A Hybrid Approach for Speaker Tracking Based on TDOA and Data-Driven Models

Bracha Laufer-Goldshtein, *Student Member, IEEE*, Ronen Talmon, *Member, IEEE* and Sharon Gannot, *Senior Member, IEEE*

***Abstract*—The problem of speaker tracking in noisy and reverberant enclosures is addressed. We present a hybrid algorithm, combining traditional tracking schemes with a new learning-based approach. A state-space representation, consisting of a propagation and observation models, is learned from signals measured by several distributed microphone pairs. The proposed representation is based on two data modalities corresponding to high-dimensional acoustic features representing the full reverberant acoustic channels as well as low-dimensional TDOA estimates. The state-space representation is accompanied by a statistical model based on a Gaussian process used to relate the variations of the acoustic channels to the physical variations of the associated source positions, thereby forming a data-driven propagation model for the source movement. In the observation model, the source positions are nonlinearly mapped to the associated TDOA readings. The obtained propagation and observation models establish the basis for employing an extended Kalman filter (EKF). Simulation results demonstrate the robustness of the proposed method in noisy and reverberant conditions.**

***Index Terms*—speaker tracking, time difference of arrival (TDOA), relative transfer function (RTF), extended Kalman filter (EKF), Gaussian process.**

I. Introduction

Speaker localization and tracking in reverberant enclosures is required in various audio applications, including: automatic camera steering in teleconferencing [1], beamforming [2], source separation [3], [4] and robot audition [5], [6]. Conventional localization methods are implemented by either a single-step optimization directly on the measured signals, or a dual-step approach. In the first category, the position is estimated for example, by a grid search over the output power of a beamformer steered to all potential locations [7], [8], or by high-resolution methods such as the multiple signal classification (MUSIC) algorithm [9]. In dual-step approaches, the first stage is estimating the TDOAs of several microphone pairs [10]–[12]. Then, in the second stage, the TDOA readings are combined to perform the actual localization [13], [14].

Bracha Laufer-Goldshtein and Sharon Gannot are with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel (e-mail: Bracha.Laufer@biu.ac.il, Sharon.Gannot@biu.ac.il); Ronen Talmon is with the Viterbi Faculty of Electrical Engineering, The Technion-Israel Institute of Technology, Technion City, Haifa 32000, Israel, (e-mail: ronen@ee.technion.ac.il).

Bracha Laufer-Goldshtein is supported by the Adams Foundation of the Israel Academy of Sciences and Humanities.

This work was supported in part by a Grant from a joint Lower Saxony-Israeli Project financially supported by the State of Lower Saxony.

In a tracking scenario, the source is moving in the enclosure in approximately continuous trajectory, implying dependence between source positions in successive time steps. Bayesian inference approaches, which model the varying source position as a stochastic process, are widely used. These methods commonly rely on estimated TDOAs, leading to nonlinear and non-Gaussian models, which can be solved, for example, using the unscented Kalman filter, the extended Kalman filter (EKF) [15], and particle filters [16]–[18].

In real environments, the presence of noise or reverberations often yields unreliable observations with spurious peaks, which may lead to severe performance degradation. Several attempts to mitigate the harmful effect of noise and reverberations, were made. In [19] an extended particle filter (EPF) solution was proposed, where an EKF is used to derive an optimal importance function for a particle filter. A multiple-hypothesis model accounting for the multipath nature of the sound propagation in reverberant enclosures was presented in [16], and was combined with an EPF in [20]. In [21], [22] a tracker was proposed based on a probability hypothesis density (PHD) filter, which is a first moment approximation of the target probability density. Robust tracking methods were also proposed using sensor networks with special structures, such as spherical microphone arrays [23] and distributed networks [24], [25]. In [26] a robust tracker based on a distributed unscented Kalman filter was proposed, in which an interacting multiple model [27] is used for accommodating the different possible motion dynamics of the speaker, yielding a smoothed trajectory of the speaker's movement in noisy and reverberant environments. Another approach to enhance the localization robustness is to fuse several observation modalities, as was demonstrated in audio-visual tracking methods [28]–[31].

Localization and tracking capabilities can be enhanced using model-based methods, assuming certain structures of either the speech signal or the acoustic channels. In [32] an autoregressive (AR) modelling for the speech components was used, and in [33], [34] the sources were modelled as sums of harmonically related sinusoids, which can describe many musical instruments and voiced speech. A model for the early reflections of the acoustic channels was presented in [35], based of which the early reflections were iteratively estimated. These models often rely on approximated physical and statistical assumptions, which do not always meet the practical conditions in complex real-world scenarios, with high levels of noise and reverberations. Recently, there is an attempt to overcome these limitations by applying data-driven models, rather than predefined physical and statistical models [36]–

[40]. In this family of methods, the central idea is to learn a mapping from high-dimensional acoustic features, extracted from the measured signals, to corresponding source positions.

Recently, we have presented several localization approaches based on the concepts of *manifold learning*. These algorithms are based on learning the mapping between the high dimensional acoustic channels to the source positions. In [40], we presented a semi-supervised source localization algorithm based on two-microphone measurements using the concept of manifold regularization in a reproducing kernel Hilbert space (RKHS). A Bayesian formulation to the localization algorithm, which is analogous to the manifold regularization approach, was presented in [41]. This Bayesian framework served as a corner stone for extending the single node (microphone pair) setup to an ad hoc network of microphone pairs, in [42]. In [43], we extended the static localization approach to tracking a moving source. The gist of the algorithm is to combine between a local interpolation of successive time steps, and a global interpolation of available prerecorded measurements. All the above methods are based on data-driven models and lead to improved localization results over TDOA-based approaches in adverse conditions [40], [44]. However, TDOA-based approaches may be superior in low reverberation and noise levels, since the TDOA readings are typically reliable under these conditions. Motivated by this observation, we propose here a hybrid tracking algorithm, which combines learning-based models with a traditional TDOA-based approach.

In this paper, we incorporate two data modalities, which are both extracted from the measured microphone signals. The first modality is the estimated TDOA, which ignores the complex reflection pattern characterizing the acoustic environment, and has an analytic known relation to the source position. The second modality comprises the full representation of the acoustic channel, which is high dimensional and has an unknown complex relation to the source position. As in [40], the relation between the acoustic channels and the source positions can be recovered by a data-driven model, built from a training set of prerecorded measurements in the enclosure of interest. We present a hybrid state-space formulation, which exploits both data modalities. The propagation of the source is expressed by a data-driven model, and translates the relations between the high-dimensional acoustic channels into a linear transition model for the source movement. The observation model is TDOA-based, and combines estimated TDOAs extracted from the measured signals and known TDOAs associated with the training set. The resulting state-space model lays the foundation for the application of an EKF. The algorithm performance is examined on simulated trajectories of a moving source in noisy and reverberant environments.

II. Problem Formulation

Consider a source moving in a reverberant enclosure. We assume that the movement of the source can be generally described by the following Markovian relation:

$$\mathbf{p}(t) = \mathbf{a}\left(\mathbf{p}(t-1), \mathcal{I}_t\right) \tag{1}$$

where $\mathbf{p}(t) = [p_x(t), p_y(t), p_z(t)]^T$ is the position of the source at time t. Here, $\mathbf{a}(\cdot)$ is the transition function, and $\mathcal{I}_t$ represents all the relevant information available at time t, such as prior information of the acoustic environment, the driving variance, etc. Traditional tracking methods usually rely on a completely heuristic propagation model, e.g., random walk or Langevin [16], [17]. Conversely, here instead of assuming a generic prior, we aim to infer a data-driven propagation model based on observations and a training set of prerecorded measurements. We assume that the training set consists of D measurements of static sources in known positions $\{\mathbf{p}_i\}_{i=1}^{D}$ in the enclosure of interest.

The estimation of the source position is based on audio signals generated by the source and measured by a set of microphones located in the enclosure. We consider a setup with M nodes, each of which consists of a pair of microphones, arbitrarily positioned in the enclosure. The source generates a speech signal $s(t)$, which is measured by all the microphones. The signal $y^{mj}(t)$ measured by the jth microphone in the mth node is given by:

$$y^{mj}(t) = \sum_{\tau} b_t^{mj}(\tau) s(t-\tau) + u^{mj}(t),\ 1 \le m \le M,\ j = 1, 2 \tag{2}$$

where b_t^{mj} is the time-varying acoustic impulse response (AIR) relating the source at position $\mathbf{p}(t)$ and the jth microphone in the mth node, and $u^{mj}(t)$ is the corresponding noise signal. Both the measured signals (2) and the training information can be exploited for estimating the source positions. For this purpose, in Sec. III, we define the propagation model (1), i.e. we specify the transition function $\mathbf{a}(\cdot)$, and define the relevant information $\mathcal{I}_t$, utilized at each time step. In Sec. IV, we define the relevant features, which are treated as noisy observations, and formulate the corresponding observation model. The two models form a state-space representation of the tracking problem, which is solved using an EKF recursion, as described in Sec.V.

