Speaker Tracking on Multiple-Manifolds with Distributed Microphones

Bracha Laufer-Goldshtein[1(✉)], Ronen Talmon[2], and Sharon Gannot[1]

[1] Bar-Ilan University, 5290002 Ramat-Gan, Israel
{bracha.laufer,sharon.gannot}@biu.ac.il
[2] Technion – Israel Institute of Technology, Technion City, 3200003 Haifa, Israel
ronen@ee.technion.ac.il

Abstract. Speaker tracking in a reverberant enclosure with an ad hoc network of multiple distributed microphones is addressed in this paper. A set of prerecorded measurements in the enclosure of interest is used to construct a data-driven statistical model. The function mapping the measurement-based features to the corresponding source position represents complex unknown relations, hence it is modelled as a random Gaussian process. The process is defined by a covariance function which encapsulates the relations among the available measurements and the different views presented by the distributed microphones. This model is intertwined with a Kalman filter to capture both the smoothness of the source movement in the time-domain and the smoothness with respect to patterns identified in the set of available prerecorded measurements. Simulation results demonstrate the ability of the proposed method to localize a moving source in reverberant conditions.

Keywords: Speaker tracking · Distributed microphones · Gaussian process · Acoustic manifold · Kalman filter

1 Introduction

Speaker localization and tracking in reverberant enclosures plays an important role in many applications, including: automatic camera steering, teleconferencing and beamforming. Conventional localization methods can be roughly divided into single- and dual-step approaches. In single-step approaches, a grid search is performed to find the position that maximizes a certain optimization criterion [5,13]. In dual-step approaches, the time difference of arrivals (TDOAs) of several microphone pairs are first estimated and then combined to perform the actual localization [2,9].

In dynamic scenarios, the measurements can be divided into short time frames, during which the source position is approximately static. Hence, in each time step the information available for the localization task is limited. However, the smoothness of the movement implies dependence across time. The temporal consistency across successive frames can be exploited by either Bayesian or

P. Tichavský et al. (Eds.): LVA/ICA 2017, LNCS 10169, pp. 59–67, 2017.
DOI: 10.1007/978-3-319-53547-0_6

non-Bayesian models. Bayesian state-space models, which are usually nonlinear and non-Gaussian, are implemented using unscented Kalman filter or extended Kalman filter [7] and particle filter [16]. In non-Bayesian approaches, the trajectory is considered as a deterministic and time-varying parameter, and a maximum likelihood criterion can be applied [14].

In realistic environments, the presence of noise or reverberation often yields spurious observations which may lead to poor localization performance. In addition, traditional localization and tracking schemes are based on approximated physical and statistical assumptions which do not always meet the practical conditions in complex real-world scenarios. Recently, there is an attempt to overcome these limitations by applying supervised or unsupervised learning-based approaches [3,4,12,15]. The idea is to form a data-driven model for the spatial characteristics of an acoustic environment, rather than using a predefined statical model.

In this paper, we derive a semi-supervised tracking algorithm based on measurements from distributed pairs of microphones. The algorithm exploits a training set of prerecorded measurements from various locations in the enclosure of interest. Capitalizing this prior information, we identify the geometrical patterns, namely the underlying manifold to which the measurement-based features are confined, and relate it to the position of the source. Recently [11], we have presented a semi-supervised localization approach, which explores the acoustic manifold associated with each microphone pair, and composes these models in the definition of a multiple-manifold Gaussian process (MMGP). Here, this data-driven statistical model is integrated into a Kalman filter scheme to impose dual-domain smoothness, both with respect to the acoustic manifold and the time domain. The algorithm performance is examined using simulated trajectories of a moving source in a reverberant room with spatially-distributed microphone pairs.

2 Problem Formulation

We consider a reverberant enclosure consisting of M nodes, where each node comprises a pair of microphones. A single source is moving in the enclosure, generating an unknown speech signal $s(n)$, which is measured by all the microphones. The received signals are contaminated by additive stationary noise sources and are given by:

$$y^{mi}(n) = \sum_k a_n^{mi}(k)s(n-k) + u^{mi}(n), \quad m = 1, \ldots, M \tag{1}$$

