Dorfan *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:2
https://doi.org/10.1186/s13636-020-00189-x

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

# Forward-backward recursive expectation-maximization for concurrent speaker tracking

Yuval Dorfan, Boaz Schwartz and Sharon Gannot*

**Abstract**

In this paper, a study addressing the task of tracking multiple concurrent speakers in reverberant conditions is presented. Since both past and future observations can contribute to the current location estimate, we propose a forward-backward approach, which improves tracking accuracy by introducing near-future data to the estimator, in the cost of an additional short latency. Unlike classical target tracking, we apply a non-Bayesian approach, which does not make assumptions with respect to the target trajectories, except for assuming a realistic change in the parameters due to natural behaviour. The proposed method is based on the recursive expectation-maximization (REM) approach. The new method is dubbed forward-backward recursive expectation-maximization (FB-REM). The performance is demonstrated using an experimental study, where the tested scenarios involve both simulated and recorded signals, with typical reverberation levels and multiple moving sources. It is shown that the proposed algorithm outperforms the regular common causal (REM).

**Keywords:** Sound source tracking, Recursive expectation-maximization, Microphone arrays, Simultaneous speakers, W-disjoint orthogonality, Forward-backward

## 1 Introduction

The task of multiple target tracking (or dynamic localization) has significant importance in civil, military and surveillance applications such as improving beamforming accuracy in speech enhancement applications, e.g. speech separation, indoor robotic assistance, and automatic steering of cameras [1–4]. Although many state-of-the-art approaches to speech separation are based on deep learning, model-based methods are preferable in cases where training data is not available. As for computer-vision-based methods for speaker tracking, in several cases, cameras are not allowed due to power consumption or privacy constraints. In addition, these methods are not suitable in cases where there is no direct line-of-sight between the sensor and the speaker being tracked. Classical tracking addresses only a subset of possible trajectories, since it usually makes assumptions, regarding e.g. the target velocity, acceleration, and jerk. In case of noncontinuous signals like speech, this might lead to tracking loss.

There are scenarios where the sensor arrays are moving [5], but usually in a room environment, the arrays are assumed to be static and only the sources are moving and should be tracked. Some tracking algorithms are based on direction of arrival (DOA) estimation. The multiple signal classification (MUSIC) algorithm [6] applies a subspace method that was later adapted to the challenges of speech processing in [7]. The steered response power with phase transform (SRP-PHAT) algorithm [8] uses generalizations of cross-correlation methods for DOA estimation. This algorithm can be applied to both 2D or 3D localization and tracking problems. Although several features were used in the literature for speaker localization and tracking,

*Correspondence: sharon.gannot@biu.ac.il
Sharon Gannot is a member of EURASIP.
Faculty of Engineering, Bar-Ilan University, 5290002 Ramat-Gan, Israel

the most commonly used features are the sub-band time difference of arrivals (TDOAs) [9] or DOAs [10].

Supervised learning methods can also be used for this task. Deep learning methods can be trained to find the DOA in different acoustic conditions. Deep learning methods have recently been proposed for sound source localization. In [11, 12], simple feed-forward deep neural networks (DNNs) were trained using generalized cross-correlation (GCC)-based audio features, demonstrating improved performance as compared with classical approaches. Yet, this method is mainly designed to deal with a single sound source at a time. In [13], the authors trained a DNN for multi-speaker DOA estimation. In [14, 15], time-domain features were used, demonstrating performance improvements in highly reverberant enclosures.

The main drawbacks of the DNN approaches for tracking applications are the need for large, usually labelled, training sets and the high sensitivity to mismatch between train and test conditions. When we are interested in ad hoc distributed networks, training is not always possible and changes of array constellation between train and test can often occur. Various tracking algorithms for distributed microphone arrays were proposed in [16–22]. In [23–26], the outdoor case is emphasized, where sensor noise is usually dominant, unlike indoor environment that is dominated by reverberation. Well-known Bayesian approaches were also applied for tracking, such as probability hypothesis density (PHD) [27, 28], particle filters, and other statistical based methods [29–33]. Several tracking and localization algorithms were tested as part of the LOCalization And TrAcking (LOCATA) challenge [34, 35].

The classical Bayesian approach is not optimal for the task of tracking concurrent speakers within a room, because subjects tend to naturally and arbitrarily move without an organized route and with many pauses. For this application, the non-Bayesian maximum likelihood (ML) approach might be more suitable, since it assumes that the speakers' trajectories are deterministic unknown parameters, bearing no statistical model. A well-known procedure for inferring the ML estimate is the expectation-maximization (EM) algorithm [36] that iteratively increases the likelihood function.

For the task of online speaker tracking, a recursive version of the EM is required, since it facilitates tracking of time-varying parameters, and enables online estimates with relatively low computational and memory loads. The first recursive version of the EM algorithm was suggested by Titterington [37], where a Newton-based method minimizes the Kullback-Leibler divergence (KLD) between the actual and parametric likelihood function, assuming that the observations are independent and identically distributed (i.i.d.). An almost surely convergence proof of Titterington recursive expectation-maximization (TREM) algorithm was given by Wang and Zhao in [38, 39], based on the results of Delyon [40]. A stochastic approximation version for the EM algorithm was proposed by Delyon et al. in [41], and its convergence was proven therein. A further study of the recursive expectation-maximization (REM) approach appeared in [42] for the problem of DOA estimation, using TREM and another recursive algorithm suggested by the authors. They showed that both algorithms converge with probability one (w.p.1.) to a stationary point of the likelihood function.

A different approach of REM was proposed by Cappé and Moulines [43], in which the model parameters and the hidden signal are estimated simultaneously. In the E-step, the sufficient statistic for the parameters is recursively updated using the latest observation. In the M-step, the parameters are optimized using the latest statistic approximation. It is shown in [43] that this series of parameters estimates converges to local minima of the KLD in the case of independent observations.

REM-based algorithms for speaker tracking in noisy and reverberant environments were presented in [44], where both Titterington's REM [37] and Cappé-Moulines' REM [43] were applied to the problem. The W-disjoint orthogonality (WDO) property was assumed to hold, as commonly used by other algorithms, in order to improve the robustness and to facilitate concurrent speaker tracking [9, 45–48].

Spatially distributed microphone nodes were used in [44], but computations were carried out in a central processing unit. In many applications/scenarios, since the communication bandwidth (BW) is limited, *distributed computation* is beneficial, where each node executes part of the computations and transmits its result rather than the entire observed data [49]. The ring-based algorithm and its modifications are based on the incremental expectation-maximization (IEM) principle suggested in [50]. The recursive distributed expectation-maximization (RDEM) method was recently proposed for distributed source localization [51], using Titterington's REM [37]. In this paper, we use the same topology from [51].

Since the methods in [44] and [51] apply only causal recursion, the localization accuracy in silent or noisy time-segments often deteriorates. Exploiting both past and future observations may improve the estimation accuracy of the current state, especially for non-stationary signals. The estimation pertaining to the past data is usually referred to as *forward* filtering, while the state estimation pertaining to the future data is referred to as *backward* filtering, which runs backwards. In online applications, the usage of future observations imposes latency to the overall system and should therefore be restricted.

A bi-directional version of recursion estimation has been presented in [52] for video signals, where the data is processed off-line in order to use the future samples

in addition to the past samples. Some of the Bayesian approaches deal with specific speech processing applications with techniques developed originally for communication applications. A maximum a posteriori (MAP) approach that exploits the Viterbi algorithm was presented in [53], where a forward-backward recursion has been used to address pauses of the speech signal.

In this paper, we propose a new tracking mechanism and use it to modify the recursive distributed expectation-maximization (RDEM) [51], resulting in the tracking forward-backward recursive expectation-maximization (TFB-REM), which is a non-Bayesian algorithm. The new algorithm allows to configure the latency according to the real-time constraints of the system and obtains an improved performance.

The rest of the paper is organized as follows. Formulation of the tracking problem and its probabilistic model is described in Section 2. Then, in Section 3, the forward-backward recursive expectation-maximization (FB-REM) is introduced and applied to the tracking problem of multiple speakers in Section 4. Experimental results are given in Section 5, and conclusions are drawn in Section 6.

## 2 Speaker tracking problem formulation

One of the major applications of the recursive algorithms is target tracking. In this work, focus is given to the tracking of concurrent speakers, a more challenging task due to the non-stationary and intermittent nature of speech signals. The problem is formulated in the time-frequency domain and in the spatial domain. Let $b = 1, \ldots, B$ denote the frequency bin index, and $S$ the number of acoustic signals captured by $M$ microphones (an even number), organized in $M/2$ independent pairs.