III. Manifold-Based Propagation Model

The relevant information in the task of source localization is reflected in the measured signals (2) through the corresponding acoustic channels b_t^{mj}, and is independent of the source signal. We claim that observed changes of the acoustic channels during the movement of the source has a direct relation to the corresponding changes in the source position. Therefore, the propagation model of the source can be inferred from the variations of the corresponding acoustic channels. However, there is no simple model that relates the acoustic channels to the source positions. In order to relate between the two, we resort to data-driven models, which are learned based on training information.

We first analyse the characteristic of the acoustic channels and their relation to the source positions. Then, we extract acoustic feature vectors, representing the acoustic channels, from the measured signals (2). Next, we derive a data-driven model attaching each feature vector to the corresponding

source position. Finally, we learn the dynamics of the movement of the source from variations in the acoustic features during the movement.

A. The Acoustic Manifold

Consider a specific reverberant enclosure, such as a conference hall, an office, or a car interior. All possible acoustic channels in this enclosure have a complex reflection pattern stemming from the different surfaces and objects characterizing the enclosure. Hence, the acoustic channels are typically modelled by a large number of coefficients, resulting in an intricate high-dimensional representation. However, in a static environment, where the enclosure characteristics and the microphone positions are approximately fixed, the difference between acoustic channels is mainly attributed to the different positions of the source [44], [45]. Thus, the true intrinsic dimension of the set of possible acoustic channels in a specific enclosure is significantly smaller than the number of variables commonly used for their representation. By virtue of this assumption, we can state that the acoustic channels in a specific enclosure pertain to a low dimensional *manifold* [40], [46]. The structure of the manifold is unknown, and can be inferred from the training information. The available training measurements in a specific enclosure can be considered as samples drawn from the manifold. By analysing the relations between the given training samples, we can form a data-driven model that represents the structure of the manifold.

In the test phase, we obtain a series of varying acoustic channels during the source movement. This series of acoustic channels can be viewed as a trajectory on the learned manifold. The trajectory on the manifold corresponds to the actual trajectory of the source in the enclosure. Our goal is to relate the observed variations in the domain of the acoustic channels to the unknown dynamics of the source movement, aiming to devise a data-driven propagation model.

B. Acoustic Feature Vectors

We would like to establish the relation between the acoustic channels and the source positions. In practice, only the measured signals are available, and the acoustic channels cannot be directly accessed. Therefore, we use the associated relative transfer function (RTF) $H^m(t,k)$, defined as the ratio between the two transfer functions of the two microphones within the mth node, i.e.

$$H^m(t,k) = \frac{B^{m2}(t,k)}{B^{m1}(t,k)} \tag{3}$$

where $B^{mj}(t,k)$ is the (unknown) transfer function of the corresponding AIR. Note that here and henceforth t is used to denote a frame index in the short-time Fourier transform (STFT) domain, and k is a frequency bin. The RTF value in the kth frequency bin is estimated in the time-frequency domain using $L+1$ frames around t, and is smoothed across time:

$$\hat{H}_0^m(t,k) \simeq \frac{\hat{\Phi}_{21}^m(t,k)}{\hat{\Phi}_{11}^m(t,k)} = \frac{\sum_{n=t-L/2}^{t+L/2} Y^{m2}(n,k)Y^{m1*}(n,k)}{\sum_{n=t-L/2}^{t+L/2} Y^{m1}(n,k)Y^{m1*}(n,k)} \tag{4}$$

$$\hat{H}^m(t,k) = \gamma \hat{H}_0^m(t,k) + (1-\gamma)\hat{H}^m(t-1,k) \tag{5}$$

where $\hat{\Phi}_{11}^m(t,k)$ and $\hat{\Phi}_{21}^m(t,k)$ are the estimated power spectral density (PSD) and cross-PSD (CPSD) of the measured signals at the mth node, $Y^{mj}(t,k)$ is the STFT of the measured signal (2), and $0 \leq \gamma \leq 1$ is a smoothing parameter. Let $\mathbf{h}(t)$ denote the concatenation of RTF values in K frequency bins and in all M nodes:

$$\begin{aligned} \mathbf{h}^m(t) &= \left[\hat{H}^m(t,k_1),\ldots,\hat{H}^m(t,k_K)\right]^T \\ \mathbf{h}(t) &= \left[\mathbf{h}^{1T}(t),\ldots,\mathbf{h}^{MT}(t)\right]^T \end{aligned} \tag{6}$$

where $\mathbf{h}^m(t)$ resides on the mth manifold $\mathcal{M}_m \subset \mathbb{R}^K$, and $\mathbf{h}(t) \in \cup_{m=1}^M \mathcal{M}_m$. Note that each node is associated with a specific manifold $\mathcal{M}_m$, which represents the underlying geometric structure of the RTFs associated with the mth node. The different nodes are assumed to be spatially distributed in the enclosure. Therefore, they represent different views, and, in general, their associated manifolds have different structures. To recover the mapping between RTFs and source positions, we combine the different relations defined by the different manifolds. Merging the information from the different manifolds increases the spatial separation and improves the ability to accurately localize the source [42].

C. Mapping the Acoustic Features to Source Positions

We define the mapping function $f_c : \cup_{m=1}^M \mathcal{M}_m \rightarrow \mathbb{R}$ which attaches to an RTF sample $\mathbf{h}(t)$ its corresponding source position, i.e. $p_c(t) = f_c(\mathbf{h}(t))$, $c \in \{x,y,z\}$. The function $f_c(\cdot)$ is modelled as a zero-mean Gaussian process [47], specified by its covariance function $\kappa : \cup_{m=1}^M \mathcal{M}_m \times \cup_{m=1}^M \mathcal{M}_m \rightarrow \mathbb{R}$:

$$f_c(\cdot) \sim \mathcal{GP}(0,\kappa). \tag{7}$$

The covariance function $\kappa(\mathbf{h}(t),\mathbf{h}(\tau))$, often termed "kernel function", translates a pairwise relation between the RTFs $\mathbf{h}(t)$ and $\mathbf{h}(\tau)$ to a pairwise similarity between the corresponding positions $p_c(t)$ and $p_c(\tau)$. However, a covariance function which is based on the standard Euclidean distance between the high-dimensional RTFs, i.e. $\|\mathbf{h}(t)-\mathbf{h}(\tau)\|_2$, reflects the physical distance only for small scales [46]. Therefore, we define a manifold-based covariance function, in which the relation between two RTFs is evaluated with respect to the manifolds of the different nodes. For this purpose, we utilize the training information.

Recall that we assume the availability of D measurements of static sources in D different locations $\{\mathbf{p}_i\}_{i=1}^D$. We estimate the RTFs $\{\mathbf{h}_i\}_{i=1}^D$, defined as in (6), for each training position. Since during training the sources are static, we estimate the RTFs according to (4) using all the associated frames, without time smoothing. The training sources are assumed to be static,

since it is simpler to acquire measurements in known positions for static sources. In addition, their corresponding RTFs can be more accurately estimated, since the estimation can utilize a higher number of frames, in contrast to a moving source for which the acoustic channels vary from frame to frame. For notational clarification, we emphasize that $\mathbf{h}_i$ is a *training* RTF sample of a static source from a known position $\mathbf{p}_i$, whereas $\mathbf{h}(t)$ is a *test* RTF sample of a moving source from an unknown position $\mathbf{p}(t)$.

The relations on the mth manifold are evaluated using a standard Gaussian kernel, with a scaling factor ε_m:

$$\kappa_m\left(\mathbf{h}_i^m, \mathbf{h}_j^m\right) = \exp\left\{-\frac{\|\mathbf{h}_i^m - \mathbf{h}_j^m\|_2^2}{\varepsilon_m}\right\}. \tag{8}$$

Note that the Gaussian kernel implicitly limits the Euclidean distance to a small range governed by ε_m. As a result, it respects the linearity of the manifold for small scales. We propose to measure the relations in each manifold separately using (8), and then combine the different perspectives of the different nodes. The purpose is to form a similarity measure between RTFs, which represents relations that are co-observed in all manifolds. By relying on the training samples $\{\mathbf{h}_i\}_{i=1}^D$, we construct a multiple-manifold covariance function [42]:

$$\kappa(\mathbf{h}_r, \mathbf{h}_l) = \frac{1}{M^2}\sum_{q,w=1}^{M}\sum_{i=1}^{D}\kappa_q(\mathbf{h}_r^q, \mathbf{h}_i^q)\kappa_w(\mathbf{h}_l^w, \mathbf{h}_i^w). \tag{9}$$

where l and r represent ascription to certain positions, whereas q and w represent ascription to certain nodes. Note that (9) can be used to evaluate the covariance of both training and test samples. The covariance of test samples of different time frames is expressed by $\kappa\left(\mathbf{h}(t), \mathbf{h}(\tau)\right)$, and the covariance of a test sample and a training sample is expressed by $\kappa(\mathbf{h}(t), \mathbf{h}_i)$.