where n is the time index, a_n^{mi}, $i = \{1, 2\}$ is the time-varying acoustic impulse response (AIR) relating the source and the ith microphone in the mth node at time n, and $u^{mi}(n)$ is the corresponding noise signal. The measured signals are partitioned into short segments of a few hundred milliseconds, which are assigned with frame index t. From each segment we constitute a feature vector $\mathbf{h}^m(t)$ that preserves the relevant information for localization which is hidden in the AIRs, and is invariant to other irrelevant factors, namely the non-stationary source signal. More specifically, we use a feature vector based on relative transfer function (RTF) estimates in a certain frequency band, which is commonly used in acoustic array processing [6]. The RTF is typically represented in high dimension with a large number of coefficients to account for the room reverberation. The observation that the RTF is controlled by a small set of parameters, such as room dimensions, reverberation time, location of the source and the sensors etc., gives rise to the assumption that it is confined to a low dimensional manifold $\mathcal{M}_m$, as was demonstrated in [10].

To track a moving source, we consider the function f^m which attaches to an RTF sample $\mathbf{h}^m(t)$ from the mth node its corresponding x, y or z coordinate of the source position $p^m(t) \equiv f^m(\mathbf{h}(t))$, for frame t. Note that although the position of the source does not depend on the specific node, the notation $p^m(t)$ is used to express that the mapping is obtained from the measurement of the mth node. The different nodes represent different views of the same acoustic scene, hence incorporating the information from the different nodes in a unifying mapping denoted by f may enrich the spatial information utilized for localization and tracking. Let $\mathbf{h}(t) = \left[[\mathbf{h}(t)^1]^T, \ldots, [\mathbf{h}(t)^M]^T\right]^T$ denote the aggregated RTF (aRTF), which is a concatenation of the RTF vectors from all the nodes. The function f associates the corresponding source position to an aRTF sample $\mathbf{h}(t)$, namely $p(t) \equiv f(\mathbf{h}(t))$. The function f, which defines an instantaneous mapping, is used here to evaluate the position of the source along its track. In the dynamic scenario, the function is used to transform the observed propagation in the RTFs domain to the physical domain of the source positions, i.e. it assists the development of a simple Markovian relation between successive positions.

To estimate the function f, we assume the availability of a training set consisting of a limited number of labelled measurements from multiple nodes, attached with corresponding source positions, and a larger amount of unlabelled measurements with unknown source locations. All the training measurements apply to static sources. The labelled set consists of n_L pairs $\{\mathbf{h}_i, \bar{p}_i\}_{i=1}^{n_L}$, and the unlabelled set consists of n_U samples $\{\mathbf{h}_i\}_{i=n_L+1}^{n_D}$, where $n_D = n_L + n_U$. Note that the microphone positions may be unknown, since they are not required for the estimation. In the test phase, we receive the measurements of a moving source, partition them into n_T short segments, and compute the corresponding RTF separately for each segment. The set $\{\mathbf{h}(t)\}_{t=1}^{n_T}$ consists of the aRTFs of all the segments, where the index t denotes their chronological order. The goal is to estimate the corresponding n_T temporary positions $\{p(t)\}_{t=1}^{n_T}$ of each sample in the set $\{\mathbf{h}(t)\}_{t=1}^{n_T}$.

3 Multiple-Manifold Gaussian Process

We first define a statistical model for each node separately and the relation between the different nodes, and then combine them in a unified model [11]. We assume that the position p^m, which is associated with the measurements of the mth node, follows a zero-mean Gaussian process, i.e. the set of all possible positions mapped from the samples of the mth node, are joint Gaussian variables. The Gaussian process is completely defined by its covariance function, which is a pairwise affinity measure between two RTF samples. We use a manifold-based covariance function, defined by:

$$\text{cov}(p_r^m, p_l^m) \equiv \tilde{k}_m(\mathbf{h}_r^m, \mathbf{h}_l^m) = \sum_{i=1}^{n_D} k_m(\mathbf{h}_r^m, \mathbf{h}_i^m) k_m(\mathbf{h}_l^m, \mathbf{h}_i^m) \tag{2}$$

where l and r represent ascription to certain positions, and k_m is a standard "kernel" function $k_m : \mathcal{M}_m \times \mathcal{M}_m \longrightarrow \mathbb{R}$. A common choice is to use a Gaussian kernel.

