortionless response; GEVD: Generalized eigenvalue decomposition; AIR: Acoustic impulse response; PSD: Power spectral density; cPSD: Cross-power spectral density; FIR: Finite-impulse response; MTF: Multiplicative transfer function; RIR: Room impulse response; LTI: Linear time invariant; DNN: Deep neural network; CNN: Convolutional neural network; MFCC: Mel-frequency cepstral coefficients; MMSE: Minimum mean square error; ASR: Automatic speech recognition; ATF: Acoustic transfer function; LCMV: Linearly constrained minimum variance; RTF: Relative transfer function; iRTF: Instantaneous relative transfer function; VAD: Voice activity detector; STSA: Short-time spectral amplitude estimator; LSAE: Log-spectral amplitude estimator; OMLSA: Optimally modified log spectral amplitude; IMCRA: Improved minima controlled recursive averaging; STFT: Short-time Fourier transform; DFT: Discrete Fourier transform; MoG: Mixture of Gaussians; MOE: Mixture of experts; MODE: Mixture of deep experts; r.v.: Random variable; p.d.f.: Probability density function; NN: Neural network; EM: Expectation-maximization; SPP: Speech presence probability; CMVN: Cepstral mean and variance normalization; NN-MM: Neural network mixture-maximum; PESQ: Perceptual evaluation of speech quality; SNR: Signal to noise ratio; SIR: Signal to interference ratio; DAE: Deep auto-encoder; LLR: Log likelihood ratio; WSS: Weighted spectral slope; Covl: Overall quality; Csig: Speech distortion; Cbak: Background distortion; WSJ: *Wall Street Journal*; WSJ1: Wall Street Journal 1; SVM: Support vector machine; IBM: Ideal binary mask; IRM: Ideal ratio mask; ReLU: Rectified linear unit; WER: Word error rate; MM: MixMax; MOS: Mean opinion score; mse: Mean squre error; pDNN: Phoneme DNN; cDNN: Classifier DNN; SGD: Stochastic gradient descent; TDOA: Time difference of arrival; CSD: Concurrent speaker detector; STOI: Short-time objective intelligibility measure; MCCSD: Multi-channel concurrent speaker detector; RIR: Room impulse response; BLSTM: Bidirectional long short-term memory; GCC: Generalized cross correlation; SRP-PHAT: Steered response power with phase transform; BIU: Bar-Ilan University; ULA: Uniform linear array; MAE: Mean absolute error; MUSIC: Multiple signal classification; w.r.t.: With respect to; CMS-DOA: CNN multi-speaker DOA

Acknowledgements

We would like to thank Mr. Pini Tandeitnik for his professional assistance during the acoustic room setup and the recordings.

Authors' contributions

Model development: HH, SC (equal contribution) and JG and SG
Experimental testing: HH, SC (equal contribution)
Writing paper: HH, SC, (equal contribution) and JG 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. The project was also supported by the Israeli Ministry of Science & Technology.

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.