In (9), the covariance is obtained by averaging over all available training samples as well as over all different nodes. Comparing the relations to other training samples yields an affinity measure, which respects the manifold structure, hence is preferable over a regular Euclidean distance between the high-dimensional RTFs. Two RTF samples which exhibit similar relations to other samples on the manifold are considered close to each other, indicating that the corresponding source positions are also in close proximity, and vice versa. In addition, we average over the relations inspected in the various nodes to fuse their different views. The defined covariance consists of all the inter-relations between the different nodes, enhancing observations which are common to pairs of nodes, and ignoring relations that appear in only one node. Further details about the derivation of (9) can be found in [42]. A nomenclature listing the different symbols and their meanings is given in Table I.

It is important to note that in the following derivation of the tracking algorithm, we regard the mapping function $f_c(\cdot)$ as a Gaussian process with a covariance function $\kappa(\cdot,\cdot)$ as stated in (7), with no restriction to a specific kernel function $\kappa(\cdot,\cdot)$. In the experimental study in Section VI, we use the kernel defined in (9). However, other definitions for the covariance of the Gaussian process can be applied as well.

$\mathbf{h}_i$	A training RTF sample, consisting of RTFs of all M nodes, associated with a static source at position $\mathbf{p}_i$, $1 \le i \le D$
$\mathbf{h}(t)$	A test RTF sample, consisting of RTFs of all M nodes, associated with a moving source at an instantaneous position $\mathbf{p}(t)$
$f_c(\cdot)$	A Gaussian process representing possible source positions, mapped from their corresponding RTFs
$\kappa(\cdot,\cdot)$	A kernel function defining the covariance of positions drawn from the Gaussian process, evaluated via the relation between the corresponding RTFs
$\mathcal{M}_m$	The manifold associated with the RTFs of the mth node

TABLE I: Nomenclature

D. Derivation of The Propagation Model

After defining the instantaneous relation between RTFs and source positions via the function $f_c(\cdot)$, we can now define the parameters of the propagation model for the source movement (1). In this model, the current source position $\mathbf{p}(t)$ is a combination of the previous source position $\mathbf{p}(t-1)$ and of relevant training positions in proximity to the source. The relations between successive positions and the chosen training positions are determined according to the relations between the observed RTF samples, formed by the manifold-based covariance terms defined in the previous section (9).

Following [48], for each test RTF sample $\mathbf{h}(t)$ we define a subset of neighboring training samples $\left\{\mathbf{h}_{t_i} \mid \| \mathbf{h}(t) - \mathbf{h}_{t_i} \| < \eta, t_i \in \{1,\ldots,D\}\right\}$, where η defines the neighborhood radius. In order to obtain fixed-size sets, we focus on N nearest-neighbors among the defined subset (assuming η is large enough to include N samples), denoted by $\bar{\mathcal{H}}_t = \{\mathbf{h}_{t_i}\}_{i=1}^N$. In this definition, the neighbors are determined based on the Euclidean distance between the corresponding RTFs. Other similarity measures can be used for this purpose, such as relying on the distance induced by the covariance in (9). Note that here, the exact extent of similarity is of secondary importance. We only need to identify nearby samples, hence, the Euclidean distance, which is meaningful for small scales, is appropriate for this task.

Let $\mathcal{H}_t = \mathbf{h}(t) \bigcup \bar{\mathcal{H}}_t$ denote an extended set of size $N+1$, which consists of the current RTF sample as well as all the chosen neighboring training samples. Let $\mathbf{f}_{t,c} = [f_c(\mathbf{h}(t)), f_c(\mathbf{h}_{t_1}), \ldots, f_c(\mathbf{h}_{t_N})]^T$ denote a concatenation of all the corresponding mappings of the function $f_c(\cdot)$ over the samples in $\mathcal{H}_t$, representing their source positions. The relation between the mappings of the t-th and the $t-1$-th sets is dictated by the Gaussian process (7). Both $\mathbf{f}_{t,c}$ and $\mathbf{f}_{t-1,c}$ are Gaussian vectors, which consist of samples from the Gaussian process $f_c(\cdot)$. Therefore, they have a joint Gaussian distribution with zero-mean and covariance matrix, which is based on the covariance terms in (9):

$$\begin{bmatrix}\mathbf{f}_{t,c}\\ \mathbf{f}_{t-1,c}\end{bmatrix}\Bigg|\mathcal{H}_{t,t-1} \sim \mathcal{N}\left(\mathbf{0}_{2(N+1)}, \begin{bmatrix}\boldsymbol{\Sigma}_{t,t} & \boldsymbol{\Sigma}_{t,t-1}\\ \boldsymbol{\Sigma}_{t,t-1}^T & \boldsymbol{\Sigma}_{t-1,t-1}\end{bmatrix}\right) \tag{10}$$

where $\mathcal{H}_{t,t-1} = \mathcal{H}_t \cup \mathcal{H}_{t-1}$, $\mathbf{0}_{2(N+1)}$ is an $2(N+1)\times 1$ vector

of all zeros, and

$$\mathbf{\Sigma}_{t,\tau} = \begin{bmatrix} \kappa(\mathbf{h}(t),\mathbf{h}(\tau)) & \kappa(\mathbf{h}(t),\mathbf{h}_{\tau_1}) & \cdots & \kappa(\mathbf{h}(t),\mathbf{h}_{\tau_N}) \\ \kappa(\mathbf{h}_{t_1},\mathbf{h}(\tau)) & \kappa(\mathbf{h}_{t_1},\mathbf{h}_{\tau_1}) & \cdots & \kappa(\mathbf{h}_{t_1},\mathbf{h}_{\tau_N}) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(\mathbf{h}_{t_N},\mathbf{h}(\tau)) & \kappa(\mathbf{h}_{t_N},\mathbf{h}_{\tau_1}) & \cdots & \kappa(\mathbf{h}_{t_N},\mathbf{h}_{\tau_N}) \end{bmatrix}. \tag{11}$$

Accordingly, the conditional distribution of $\mathbf{f}_{t,c}$ given $\mathbf{f}_{t-1,c}$ is also Gaussian:

$$\Pr\left(\mathbf{f}_{t,c}|\mathbf{f}_{t-1,c},\mathcal{H}_{t,t-1}\right) = \mathcal{N}\left(\mathbf{A}(\mathcal{H}_{t,t-1})\mathbf{f}_{t-1,c},\mathbf{Q}(\mathcal{H}_{t,t-1})\right) \tag{12}$$

where

$$\begin{aligned} \mathbf{A}(\mathcal{H}_{t,t-1}) &= \mathbf{\Sigma}_{t,t-1}\mathbf{\Sigma}_{t-1,t-1}^{-1} \\ \mathbf{Q}(\mathcal{H}_{t,t-1}) &= \mathbf{\Sigma}_{t,t} - \mathbf{\Sigma}_{t,t-1}\mathbf{\Sigma}_{t-1,t-1}^{-1}\mathbf{\Sigma}_{t,t-1}^{T}. \end{aligned} \tag{13}$$

Conveniently, the Gaussian conditional distribution induces linear dependence between the positions of the current set $\mathbf{f}_{t,c}$ and the positions of the preceding set $\mathbf{f}_{t-1,c}$. Thus, the propagation of the source positions in (1) can be formulated by a linear equation of successive time steps, with a Gaussian driving noise:

$$\mathbf{f}_{t,c} = \mathbf{A}(\mathcal{H}_{t,t-1})\mathbf{f}_{t-1,c} + \boldsymbol{\xi}(\mathcal{H}_{t,t-1}) \tag{14}$$

where the noise characteristics are directly inferred from the conditional probability (12), i.e. $\boldsymbol{\xi}(\mathcal{H}_{t,t-1}) \sim \mathcal{N}\left(\mathbf{0}_{N+1},\mathbf{Q}(\mathcal{H}_{t,t-1})\right)$. Here, we extend the commonly used random walk model [20], by interpolating over both the previous position and close positions from the training set. The model parameters, i.e. the state-transition matrix and the variance of the process noise, are computed using $\mathcal{I}_t \equiv \mathcal{H}_{t,t-1}$, based on the relations between the corresponding RTFs with respect to the different manifolds. This way a nonlinear regression in the high-dimensional space of the RTFs results in a linear (time-varying) propagation model for the source movement.