The signal captured by the $i$th microphone, $i = 1, 2$ of the $m$th pair, $m = 1, \ldots, M/2$, is given by

$$z_m^i(t,b) = \sum_{s=1}^{S} a_{sm}^i(t,b) v_s(t,b) + n_m^i(t,b), \tag{1}$$

where $s = 1, \ldots, S$ is the source index, $v_s(t,b)$ denotes the $s$th source signal, $n_m^i(t,b)$ denotes an additive noise, and $a_{sm}^i(t,b)$ denotes the time-variant room impulse response (RIR).

For each time-frequency (T-F) bin, the pair-wise TDOAs, denoted by $\tau_{m,b}(t)$, are calculated from the cross-correlation. Under the assumption of speech sparsity [54], each time-frequency bin vector is dominated by a single source, namely that the WDO assumption is satisfied. This implies that for each T-F bin, the summation in (1) is dominated by only a single source.

Let $\mathbf{p} = [x, y, z]$ be the 3D Cartesian coordinates within a given indoor space and further let $\mathcal{P}$ be the set of all candidate speakers' positions, without assuming any prior knowledge about the number of the sources and their dynamics. Multiple positions may receive high values according to the number of active speakers. The grid of candidate positions is defined similarly to [51]. The noiseless TDOAs associated with each grid point can be calculated in advance from geometrical considerations:

$$\tilde{\tau}_{m,b}(\mathbf{p}) \triangleq \frac{||\mathbf{p} - \mathbf{p}_m^1|| - ||\mathbf{p} - \mathbf{p}_m^2||}{c}\ ;\ \forall \mathbf{p} \in \mathcal{P}, \tag{2}$$

where $\mathbf{p}_m^1$ and $\mathbf{p}_m^2$ are the locations of the microphones assumed to be perfectly known, $||\cdot||$ denotes the Euclidean norm, and $c$ is the sound velocity.

We then attribute a mixture of Gaussians (MoG) statistical model to the TDOAs:

$$\tau_{m,b}(t) \sim \sum_{\mathbf{p}} w_{\mathbf{p}}(t) \mathcal{N}\left(\tau_{m,b}(t); \tilde{\tau}_{m,b}(\mathbf{p}), \sigma^2\right), \tag{3}$$

where $\sigma^2$ is the Gaussians' variance, which is assumed to be a known, constant, parameter.

The weights $w_{\mathbf{p}}(t)$ are unknown parameters, designating the probability of a speaker to be located at position $\mathbf{p}$ at time $t$ that should satisfy the following constraints:

$$\sum_{\mathbf{p}\in\mathcal{P}} w_{\mathbf{p}}(t) = 1\ ;\ 0 \le w_{\mathbf{p}}(t) \le 1. \tag{4}$$

The set of unknown parameters to be found by the EM algorithm consists of the Gaussian weights, $\boldsymbol{w}(t) = \mathrm{vec}_{\mathbf{p}}\left(w_{\mathbf{p}}(t)\right)$, since the mean of each Gaussian is calculated in advance over the grid of positions and its variance is assumed to be known. This set of unknown parameters can be thought of as a set of hypotheses for potential speakers' positions. Each grid point with sufficiently high weight will be marked as a potential speaker's position.

The probability density function (p.d.f.) of all augmented measurements at each node at time $t$ is given by

$$f\left(\boldsymbol{\tau}_m(t)\ ;\ \boldsymbol{w}(t)\right) = \prod_b \sum_{\mathbf{p}\in\mathcal{P}} w_{\mathbf{p}}(t) \cdot \mathcal{N}\left(\tau_{m,b}(t)\ ;\tilde{\tau}_{m,b}(\mathbf{p}), \sigma^2\right). \tag{5}$$

In the following sections, we propose a forward-backward method for the estimation of this model's parameters, $\boldsymbol{w}(t)$. First, in Section 3, we derive the general scheme of the forward-backward version of the REM. Then, in Section 4, the general scheme is applied to the above model for speaker tracking.

## 3 FB-REM—general derivation

We begin the derivation in defining the criterion for online parameter estimation and present the general algorithm proposed in this paper, namely the FB-REM scheme. The FB-REM approach for parameter estimation is, similarly to the more general REM approach, a non-Bayesian method that does not rely on a statistical model for the parameters. In this section, we use $\boldsymbol{\theta}$ for the set of

parameters. In Section 3.1, we discuss the classical REM, and in Section 3.2 the FB-REM.

### 3.1 REM

The REM algorithm is first derived for i.i.d. observations, denoted by $\boldsymbol{\tau}(t)$. The true, unknown p.d.f. of the observations is denoted by $h(\boldsymbol{\tau})$, and the parametric p.d.f. is $f(\boldsymbol{\tau};\boldsymbol{\theta})$. A common criterion for online parameter estimation is stated in terms of the KLD, defined as

$$k(\boldsymbol{\theta}) = \mathbb{E}\left\{\log h(\boldsymbol{\tau}) - \log f(\boldsymbol{\tau};\boldsymbol{\theta})\right\}\ , \tag{6}$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation taken with respect to (w.r.t.) $h(\boldsymbol{\tau})$, the time index $t$ does not appear in $\boldsymbol{\tau}(t)$ since time dependency is cancelled out by the expectation operation and due to the i.i.d. nature of the observations. In our case, we write the optimization criterion as a minimization of the KLD:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{argmin}\, k(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{argmax}\ \left\{\mathbb{E}\{\log f(\boldsymbol{\tau};\boldsymbol{\theta})\}\right\}\ . \tag{7}$$

It is noteworthy that $\boldsymbol{\theta}^*$ and the ML (batch) estimation of $\boldsymbol{\theta}$ are asymptotically equivalent, i.e. they converge for growing number of observations. The KLD criterion is commonly used in online procedures [43], since it fits dynamic cases, where the parameters are time-varying.

As discussed in Section 1, the EM is a common approach for maximizing the likelihood in problems involving hidden data. The hidden data at time $t$ is defined as $\mathbf{y}(t)$, and the respective *complete data* p.d.f. is assumed to be i.i.d. and denoted by $f\left(\boldsymbol{\tau}(t),\mathbf{y}(t);\boldsymbol{\theta}\right)$.

The specific definition of $\mathbf{y}(t)$ for our tracking problem is given in Section 4.1. The EM can also be applied recursively, e.g. as presented by Titterington [37]. The TREM maximizes the KLD using the recursion

$$\widehat{\boldsymbol{\theta}}^{\mathrm{Ti}}_{t+1} = \widehat{\boldsymbol{\theta}}^{\mathrm{Ti}}_{t} + \tfrac{1}{t}\cdot\mathbf{I}_C^{-1}\left(\widehat{\boldsymbol{\theta}}^{\mathrm{Ti}}_{t}\right)\cdot\boldsymbol{s}\left(\boldsymbol{\tau}(t);\widehat{\boldsymbol{\theta}}^{\mathrm{Ti}}_{t}\right)\ , \tag{8}$$

where $\widehat{\boldsymbol{\theta}}^{\mathrm{Ti}}_{t}$ is the previous estimate of $\boldsymbol{\theta}$, $\boldsymbol{s}\big(\boldsymbol{\tau}(t);\boldsymbol{\theta}\big)$ denotes the *scoring* function, and $\mathbf{I}_C(\boldsymbol{\theta})$ denotes the Fisher information matrix (FIM) of the complete data,

$$\boldsymbol{s}(\boldsymbol{\tau};\boldsymbol{\theta}) = \mathbb{E}_C\left\{\nabla_{\boldsymbol{\theta}}\log f\left(\boldsymbol{\tau},\mathbf{y};\boldsymbol{\theta}\right)|\boldsymbol{\tau};\boldsymbol{\theta}\right\}\ , \tag{9}$$

$$\mathbf{I}_C(\boldsymbol{\theta}) = -\mathbb{E}_C\left\{\nabla_{\boldsymbol{\theta}^2}\log f(\boldsymbol{\tau},\mathbf{y};\boldsymbol{\theta})\right\}\ , \tag{10}$$

where $\mathbb{E}_C$ denotes the expectation taken w.r.t. the complete data p.d.f $f(\boldsymbol{\tau},\mathbf{y};\boldsymbol{\theta})$. The computation of Eqs. (9) and (10) requires the estimation of the hidden data, which is derived in Section 4.