Considering multiple nodes, we similarly define the correlation between two source positions p_r^q and p_l^w associated with nodes q and w, respectively. We assume that p_r^q and p_l^w are jointly Gaussian and that their covariance is defined by:

$$\text{cov}(p_r^q, p_l^w) \equiv \tilde{k}_{qw}(\mathbf{h}_r^q, \mathbf{h}_l^w) = \sum_{i=1}^{n_D} k_q(\mathbf{h}_r^q, \mathbf{h}_i^q) k_w(\mathbf{h}_l^w, \mathbf{h}_i^w). \tag{3}$$

Note that in both (2) and (3), the covariance is constituted by an average over all the available training samples. This averaging implies that the similarity between two samples from the manifold can be determined according to the way they are viewed by other samples residing on the same manifold. When two samples convey similar connections (i.e. proximity or remoteness) to other samples, it indicates that they are closely related with respect to the manifold. In (3), we cannot directly compute the distance between the corresponding RTF samples since they present different views of two nodes. Thus, we choose another sample $\mathbf{h}_i$, and compare the distances with respect to $\mathbf{h}_i$ as it is viewed by the different nodes. The inter-relations in the qth and wth manifolds are computed separately, and then they are composed by multiplying the corresponding kernels.

To fuse the different perspectives presented by the different nodes, we define the multiple-manifold Gaussian process (MMGP) p as the mean of the Gaussian processes of all the nodes:

$$p = \frac{1}{M}(p^1 + p^2 + \ldots + p^M). \tag{4}$$

Due to the assumption that the processes are jointly Gaussian, the process p is also Gaussian with zero-mean and a covariance function given by:

$$\text{cov}(p_r, p_l) = \frac{1}{M^2}\text{cov}\left(\sum_{q=1}^{M} p_r^q, \sum_{w=1}^{M} p_l^w\right) = \frac{1}{M^2}\sum_{q,w=1}^{M} \text{cov}(p_r^q, p_l^w). \tag{5}$$

Using the definitions of (2) and (3) we get the covariance for p_r and p_l:

$$\text{cov}(p_r, p_l) \equiv \tilde{k}(\mathbf{h}_r, \mathbf{h}_l) = \frac{1}{M^2}\sum_{i=1}^{n_D}\sum_{q,w=1}^{M} k_q(\mathbf{h}_r^q, \mathbf{h}_i^q) k_w(\mathbf{h}_l^w, \mathbf{h}_i^w). \tag{6}$$

Here, the covariance is defined by averaging over all the available training samples as well as over all pairs of nodes. The induced kernel $\tilde{k}(\mathbf{h}_r, \mathbf{h}_l)$, can be considered as a *composition of kernels*, which, in addition to connections acquired in each node separately, incorporates the extra spatial information manifested in the mutual relationship between RTFs from different nodes.

4 Multiple-Manifold Speaker Tracking

The tracking is performed by a state-space representation, formulated according to the statistical relations implied by the MMGP. In the dynamic scenario, the test aRTF samples $\{\mathbf{h}(t)\}_{t=1}^{n_T}$, and their associated unknown source positions $\{p(t)\}_{t=1}^{n_T}$ are treated as two time-series, which are mutually-related through the mapping f. The propagation model, specifying the relation between the source positions in successive time steps, is defined according to similarities between the corresponding aRTFs, as induced by the covariance of the MMGP. This way the movement of the source is constrained to vary smoothly with respect to the manifolds of the different nodes. The measurement model relates the current sample $\mathbf{h}(t)$ to all the other available training samples. The resulting state-space representation is solved by a Kalman-filter, in which the source position, predicted through the local interpolation devised by successive samples, is updated by a global interpolation formed by all the training information.

We first define the propagation model. The position $p(t-1)$ at time $t-1$ and the current position $p(t)$ are two samples from the MMGP defined in the previous section. Hence, the random variables $p(t)$ and $p(t-1)$ have a joint normal distribution, and their conditional probability is given by:

$$p(t)|p(t-1) \sim \mathcal{N}\left(\frac{\tilde{\Sigma}_{t,t-1}}{\tilde{\Sigma}_{t-1}}p(t-1), \tilde{\Sigma}_t - \frac{\tilde{\Sigma}_{t,t-1}^2}{\tilde{\Sigma}_{t-1}}\right) \quad (7)$$

where $\tilde{\Sigma}_t$ and $\tilde{\Sigma}_{t-1}$ are the variances of $p(t)$ and $p(t-1)$ respectively, and $\tilde{\Sigma}_{t,t-1}$ is the covariance of $p(t)$ and $p(t-1)$. For a Gaussian process, the propagated probabilities in (7) can be equivalently represented by a linear propagation equation with an additive Gaussian noise ξ_t