The model applies to each of the coordinates, x, y or z, independently. Here, we assume that the variations of the RTFs reflect an independent movement of the source in either direction. We use this independence assumption to simplify the derived Gaussian process mapping. However, the observation model presented in Section IV, which accounts on TDOA readings, uses the full 3D location as required by the physical model. Consequently, the full propagation model for the 3-D position $\mathbf{f}_t = \left[\mathbf{f}_{t,x}^T, \mathbf{f}_{t,y}^T, \mathbf{f}_{t,z}^T\right]^T$ is given by:

$$\mathbf{f}_t = \mathbf{A}_3(\mathcal{H}_{t,t-1})\mathbf{f}_{t-1} + \boldsymbol{\xi}_3(\mathcal{H}_{t,t-1}) \tag{15}$$

where $\mathbf{A}_3(\mathcal{H}_{t,t-1}) = \mathbf{A}(\mathcal{H}_{t,t-1}) \otimes \mathbf{I}_3$ and $\boldsymbol{\xi}_3(\mathcal{H}_{t,t-1}) \sim \mathcal{N}\left(\mathbf{0}_{3(N+1)},\mathbf{Q}_3(\mathcal{H}_{t,t-1})\right)$ with $\mathbf{Q}_3(\mathcal{H}_{t,t-1}) = \mathbf{Q}(\mathcal{H}_{t,t-1}) \otimes \mathbf{I}_3$. Here, $\otimes$ is the Kronecker product and $\mathbf{I}_3$ is the 3×3 identity matrix.

IV. TDOA-Based Observation Model

The observations are formed by the range differences $\boldsymbol{r} = \left[r^1,\ldots,r^M\right]^T$ of each of the nodes. The range differences have a known nonlinear relation to the source position:

$$r^m = g\left(\mathbf{p}\right) = \left\|\mathbf{p} - \mathbf{q}^{m2}\right\|_2 - \left\|\mathbf{p} - \mathbf{q}^{m1}\right\|_2 \tag{16}$$

where $\mathbf{q}^{mj}$ is the position of the jth microphone in the mth node (assumed to be known). The range differences attached with the current time step, can be estimated using the generalized cross-correlation (GCC) method [10], or they can be extracted from the estimated RTFs [49]:

$$\hat{r}^m(t) = \frac{1}{v}\operatorname*{argmax}_{\tau} \hat{h}^m(t,\tau) \equiv \text{IDFT}\left\{\hat{H}^m(t,k)\right\} \tag{17}$$

where v is the sound velocity. For the subset $\bar{\mathcal{H}}_t$ of the chosen neighbors, the range differences $\{\hat{\mathbf{r}}_{t_i}\}_{i=1}^N$ can be computed by (16), using the corresponding measured positions $\{\mathbf{p}_{t_i}\}_{i=1}^N$.

Let $\hat{\mathbf{r}}_t = \left[\hat{\mathbf{r}}^T(t), \hat{\mathbf{r}}_{t_1}^T, \ldots, \hat{\mathbf{r}}_{t_N}^T\right]^T$ be the concatenation of $M(N+1)$ values/estimates of the range differences associated with the set $\mathcal{H}_t$. A nonlinear observation model is formed by:

$$\hat{\mathbf{r}}_t = \mathbf{g}(\mathbf{f}_t) + \boldsymbol{\zeta}_t \tag{18}$$

where $\mathbf{g}(\mathbf{f}_t) = [\mathbf{g}^T(\mathbf{p}(t)), \mathbf{g}^T(\mathbf{p}_{t_1}), \ldots, \mathbf{g}^T(\mathbf{p}_{t_N})]^T$ and

$$\mathbf{g}(\mathbf{p}) = \begin{bmatrix} \left\|\mathbf{p} - \mathbf{q}^{12}\right\|_2 - \left\|\mathbf{p} - \mathbf{q}^{11}\right\|_2 \\ \vdots \\ \left\|\mathbf{p} - \mathbf{q}^{M2}\right\|_2 - \left\|\mathbf{p} - \mathbf{q}^{M1}\right\|_2 \end{bmatrix}. \tag{19}$$

Here, $\boldsymbol{\zeta}_t \sim \mathcal{N}\left(\mathbf{0}_{M(N+1)},\mathbf{R}_t\right)$ is the observation error, with a diagonal covariance matrix $\mathbf{R}_t$:

$$\mathbf{R}_t = \text{blkdiag}\left\{\mathbf{R}(t),\mathbf{R}_{t_1},\ldots,\mathbf{R}_{t_N}\right\} \tag{20}$$

where $\mathbf{R}(t)$ and $\mathbf{R}_{t_i}$ are the covariance matrices of the observation noise associated with the current sample and with the ith sample in $\bar{\mathcal{H}}_t$, respectively:

$$\begin{aligned} \mathbf{R}(t) &= \text{diag}\left\{(\sigma_t^1)^2,\ldots,(\sigma_t^M)^2\right\} \\ \mathbf{R}_{t_i} &= \text{diag}\left\{(\sigma_{t_i}^1)^2,\ldots,(\sigma_{t_i}^M)^2\right\}. \end{aligned} \tag{21}$$

Typically, $(\sigma_{t_i}^m)^2 \ll (\sigma_t^m)^2$ for all $1 \leqslant m \leqslant M, 1 \leqslant i \leqslant N$, since $\hat{r}_{t_i}^m$ is computed in (16) using the corresponding measured position $\mathbf{p}_{t_i}$, while $\hat{r}^m(t)$ is estimated by (17). In addition, the variance $(\sigma_t^m)^2$ is influenced by reverberation and noise levels, as well as by the microphone positions with respect to the speaker. Conversely, the variance $(\sigma_{t_i}^m)^2$ is independent of the acoustic conditions, reflecting the reliability of the measured training positions.

V. Extended Kalman Filter Tracking

A state-space representation is formed by combining the propagation model with the observation model. In our case, we combine the manifold-based propagation model (15), which is derived based on the RTFs, and the TDOA-based observation model (18), which relies on the TDOA readings. Both models take advantage of the training information. Further discussion about this combination, as a special case of a more general scheme, is given at the end of this section. We obtain the following state-space representation:

$$\begin{aligned} \mathbf{f}_t &= \mathbf{A}_3(\mathcal{H}_{t,t-1})\mathbf{f}_{t-1} + \boldsymbol{\xi}_3(\mathcal{H}_{t,t-1}) \\ \hat{\mathbf{r}}_t &= \mathbf{g}(\mathbf{f}_t) + \boldsymbol{\zeta}_t. \end{aligned} \tag{22}$$

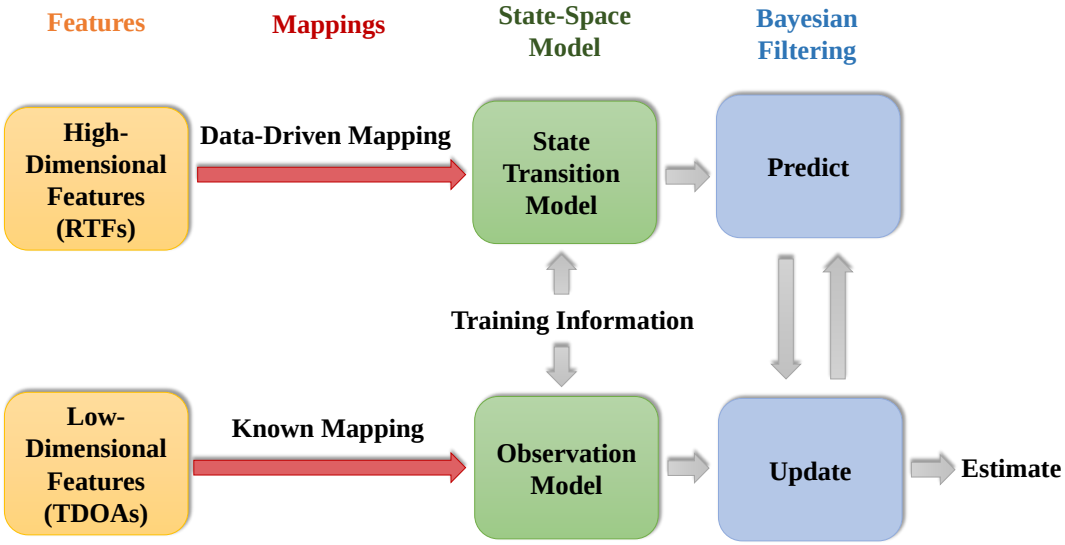


Fig. 1: An illustration of the proposed general scheme, which combines two data modalities of different types: The first type is high-dimensional features (RTFs), which are mapped to the hidden states (source positions) by a data-driven model using the training information, and form the state transition model (prorogation model for the source movement (15)). The second type is low dimensional features (TDOAs), which have known relation to the hidden states and form the observation model (TDOA-based observation model (18)). The derived state-space representation is iteratively solved using Bayesian filtering in two steps: prediction step and update step (EKF recursion (25)).

Due to the nonlinearity of the observation model a nonlinear Bayesian filtering technique should be applied. The particle filter [50], which is based on random sampling, cannot be directly applied here, since the evaluation of the covariance terms in (22) is based on the corresponding RTFs, which are given only for the pre-generated training samples, and are unavailable for other random samples. Either the unscented Kalman filter [51] or the extended Kalman filter (EKF) may be applied. In [15], the trackers based on these two nonlinear filters were shown to yield comparable results, hence we adopt here the EKF recursion due to its simplicity.

