# Deep Ranking-Based DOA Tracking Algorithm

Renana Opochinsky*, Gal Chechik† and Sharon Gannot*
* The Alexander Kofkin Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel.
Email: {renana.klainman,sharon.gannot}@biu.ac.il
† Gonda brain research center, Bar-Ilan University, Ramat-Gan, Israel and NVIDIA Research
Email: gal.chechik@biu.ac.il

***Abstract*—In this study, we present a weak-supervised deep neural network-based tracking algorithm for a moving source. A triplet-loss network is trained with instantaneous spatial features to estimate the time-varying DOA. The core idea is to minimize the use of labeled samples (i.e. samples which are accurately localized, and difficult to acquire) by using instead partial knowledge drawn from an unlabeled, and much larger, dataset in which only the relative spatial ordering between the samples is known. We use a deep learning architecture that stochastically combines a triplet-ranking loss for the unlabeled samples and a spatial loss for the labelled samples and learns a nonlinear deep embedding that maps acoustic features to an azimuth angle of the source. We show that it is unnecessary to train the network with a large number of random trajectories of a moving source, and that triplets of static sources from the same locus, which can be more easily acquired, are sufficient. A simulation study demonstrates the applicability of the proposed method to dynamic problems.**

***Index Terms*—acoustic source tracking, deep embedding learning, triplet-loss, relative transfer function**

## I. Introduction

Speaker localization (specifically, direction of arrival (DOA) estimation) in acoustic environments using a microphone array is a basic building block in various audio applications, including smart home devices, automatic camera steering, beamforming, source separation, and robot audition. The problem further complicates when the sound source is free to move. In this case, the source should be dynamically localized, necessitating online tracking algorithms.

The problem of DOA estimation of sound sources has attracted the attention of the research community for more than four decades. Among the most common methods are those based on the analysis of the cross-correlation between microphones, namely the generalized cross-correlation phase transform (GCC-PHAT) [1] and its multichannel extension, the steered response power phase-transform (SRP-PHAT) [2]. Although mainly addressing non-reverberant environments, methods based on the analysis of the spatial correlation matrix of the received signals, namely the multiple signal classification (MUSIC) algorithm [3] and its extensions, are also widely used in audio processing applications [4]. These techniques were developed for static scenarios, and are not directly addressing the dynamic case.

Several recursive DOA tracking methods, based on recursive least squares, are presented in [5]. Probabilistic approaches, directly addressing the time-varying tracking problem, including Bayesian inference methods, e.g. particle filters [6]–[9]; probability hypothesis density (PHD) filtering, mainly used in multi-modal processing [10], [11]; factor graphs [12]; and non-Bayesian methods based on the recursive expectation maximization (EM) procedure [13]–[15].

Training-based localization methods have recently gathered momentum. Under this paradigm, acoustic features are first extracted from the measured signals and then mapped to the corresponding source positions by applying a learned nonlinear function. Training-based source localization algorithms can be either fully-supervised, i.e. all data points in the training corpus are attached with accurate location labels, or semi/weakly-supervised, i.e. only a small percentage of the training set is labelled and the rest is unlabelled or attributed with weak labels. Examples for fully-supervised acoustic localization algorithms using convolution neural networks (CNNs) can be found in [16], [17]. In these works, the tracking of moving sources is facilitated by the utilization of instantaneous features. Tracking is explicitly addressed in [18], [19] by the application of convolutional recurrent neural networks (CRNNs). Among the semi-supervised methods, we can list the manifold-learning based methods [20], [21] or methods based on variational autoencoder (VAE) [22].

Recently, we have presented a weakly-supervised deep-learning localization method that only utilizes a few labeled samples attributed with accurate position labels (referred to as *anchors*), together with a larger set of unlabeled samples, for which only their relative spatial ordering is known [23]. This method has shown to yield high localization accuracy in static scenarios for various reverberation levels.

In this paper, we present an extension of our previously proposed method, which is applicable to moving speakers. The algorithm's performance is examined using simulated trajectories of a moving source in a reverberant room. The main contribution of this paper is the demonstration of the ability of the proposed architecture to extract relevant information from the weak spatial ordering of static sources enabling source tracking in dynamic scenarios.

## II. Problem Formulation

We consider a single moving speaker recorded by a pair of microphones. In this section, we will first discuss the measurement model and then describe the feature vector.