It was shown in [38] that under certain regularity conditions, the TREM converges w.p.1. to a stationary point of the KLD. However, a w.p.1. convergence is impossible in cases of time-varying parameters, and we therefore substitute $t^{-1}$ in (8) by a constant smoothing coefficient $(1-\gamma_F)$, facilitating tracking capabilities for dynamic cases. Denote by $\widehat{\boldsymbol{\theta}}^{\mathrm{F}}_{t}$ the parameter estimation at time $t$, obtained by the following forward recursion,

$$\begin{aligned}&\widehat{\boldsymbol{\theta}}^{\mathrm{F}}_{t+1} = \gamma_F\cdot\widehat{\boldsymbol{\theta}}^{\mathrm{F}}_{t} + (1-\gamma_F)\cdot\mathbf{I}_C^{-1}\left(\widehat{\boldsymbol{\theta}}^{\mathrm{F}}_{t}\right)\cdot\boldsymbol{s}\left(\boldsymbol{\tau}(t);\widehat{\boldsymbol{\theta}}^{\mathrm{F}}_{t}\right),\\ &\quad 0<\gamma_F<1\ .\end{aligned} \tag{11}$$

Note that unlike (8), we also normalized $\widehat{\boldsymbol{\theta}}^{\mathrm{F}}_{t}$ by $\gamma_F$, which is equivalent to a constant attenuation and does not affect the asymptotic behaviour of the algorithm. The procedure (11) also converges to a stationary point of the KLD, but in a weak sense.

Further note that when the parameters are time-varying, the observations are no longer identically distributed, and asymptotic convergence is not relevant anymore. For an extensive theoretical examination of this recursive scheme for parameter estimation, please refer to [55] and Chapter 8 in [56]. However, an intuitive yet accurate description of the convergence can be given as follows. For higher values of $\gamma_F$, the update rate of the algorithm is slower, meaning that the estimation is biased by the past values of the parameter, but the estimation variance is lower, due to the longer averaging window. For lower $\gamma_F$ values, the convergence speed is higher, which means a lower bias, but this comes at the expense of higher variance due to the shorter averaging window.

In the following section, we propose a method that improves the performance of (11) at the expense of higher latency. This is done by utilizing future observations in the estimation procedure and may reduce both the bias and the variance of the estimator. Reduction of the bias induced by past information is very important in the non-stationary case of moving speakers.

### 3.2 The proposed FB-REM approach

To use the near-future observations, we propose the FB-REM algorithm, defined by

$$\widehat{\boldsymbol{\theta}}^{\mathrm{FB}}_{t+1} = \alpha_{FB}\cdot\widehat{\boldsymbol{\theta}}^{\mathrm{F}}_{t+1} + (1-\alpha_{FB})\cdot\widehat{\boldsymbol{\theta}}^{\mathrm{B}}_{t+1}\ , \tag{12}$$

where $\widehat{\boldsymbol{\theta}}^{\mathrm{F}}_{t}$ was defined in (11), and $0\le\alpha_{FB}\le 1$ is a weighting factor of the past and the future terms. The backward estimator $\widehat{\boldsymbol{\theta}}^{\mathrm{B}}_{t}$ is calculated by the backward recursion

$$\begin{aligned}&\widehat{\boldsymbol{\theta}}^{\mathrm{B}}_{k} = \gamma_B\cdot\widehat{\boldsymbol{\theta}}^{\mathrm{B}}_{k+1} + (1-\gamma_B)\cdot\mathbf{I}_C^{-1}\left(\widehat{\boldsymbol{\theta}}^{\mathrm{B}}_{k+1}\right)\cdot\boldsymbol{s}\left(\boldsymbol{\tau}_k;\widehat{\boldsymbol{\theta}}^{\mathrm{B}}_{k+1}\right);\\ &\quad k=t+D,\ldots,t+1,\ \ 0<\gamma_B<1\ ,\end{aligned} \tag{13}$$

where $D$ is the number of future samples used for the current estimate.

In a Bayesian framework, there would have been an optimal choice of the smoothing factors $\gamma_F$, $\gamma_B$, and $\alpha_{FB}$, obtained by a Bayesian statistical model. The gain of the Kalman smoother [57] and the coefficient used in Viterbi algorithm [58] are determined this way. However, since we intentionally adopted a non-Bayesian approach, the

values of $\gamma_F$, $\gamma_B$, and $\alpha_{FB}$ are determined according to the required dynamics of the algorithm, and the nature of the stochastic processes. We previously discussed how $\gamma_F$ trades-off the update speed versus the accuracy of the algorithm. Similarly, high $\gamma_B$ reduces the variance of the algorithm in the expense of higher bias. Finally, $\alpha_{FB}$ determines the weight of the past and the future observations on the current estimate.

In practice, $\alpha_{FB}$ is mainly determined by the number of the future observations that are actually used to update the estimation, which in turn is determined by the latency constraints of the application. Rewriting (13) as

$$\widehat{\boldsymbol{\theta}}^{\mathrm{B}}_{t+1} = (1-\gamma_B)\cdot \sum_{k=t+1}^{t+D} {\gamma_B}^{k-t}\cdot \mathbf{I}_C^{-1}\left(\widehat{\boldsymbol{\theta}}^{\mathrm{B}}_k\right)\cdot \boldsymbol{s}\left(\boldsymbol{\tau}_k;\widehat{\boldsymbol{\theta}}^{\mathrm{B}}_k\right). \quad (14)$$

It can be seen that for every value of $\gamma_B$ and every computing precision requirement, there exists an integer $D_{\max}$ such that ${\gamma_B}^{D_{\max}} \approx 0$. If $D = D_{\max}$ is chosen, the backward recursion is equivalent to an infinite backward recursion, similarly to the forward recursion for large enough $t$ values. However, in practical applications, using $D_{\max}$ future observations might introduce unacceptable latency, and a lower value of $D$ should be chosen instead. This choice deteriorates the accuracy of the backward recursion (13), which should be compensated by increasing the value of $\alpha_{FB}$.

In the next section, the algorithm (12) will be applied to the problem of multiple speaker tracking.

## 4 Tracking forward-backward REM (TFB-REM) for concurrent dynamic speakers

The concept developed above is applied to the multiple speaker tracking problem. We first briefly mention the tracking forward-recursive expectation-maximization (TF-REM) for the forward direction, as developed in [51] for localization using only past and present samples. Then, the TF-REM algorithm is extended by adding the backward recursion using the FB-REM that was developed in the previous section. The combination of these two directions is dubbed tracking forward-backward recursive expectation-maximization (TFB-REM). The rest of this section is organized as follows. In Section 4.1, we define the hidden data model we use in the tracking problem, and in Sections 4.2 and 4.3, the REM and the forward-backward REM algorithms are applied to this model, respectively.

### 4.1 Hidden data statistical model

The TF-REM was already presented in [51] and is briefly repeated here, since it is the basis for developing the TFB-REM in the next section.

First, we present the hidden data model that complements the observation p.d.f. in (5). Following [51], we let $y_m(t,b,\mathbf{p})$ be an indicator random variable that equals to one if a speaker located at $\mathbf{p}$ is active in the observed variable $\tau_{m,b}(t)$ and zero otherwise. In other words, $y_m(t,b,\mathbf{p}) = 1$ means that a speaker in $\mathbf{p}$ is present in the $(t,b)$ bin of the $m$th node. Intuitively, the tracking challenge becomes much simpler to solve given this additional information. This formulation also allows a distributed implementation of the algorithm, which is useful in many cases of ad hoc networks. Each pair of microphones may describe an electronic device with independent processing unit and communication capability.

The conditional p.d.f. of the observations, assuming independence between nodes is

$$\begin{aligned} f\left(\boldsymbol{\tau}(t)|\mathbf{y}(t);\boldsymbol{w}(t)\right) &= \prod_m f(\boldsymbol{\tau}_m(t)|\mathbf{y}_m(t);\boldsymbol{w}(t)) \\ &= \prod_{m,b}\sum_{\mathbf{p}\in\mathcal{P}} y_m\left(t,b,\mathbf{p}\right)\cdot\mathcal{N}\left(\tau_{m,b}(t);\tilde{\tau}_{m,b}(\mathbf{p}),\sigma^2\right), \end{aligned} \quad (15)$$

where $\mathbf{y}(t) = \mathrm{vec}_{m,b,\mathbf{p}}\left(y_m(t,b,\mathbf{p})\right)$ is the hidden data set. Thus, the *complete data* p.d.f. is given by

$$\begin{aligned} f(\boldsymbol{\tau}(t),\mathbf{y}(t);\boldsymbol{w}(t)) = \prod_{m,b}\sum_{\mathbf{p}\in\mathcal{P}} w_{\mathbf{p}}(t) y_m(t,b,\mathbf{p}) \\ \cdot\mathcal{N}\left(\tau_{m,b}(t);\tilde{\tau}_{m,b}(\mathbf{p}),\sigma^2\right). \end{aligned} \quad (16)$$

In Section 4.2, $y_m(t,b,\mathbf{p})$ is estimated by the forward recursion, and in Section 4.3, its backward estimator is given. These estimates are denoted by $\hat{y}^{(\mathrm{F})}_m(t,b,\mathbf{p})$ and $\hat{y}^{(\mathrm{B})}_m(t,b,\mathbf{p})$, respectively. Correspondingly, the forward and backward estimates of $w_{\mathbf{p}}(t)$ will be denoted by $\hat{w}_{\mathrm{F},\mathbf{p}}(t)$ and $\hat{w}_{\mathrm{B},\mathbf{p}}(t)$, respectively.