The EKF algorithm applies a linearization of the observation model (which is given here in a closed-form, in contrast to the inferred propagation model), using the following Jacobian:

$$\nabla_{\mathbf{f}}\mathbf{g}(\mathbf{f}_t) = \text{blkdiag}\left\{ \nabla_{\mathbf{p}}\,\mathbf{g}(\mathbf{p}(t)), \nabla_{\mathbf{p}}\mathbf{g}(\mathbf{p}_{t_1}), \ldots, \nabla_{\mathbf{p}}\mathbf{g}(\mathbf{p}_{t_N}) \right\} \tag{23}$$

where

$$\nabla_{\mathbf{p}}\mathbf{g}(\mathbf{p}) = \begin{bmatrix} \left(\frac{\mathbf{p}-\mathbf{q}^{12}}{\|\mathbf{p}-\mathbf{q}^{12}\|_2} - \frac{\mathbf{p}-\mathbf{q}^{11}}{\|\mathbf{p}-\mathbf{q}^{11}\|_2}\right)^T \\ \vdots \\ \left(\frac{\mathbf{p}-\mathbf{q}^{M2}}{\|\mathbf{p}-\mathbf{q}^{M2}\|_2} - \frac{\mathbf{p}-\mathbf{q}^{M1}}{\|\mathbf{p}-\mathbf{q}^{M1}\|_2}\right)^T \end{bmatrix}. \tag{24}$$

Accordingly, the EKF recursion takes the following form:

$$\begin{aligned} \hat{\mathbf{f}}(t|t-1) &= \mathbf{A}_{3t}\hat{\mathbf{f}}(t-1|t-1) \\ \mathbf{\Pi}(t|t-1) &= \mathbf{A}_{3t}\mathbf{\Pi}(t-1|t-1)\mathbf{A}_{3t}^T + \mathbf{Q}_{3t} \\ \hat{\mathbf{f}}(t|t) &= \hat{\mathbf{f}}(t|t-1) + \mathbf{\Gamma}(t)\left(\hat{\mathbf{r}}_t - \mathbf{g}\left(\hat{\mathbf{f}}(t|t-1)\right)\right) \\ \mathbf{\Pi}(t|t) &= \left(\mathbf{I}_{3(N+1)} - \mathbf{\Gamma}(t)\mathbf{G}_t\right)\mathbf{\Pi}(t|t-1) \end{aligned} \tag{25}$$

where $\mathbf{A}_{3t} \equiv \mathbf{A}_{3t}(\mathcal{H}_{t,t-1})$, $\mathbf{Q}_{3t} \equiv \mathbf{Q}_{3t}(\mathcal{H}_{t,t-1})$. Here, $\mathbf{\Pi}(t|t-1)$ is the predicted covariance, $\mathbf{\Pi}(t|t)$ is the posteriori covariance, and $\mathbf{\Gamma}(t)$ is the Kalman gain, defined as:

$$\mathbf{\Gamma}(t) = \mathbf{\Pi}(t|t-1)\mathbf{G}_t^T\left(\mathbf{G}_t\mathbf{\Pi}(t|t-1)\mathbf{G}_t^T + \mathbf{R}_t\right)^{-1}. \tag{26}$$

where $\mathbf{G}_t \equiv \nabla_{\mathbf{f}}\mathbf{g}\left(\hat{\mathbf{f}}(t|t-1)\right)$. The proposed tracking scheme is summarized in Algorithm 1.

The resulting estimator $\hat{\mathbf{f}}(t|t)$ is a combination of a predicted position $\hat{\mathbf{f}}(t|t-1)$, and a correction term $\mathbf{\Gamma}(t)\left(\hat{\mathbf{r}}_t - \mathbf{g}(\mathbf{f}(t|t-1))\right)$. The predicted position is constructed by a local interpolation of the previous position and adjacent training positions, using manifold-based models. The correction term is devised from the observed range differences. Note that the proposed hybrid estimator consists of two estimates based on two data modalities with different characteristics, namely the RTFs and the TDOAs. On the one hand, the TDOAs represent low-dimensional observations with known relation to the source positions. The TDOA readings suffer from two major disadvantages. First, the TDOA estimation based on the measured signals may be unreliable and its accuracy degrades as reverberation level increases. Second, some of the relevant information is lost when instead of considering the entire acoustic channels, the TDOA, which represents only the direct arrival of the response, is extracted. On the other hand, the RTFs represent high-dimensional features with unknown complex relations to the source positions. The problem is alleviated by the assumption that the RTFs are confined to a manifold of much lower dimensions. The mapping to the corresponding source positions is modelled by a Gaussian process defined with respect to the manifolds of the different nodes. The unknown relations are recovered by a data-driven

model deduced from the training information.

Algorithm 1: Hybrid Tracking

Input :

- A training set consisting of D RTF samples $\{\mathbf{h}_i\}_{i=1}^{D}$ of static sources located at known positions $\{\mathbf{p}_i\}_{i=1}^{D}$.
- New test measurements of a moving source along an unknown trajectory.

Output:

- Estimated source positions $\{\hat{\mathbf{p}}(t)\}_{t=1}^{T}$ corresponding the test measurements.

For each time segment t:

1) Estimate the concatenated RTF vector $\mathbf{h}(t)$ of (6), using (4) and (5).
2) Search for N nearest neighbors of $\mathbf{h}(t)$ among the training samples, and form the set $\bar{\mathcal{H}}_t = \{\mathbf{h}_{t_i}\}_{i=1}^{N}$.
3) Form the sets $\mathcal{H}_t = \mathbf{h}(t) \bigcup \bar{\mathcal{H}}_t$ and $\mathcal{H}_{t,t-1} = \mathcal{H}_t \cup \mathcal{H}_{t-1}$.
4) Compute the correlation terms between the current sample $\mathbf{h}(t)$ and the sets $\mathcal{H}_t$ and $\mathcal{H}_{t-1}$, using (8) and (9).
5) Compute the matrices $\mathbf{A}(\mathcal{H}_{t,t-1})$ and $\mathbf{Q}(\mathcal{H}_{t,t-1})$ according to (13).
6) Estimate the range differences $\{\hat{r}^m(t)\}_{m=1}^{M}$, according to (17).
7) Apply EKF recursion according to (25) and (26).

The proposed hybrid scheme combines two data modalities with different properties, aiming to inherit the advantages of both, in order to improve the localization accuracy. The core idea of combining different types of data modalities can be generalized in various ways. Both the RTFs and the TDOAs can be substituted by other relevant observations with similar properties, which can be extracted from the measured signals (2), or from other available measurements. Following the concepts of the derived estimator, a large variety of hybrid trackers can be derived, by the general scheme illustrated in Fig. 1.

VI. Experimental Study

We carried out a simulation study to examine the ability of the proposed method to track a moving source in 2-D. In Section VI-A, we describe the setup and present initial tracking results with fixed and varying velocity movements. In Section VI-B, we present the reference algorithms and discuss their computational complexity. A comparison of the performance for different reverberation and noise levels is provided in Section VI-C. Additional aspects of the proposed method are examined in Section VI-D.

A. Experimental Setup and Initial Results

We simulated a $5.2\times6.2\times3$m room with 3 pairs of microphones mounted next to the room walls, using MCROOMSIM, a multichannel room acoustics simulator [52]. To simulate the measured signal corresponding to a moving source along a specific trajectory, we filtered different parts of the speech signal with AIRs corresponding to different positions along the specified trajectory. The filtering was performed in the STFT domain using frames of 341ms, and 93.75% overlap, where each frame was multiplied by the corresponding transfer-function. Using a proper inverse-STFT, we obtained the time-domain signals.

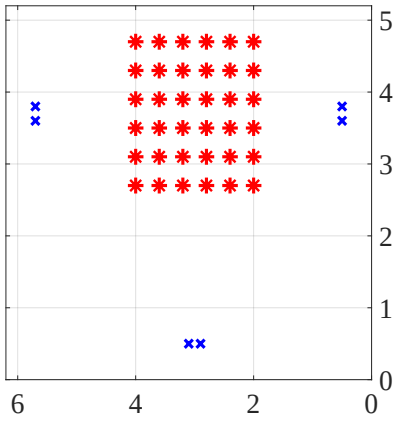


Fig. 2: The room setup: blue x-marks denote the microphone positions and red asterisks denote the training positions.

The locations of the source, both in the training and in the test, were confined to a 2×2m rectangular region, in a fixed height of 1.5m (the same height of all the microphones). The sampling rate was set to 16kHz. We generated a training set with $D = 36$ samples, forming a regular grid with a resolution of 0.4m. The samples were generated using 3s long speech signals, in noiseless conditions. The room setup and the training positions are illustrated in Fig. 2.