Received: 1 December 2020 Accepted: 8 March 2021
Published online: 08 April 2021

References

1. R. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
2. J. P. Dmochowski, J. Benesty, S. Affes, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Broadband MUSIC: opportunities and challenges for multiple source localization, (2007), pp. 18–21
3. M. S. Brandstein, H. F. Silverman, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A robust method for speech signal time-delay estimation in reverberant rooms, vol. 1, (1997), pp. 375–378
4. C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, W. Kellermann, The LOCATA challenge: acoustic source localization and tracking. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1620–1643 (2020)
5. S. Chakrabarty, E. A. P. Habets, Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. IEEE J. Sel. Top. Signal Process. **13**(1), 8–21 (2019)
6. X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, H. Li, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A learning-based approach to direction of arrival estimation in noisy and reverberant environments, (2015), pp. 2814–2818
7. F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, F. Piazza, in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. A neural network based algorithm for speaker localization in a multi-room environment, (2016), pp. 1–6
8. R. Takeda, K. Komatani, in *IEEE Spoken Language Technology Workshop (SLT)*. Discriminative multiple sound source localization based on deep neural networks using independent location model, (2016), pp. 603–609
9. H. Pujol, E. Bavu, A. Garcia, in *International Congress on Acoustics (ICA)*. Source localization in reverberant rooms using deep learning and microphone arrays, (2019), pp. 1–8
10. J. M. Vera-Diaz, D. Pizarro, J. Macias-Guarasa, Towards end-to-end acoustic localization using deep learning: from audio signals to source position coordinates. Sensors. **18**(10), 3418 (2018)
11. S. Chakrabarty, E. A. P. Habets, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Broadband DOA estimation using convolutional neural networks trained with noise signals, (2017), pp. 136–140
12. S. Rickard, O. Yilmaz, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. On the approximate W-disjoint orthogonality of speech, (2002), pp. 3049–3052
13. O. Ronneberger, P. Fischer, T. Brox, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. U-net: convolutional networks for biomedical image segmentation (Springer, 2015), pp. 234–241
14. O. Ernst, S. E. Chazan, S. Gannot, J. Goldberger, in *The 26th European Signal Processing Conference (EUSIPCO)*. Speech dereverberation using fully convolutional networks, (2018), pp. 390–394
15. S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, S. Gannot, in *European Signal Processing Conference (EUSIPCO)*. Multi-microphone speaker separation based on deep DOA estimation, (2019)
16. S. D. Grechkov, V. P. Semenov, A. A. Bezrukov, in *IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. Comparative analysis of the usage of neural networks for sound processing, (2020), pp. 1389–1391
17. Y. Zhang, Q. Duan, Y. Liao, J. Liu, R. Wu, B. Xie, in *The 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*. Research on speech enhancement algorithm based on SA-Unet, (2019), pp. 818–8183
18. R. Giri, U. Isik, A. Krishnaswamy, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Attention wave-U-net for speech enhancement, (2019), pp. 249–253
19. E. Hadad, F. Heese, P. Vary, S. Gannot, in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. Multichannel audio database in various acoustic environments, (2014), pp. 313–317
20. S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001)
21. S. Stenzel, J. Freudenberger, G. Schmidt, in *4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. A minimum variance beamformer for spatially distributed microphones using a soft reference selection, (2014), pp. 127–131
22. O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004)
23. V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
24. O. Ronneberger, P. Fischer, T. Brox, in *International Conference on Medical Image Computing and Computer-assisted Intervention*. U-net: convolutional networks for biomedical image segmentation, (2015), pp. 234–241
25. D. Wang, J. Chen, Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(10), 1702–1726 (2018)
26. J. B. Allen, D. A. Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
27. D. B. Paul, J. M. Baker, in *Workshop on Speech and Natural Language*. The design for the Wall Street Journal-based CSR corpus, (1992), pp. 357–362. https://www.aclweb.org/anthology/H92-1073
28. D. P. Kingma, J. Ba, Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
29. S. Markovich-Golan, 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). https://doi.org/10.1109/TASL.2009.2016395
30. I. Dokmanić, R. Scheibler, M. Vetterli, Raking the cocktail party. IEEE J Sel. Top. Signal Process. **9**(5), 825–836 (2015)
31. A. Deleforge, F. Forbes, R. Horaud, Acoustic space learning for sound-source separation and localization on binaural manifolds. Int. J. Neural Syst. **25**(01), 1440003 (2015)
32. B. Laufer-Goldshtein, R. Talmon, S. Gannot, Semi-supervised sound source localization based on manifold regularization. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(8), 1393–1407 (2016)
33. S. E. Chazan, L. Wolf, E. Nachmani, Y. Adi, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Single channel voice separation for unknown number of speakers under reverberant and noisy settings, (2021). https://doi.org/http://arxiv.org/abs/2011.02329
34. 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). Invited tutorial paper
35. S. Gannot, T. G. Dvorkind, Microphone array speaker localizers using spatial-temporal information. EURASIP J. Adv. Signal Process. **2006**, 1–17 (2006)
36. B. Laufer-Goldshtein, R. Talmon, S. Gannot, A hybrid approach for speaker tracking based on TDOA and data-driven models. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(4), 725–735 (2018)
37. A. K. Das, T. T. Lai, C. W. Chan, C. K. Y. Leung, A new non-linear framework for localization of acoustic sources. Struct. Health Monit. **18**(2), 590–601 (2019)

Publisher's Note