### 4.2 RDEM applied in the forward direction

In [44] and [51], the TF-REM was derived for the general algorithm in (11) in detail, and only the resulting formulae are given in this section. In the E-Step, estimation of the hidden data is given by

$$\hat{y}^{(\mathrm{F})}_m(t,b,\mathbf{p}) \triangleq \frac{\hat{w}_{\mathrm{F},\mathbf{p}}(t-1)\cdot\mathcal{N}\left(\tau_{m,b}(t);\tilde{\tau}_{m,b}(\mathbf{p}),\sigma^2\right)}{\sum_{\tilde{\mathbf{p}}\in\mathcal{P}}\hat{w}_{\mathrm{F},\tilde{\mathbf{p}}}(t-1)\cdot\mathcal{N}\left(\tau_{m,b}(t);\tilde{\tau}_{m,b}(\tilde{\mathbf{p}}),\sigma^2\right)}. \quad (17)$$

Now, define the aggregation of the hidden variables along the frequency axis as:

$$\hat{y}^{(\mathrm{F})}_m(t,\mathbf{p}) \triangleq \frac{1}{B}\sum_{b=1}^{B}\hat{y}^{(\mathrm{F})}_m(t,b,\mathbf{p}). \quad (18)$$

The results of the E-step are utilized for weight estimation per position:

$$\hat{w}_{\mathrm{F},\mathbf{p}}(t) \triangleq \frac{1}{M/2}\sum_{m=1}^{M/2}\hat{y}^{(\mathrm{F})}_m(t,\mathbf{p}). \quad (19)$$

**Algorithm 1:** Acoustic source TF-REM algorithm.

1) **Initialize** $\hat{\boldsymbol{w}}_{\mathrm{F},R}(0) = 1/|\mathcal{P}|$
2) For each time-instant, $t$ run:
**E-step**
Calculate simultaneously and locally $\bar{y}_m^{(\mathrm{F})}(t, \mathbf{p})$
$\forall m = 1, \ldots, M/2$
**M-step**
Aggregate results and calculate $\hat{\boldsymbol{w}}_{\mathrm{F},R}(t)$ using (19)
3) Find sources location: Apply a threshold to $\hat{\boldsymbol{w}}_{\mathrm{F},R}(t)$
4) Go back to 2) for next time point.

As was shown in [44] and [51], the M-Step reduces to a compact recursive equation:

$$\hat{\boldsymbol{w}}_{\mathrm{F},R}(t) = \hat{\boldsymbol{w}}_{\mathrm{F},R}(t-1) + \gamma_{\mathrm{F}} \left( \hat{\boldsymbol{w}}_{\mathrm{F}}(t) - \hat{\boldsymbol{w}}_{\mathrm{F},R}(t-1) \right), \tag{20}$$

where $\hat{\boldsymbol{w}}_F(t)$ is a vector consisting of the instantaneous version of the parameters,

$$\hat{\boldsymbol{w}}_{\mathrm{F}}(t) = \mathrm{vec}_{\mathbf{p}} \left( \hat{w}_{\mathrm{F},\mathbf{p}}(t) \right). \tag{21}$$

The TF-REM procedure is summarized in Algorithm 1. Note that the local hidden set facilitates a wide range of network topology alternatives. These distributed computation aspects are left for future study, as the main contribution of this manuscript is independent of the network topology.

### 4.3 TFB-REM for multiple speakers

In order to develop the TFB-REM, we first derive the backward recursion. Its derivation follows the steps of the TF-REM replacing the notation F with B and the time index $(t-1)$ with $(t+1)$ in Eqs. (17)–(21) above. In the backward recursion, the time samples are processed in the anti-causal direction, and the following formula is obtained.

$$\hat{\boldsymbol{w}}_{\mathrm{B},R}(t) = \hat{\boldsymbol{w}}_{\mathrm{B},R}(t+1) + \gamma_{\mathrm{B}} \left( \hat{\boldsymbol{w}}_{\mathrm{B}}(t) - \hat{\boldsymbol{w}}_{\mathrm{B},R}(t+1) \right) \quad t \in \{t+D-1,\ t+D-2,\ \ldots,\ t\} \tag{22}$$

Due to this anti-causal process, we have to choose a processing delay, defined above as $D$. The algorithm tracks the speakers along time using relevant past and future samples according to this processing delay and the speaker dynamics.

Finally, the FB-REM approach in (12) for the problem of speaker tracking can be written as,

$$\hat{\boldsymbol{w}}_{\mathrm{FB},R}(t) = \alpha_{FB} \cdot \hat{\boldsymbol{w}}_{\mathrm{F},R}(t) + (1 - \alpha_{FB}) \cdot \hat{\boldsymbol{w}}_{\mathrm{B},R}(t). \tag{23}$$

This is a weighted superposition of two separate localization maps created independently by the two RDEM processes that build jointly a single localization map for each relevant time point. Note that when $\alpha_{FB} = 1$, there is no use of future data and we obtain the TF-REM derived in the previous subsection.

**Algorithm 2:** Acoustic source TFB-REM algorithm.

1) **Initialize** $\hat{\boldsymbol{w}}_{\mathrm{F},R}(0) = 1/|\mathcal{P}|$
2) For each time-instant, $t$ run:
a) **Forward estimation:**
**E-step**
Calculate simultaneously and locally $\bar{y}_m^{(\mathrm{F})}(t, \mathbf{p})$
$\forall m = 1, \ldots, M/2$
**M-step**
Aggregate results and calculate $\hat{\boldsymbol{w}}_{\mathrm{F},R}(t)$ using (19)
b) **Backward estimation:**
**Initialize** $\hat{\boldsymbol{w}}_{\mathrm{B},R}(t+D) = 1/|\mathcal{P}|$
**for** $t' = t+D-1$ **to** $t$ **do**
  **E-step**
  Calculate simultaneously and locally $\bar{y}_m^{(\mathrm{B})}\left(t', \mathbf{p}\right)$
  $\forall m = 1, \ldots, M/2$
  **M-step**
  Aggregate results and calculate $\hat{\boldsymbol{w}}_{\mathrm{B},R}\left(t'\right)$ using (22)
**end**
c) Calculate $\hat{\boldsymbol{w}}_{\mathrm{FB},R}(t)$ according to (23)
3) Find sources location: Apply a threshold to $\hat{\boldsymbol{w}}_{\mathrm{FB},R}(t)$
4) Set $\hat{\boldsymbol{w}}_{\mathrm{F},R}(t) = \hat{\boldsymbol{w}}_{\mathrm{FB},R}(t)$
5) Go back to 2) for next time step.

As in many other EM-based algorithms, initialization is an important and non-trivial task. The TF-REM proceed from one time step to another according to the recursion. The backward RDEM at time $t$ is initialized with a uniform position distribution at the future time $t+D$.

The localization map in (23) consists of soft values representing the probability of an acoustic activity at each specific position in the room. After estimating $\hat{\boldsymbol{w}}_{\mathrm{FB},R}(t)$ at each time step, we apply a threshold for all values to determine the active positions, meaning the number of active speakers and their current positions.

Before switching to the next time point estimation, we set

$$\hat{\boldsymbol{w}}_{\mathrm{F},R}(t) = \hat{\boldsymbol{w}}_{\mathrm{FB},R}(t), \tag{24}$$

which is assumed to be an improved estimate of the current acoustic position map. We then start the entire procedure for both RDEM and their weighted combination described above for the next time step. The entire TFB-REM procedure is summarized in Algorithm 2.

## 5 Results and discussion

This section evaluates the performance of the TFB-REM algorithm and compare it to TF-REM and to a generalization of the SRP-PHAT algorithm [8] for the 2D case [59]. The study comprises four subsections. In Section 5.1,

we discuss the parameters of the algorithms and the considerations taken when tuning them. In Section 5.2, we describe the room setup for simulation and recordings, and Section 5.3 compares the various algorithms using Monte Carlo simulations and then examines the sensitivity to array perturbations. Finally, in Section 5.4, we compare the performance of the algorithms for real-life recordings of moving speakers acquired at the Bar-Ilan acoustic lab.

### 5.1 Parameter choice

There are a few important parameters mentioned above that should be chosen to improve the tracking algorithm performance.

Two important parameters are the TREM smoothing factors $\gamma_F$ and $\gamma_B$. The parameter $\gamma_F$ was already tuned experimentally for the regular RDEM [44]. We tuned both values empirically for our model. We tried values in the range $0.01 - 0.9$ and eventually chose $\gamma_F = 0.015$ and $\gamma_B = 0.04$.