An initial examination was carried out to track a source, moving along both a straight line and a sinusoidal trajectory in the designated region. The reverberation time was set to 200ms. The duration of the entire movement of the source was 3s along the straight line, and 5s along the sinusoidal movement. For both movement types, the source average velocity was approximately 1m/s. The measured signals were split into frames of 128ms, with 75% overlap between successive frames. For each frame, the RTF was estimated according to (4) using 3 successive time frames, and smoothed across time as in (5). Each RTF sample consisted of $K = 250$ frequency bins, corresponding to the $0 - 2$kHz frequency band, where most of the speech components are concentrated. Figure 3 depicts the two trajectories along the x and the y axes and the obtained tracking results. It can be observed that the proposed method is able to track the source for both trajectories.

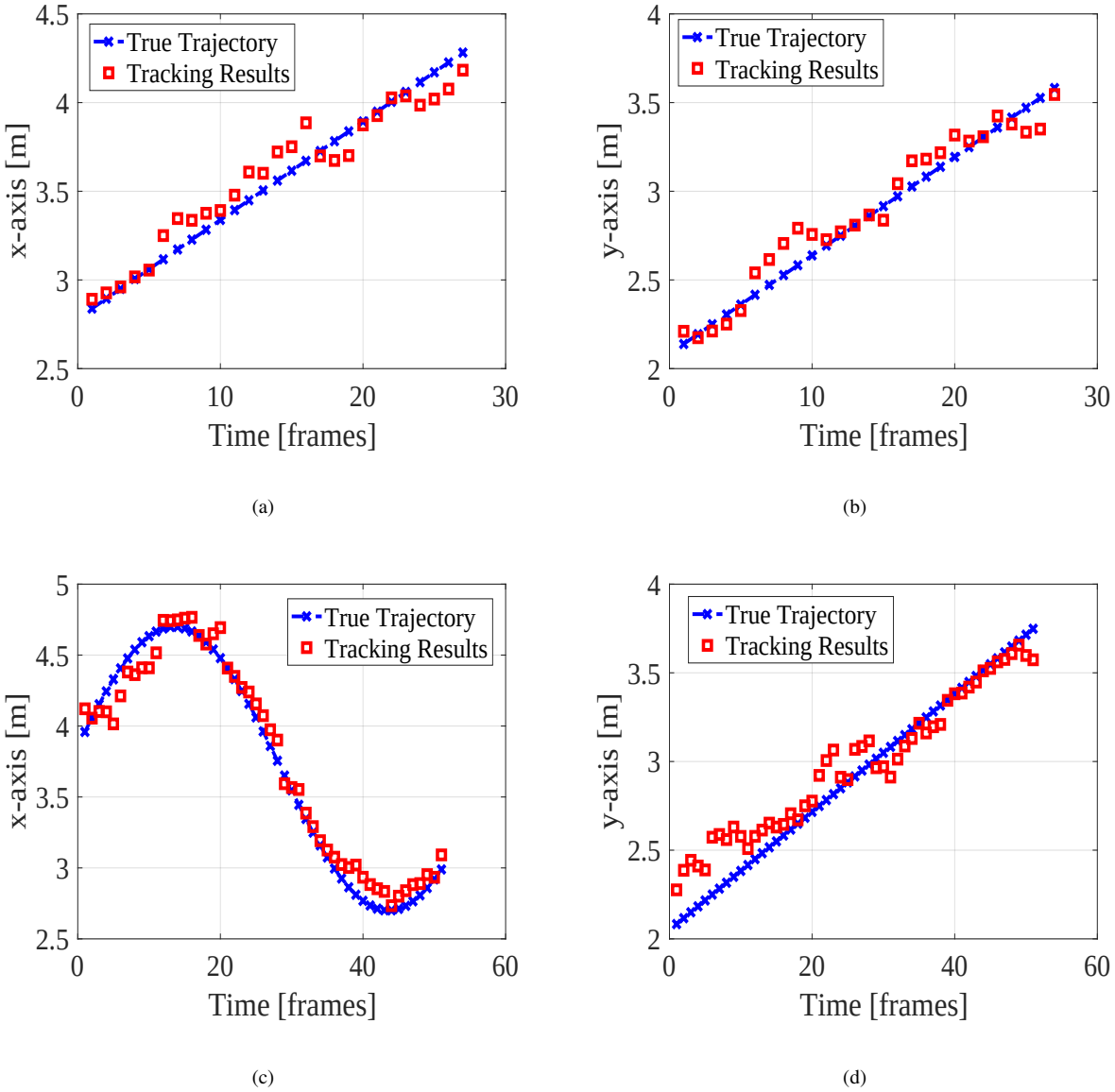


Fig. 3: True path and estimated path for (a)-(b) a straight line movement and for (c)-(d) a sinusoidal movement.

B. Compared Algorithms and Computational Complexity

A more comprehensive examination was carried out to evaluate the performance of the proposed method for different reverberation and noise levels. The results were compared with a TDOA-based tracker ('TDOA-EKF') as in [15], in which the manifold-based propagation model (14) is substituted by a simple random walk model. In addition, we compared the proposed method to a learning-based approach ('KNN-KF') adapted from [48], in which the TDOA-based observation model (18) is substituted by a linear model that links between the predicted positions of the subset $\bar{\mathcal{H}}_t$ to their known positions. These two competing methods represent two opposite extremes, which are combined in the proposed hybrid approach. The algorithms' parameters are summarized in Table II.

We first discuss the computational complexity of the proposed method as compared to the KNN-KF algorithm and the TDOA-EKF algorithm. For simplicity, we equally weight multiplications, divisions, additions, subtractions and exponentiations. The major factors that influence the complexity of the implementation are: (i) the number of nodes M, (ii) the number of training samples D, (iii) the number of training neighbors N, (iv) the FFT length F and (v) the number of concatenated frequency bins K. The orders of magnitude of the different operations performed by the algorithms are summarized in Table III. The computations required in the training stage (can be found in [42]) are performed off-line, hence are omitted from Table III. The complexity of computing the covariance terms (9) for constructing the matrices $\mathbf{A}(\mathcal{H}_{t,t-1})$ and $\mathbf{Q}(\mathcal{H}_{t,t-1})$ according to (13), is based on the analysis in [42]. The EKF recursion (25) requires several

TABLE II: Algorithms' Parameters

	TDOA-EKF	KNN-KF	Hybrid
Trans. Mat.	$0.99 \cdot \mathbf{I}_3$	data-driven	data-driven
Trans. Noise Cov.	$10^{-4} \cdot \mathbf{I}_3$	data-driven	data-driven
Obs. Noise Cov.	$(\sigma_t^m)^2 = \begin{cases} 10^{-7} & \text{for } 200-300\text{ms} \ / \ 15-25\text{dB} \\ 10^{-5} & \text{for } 400-600\text{ms} \ / \ 0-10\text{dB} \end{cases}$	$10^{-3} \cdot \mathbf{I}_N$	$(\sigma_t^m)^2 = \begin{cases} 10^{-3} & \text{for } 200-500\text{ms} \ / \ 15-25\text{dB} \\ 10^{-2} & \text{for } 600\text{ms} \ / \ 0-10\text{dB} \end{cases}$ $(\sigma_{t_i}^m)^2 = 10^{-6}, \forall 1 \leq i \leq N$

matrix multiplications and one matrix inversion, which at most require $\mathcal{O}(M^3)$ or $\mathcal{O}(M^3(N+1)^3)$ operations for either of the algorithms, assuming $M > 2$.

C. Algorithms' Performance

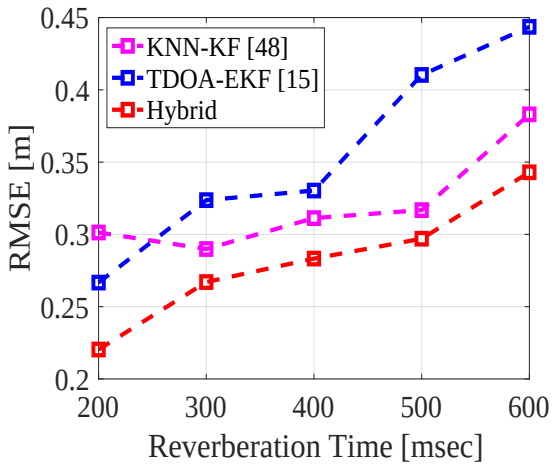
Fifty Monte-Carlo trials were carried out with different speakers moving along the defined straight line for 3s. The same room and the same array geometry were used for both training and testing. The test signals were generated using different utterances than the ones used for training. Diffuse noise signals with white spectrum were added to the measurements. The average root mean square errors (RMSEs) of all three algorithms are depicted in Fig. 4 as a function of the reverberation level (noiseless) and as a function of the noise level (200ms reverberation time).

It can be observed that the TDOA-based approach [15] preforms well in low reverberation and noise levels. However its performance degrades in reverberant and noisy conditions, most likely due to inaccurate TDOA estimates. Conversely, the learning-based approach [48], which relies on the training information and takes into consideration the representation of the acoustic channels, is more robust to reverberation and noise. The proposed hybrid algorithm outperforms both competing algorithms for all reverberation and noise levels. We conclude that the hybrid method inherits the benefits of both approaches, yielding an improved performance in various conditions.

