Online Clustering of Narrowband Position Estimates with Application to Multi-speaker Detection and Tracking

Maja Taseska, Gleni Lamani and Emanuël A.P. Habets

Abstract Speaker detection, localization and tracking are required in systems that involve e.g. hands-free speech acquisition, or blind source separation. Localization can be done in the (TF) domain, where location features extracted using microphone arrays are used to cluster the TF bins corresponding to the same source. The TF clustering approaches provide an alternative to the Bayesian tracking approaches that are based on Kalman and particle filters. In this work, we propose a maximum-likelihood approach where detection, localization, and tracking are achieved by online clustering of narrowband position estimates, while incorporating the speech presence probability at each TF bin in a unified manner.

Keywords Multi-speaker tracking · Maximum likelihood · Number of source estimation · Distributed arrays

1 Introduction

To provide high-quality capture of speech in communication and entertainment systems without requiring close-talking microphones, the spatial diversity of microphone arrays is exploited to extract sources of interest. Such systems need to localize and track a desired source, and use the location information to compute a spatial filter (beamformer) that extracts the source signal and reduces interferers. A multitude state-of-the-art tracking approaches estimate the evolution of the source positions using Kalman or particle filters (see [1, 2] and references therein). Although these approaches are elegantly formulated within a Bayesian framework and provide excellent tracking performance, the estimated source positions can only be used to steer data-independent beamformers to the estimated source locations.

A joint institution of the University Erlangen-Nuremberg and Fraunhofer IIS.

M. Taseska (✉) · G. Lamani · E.A.P. Habets
International Audio Laboratories Erlangen, Erlangen, Germany
e-mail: maja.taseska@audiolabs-erlangen.de

P.J. Soh et al. (eds.), *Advances in Machine Learning and Signal Processing*,
Lecture Notes in Electrical Engineering 387, DOI 10.1007/978-3-319-32213-1_6

From the vast literature on spatial filters, it is known that such filters often provide insufficient performance in reverberant environments.

The localization can alternatively be done by exploiting the speech sparsity in the TF domain [3]. Such approaches often involve clustering of location features, such as phase differences between sensors [4, 5], (DOAs) [6, 7], or narrowband positions [8]. As a by-product of the localization, each TF bin is classified to the dominant source providing means to track the second-order statistics of the sources [7–9]. The latter can be used to compute data-dependent beamformers which offer better performance in reverberant and noisy environments than data-independent beamformers. Note that after clustering of DOAs or phase differences, an additional step is required to obtain the Cartesian coordinates of the sources, in case they are required. Due to the non-linear and non-injective relation between the DOA and the position, this step is non-trivial. Approaches that estimate the positions have been proposed in [5], by deriving the position from clustered phase differences, and in [8], by clustering narrowband position estimates. However, these approaches assume that the number of sources is fixed, which is restrictive in practice. DOA-based clustering that detects the number of sources online has been proposed in [7, 10].

For the application of source localization, source tracking, and clustering of the TF bins, we propose a (ML) framework based on narrowband position estimates obtained from distributed arrays. Narrowband positions were used in our previous work for localization [8] and tracking of a known number of sources [11]. In this paper, we address dynamic scenarios where the number of sources is unknown and time-varying. The proposed framework consists of (i) estimating the parameters of a mixture model by an online (EM) algorithm (Sects. 2 and 3) and (ii) a data-driven mechanism to detect appearing/disappearing sources and add/remove the corresponding clusters (Sect. 4). A further contribution is the unified treatment of speech presence uncertainty in the model, which increases the robustness to noise and reverberation without requiring voice activity detection in a pre-processing stage as in [11, 12]. The cluster information can be used for multi-speaker tracking, as well as for computation of data-dependent spatial filters for (BSS). The evaluation in Sect. 5 focuses on the tracking application, whereas the evaluation of a BSS system is an ongoing work.

2 Probabilistic Model of Narrowband Position Estimates

The signals from S sources in a noisy and reverberant enclosure are captured by A microphone arrays. The signal at microphone m from the a-th array, at time index n and frequency index k in the (STFT) domain is given by $Y_m^{(a)}(n,k) = \sum_{s=1}^{S} X_{m,s}^{(a)}(n,k) + V_m^{(a)}(n,k)$, where $X_{m,s}^{(a)}$ and $V_m^{(a)}$ denote the signal of the s-th source and the noise, respectively. The different signals represent realizations of mutually uncorrelated random processes. Assuming far field conditions and selecting an arbitrary reference microphone m', a DOA θ_a at array a can be obtained

at each TF bin by a phase difference-based estimator [13] (TF indices omitted for brevity)

$$n_a = \begin{bmatrix} \cos(\theta_a) \\ \sin(\theta_a) \end{bmatrix} = \frac{c}{2\pi f} [\boldsymbol{D}_{m'}^{(a)}]^{\dagger} \arg \frac{\boldsymbol{y}^{(a)}}{Y_{m'}^{(a)}}, \tag{1}$$