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

### A. Measurements

The position of the speaker varies over time and will be described by $\mathbf{p(t)} = [r(t), \theta(t), \phi(t)]^\top$. In this paper, we consider the DOA estimation problem and consequently the azimuth angle $\phi(t)$ will solely describe the source position. Let the uttered speech signal be $s(n)$ and its respective short-time Fourier transform (STFT) representation $s(l,k)$, where $l$ is the frame index and $k$ is the frequency index. The speech signal is captured by two microphones located in the room, and the respective measured signals in the STFT domain are approximately given by:

$$x_{i\phi}(l,k) = h_{i\phi}(l,k)s(l,k) + u_i(l,k), \tag{1}$$

where $i = 1, 2$ is the microphone index, $h_{i\phi}(l,k)$ is the acoustic transfer function (ATF) relating the speaker located at azimuth angle $\phi$ and the $i$-th microphone and $u_i(l,k)$ is an additive, assumed for simplicity to be spatially-white and stationary, noise signal.

### B. Time-frequency features

In this work, we have chosen to use an instantaneous version of the relative transfer function (RTF) [24] as the feature vector. The RTF is known to encapsulate the *spatial fingerprint* of a sound source [25]. As the instantaneous relative transfer function (iRTF) version uses the current frame and a few context frames, it may facilitate source tracking in dynamic scenarios. Define the iRTF as:

$$\text{iRTF}(l,k) = \frac{\sum_{i=l-n_c}^{l+n_c} x_{2\phi}(i,k)x_{1\phi}^*(i,k)}{\sum_{i=l-n_c}^{l+n_c} x_{1\phi}(i,k)x_{1\phi}^*(i,k)} \tag{2}$$

with $2n_c$ the number of context frames. A small value of $n_c$ promotes fast tracking at the expense of high estimation variance, and vice versa.

The frame-dependant feature vector $\hat{\mathbf{h}}(l)$ is constructed in the following way. First, we concatenate the iRTF components over frequencies in the interesting band, namely, only frequencies in the range $k_1, \ldots, k_D$ where most of the speech signal power is concentrated. Next, we split the complex-valued iRTF to the real and imaginary parts. The feature vector is thus given by:

$$\hat{\mathbf{h}}(l) = [\Re(\text{iRTF}(l,k_1), \ldots, \text{iRTF}(l,k_D)), \\ \Im(\text{iRTF}(l,k_1), \ldots, \text{iRTF}(l,k_D))]. \tag{3}$$

## III. DOA TRACKING ALGORITHM

In this work, we follow the network architecture proposed in [23]. In the training phase, the loss function involves two terms. A *spatial loss*, reflecting our knowledge on the azimuth angle of some labelled samples, and a *ranking loss*, reflecting our knowledge on the relative proximity of the unlabelled samples. Recently, ranking-loss and triplet-loss training schemes were applied in audio processing problems [26]–[29]. In the proposed scheme, the network is trained with static sources and long speech utterances, and in the test phase it is applied to moving sources. The other acoustic conditions remain intact.

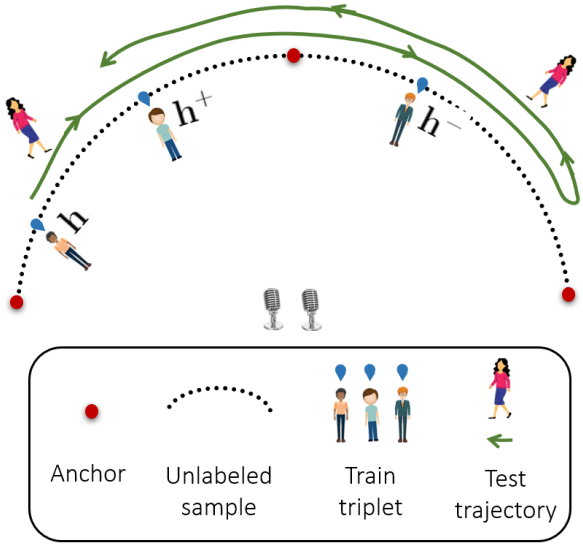



Fig. 1. An illustration of the experimental setup. The source position, both in the training and test stages, is confined to the arc, with $\phi \in [0°, 180°]$. The black dots mark the positions of the unlabeled data and blue triplet of dots are example of train triplet. The red dots mark the labeled examples, green trajectory represents the test speaker trajectory.