$$p(t) = g_t \cdot p(t-1) + \xi_t \quad (8)$$

where $g_t = \frac{\tilde{\Sigma}_{t,t-1}}{\tilde{\Sigma}_{t-1}}$ and $\xi_t \sim \mathcal{N}\left(0, \sigma_\xi^2\right)$ with $\sigma_\xi^2 = \tilde{\Sigma}_t - \frac{\tilde{\Sigma}_{t,t-1}^2}{\tilde{\Sigma}_{t-1}}$. Since there is no prior information on the actual trajectory of the speaker, it is reasonable to use a simplified random walk model as in (8). However, it should be noted that the proposed model is data-driven, in the sense that both the transition factor g_t and the driving noise variance σ_ξ^2 are determined based on the relation between the current aRTF sample $\mathbf{h}(t)$ and the preceding one $\mathbf{h}(t-1)$. When the aRTF samples are close to each other, namely that the acoustic characteristics have hardly changed, it is assumed that only a slight movement of the source has occurred. In this case, we receive $\tilde{\Sigma}_{t,t-1} \approx \tilde{\Sigma}_t \approx \tilde{\Sigma}_{t-1}$, yielding $g_t \approx 1$ and $\sigma_\xi^2 \approx 0$, which implies that $p(t) \approx p(t-1)$ as desired. Overall, the proposed propagation model imposes a smooth variation of the position with respect to the manifolds associated with the different nodes, and reflects the strong relation between the physical domain and the aRTFs domain.

As for the measurement model, we can form an observation q_t that represents the estimated position based on the available training samples. Let $\bar{\mathbf{p}}_L = [\bar{p}_1, \ldots, \bar{p}_{n_L}]^T$ be a concatenation of the measured positions of the labelled set. We assume that the measured positions $\bar{p}_i = p_i + \eta_i$, are noisy versions of the actual position p_i, due to imperfections in the measurements while acquiring the labelled set. Assuming that η_i is an independent Gaussian noise with variance σ^2 yields that $p(t)$ and $\bar{\mathbf{p}}_L$ are jointly Gaussian, and their conditional distribution is given by:

$$p(t)|\bar{\mathbf{p}}_L \sim \mathcal{N}\left(\tilde{\boldsymbol{\Sigma}}_{Lt}^H \left(\tilde{\boldsymbol{\Sigma}}_L + \sigma^2 \mathbf{I}_{n_L}\right)^{-1} \bar{\mathbf{p}}_L, \tilde{\Sigma}_{t,t} - \tilde{\boldsymbol{\Sigma}}_{Lt}^H \left(\tilde{\boldsymbol{\Sigma}}_L + \sigma^2 \mathbf{I}_{n_L}\right)^{-1} \tilde{\boldsymbol{\Sigma}}_{Lt}\right) \quad (9)$$

where $\tilde{\boldsymbol{\Sigma}}_L$ is an $n_L \times n_L$ covariance matrix defined over the function values at the labelled samples, $\tilde{\boldsymbol{\Sigma}}_{Lt}$ is an $n_L \times 1$ covariance vector between the function values at the labelled samples and $p(t)$, and $\mathbf{I}_{n_L}$ is the $n_L \times n_L$ identity matrix. Accordingly, we define the observation as $q_t = \mathbf{Q}_t \bar{\mathbf{p}}_L$, where $\mathbf{Q}_t = \tilde{\boldsymbol{\Sigma}}_{Lt}^H \left(\tilde{\boldsymbol{\Sigma}}_L + \sigma^2 \mathbf{I}_{n_L}\right)^{-1}$. The corresponding measurement model can be expressed as:

$$q_t = p(t) + \zeta_t \quad (10)$$

where $\zeta_t \sim \mathcal{N}\left(0, \sigma_\zeta^2\right)$ with $\sigma_\zeta^2 = \tilde{\Sigma}_{t,t} - \tilde{\boldsymbol{\Sigma}}_{Lt}^H \left(\tilde{\boldsymbol{\Sigma}}_L + \mathbf{I}_{n_L}\right)^{-1} \tilde{\boldsymbol{\Sigma}}_{Lt}$. Here as well, the covariance terms are calculated using the kernel $\tilde{k}$ defined in the previous section. Since an *actual* noisy measurement of the position does not exist, we use instead an *artificial* data-driven measurement q_t, formed by the current sample $\mathbf{h}(t)$ and the entire training set. This measurement is, in fact, an estimated position, which is obtained by a global interpolation of the labelled samples, based on the learnt manifold-based model. The aRTF sample $\mathbf{h}(t)$ allows the calculation of the covariance $\tilde{\boldsymbol{\Sigma}}_{Lt}$ that is essential for the evaluation of the artificial position measurement. Our confidence in the artificial measurement is determined according to the variance of the estimator, and is expressed in the model by the variance of the measurement noise ζ_t.

To summarize, the proposed state-space model is given by:

$$\begin{aligned} p(t) &= g_t \cdot p(t-1) + \xi_t \\ q_t &= p(t) + \zeta_t. \end{aligned} \quad (11)$$

Since both the process and the observation models are linear, a standard Kalman filter can be applied for recursively solving (11). The Kalman filter recursion takes the following form:

$$\begin{aligned}\hat{p}(t|t-1) &= g_t \cdot \hat{p}(t-1|t-1)\\ \gamma(t|t-1) &= g_t^2 \gamma(t-1|t-1) + \sigma_\xi^2\\ \hat{p}(t|t) &= \hat{p}(t|t-1) + \kappa(t)\left(q_t - \hat{p}(t|t-1)\right)\\ \gamma(t|t) &= \left(1-\kappa(t)\right)\gamma(t|t-1)\end{aligned} \tag{12}$$

where $\gamma(t|t-1)$ is the predicted covariance, $\gamma(t|t)$ is the posteriori covariance, and $\kappa(t)$ is the Kalman gain, defined as:

$$\kappa(t) = \frac{\gamma(t|t-1)}{\gamma(t|t-1) + \sigma_\zeta^2}. \tag{13}$$

Note that the measurements of the moving source, accumulated through run-time, can be considered as additional unlabelled data. Thus, the current measurements can be used to update the manifold-based covariance terms in (7) and (9). An efficient recursive adaptation for the MMGP was presented in [11].

5 Experimental Study

We conducted a simulation of a 2-D tracking of a moving source. We simulated a 5.2 × 6.2 × 3 room with 4 pairs of microphones mounted next to the room walls, using an efficient implementation [8] of the image method [1]. All the measurements were confined to a 2 × 2 m rectangular region, at a fixed height of 2 m (the same height of all the microphones). We generated a training set with $n_L = 36$ labelled samples, without additional unlabelled samples ($n_U = 0$). The labelled samples form a fixed grid with resolution of 0.4 m, and were generated using 10 s long speech signals. The room setup is presented in Fig. 1.

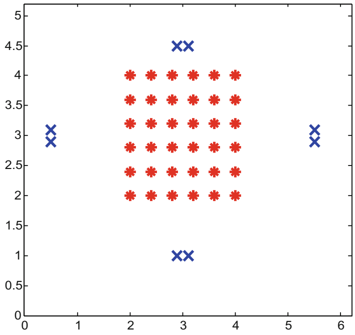

Fig. 1. Room setup: the blue x-marks denote the microphones and the red asterisks denote the labelled samples. (Color figure online)

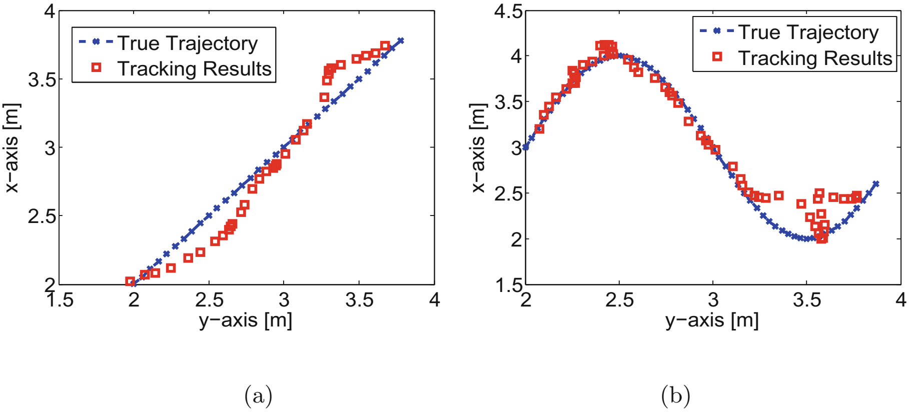

Fig. 2. True path and estimated path for (a) straight line movement and for (b) sinusoidal movement.