The next parameter to be discussed is the past-future weighting coefficient, $\alpha_{FB}$. For off-line applications, we might expect the past and the future samples to be equally weighted, meaning $\alpha_{FB} = 0.5$, since the relevance of past or future observations to the present observation is similar. However, since we apply the proposed algorithms in an online scenario, the overall latency was restricted, thus rendering the backward recursion less accurate than the forward recursion. In this case, the forward recursion should be weighted more than the backward recursion. The results show that the best choice was 0.65 (we tried values in the range of $0.1 - 0.9$).

An important parameter, which has close relations with the smoothing factors, is the latency, $D$. Assuming slow dynamics for indoor speaker tracking, it is reasonable to set $D$ to approximately 1 s. It is obvious that for online applications, as addressed here, we might set $D$ to a smaller value.

The MoG variance, $\sigma^2$ was set to 2[Samples$^2$]. The unit [Samples] is a function of the sampling frequency, which is set to 16 kHz. In this experimental study, the variance is fixed over time and frequency bin, but we found that it can significantly influence the performance. Therefore, an adaptive mechanism for setting $\sigma^2$ is left for a future study.

The size of a single unit in the localization grid was set to $0.10 \times 0.10$ m$^2$, which was shown in [44] to be sufficient for tracking real speakers that should not be treated as point sources, due to their body volume. Preliminary tests showed that the chosen resolution best fits the trade-off between computational complexity and required accuracy.

The number of frequency bins was also examined experimentally within the range that is reported in other narrow-band approaches [9], and was set to $B = 16$ bins. The choice of the short-time Fourier transform (STFT) frame length is 64 ms, similarly to previous localization and tracking algorithms [44, 49, 51, 60–63].

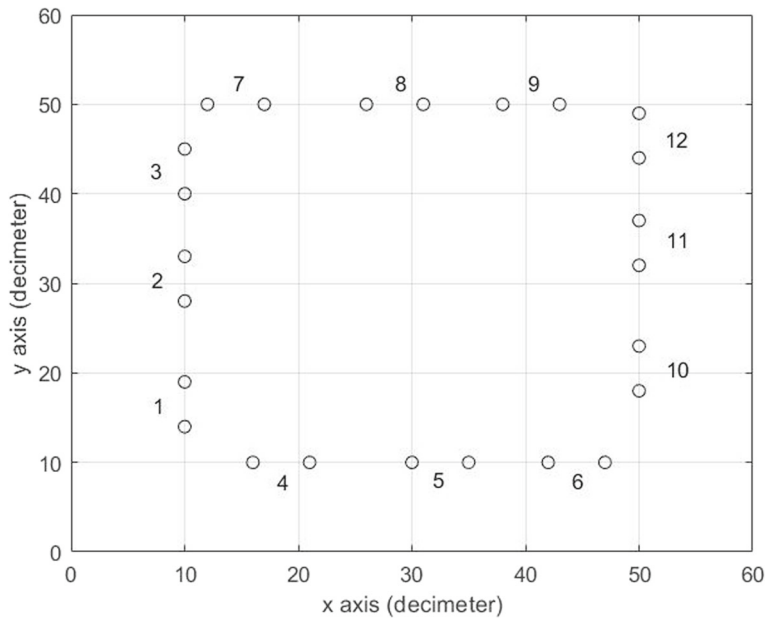


**Fig. 1** Microphone pairs map for the simulated room. The 'o' stands for a microphone positions in the two-dimensional plane. The nodes are numbered 1 : 12 containing two microphones each

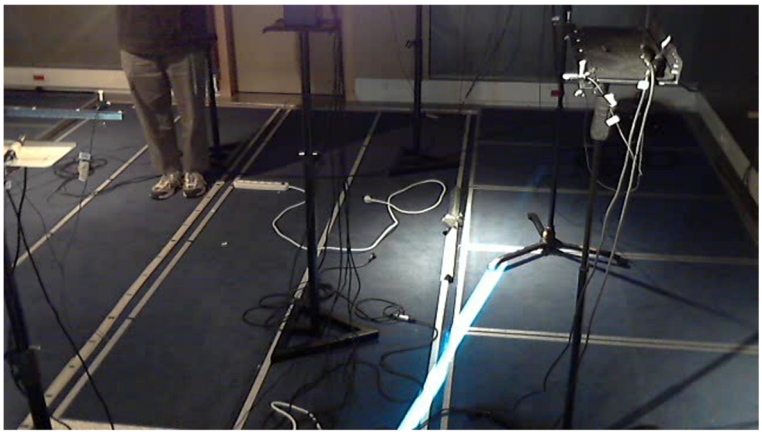

**Fig. 2** Recording room setup with a single speaker and $T_{60} = 250$ ms

### 5.2 Room setup

To evaluate the tracking capabilities of the algorithms, we tested them in both simulated data and recordings of real human speaker setups, as described in the sequel.

The dimensions of the simulated room were $6 \times 6 \times 2.4$ m, with twelve pairs of microphones encompassing the acoustic scene in the room. The number of microphone pairs (or nodes) is very important in the presence of high noise and high reverberation levels. A setup of twelve microphones pairs was chosen following the experimental study in [44]. A map of the simulated room and microphones is given in Fig. 1, where the microphone positions are marked by 'o'-s, and every pair is numbered from 1 to 12. In the examined scenario, all speakers and microphones are positioned at the same height, for simplicity.

Simulation of moving sources was carried out by the image method [64], with an efficient implementation of the RIRs computation in [65]. In this experiment, one, two, or three sources moved along randomly chosen trajectories within the room. The RIRs along these trajectories were sampled every 0.04 s to generate moving sources in reverberant environment.

Recording of real human speakers was carried out as described in [44, 51]. The recordings were carried out in the speech and acoustic lab of Bar-Ilan University. This is a $6 \times 6 \times 2.4$ m room with a reverberation time controlled by 60 interchangeable panels covering the room facets. The measurement equipment includes a RME Hammerfall DSP Digiface sound-card and a RME Octamic (for Microphone Pre Amp and digitization (A/D)). AKG type CK-32 omnidirectional microphones were used. All measurements were carried out with a sampling frequency of 48 kHz and a resolution of 24 bits. The multi-microphone signals were acquired using Matlab©. A snapshot of the room tuned for low reverberation level ($T_{60} = 250$ ms) is shown in Fig. 2.

In the real recordings, only seven pairs of microphones were used. As in the simulated experiment, we focused on a two-dimensional tracking, while the height estimation can be derived as an extension of the algorithms. While we assume a two-dimensional setup with all the microphones positioned at height, 130 cm from the floor, we note that real human speakers are recorded and their height vary a bit.

For controlling the recordings and determining an accurate ground-truth for the tracking algorithms, we had a video camera in the room as well as very precise paths marked on the floor of the room. The speakers were walking along the paths and were recorded with all microphones and video in a synchronized way.

### 5.3 Performance in simulated recordings

Using simulated recordings enables a comprehensive statistical investigation and testing of a large variety of trajectories in the room. Three different scenarios were examined. The first scenario consist of a single speaker and reverberation time $T_{60} = 400$ ms. The two other scenarios consist of either two or three speakers and reverberation time $T_{60} = 120$ ms.

To compare the performance of the different algorithms, we followed the procedure described in [49, 51], updated for a dynamic speaker scenario. We executed 100 Monte Carlo trials and calculated the root mean square error

**Table 1** RMSE for tracking scenarios (100 Monte Carlo trials) for one, two, or three speakers. The error is calculated in meters. Reverberation time for the single-speaker case $T_{60} = 400$ ms and for the two- and three-speaker case $T_{60} = 120$ ms

| Algorithm | 1 speaker [m] | 2 speakers [m] | 3 speakers [m] |
|---|---|---|---|
| SRP-PHAT | 0.32 | 0.42 | 0.54 |
| TF-REM | 0.32 | 0.30 | 0.41 |
| TFB-REM | 0.30 | 0.23 | 0.36 |

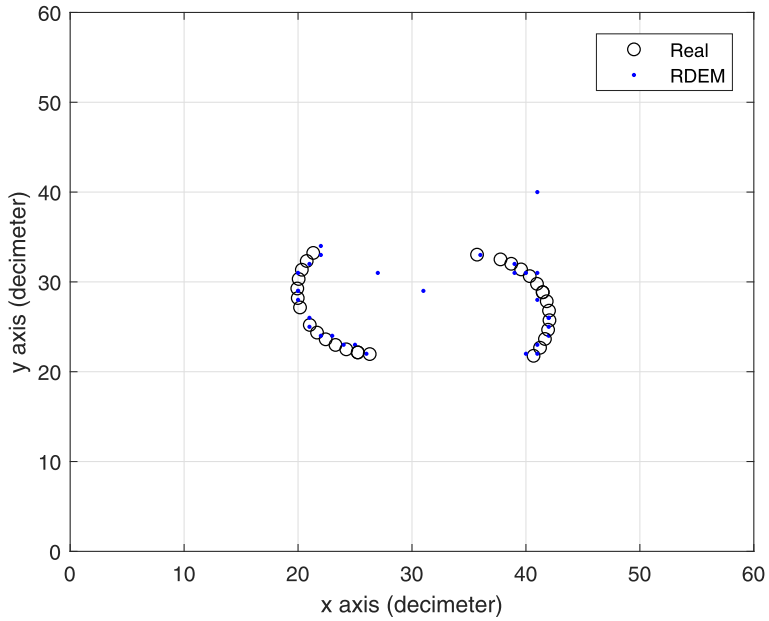


**Fig. 3** Trajectories of two speakers: real and TF-REM

(RMSE), and the average results (using the entire ensemble and along all possible trajectories) are presented in Table 1. The performance of the SRP-PHAT, TF-REM, and TFB-REM are compared for one-, two-, and three-speaker scenarios.