Due to the source movement, the iRTF becomes time-varying and can only be estimated from short speech segments. In this section we summarize the training procedure and discuss in details the data arrangement, stressing the mismatch between the train and test phases.

### A. Overview

We address the problem of inferring the time-varying source azimuth angle $\phi(l)$ from the instantaneous feature vector $\mathbf{h}(l)$, defined in (3), by learning a nonlinear function $\phi = f(\mathbf{h})$ that maps an iRTF sample onto an embedding space that corresponds to $\phi$. The mapping will be learned using a combination of weakly-labeled and labeled samples. We use two sources of information during training: first, strong supervision - in the form of a small number of samples with known azimuth angles; second, weak supervision - in the form of a large set of samples for which only their relative proximity is known.

Figure 1 illustrates the proposed scheme. Consider a set of $n$ sound recordings sampled along an arc in the training stage.[1] We assume that a small number, $n_a$, of these samples termed *anchors* are labelled, namely their corresponding azimuth angle is accurately known. These samples are represented in the figure as red dots on the arc. The angles of the remaining $n - n_a$ samples, represented by black dots on the arc, are unknown. Instead, their relative proximity, which can be more easily obtained, is available. We refer to these samples as *weakly-labelled*. From these points, we form triplets of samples, represented as blue points in the figure. Each triplet comprises one *query* sample, one *positive* sample and one

[1]The problem complicates if 2D or 3D source location should be inferred, necessitating a larger number of microphones and spatially distributed training points.

*negative* sample. The only available information is that the positive sample is closer to the query sample than the negative sample. In the test phase, one speaker is moving along the arc in an arbitrarily unknown trajectory (depicted as a green curve in the figure) while uttering speech. The goal of the algorithm is to track the time-varying angle of the speaker.

A deep architecture is used to infer the nonlinear function $f(\mathbf{h})$. It is implemented as a combined system comprising three identical building blocks with shared weights, each of which is a simple *fully-connected* network as described in [23]. The input to the combined system is either a triplet or an anchor, according to the training routine that will be explained in the sequel. Note that in the test phase, only a single replica of the network is used to estimate the angle of the source.

### B. Training with a combined loss

The overall loss stochastically combines *spatial loss* component with *ranking loss* component. This term encourages the smoothness and continuity of the mapping $f(\mathbf{h})$, encouraging RTFs from close positions to be located closer in the embedded space. However, the learned mapping is entirely free to learn any monotonic function of the true predictions. Therefore, to anchor the predicted location to the true spatial space, we further enforce a correspondence of some RTF with specific spatial positions, as follows:

$$\text{loss} = \begin{cases} l_{\text{ranking}}(\mathbf{h}, \mathbf{h}^+, \mathbf{h}^-) & \text{with probability } \alpha \\ l_{\text{spatial}}(\mathbf{h}^{\text{anchor}}, \phi^{\text{anchor}}) & \text{with probability } 1-\alpha \end{cases}, \tag{4}$$

where $\alpha \in [0, 1]$ controls the relative weight of the two loss terms. The ranking loss $l_{\text{ranking}}(\mathbf{h}, \mathbf{h}^+, \mathbf{h}^-)$, which was chosen to be implemented in the format of triplet loss, is defined as

$$l_{\text{ranking}}(\mathbf{h}, \mathbf{h}^+, \mathbf{h}^-) = \max\left(0, 1 + |f(\mathbf{h}) - f(\mathbf{h}^+)| - |f(\mathbf{h}) - f(\mathbf{h}^-)|\right) \tag{5}$$

and the spatial loss $l_{\text{spatial}}(\mathbf{h}^{\text{anchor}}, \phi^{\text{anchor}})$ is defined as

$$l_{\text{spatial}}(\mathbf{h}^{\text{anchor}}, \phi^{\text{anchor}}) = |f(\mathbf{h}^{\text{anchor}}) - \phi^{\text{anchor}}|. \tag{6}$$

The training routine includes alternating activation of both loss components. With probability $\alpha$ we sample a triplet of unlabelled samples and compute the ranking loss. Similarly, with probability $1-\alpha$ we sample an anchor, i.e. a sample with known location, and compute the spatial loss w.r.t. its true location $\phi^{\text{anchor}}$.