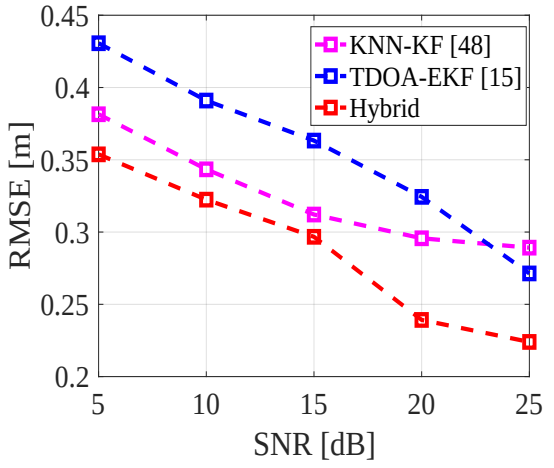
D. Performance Analysis

We further investigated several aspects concerning the proposed method. First, we evaluated the tracking performance with respect to the number of nodes in the network. Second, we examined the ability of the proposed method to track to changes in a switching scenario with several nonconcurrent moving speakers.

We examined the influence of the number of nodes on the tracking performance. We used a network with five distributed microphone pairs consisting of the three original nodes depicted in Fig. 2, and two additional nodes located at: $[1.5, 4.5, 1.5]$ and $[1.5, 1.5, 1.5]$. For each fixed number of nodes $\tilde{M} \leq 5$, we randomly chose $\tilde{M}$ nodes out of five. We used the chosen nodes for tracking a source moving along the defined straight line for $T_{60} = 300$ms. The error was averaged over 50 trials with different speakers and different network



(a)



(b)

Fig. 4: The RMSE for (a) various reverberation levels and for (b) various noise levels.

constellations. The average RMSE is depicted in Fig. 5 as a function of the number of nodes $\tilde{M} \in \{1, \ldots, 5\}$. We observe a gradual performance improvement as the number of nodes increases. By adding more nodes, we gain more information representing different perspectives. Merging all the view points together facilitates the identification of interfering factors and

TABLE III: Computational Complexity

	TDOA-EKF	KNN-KF	Hybrid
RTF/Spectrum Estimation (4),(5)	$\mathcal{O}(MF\log_2 F)$	$\mathcal{O}(MF\log_2 F)$	$\mathcal{O}(MF\log_2 F)$
TDOA Estimation (17)	$\mathcal{O}(MF\log_2 F)$	-	$\mathcal{O}(MF\log_2 F)$
N Nearest-Neighbors Search	-	$\mathcal{O}(D\log_2 D)$	$\mathcal{O}(D\log_2 D)$
Covariance Computation (9),(13)	-	$\mathcal{O}(KMD+M^2DN)$	$\mathcal{O}(KMD+M^2DN)$
EKF Recursion (25)	$\mathcal{O}(M^3)$	$\mathcal{O}(M^3(N+1)^3)$	$\mathcal{O}(M^3(N+1)^3)$

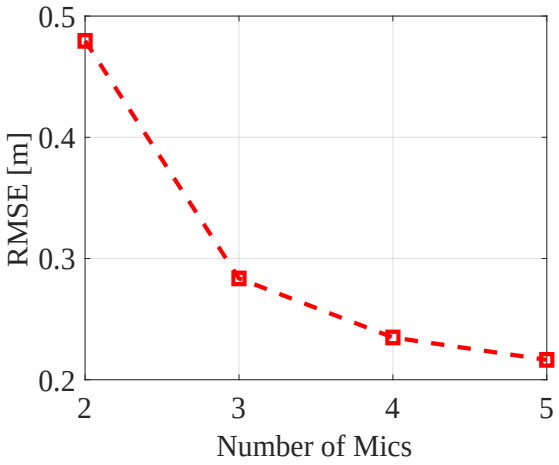
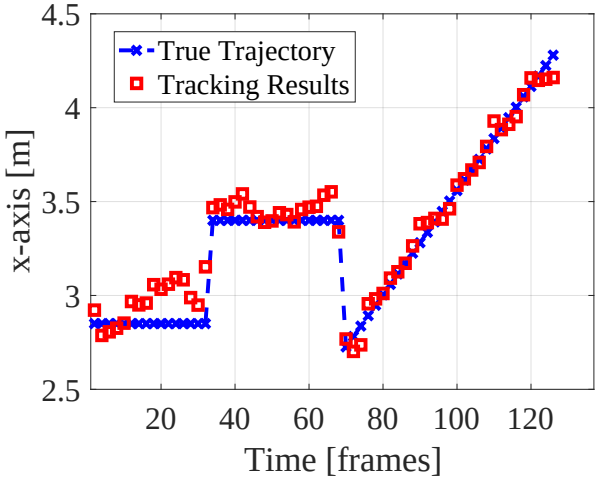
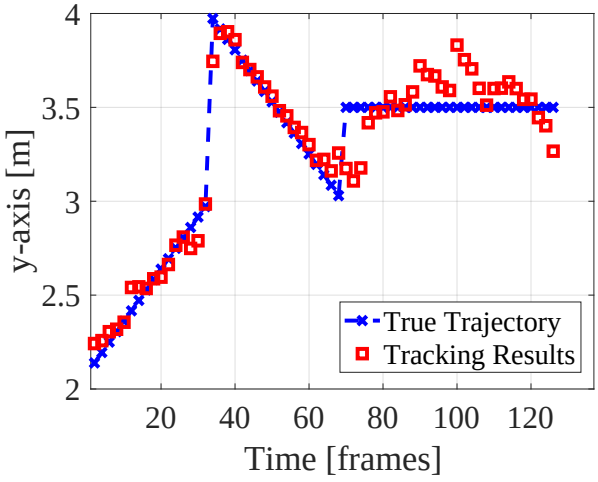


Fig. 5: The RMSE as a function of the number of nodes



(a)



(b)

Fig. 6: Tracking results in a switching scenario of three nonconcurrent speakers

the accurate localization of the source position by appropriately matching the observations made by the nodes.

Finally, we examined a switching scenario, which consists of three nonconcurrent speakers with three different trajectories starting from different points. The trajectory of the first speaker starts from $[2.85, 2, 1.5]$, continues along the y-axis for 1m during 1.5s, and ends at $[2.85, 3, 1.5]$. The trajectory of the second speaker starts from $[3.4, 4, 1.5]$, continues along the y-axis in the opposite direction for 1m during 1.5s, and ends at $[3.4, 3, 1.5]$. The trajectory of the third speaker starts from $[2.7, 3.5, 1.5]$, continues along the x-axis for 2m during 3s, and ends at $[4.7, 3.5, 1.5]$. The reverberation time was set to 200ms. The estimated trajectories evaluated by the proposed method are depicted in Fig. 6 along the x and the y axes. We observe that the proposed method is able to track changes after a short adaptation time.

We conclude this section by highlighting the main points demonstrated in the experimental results. We have shown that the proposed method can track a moving source in various noisy and reverberant conditions. In addition, we have seen that the proposed method is superior over either a traditional TDOA-based approach or a pure learning-based approach, stressing the advantage of the combination of both methods in the proposed hybrid algorithm. It was also shown that the performance is improved by increasing the number of nodes in the network, and that the proposed method successfully adapt to changes in a switching scenario.

VII. Conclusions

A hybrid tracking algorithm is presented using a learning-based model combined with a TDOA-based model. The source propagation in the physical domain is learned from the variations of the RTF samples with respect to an acoustic manifold.

The structure of the manifold is inferred in a data-driven manner from the training information. The source position is nonlinearly related to the estimated TDOAs, which constitute the observation model. The resulting state-space formulation exploits both the characteristics of the full acoustic channels represented by the RTFs, and the direct arrival information represented by the TDOA readings. Simulation results demonstrate the ability of the proposed method to locate the source in challenging noisy and reverberant conditions. The algorithm exploits the high accuracy of TDOA-based methods in optimal conditions, while maintaining robustness, which characterizes learning-based approaches.

REFERENCES

[1] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2000, pp. 909–912.

[2] O. Thiergart, M. Taseska, and E. A. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 2182–2196, 2014.

[3] K. H. Knuth, "Bayesian source separation and localization," in *SPIE: Bayesian Inference for Inverse Problems*, 1998, pp. 147–158.

[4] Y. Dorfan, D. Cherkassky, and S. Gannot, "Speaker localization and separation using incremental distributed expectation-maximization," in *IEEE 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1256–1260.

[5] K. Nakadai, H. G. Okuno, H. Kitano, *et al.*, "Real-time sound source localization and separation for robot audition," in *IEEE International Conference on Spoken Language Processing*, 2002, pp. 193–196.