As evident from the table, the reference algorithm SRP-PHAT has the highest RMSE for more than one speaker. The TF-REM has higher RMSE than the TFB-REM, as expected by the incorporation of future data.

The trajectories of one trial of the simulated scenario with two speakers for the TF-REM algorithm is shown in Fig. 3. The real locations are marked with black circles ('o') and the estimates with blue dots ('·'). The same scenario with the TFB-REM algorithm is shown in Fig. 4. The real

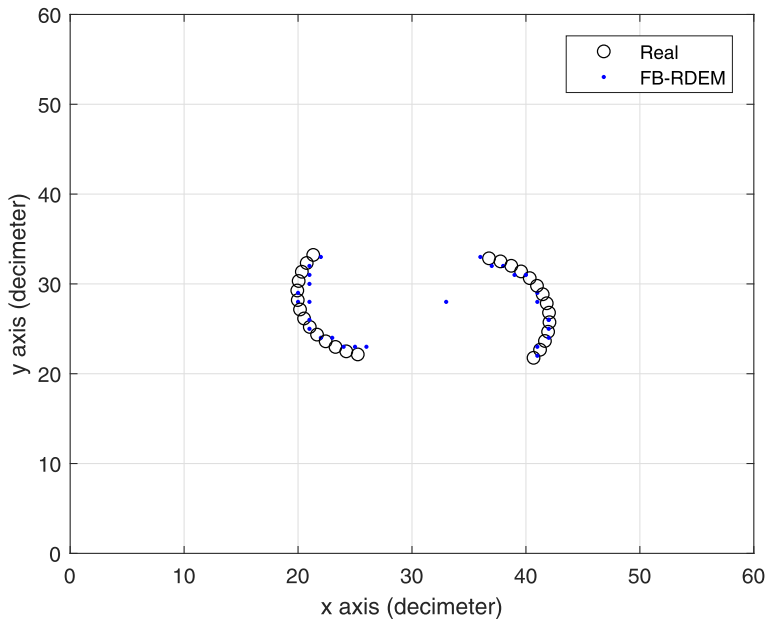


**Fig. 4** Trajectories of two speakers: real and TFB-REM

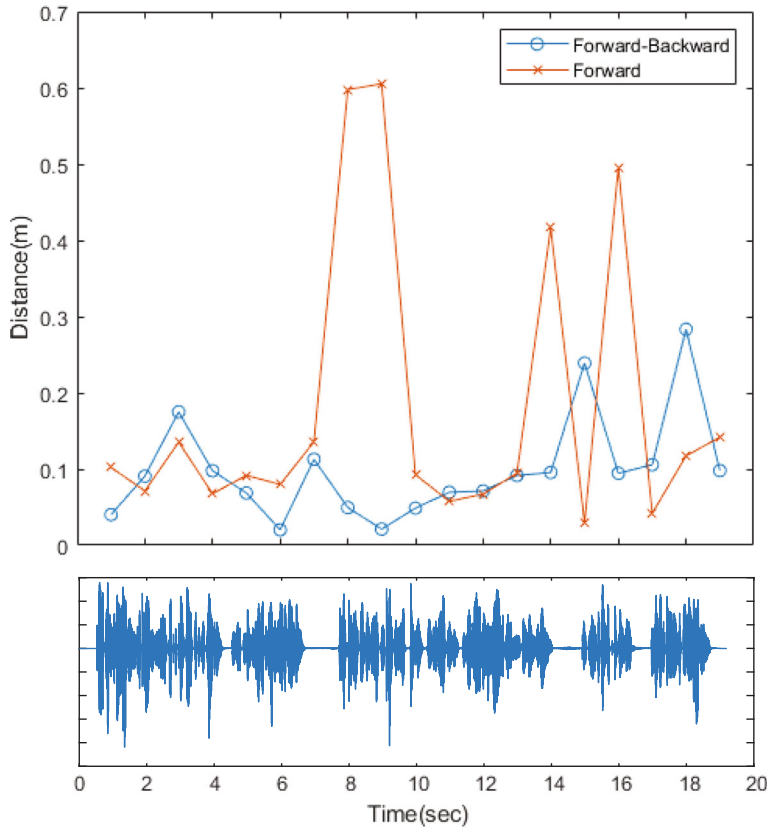


**Fig. 5** Tracking error for two speaker with inactive gaps

trajectories of the speakers are sampled in slightly different time stamps, due to the latency that is introduced to the TFB-REM algorithm.

It can be observed that for both algorithms the majority of estimates are very close to the real locations. The TF-REM produces three outliers, while the TFB-REM has only a single outlier.

A profound advantage of integrating information from various time slots is to bridge gaps of inactive time periods. We examined such a case and computed the tracking error for both algorithms as shown in Fig. 5. It can be observed that the TFB-REM outperforms the TF-REM in this case for the relevant time slots (around 8 s) and later around 14 (smaller differences). The estimation is applied once every second, since dynamic is slow indoor and the complexity of the tracking highly depends on the resolution used.

We have also evaluated the effect of imperfect microphone positions on the performance. The results for one scenario with two speakers is presented in Table 2. The performance of the SRP-PHAT, TF-REM, and proposed TFB-REM are evaluated for a few random misplacement of the microphones. The proposed algorithm outperforms the other two algorithms for standard deviation of up to 50 mm. The highest performance advantage is in the case of perfect array calibration.

### 5.4 Performance for real human recordings

The real recordings include two experiments with human speakers. The first recording was executed with reverberation time $T_{60} = 250$ ms and a single speaker standing for 9 s and then walking on a 2.10-m-long straight line for 33 s. From this point, we focus on the two recursive algorithms in order to illustrate the influence of using future samples. The tracking errors for the two algorithms are shown in Fig. 6. It can be observed that the TFB-REM

**Table 2** RMSE for tracking scenarios for various array position shift (measured in mm). The error is calculated in meters

| Array shift std [mm] | xSRP-PHAT [m] | TF-REM [m] | TFB-REM [m] |
|---|---|---|---|
| 0 | 0.561 | 0.181 | 0.099 |
| 1 | 0.761 | 0.185 | 0.146 |
| 10 | 0.576 | 0.186 | 0.131 |
| 50 | 0.701 | 0.187 | 0.152 |

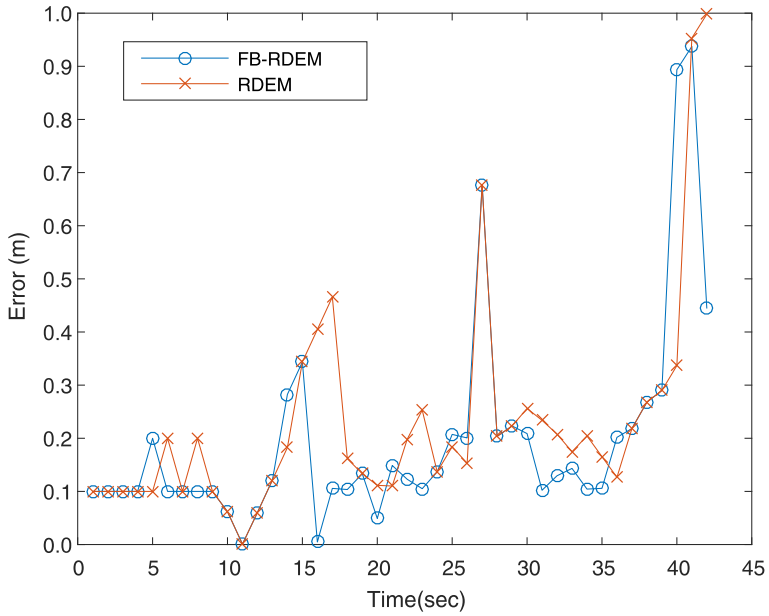


**Fig. 6** Tracking error for a single speaker

slightly outperforms the TF-REM in this case. The RMSE averaged over the trajectory for TFB-REM is 0.20 m and for TF-REM 0.23 m.