### C. Data arrangement

Organization of the data in triplets is required for implementing the triplet loss. The input to the network is a triplet, each of which comprises three iRTF samples $(\mathbf{h}, \mathbf{h}^+, \mathbf{h}^-)$; namely a query sample $\mathbf{h}$, a positive sample $\mathbf{h}^+$, and a negative sample $\mathbf{h}^-$. After randomly drawing $\mathbf{h}$, two other samples are drawn from the unlabeled set and assigned to $\mathbf{h}^+$ and $\mathbf{h}^-$, based on their proximity to $\mathbf{h}$. Note that the exact positions of $\mathbf{h}^+$ and $\mathbf{h}^-$ are not required. The triplets are monotonically sampled in the range $[0^\circ - 180^\circ]$ on the predefined source locus, an arc in our case. During the training phase, the speech sources are static and the iRTFs are estimated using long utterances to guarantee accurate estimation.

The test set comprises observations from *unknown* angles of a moving speaker on the same arc. We assume a perfect match between the training phase and the test phase in terms of the acoustic conditions, namely the reverberation time and the signal to noise ratio (SNR) level, and of the microphone array constellation within the room. Yet, due to the speaker's movement in the test phase, the acoustic paths from the source to the microphone rapidly change over time, resulting in respective time-variations in the iRTF. This implies a significant mismatch between the training and test conditions. Despite this mismatch in the generation of the signals, the proposed algorithm applies the mapping function that was inferred in the training phase $\hat{\phi}^{\text{test}} = f(\mathbf{h}^{\text{test}})$, to dynamically obtain the source angle.

## IV. Simulation Study

In this section, we evaluate the performance of the proposed algorithm and compare it with a nearest neighbor (NN) approach.

### A. Training set

To generate the training data we simulated a two-microphone array with 8 cm inter-distance. The array was randomly positioned at 20 different positions in a $6 \times 6.2 \times 3$ simulated room. Three reverberation times, 200 ms, 400 ms and 600 ms were simulated. We used the room impulse response (RIR) generator,[2] efficiently implementing the image method [30]. The source is known to be *statically* positioned on an arc with 2 m radius with respect to the microphone pair center, with an azimuth angle $\phi$ in the range $[0^\circ - 180^\circ]$.

For each source position, clean anechoic speech signals were drawn from the TIMIT dataset [31]. The speech signals were convolved with acoustic impulse responses (AIRs) relating the source position and the microphones' positions. These reverberant signals are further contaminated by spatially-white noise signals with SNR levels of either 15, 20 or 30 dB. The microphones and the source are arranged in the same plane with identical heights.

The sampling rate was set to 16KHz and the frame-length of the STFT to $K = 1024$, with an overlap of $75\%$ between two successive frames. The feature vector comprises $D = 74$ frequency bins corresponding to the frequency range $[0.2 - 2.5]$ kHz, where most of the speech power is concentrated. The dimensions of the feature vector are therefore $148 \times 1$ per time-frame.

The training set is a collection of speech samples as described above, uttered by static speakers located at different positions on the arc. For each acoustic condition, $n = 1440$ samples, uniformly distributed in the range $[0^\circ - 180^\circ]$, were used to train the network parameters. These include $n_a$ anchors. The effect of the number of anchors on the performance of the proposed and the competing algorithms will be analyzed

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

in Sec. IV-D. The exact positions of the other $1440-n_a$ speech utterances are not known during training but rather the relative order of triplets of samples, as explained in Sec. III-B. The weight parameter $\alpha$ was set to $0.95$.

### B. Test set

As explained in Sec. III-C, the acoustic conditions and the microphone constellation in each test case match a corresponding condition in the training set. The main and crucial difference between the training and test phases lies in the source dynamics. While the sources in the training phase are static, they are free to move in the test phase, with an arbitrarily unknown trajectory on the same arc. Hence, the azimuth angle $\phi$ becomes a *time-varying* unknown value in the range $[0^\circ - 180^\circ]$.

To simulate moving sources, we used the signal generator package.[3] The performance of the proposed algorithm is exemplified in this paper with a sinusoidal trajectory with amplitude of $60^\circ$ with velocities in the range $[0.14-1.1]$ m/s. Other trajectories were also tested but are not presented here due to space constraints. Overall, $n_{\text{val}} = 420$ samples were used as a validation set for tuning the hyper-parameters of the algorithm. Other $n_{\text{test}} = 420$ samples were used as a test set for evaluating the model accuracy. Both validation and test phases are using moving sources.