[6] J. Hörnstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots-building audio-motor maps based on the HRTF," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006, pp. 1170–1176.

[7] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer Berlin Heidelberg, 2001, pp. 157–180.

[8] J. P. Dmochowski and J. Benesty, "Steered beamforming approaches for acoustic source localization," *I. Cohen, J. Benesty, and S. Gannot (Eds.), Speech Processing in Modern Communication, Springer*, pp. 307–337, 2010.

[9] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[10] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[11] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1997, pp. 375–378.

[12] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[13] H. Schau and A. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 8, pp. 1223–1225, 1987.

[14] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[15] S. Gannot and T. G. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 1–17, 2006.

[16] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, 2001, pp. 3021–3024.

[17] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 826–836, 2003.

[18] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3291–3304, 2006.

[19] X. Zhong and J. R. Hopgood, "Nonconcurrent multiple speakers tracking based on extended Kalman particle filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 293–296.

[20] A. Levy, S. Gannot, and E. A. Habets, "Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1540–1555, 2011.

[21] M. Beard and S. Arulampalam, "Performance of PHD and CPHD filtering versus jipda for bearings-only multi-target tracking," in *15th International Conference on Information Fusion (FUSION)*. IEEE, 2012, pp. 542–549.

[22] C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely, "Bearing-only acoustic tracking of moving speakers for robot audition," in *IEEE International Conference on Digital Signal Processing (DSP)*, 2015, pp. 1206–1210.

[23] L. Kumar, K. Singhal, and R. M. Hegde, "Robust source localization and tracking using music-group delay spectrum over spherical arrays," in *IEEE 5th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2013, pp. 304–307.

[24] X. Zhong, A. Mohammadi, A. B. Premkumar, and A. Asif, "A distributed particle filtering approach for multiple acoustic source tracking using an acoustic vector sensor network," *Signal Processing*, vol. 108, pp. 589–603, 2015.

[25] Q. Zhang, Z. Chen, and F. Yin, "Speaker tracking based on distributed particle filter in distributed microphone networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 9, pp. 2433–2443, 2017.

[26] Y. Tian, Z. Chen, and F. Yin, "Distributed IMM-unscented Kalman filter for speaker tracking in microphone array networks," *IEEE transactions on audio, speech, and language processing*, vol. 23, no. 10, pp. 1637–1647, 2015.

[27] H. A. Blom and Y. Bar-Shalom, "The interacting multiple model algorithm for systems with markovian switching coefficients," *IEEE transactions on Automatic Control*, vol. 33, no. 8, pp. 780–783, 1988.

[28] D. Zotkin, R. Duraiswami, and L. S. Davis, "Multimodal 3-D tracking and event detection via the particle filter," in *IEEE Workshop on Detection and Recognition of Events in Video*, 2001, pp. 20–27.

[29] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 1, pp. 1154–1164, 2002.

[30] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio–visual active speaker tracking in cluttered indoors environments," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 3, pp. 799–807, 2008.

[31] I. D. Gebru, X. Alameda-Pineda, R. Horaud, and F. Forbes, "Audio-visual speaker localization via weighted clustering," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[32] J. Dmochowski, J. Benesty, and S. Affes, "Linearly constrained minimum variance source localization and spectral estimation," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 8, pp. 1490–1502, 2008.

[33] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 923–933, 2013.

[34] J. R. Jensen, J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "On frequency domain models for TDOA estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 11–15.

[35] J. R. Jensen, J. K. Nielsen, R. Heusdens, and M. G. Christensen, "DOA estimation of audio sources in reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 176–180.

[36] P. Smaragdis and P. Boufounos, "Position and trajectory learning for microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 358–368, 2007.

[37] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International journal of neural systems*, vol. 25, no. 01, p. 1440003, 2015.

[38] N. Bertin, S. Kitić, and R. Gribonval, "Joint estimation of sound source location and boundary impedance with physics-driven cosparse regularization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6340–6344.

[39] D. Salvati, C. Drioli, and G. L. Foresti, "A weighted MVDR beamformer based on SVM learning for sound source localization," *Pattern Recognition Letters*, vol. 84, pp. 15–21, 2016.

[40] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1393–1407, 2016.

[41] ——, "Manifold-based Bayesian inference for semi-supervised source localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6335–6339.

[42] ——, "Semi-supervised source localization on multiple-manifolds with distributed microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1477–1491, 2017.

[43] ——, "Speaker tracking on multiple-manifolds with distributed microphones," in *Interntional Conference on Latent Variable Analysis and Signal Seperation (LVA/ICA)*, 2017, pp. 59–67.

[44] ——, "Relative transfer function modeling for supervised source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[45] R. Talmon, I. Cohen, and S. Gannot, "Supervised source localization using diffusion kernels," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 245–248.

[46] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Study on manifolds of acoustic responses," in *Interntional Conference on Latent Variable Analysis and Signal Seperation (LVA/ICA)*, 2015, pp. 203–210.

[47] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[48] Y. Wang and B. Chaib-Draa, "A KNN based Kalman filter Gaussian process regression." in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

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

[50] N. J. Gordon, D. J. Salmond, and A. F. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," in *IEE Proceedings F - Radar and Signal Processing*, vol. 140, no. 2, 1993, pp. 107–113.

[51] S. J. Julier and J. K. Uhlmann, "New extension of the Kalman filter to nonlinear systems," in *AeroSense: 11th International Symposium on Aerospace/Defense Sensing, Simulation and Control*, 1997, pp. 182–193.

[52] A. Wabnitz, N. Epain, C. Jin, and A. Van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *International Symposium on Room Acoustics*, 2010, pp. 1–6.

Bracha Laufer-Goldshtein received the B.Sc. (summa cum laude) and M.Sc. (cum laude) degrees in electrical engineering in 2013 and 2015, respectively, from Bar-Ilan University, Ramat Gan, Israel, where she is currently working toward the Ph.D. degree with the Speech and Signal Processing Laboratory, Faculty of Engineering. She was awarded the Adams Fellowship by the Israel Academy of Sciences and Humanities for the year 2017-2018. Her research interests include statistical signal processing, speaker localization, array processing, and geometric methods for data analysis.

Ronen Talmon is an Assistant Professor of electrical engineering at the Technion – Israel Institute of Technology, Haifa, Israel. He received the B.A. degree (Cum Laude) in mathematics and computer science from the Open University in 2005, and the Ph.D. degree in electrical engineering from the Technion in 2011.

From 2000 to 2005, he was a software developer and researcher at a technological unit of the Israeli Defense Forces. From 2005 to 2011, he was a Teaching Assistant at the Department of Electrical Engineering, Technion. From 2011 to 2013, he was a Gibbs Assistant Professor at the Mathematics Department, Yale University, New Haven, CT. In 2014, he joined the Department of Electrical Engineering of the Technion.

His research interests are statistical signal processing, analysis and modelling of signals, speech enhancement, biomedical signal processing, applied harmonic analysis, and diffusion geometry.

Dr. Talmon is the recipient of the Irwin and Joan Jacobs Fellowship, the Andrew and Erna Fince Viterbi Fellowship, and the Horev Fellowship.

Sharon Gannot (S'92-M'01-SM'06) received the B.Sc. degree (summa cum laude) from the Technion Israel Institute of Technology, Haifa, Israel, in 1986, and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Tel Aviv, Israel, in 1995 and 2000, respectively, all in electrical engineering. In 2001, he held a Postdoctoral position at the Department of Electrical Engineering, KULeuven, Belgium. From 2002 to 2003, he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. Currently, he is a Full Professor at the Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel, where he is heading the Speech and Signal Processing laboratory and the Signal Processing Track. His research interests include multi-microphone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation, dereverberation, single microphone speech enhancement, and speaker localization and tracking. Prof.. Gannot has served as an Associate Editor of the EURASIP Journal of Advances in Signal Processing in 2003-2012, and as an Editor of several special issues on multi-microphone speech processing of the same journal. He has also served as a Guest Editor of the ELSEVIER Speech Communication and Signal Processing journals. He has served as an Associate Editor of the IEEE Transactions on Audio, Speech, and Language Processing in 2009-2013, and an area chair for the same journal 2013-2017. Currently, he serves as a moderator for arXiv in the field of audio and speech processing. He also serves as a reviewer of many IEEE journals and conferences. He is a member of the Audio and Acoustic Signal Processing technical committee of the IEEE since January 2010. Since January 2017, he serves as the committee chair. He is also a member of the technical and steering committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) since 2005 and was the General Co-chair of IWAENC held at Tel-Aviv, Israel in August 2010. He has served as the General Co-chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in October 2013. He was selected (with colleagues) to present tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013, and EUSIPCO 2013 and a keynote speaker for IWAENC 2012 and LVA/ICA 2017. He received the Bar-Ilan University outstanding lecturer award for 2010 and 2014. Prof. Gannot is also a co-recipient of eight best paper awards.