The second recording involved two concurrent speakers standing for 2 s and then walking on parallel straight lines towards each other for 21 s. The averaged tracking errors for both algorithms are shown in Fig. 7. It can be observed that the TFB-REM significantly outperforms the TF-REM. The RMSE for the TFB-REM is 0.48 m, and for the TF-REM, it is 0.83 m. The additional speaker naturally degrades the tracking results of both algorithms.

## 6 Conclusions

We have developed an online, non-Bayesian, algorithm for multiple concurrent speaker tracking in reverberant environments. We have first introduced a backward recursive version of the RDEM algorithm. Then, the TFB-REM, a combined forward-backward RDEM algorithm, was developed. The TF-REM and the TFB-REM algorithms are evaluated using both simulated environment and real recordings of walking humans and compared with a baseline method, a variant of the SRP-PHAT method for the 2D case.

We demonstrated that the introduction of the short latency in the TFB-REM indeed improves performance, especially by facilitating the tracking of intermittent sources by bridging over short silence periods. Specifically, we have compared the tracking capabilities of the forward and the forward-backward schemes in single-, two-, and three-speaker scenarios. While the TFB-REM only slightly outperformed the TF-REM in the single-speaker scenario, it was significantly better for two or three concurrent speaker scenarios.

Unlike other approaches (mainly Bayesian approaches), the recursive algorithms neither make assumptions regarding the speakers' dynamics nor require training data, as most DNN-based approaches. They are also shown to be rather robust to inaccuracy of array location, which is a very common imperfection in ad hoc networks.

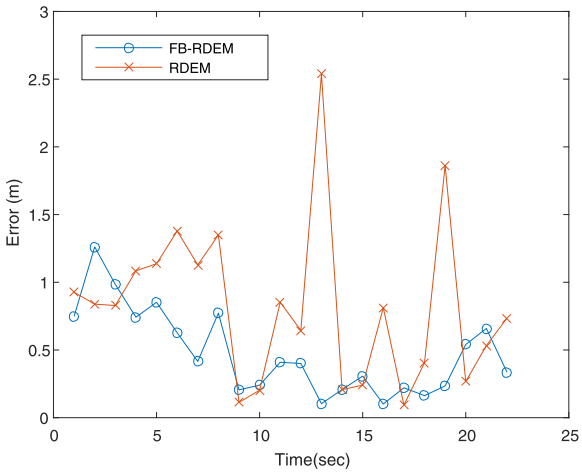


**Fig. 7** Tracking error for two speakers

**Abbreviations**
BW: Bandwidth; FB-REM: Forward-backward recursive expectation-maximization; DOA: Direction of arrival; EM: Expectation-maximization; FIM: Fisher information matrix; IEM: Incremental expectation-maximization; i.i.d.: Independent and identically distributed; KLD: Kullback-Leibler divergence; LOCATA: LOCalization And TrAcking; MAP: Maximum a posteriori; ML: Maximum likelihood; MLE: Maximum likelihood estimator; MoG: Mixture of Gaussians; p.d.f.: Probability density function; PHAT: Phase transform; PHD: Probability hypothesis density; RDEM: Recursive distributed expectation-maximization; REM: Recursive expectation-maximization; RIR: Room impulse response; RMSE: Root mean square error; SNR: Signal to noise ratio; STFT: Short-time Fourier transform; TDOA: Time difference of arrival; T-F: Time-frequency; TREM: Titterington recursive expectation-maximization; TFB-REM: Tracking forward-backward recursive expectation-maximization; TF-REM: Tracking forward-recursive expectation-maximization; w.p.1.: With probability one; w.r.t.: With respect to; WDO: W-disjoint orthogonality FCN: Fully convolutional network; DNN: Deep neural network; CNN: Convolutional neural network; GCC: Generalized cross-correlation; SRP-PHAT: Steered response power with phase transform; MUSIC: Multiple signal classification;

**Acknowledgements**
We would like to thank Mr. Pini Tandeitnik for his professional assistance during the acoustic room setup and the recordings. We would also like to thank our three real speakers recorded in the acoustic room. First, Dr. Dovid Levin for the recording campaign and some professional advice. Second, Miya and Mannie Dorfan for the recording campaign.

**Authors' contributions**
Model development: YD, BS, and SG. Experimental testing: YD and BS. Writing the paper: YD, BS, and SG. All authors read and approved the final manuscript.

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

**Availability of data and materials**
N/A.

**Consent for publication**
All authors agree to the publication in this journal.

**Competing interests**
The authors declare that they have no competing interests.

Received: 8 April 2020 Accepted: 23 November 2020
Published online: 09 January 2021

**References**

1. S. S. Blackman, *Multiple-Target Tracking with Radar Applications*. (Artech House, Inc., Dedham, MA, 1986), p. 463
2. Y. Bar-Shalom, *Multitarget-Multisensor Tracking: Advanced Applications*. (Artech House, Norwood, MA, 1990), p. 391
3. M. E. Liggins, C.-Y. Chong, I. Kadar, M. G. Alford, V. Vannicola, S. Thomopoulos, Distributed fusion architectures and algorithms for target tracking. Proc. IEEE. **85**(1), 95–107 (1997)
4. C. Gui, P. Mohapatra, in *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking*. Power conservation and quality of surveillance in target tracking sensor networks (ACM, 2004), pp. 129–143
5. C. E. Y. Dorfan, S. Gannot, P. A. Naylor, C. Evers, in *2016 24th European Signal Processing Conference (EUSIPCO)*. Speaker localization with moving microphone arrays (IEEE, Budapest), pp. 1003–1007
6. R. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
7. J. P. Dmochowski, J. Benesty, S. Affes, in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Broadband music: opportunities and challenges for multiple source localization (IEEE, 2007), pp. 18–21
8. M. S. Brandstein, H. F. Silverman, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A robust method for speech signal time-delay estimation in reverberant rooms (IEEE, 1997)
9. Y. Dorfan, A. Plinge, G. Hazan, S. Gannot, Distributed expectation-maximization algorithm for speaker localization in reverberant environments. IEEE/ACM Trans Audio Speech Lang. Process. **26**(3), 682–695 (2017)
10. W. Xue, W. Liu, in *Interspeech*. Direction of arrival estimation based on subband weighting for noisy conditions (Conference of the International Speech Communication Association (ISCA), Portland, 2012)
11. X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, H. Li, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A learning-based approach to direction of arrival estimation in noisy and reverberant environments (IEEE, Brisbane, 2015), pp. 2814–2818
12. F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, F. Piazza, in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. A neural network based algorithm for speaker localization in a multi-room environment (IEEE, Salerno, 2016), pp. 1–6
13. R. Takeda, K. Komatani, in *IEEE Spoken Language Technology Workshop (SLT)*. Discriminative multiple sound source localization based on deep neural networks using independent location model (IEEE, San Juan, 2016), pp. 603–609
14. H. Pujol, E. Bavu, A. Garcia. Source localization in reverberant rooms using Deep Learning and microphone arrays (The International Commission on Acoustics (ICA), Aachen, 2019)
15. J. M. Vera-Diaz, D. Pizarro, J. Macias-Guarasa, Towards end-to-end acoustic localization using deep learning: from audio signals to source position coordinates. Sensors. **18**(10), 3418 (2018)
16. M. Azimi-Sadjadi, Y. Jiang, G. Wichern, in *Defense and Security Symposium*. Properties of randomly distributed sparse acoustic sensors for ground vehicle tracking and localization (International Society for Optics and Photonics, 2006)
17. I. Marković, I. Petrović, Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering. Robot. Auton Syst. **58**(11), 1185–1196 (2010)
18. X. Zhong, A. B. Premkumar, in *2012 15th International Conference on Information Fusion*. A random finite set approach for joint detection and tracking of multiple wideband sources using a distributed acoustic vector sensor array (IEEE, Singapore, 2012), pp. 519–526
19. W. Li, Y. Jia, J. Du, J. Zhang, Distributed multiple-model estimation for simultaneous localization and tracking with NLOS mitigation. IEEE Trans. Veh. Technol. **62**(6), 2824–2830 (2013)
20. A. Brutti, F. Nesta, Tracking of multidimensional TDOA for multiple sources with distributed microphone pairs. Comput. Speech Lang. **27**(3), 660–682 (2013)
21. W.-P. Chen, J. C. Hou, L. Sha, Dynamic clustering for acoustic target tracking in wireless sensor networks. IEEE Trans. Mob. Comput. **3**(3), 258–271 (2004)
22. X. Sheng, Y.-H. Hu, P. Ramanathan, in *Fourth International Symposium on Information Processing in Sensor Networks (IPSN)*. Distributed particle filter with GMM approximation for multiple targets localization and tracking in wireless sensor network (IEEE, Los Angeles California, 2005), pp. 181–188
23. L. E. Parker, B. Birch, C. Reardon, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Indoor target intercept using an acoustic sensor network and dual wavefront path planning, vol. 1 (IEEE, Las Vegas, 2003), pp. 278–283
24. F. Talantzis, A. Pnevmatikakis, A. G. Constantinides, Audio–visual active speaker tracking in cluttered indoors environments. IEEE Trans. Syst. Man Cybern. Part B Cybern. **38**(3), 799–807 (2008)
25. B. Laufer-Goldshtein, R. Talmon, S. Gannot, in *International Conference on Latent Variable Analysis and Signal Separation*. Speaker tracking on multiple-manifolds with distributed microphones (Springer, 2017), pp. 59–67
26. M. Taseska, G. Lamani, E. A. Habets, in *Advances in Machine Learning and Signal Processing*. Online clustering of narrowband position estimates with application to multi-speaker detection and tracking (Springer, Cham, 2016), pp. 59–69
27. P. J. Walmsley, S. J. Godsill, P. J. Rayner, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Polyphonic pitch tracking using joint bayesian estimation of multiple frame parameters (IEEE, New Paltz, 1999), pp. 119–122
28. L. D. Stone, R. L. Streit, T. L. Corwin, K. L. Bell, *Bayesian Multiple Target Tracking*. (Artech House, Boston, US, 2013)