A note on the the number of context frame is in place. While better estimation accuracy may be expected if $n_{\text{c}}$ increases, its value is upper-bounded by the "smearing" effect of the source dynamics, which becomes more pronounced in high velocities. The performance of the algorithm as a function of $n_{\text{c}}$ will be examined in Sec. IV-D.

### C. Performance measure and competing method

The metric used to assess the localization accuracy is the root mean square error (RMSE) over the test set between the true test angles and the inferred angles using the trained mapping function, as defined below:

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(\phi_i - f(\mathbf{h}_i)\right)^2}. \tag{7}$$

The proposed algorithm was compared with a nearest neighbor (NN) approach. In the latter, the source is localized by evaluating the distance between the measured iRTF and the set of all anchor iRTFs, obtained in the training phase as labeled data. The position estimate is obtained by selecting the closest anchor point using the Euclidean distance between the iRTFs. We report in Sec. IV-D the mean RMSE averaged over the best 15 trials out of 20 constellations and the corresponding standard deviation.

### D. Results

The RMSE results of the proposed and competing algorithms are evaluated, as a function of various parameters. Table I depicts the RMSE results as a function of $n_a$. It is evident that increasing the number of anchors improves performance for both methods. The proposed approach significantly outperforms the NN approach. As can be deduced from Table II, higher SNR levels also improves performance. Investigating the RMSE performance as a function of $T_{60}$ reveals an interesting behaviour. In most cases, the results for $T_{60} = 400$ ms outperforms the results obtained for both the low and high reverberation levels. This can be explained by the iRTF characteristics. As shown in several studies, see [25] and references thereof, this feature acts as an *acoustic fingerprint* of the source position. Hence, when reverberation level increases the structure of the feature becomes more intricate and consequently the differences between adjacent positions more pronounced. In the high reverberation case, $T_{60} = 600$ ms, a slight performance degradation occurs due to various factors, such as STFT frame-size, etc.

We also evaluate performance as a function of the context length, from $n_{\text{c}} = 2$ frames (corresponding to 128 ms), to $n_{\text{c}} = 20$ frames (corresponding to 1.28 s).

As evident from Table III, the context length is a significant factor in the tracking accuracy, with longer context length implying better results. As this is true for all tested velocities in the range $0.14 - 1.1$ m/s, only averaged results are presented. At higher velocities (not presented here), performance degradation is encountered due to the smearing effect of the dynamic scenario. Finally, we exemplify in Fig. 2 the estimated trajectories for $T_{60} = 400, 600$ ms. Excellent tracking capabilities are demonstrated for $T_{60} = 400$ ms. In the more challenging $T_{60} = 600$ ms case, a performance degradation is observed, especially when the source acceleration is high.

TABLE I
MEAN RMSE (AND STANDARD DEVIATION) FOR THREE VALUES OF $n_a$ AND $T_{60} = 200, 400, 600$ MS. SNR=30 DB, $n_c = 20$ FRAMES, VELOCITY=$0.55$ M/S.

| $n_a$ | 2 anchors | | | 3 anchors | | | 5 anchors | | |
|---|---|---|---|---|---|---|---|---|---|
| $T_{60}$ | 200 | 400 | 600 | 200 | 400 | 600 | 200 | 400 | 600 |
| NN | 47.6(0.3) | 47.4(0.3) | 47.7(0.1) | 30.1(2.2) | 28.9(1.1) | 29.3(1.3) | 21.3(2.8) | 21.1(3.1) | 24.1(2.7) |
| Proposed | 16.3(3.5) | 14.4(5.0) | 19.4(4.0) | 14.8(6.1) | 10.5(3.4) | 10.7(0.8) | 13.3(5.2) | 8.9(2.7) | 10.5(1.1) |

TABLE II
MEAN RMSE (AND STANDARD DEVIATION) FOR THREE SNR VALUES AND $T_{60} = 200, 400, 600$ MS. $n_a = 5$, $n_c = 20$ FRAMES, VELOCITY=$0.55$ M/S.