We examined two types of trajectories: a straight line along the diagonal of the rectangular region and a sinusoidal trajectory. The duration of the entire movement of the source was 3 s and 5 s for the straight line movement and for the sinusoidal movement, respectively. For both movement types, the source average velocity was approximately 1 m/s, and the measured signals were divided into segments of 330 ms with 75% overlap. For each segment, the corresponding RTF was estimated in 2048 frequency bins. In Fig. 2, we plot the two movements and the tracking results received for 300 ms reverberation time in noiseless conditions.

It can be observed that the proposed method is able to track the source for both types of trajectories. The root mean square errors (RMSEs) were 13 cm and 17 cm for the straight line movement and for the sinusoidal movement, respectively. The error is larger for the sine path compared to the straight path, since it is more complicated and neither the velocity nor the acceleration are fixed. In addition, for regions closer to the microphone positions we receive lower error compared to remote regions, as can be observed by comparing the tracking results around the two peaks of the sine path. We conclude that the proposed algorithm is capable of accurately tracking the source in a reverberant environment.

6 Conclusions

A semi-supervised tracking algorithm based on measurements from distributed pairs of microphones is presented. The tracking is carried out by Kalman filtering which exploits smoothness in two domains. The first is the commonly assumed smoothness of the source trajectory in the time domain. The second is related to the data-driven model inferred from the prerecorded measurements. The source position is assumed to vary smoothly with respect to the multiple acoustic manifolds associated with the different nodes. The resulting tracker is shown to accurately track a moving source in a simulated reverberant room. In future work, we intend to examine a more sophisticated modelling of the source movement.

References

1. Allen, J., Berkley, D.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
2. Benesty, J.: Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. J. Acoust. Soc. Am. **107**(1), 384–391 (2000)
3. Bertin, N., Kitić, S., Gribonval, R.: 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), pp. 6340–6344 (2016)
4. Deleforge, A., Forbes, F., Horaud, R.: Acoustic space learning for sound-source separation and localization on binaural manifolds. Int. J. Neural Syst. **25**(1), 1440003 (2015)
5. Dmochowski, J.P., Benesty, J.: Steered beamforming approaches for acoustic source localization. In: Cohen, I., Benesty, J., Gannot, S. (eds.) Speech Processing in Modern Communication, pp. 307–337. Springer, Heidelberg (2010)
6. Gannot, S., Burshtein, D., Weinstein, E.: Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Sign. Process. **49**(8), 1614–1626 (2001)
7. Gannot, S., Dvorkind, T.G.: Microphone array speaker localizers using spatial-temporal information. EURASIP J. Adv. Sig. Process. **2006**(1), 1–17 (2006)
8. Habets, E.A.P.: Room impulse response (RIR) generator, July 2006. http://home.tiscali.nl/ehabets/rir_generator.html
9. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Sig. Process. **24**(4), 320–327 (1976)
10. Laufer-Goldshtein, B., Talmon, R., Gannot, S.: A study on manifolds of acoustic responses. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) LVA/ICA 2015. LNCS, vol. 9237, pp. 203–210. Springer, Heidelberg (2015). doi:10.1007/978-3-319-22482-4_23
11. Laufer-Goldshtein, B., Talmon, R., Gannot, S.: Semi-supervised source localization on multiple-manifolds with distributed microphones. pre-print arXiv:1610.04770v1, September 2016
12. Salvati, D., Drioli, C., Foresti, G.L.: A weighted MVDR beamformer based on SVM learning for sound source localization. Pattern Recogn. Lett. **84**, 15–21 (2016)
13. Schmidt, R.O.: Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
14. Schwartz, O., Gannot, S.: Speaker tracking using recursive EM algorithms. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(2), 392–402 (2014)
15. Smaragdis, P., Boufounos, P.: Position and trajectory learning for microphone arrays. IEEE Trans. Audio Speech Lang. Process. **15**(1), 358–368 (2007)
16. Ward, D.B., Lehmann, E.A., Williamson, R.C.: Particle filtering algorithms for tracking an acoustic source in a reverberant environment. IEEE Trans. Speech Audio Process. **11**(6), 826–836 (2003)