29. C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, B. Rafaely, in *IEEE International Conference on Digital Signal Processing (DSP)*. Bearing-only acoustic tracking of moving speakers for robot audition (IEEE, Singapore, 2015), pp. 1206–1210
30. A. Levy, S. Gannot, E. A. P. Habets, Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments. IEEE Trans. Audio Speech Lang. Process. **19**(6), 1540–1555 (2011)
31. X. Zhong, A. Premkumar, A. Madhukumar, Particle filtering for acoustic source tracking in impulsive noise with alpha-stable process. IEEE Sensors J. **13**(2), 589–600 (2013)
32. C. Evers, Y. Dorfan, S. Gannot, P. Naylor, in *International Conference on Audio and Acoustic Signal Processing (ICASSP)*. Source tracking using moving microphone arrays for robot audition (IEEE, New-Orleans, LA, USA, 2017)
33. N. Roman, D. Wang, Binaural tracking of multiple moving sources. IEEE Trans. Audio Speech Lang. Process. **16**(4), 728–739 (2008)
34. H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, W. Kellermann, in *IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. The LOCATA challenge data corpus for acoustic source localization and tracking (IEEE, Sheffield, 2018), pp. 410–414
35. C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, W. Kellermann, in *16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. LOCATA challenge-evaluation tasks and measures (IEEE, Tokyo, 2018), pp. 565–569
36. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. **39**(1), 1–38 (1977)
37. D. M. Titterington, Recursive parameter estimation using incomplete data. J. R. Stat. Soc. Ser. B Methodol. **46**(2), 257–267 (1984)
38. S. Wang, Y. Zhao, Almost sure convergence of Titterington's recursive estimator for mixture models. Stat. Probab. Lett. **76**(18), 2001–2006 (2006)
39. B. Schwartz, S. Gannot, E. A. Habets, Y. Noam, Recursive maximum likelihood algorithm for dependent observations. IEEE Trans. Sig. Process. **67**(5), 1366–1381 (2019)
40. B. Delyon, General results on the convergence of stochastic algorithms. IEEE Trans. Autom. Control. **41**(9), 1245–1255 (1996)
41. B. Delyon, M. Lavielle, E. Moulines, Convergence of a stochastic approximation version of the EM algorithm. Ann. Stat. **27**(1), 94–128 (1999)
42. P.-J. Chung, J. F. Bohme, Recursive EM and SAGE-inspired algorithms with application to DOA estimation. IEEE Trans. Sig. Process. **53**(8), 2664–2677 (2005)
43. O. Cappé, E. Moulines, On-line expectation maximization algorithm for latent data models. J. R. Stat. Soc. Ser. B Methodol. **71**(3), 593–613 (2009)
44. O. Schwartz, S. Gannot, Speaker tracking using recursive EM algorithms. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(2), 392–402 (2014)
45. T. Wolff, M. Buck, G. Schmidt, in *ITG Conference on Voice Communication (SprachKommunikation)*. A subband based acoustic source localization system for reverberant environments (VDE, 2008)
46. A. D. Firoozabadi, H. R. Abutalebi, in *Sixth IEEE International Symposiumon Telecommunications (IST)*. Combination of nested microphone array and subband processing for multiple simultaneous speaker localization (IEEE, 2012), pp. 907–912
47. A. D. Firoozabadi, H. R. Abutalebi, in *21st Iranian Conference on Electrical Engineering (ICEE)*. Localization of multiple simultaneous speakers by combining the information from different subbands (IEEE, 2013), pp. 1–6
48. O. Schwartz, Y. Dorfan, E. A. P. Habets, S. Gannot, in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. Multi-speaker doa estimation in reverberation conditions using expectation-maximization (IEEE, 2016), pp. 1–5
49. Y. Dorfan, G. Hazan, S. Gannot, in *the 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. Multiple acoustic sources localization using distributed expectation-maximization algorithm (IEEE, Nancy, 2014), pp. 72–76
50. R. Neal, G. E. Hinton, in *Learning in Graphical Models*. A view of the EM algorithm that justifies incremental, sparse, and other variants (Kluwer Academic Publishers, MIT Press, Cambridge, MA, USA, 1998), pp. 355–368
51. Y. Dorfan, S. Gannot, Tree-based expectation-maximization algorithms for localization of acoustic sources. IEEE Trans.Audio Speech Lang. Process. **23**(10), 1692–1703 (2015)
52. T. Caljon, V. Enescu, P. Schelkens, H. Sahli, in *International Conference on Advanced Concepts for Intelligent Vision Systems*. An offline bidirectional tracking scheme (Springer, Berlin, Heidelberg, 2005), pp. 587–594
53. A. Brutti, M. Omologo, P. Svaizer, in *International Workshop on Acoustic Signal Enhancement*. Maximum a posteriori trajectory estimation for acoustic source tracking (VDE, 2012), pp. 1–4
54. O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Sig. Process. **52**(7), 1830–1847 (2004)
55. A. Benveniste, M. Metivier, P. Priouret, *Adaptive Algorithms and Stochastic Approximations, vol. 22*. (Springer, Heidelberg, Germany, 2012)
56. H. J. Kushner, G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd edn. (Springer, Heidelberg, Germany, 2003)
57. R. E. Kalman, *et al*, A new approach to linear filtering and prediction problems. J Basic Eng. **82**(1), 35–45 (1960)
58. G. D. Forney, The viterbi algorithm. Proc. IEEE. **61**(3), 268–278 (1973)
59. H. Do, H. F. Silverman, Y. Yu, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array, vol. 1 (IEEE, Honolulu, 2007), p. 121
60. Y. Dorfan, D. Cherkassky, S. Gannot, in *European Signal Processing Conference (EUSIPCO)*. Speaker localization and separation using incremental distributed expectation-maximization (IEEE, Nice, 2015), pp. 1256–1260
61. O. Schwartz, Y. Dorfan, E. A. P. Habets, S. Gannot, in *Proc. Intl. Workshop for Acoustic Echo Cancellation and Noise Control (IWAENC)*. Multiple DOA estimation in reverberant conditions using EM, (Xi'an, China, 2016)
62. Y. Dorfan, O. Schwartz, B. Schwartz, E. A. P. Habets, S. Gannot, in *IEEE Science of Electrical Engineering (ICSEE)*. Multiple DOA estimation and blind source separation using estimation-maximization, (Eilat, Israel, 2016), pp. 1–5
63. O. Schwartz, Y. Dorfan, M. Taseska, E. A. Habets, S. Gannot, in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*. DOA estimation in noisy environment with unknown noise power using the EM algorithm (IEEE, 2017), pp. 86–90
64. J. B. Allen, D. A. Berkley, Image method for efficiently simulating small-room acoustics. J Acoust. Soc. Am. **65**(4), 943–950 (1979)
65. E. A. P. Habets, Room impulse response (RIR) generator (2010). http://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator. Accessed 1 Aug 2018


## Publisher's Note