| SNR | 15 dB | | | 20 dB | | | 30 dB | | |
|---|---|---|---|---|---|---|---|---|---|
| $T_{60}$ | 200 | 400 | 600 | 200 | 400 | 600 | 200 | 400 | 600 |
| NN | 16.6(2.6) | 23.5(1.8) | 24.4(2.2) | 21.8(3.0) | 17.7(2.0) | 20.3(1.2) | 21.3(2.8) | 21.1(3.1) | 24.1(2.7) |
| Proposed | 19.0(1.4) | 10.7(2.3) | 10.5(1.1) | 14.3(3.9) | 10.7(3.6) | 10.0(0.9) | 13.3(5.2) | 8.9(2.7) | 10.5(1.1) |

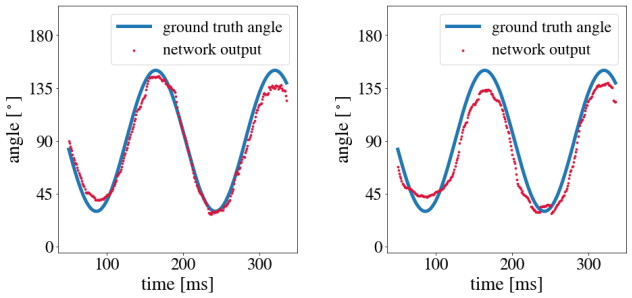
TABLE III
MEAN RMSE (AND STANDARD DEVIATION) FOR DIFFERENT VALUES OF $n_c$ AND $T_{60} = 400$ MS. $n_a$=5 AND SNR= 30 DB AVERAGED ON ALL VELOCITIES.

| $n_c$ | 2 | 5 | 10 | 20 |
|---|---|---|---|---|
| NN | 27.0(1.9) | 21.1(3.4) | 22.1(1.5) | 21.2(3.2) |
| Proposed | 17.7(1.0) | 13.4(0.6) | 9.9(0.5) | 8.3(0.6) |

## V. CONCLUSIONS AND DISCUSSION

In this paper, we introduced a weakly-supervised approach for tracking a moving speaker using deep neural networks.

[3] Available online at `github.com/ehabets/Signal-Generator`




(a) $T_{60} = 400$ ms (b) $T_{60} = 600$ ms

Fig. 2. Estimated (red dots) and ground-truth (solid blue lines) azimuth angle vs. time using $n_a = 5$ anchors. Data parameters were SNR=30 dB, $n_c = 20$ frames, velocity=0.55 m/s.

The algorithm reduces the dependence on labeled, expensive to acquire, data by utilizing the information on the relative proximity of samples. The method learns a mapping between an iRTF feature and the source position. Simulation results demonstrate that the proposed ranking-based method can accurately track a moving source in a wide range of reverberation conditions and SNR levels, and its superiority over an NN-based approach. Our network is geared towards learning known acoustic environments, e.g. a robot that learns to act in a specific room. Accordingly, we assume a perfect match between the acoustic conditions and the microphones-room constellation in the training and test phases. However, the train and test phases significantly differ in the dynamic characteristics of the problem. While the sources are static in the training phase, they are free to move in the test phase.

## REFERENCES


[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Tran. on Acoustics, Speech, and Signal Proc.*, vol. 24, no. 4, pp. 320–327, 1976.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays*. Springer, 2001, pp. 157–180.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Tran. on Anten. and Prop.*, vol. 34, no. 3, pp. 276–280, 1986.

[4] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Tran. on Audio, Speech, and Language Proc.*, vol. 28, pp. 1620–1643, 2020.

[5] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Proc.*, vol. 85, no. 1, pp. 177–204, 2005.

[6] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund, "Particle filters for positioning, navigation, and tracking," *IEEE Tran. on signal Proc.*, vol. 50, no. 2, pp. 425–437, 2002.

[7] X. Zhong, A. Premkumar, and A. Madhukumar, "Particle filtering for acoustic source tracking in impulsive noise with alpha-stable process," *IEEE Sensors J.*, vol. 13, no. 2, pp. 589–600, 2012.

[8] A. Levy, S. Gannot, and E. A. Habets, "Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments," *IEEE Tran. on Audio, Speech, and Language Proc.*, vol. 19, no. 6, pp. 1540–1555, 2010.

[9] C. Evers, E. A. Habets, S. Gannot, and P. A. Naylor, "DoA reliability for distributed acoustic tracking," *IEEE Signal Proc. Letters*, vol. 25, no. 9, pp. 1320–1324, 2018.

[10] Z. Fu, F. Angelini, J. Chambers, and S. M. Naqvi, "Multi-level cooperative fusion of GM-PHD filters for online multiple human tracking," *IEEE Tran. on Multimedia*, vol. 21, no. 9, pp. 2277–2291, 2019.

[11] Q. Liu, W. Wang, T. de Campos, P. Jackson, and A. Hilton, "Multiple speaker tracking in spatial audio via PHD filtering and depth-audio fusion," *IEEE Tran. on Multimedia*, vol. 20, no. 7, pp. 1767–1780, 2018.

[12] K. Weisberg, B. Laufer-Goldshtein, and S. Gannot, "Simultaneous tracking and separation of multiple sources using factor graph model," *IEEE/ACM Tran. on Audio, Speech, and Language Proc.*, vol. 28, pp. 2848–2864, 2020.

[13] O. Schwartz and S. Gannot, "Speaker tracking using recursive em algorithms," *IEEE/ACM Tran. on Audio, Speech, and Language Proc.*, vol. 22, no. 2, pp. 392–402, 2013.

[14] Y. Dorfan, A. Plinge, G. Hazan, and S. Gannot, "Distributed expectation-maximization algorithm for speaker localization in reverberant environments," *IEEE/ACM Tran. on Audio, Speech, and Language Proc.*, vol. 26, no. 3, pp. 682–695, 2017.

[15] Y. Dorfan, B. Schwartz, and S. Gannot, "Forward-backward recursive expectation-maximization for concurrent speaker tracking," *EURASIP J. on Audio, Speech, and Music Proc.*, vol. 2021, no. 1, pp. 1–13, 2021.

[16] S. Chakrabarty and E. A. Habets, "Multi-speaker doa estimation using deep convolutional networks trained with noise signals," *IEEE J. of Selected Topics in Signal Proc.*, vol. 13, no. 1, pp. 8–21, 2019.

[17] H. Hammer, S. E. Chazan, J. Goldberger, and S. Gannot, "Dynamically localizing multiple speakers based on the time-frequency domain," *EURASIP Journal on Audio, Speech and Music*, Mar. 2021.

[18] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *26th European Signal Proc. Conference (EUSIPCO)*, 2018, pp. 1462–1466.

[19] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings," *IEEE J. of Selected Topics in Signal Proc.*, vol. 13, no. 1, pp. 22–33, 2019.

[20] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised source localization on multiple manifolds with distributed microphones," *IEEE/ACM Tran. on Audio, Speech, and Language Proc.*, vol. 25, no. 7, pp. 1477–1491, 2017.

[21] ——, "A hybrid approach for speaker tracking based on TDOA and data-driven models," *IEEE/ACM Tran. on Audio, Speech, and Language Proc.*, vol. 26, no. 4, pp. 725–735, 2018.

[22] M. J. Bianco, S. Gannot, E. Fernandez-Grande, and P. Gerstoft, "Semi-supervised source localization in reverberant environments with deep generative modeling," *arXiv preprint arXiv:2101.10636*, 2021.

[23] R. Opochinsky, B. Laufer-Goldshtein, S. Gannot, and G. Chechik, "Deep ranking-based sound source localization," in *IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics (WASPAA)*, 2019, pp. 283–287.

[24] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Tran. on Signal Proc.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[25] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Data-driven multi-microphone speaker localization on manifolds," *Foundations and Trends in Signal Proc.*, vol. 14, no. 1–2, pp. 1–161, 2020.

[26] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances." in *Interspeech*, 2017, pp. 1487–1491.

[27] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech emotion recognition from variable-length inputs with triplet loss function," in *Interspeech*, 2018, pp. 3673–3677.

[28] Z. Lian, Y. Li, J. Tao, and J. Huang, "Speech emotion recognition via contrastive loss under siamese networks," in *Joint Workshop on Affective Social Multimedia Computing and Multi-Modal Affective Computing of Large-Scale Multimedia Data*, 2018, pp. 21–26.

[29] N. Turpault, R. Serizel, and E. Vincent, "Semi-supervised triplet loss based learning of ambient audio embeddings," in *IEEE Inter. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2019, pp. 760–764.

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

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Tech. Report N*, vol. 93, p. 27403, 1993.