where $\boldsymbol{D}_{m'}^{(a)} = [\boldsymbol{d}_1^{(a)} - \boldsymbol{d}_{m'}^{(a)}, \ldots, \boldsymbol{d}_{M_a}^{(a)} - \boldsymbol{d}_{m'}^{(a)}]^{\mathrm{T}}$, $\boldsymbol{d}_m^{(a)}$ for $m = 1, \ldots M_a$ denote the positions of the microphones from array a, $\boldsymbol{y}^{(a)}$ contains the signals from array a stacked in a vector, c and f are the speed of sound and the frequency in Hz, $\dagger$ denotes the Moore-Penrose pseudoinverse, and $\arg(\cdot)$ is taken element-wise. Although the DOAs consider only azimuth θ_a, an extension to elevation is possible. By triangulating two vectors $\boldsymbol{n}_{a_1}$ and $\mathbf{n}_{a_2}$ from different arrays, a position $\boldsymbol{\Theta}_{nk}$ is obtained for each bin (n, k). As the signals at each TF bin represent (RV), $\boldsymbol{\Theta}_{nk}$ is also an RV. The position estimates are used to cluster the TF bins based on their respective dominant source. If at TF bin (n, k), the energy of a given source is dominant over the other sources and the noise, the position $\boldsymbol{\Theta}_{nk}$ represents a good estimate of the source location. Note that the DOA estimates can be obtained with any narrowband estimator and the choice of an estimator influences the accuracy of the clustering.

Due to the speech sparsity in the STFT domain [3], it can be assumed that there is at most one dominant source at each TF bin. Let z_{nk} be a discrete RV that takes values from 0 to S, indicating the dominant source as follows

$$z_{nk} = 0 \quad \text{if} \quad \mathbf{y}(n,k) = \mathbf{v}(n,k) \,(\text{i.e. only noise present}), \tag{2a}$$

$$z_{nk} = s \quad \text{if} \quad \mathbf{y}(n,k) \approx \mathbf{x}_s(n,k) + \mathbf{v}(n,k), \tag{2b}$$

where the entries of $\mathbf{y}$, $\mathbf{x}_s$, and $\mathbf{v}$ contain the respective signals from all microphones. Marginalizing over the unobservable RV z, the distribution of the observable RV $\boldsymbol{\Theta}$ is given by $p(\boldsymbol{\Theta}) = \sum_z p(z) p(\boldsymbol{\Theta}|z)$ (subscript nk omitted for brevity). We propose the following parametric model for the likelihood $p(\boldsymbol{\Theta}|z)$

$$p(\boldsymbol{\Theta} \,|\, z) = \mathcal{N}(\boldsymbol{\Theta}; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \quad \text{for} \quad z \neq 0, \tag{3a}$$

$$p(\boldsymbol{\Theta} \,|\, z) = \mathcal{U}(\boldsymbol{\Theta}), \quad \text{for} \quad z = 0, \tag{3b}$$

where $\mathcal{U}$ is a uniform distribution (noise is localized across the room with equal probability) and $\mathcal{N}(\boldsymbol{\Theta}; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ is a two-dimensional Gaussian distribution with mean $\boldsymbol{\mu}_z$ and covariance matrix $\boldsymbol{\Sigma}_z$. The mean $\boldsymbol{\mu}_z$ represents the true location of the z-th source in the xy plane. By writing the marginal distribution of z as

$$p(z) = p(z \,|\, z \neq 0) p(z \neq 0) + p(z \,|\, z = 0)\, p(z = 0), \tag{4}$$

and introducing $\pi_z = p(z \,|\, z \neq 0)$ and $\pi_0 = p(z \neq 0)$, where π_0 is the (SPP), the following holds

$$p(z) = \begin{cases} \pi_z \, \pi_0, & \text{if } z \neq 0 \\ (1 - \pi_0) & \text{if } z = 0. \end{cases} \tag{5}$$

The parametrized distribution of the observable RV $\boldsymbol{\Theta}$ can now be written as

$$p(\boldsymbol{\Theta}; \mathcal{P}, \pi_0) = \sum_{z \neq 0} \pi_z \pi_0 \mathcal{N}(\boldsymbol{\Theta}; \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) + (1 - \pi_0) \, \mathcal{U}(\boldsymbol{\Theta}), \tag{6}$$

where $\mathcal{P} = \{\pi_z, \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z\}_{z \in [1,S]}$ are unknown parameters. The SPP π_0 is treated as a known parameter as it can be estimated at each bin, independently of the clustering framework. The interested reader is referred to [14] and references therein for details on the computation of the SPP π_0.

The goal in this work consists of inferring and tracking the time-varying parameters $\mathcal{P}_n$ and the number of speakers S_n online as blocks of narrowband position estimates become available. The framework should be capable of removing and adding clusters for sources that disappear and appear, respectively.

3 Maximum Likelihood-Based Online Clustering

In this section, we derive the update of the parameters $\widehat{\mathcal{P}}_{n-1}$ at frame $n-1$ to the new estimates $\widehat{\mathcal{P}}_n$ in light of the input data at frame n. For ease of exposition, we assume that the number of speakers is equal in the two successive frames. A mechanism that determines the number of speakers is detailed in Sect. 4.

The observable data at frame n is the set of position estimates and the corresponding SPP from the most recent L frames, across all K frequency bins

$$D_n = \{(\boldsymbol{\Theta}_{ik}, \pi_{0,ik}) | i \in [n - L + 1, n], k \in [1, K]\}. \tag{7}$$

Note that depending on the accuracy of the DOA estimates at different frequencies and on the spatial aliasing limit for the microphone array, data only from a subset of frequencies can be used as done e.g. in [7]. The parameters $\widehat{\mathcal{P}}_n$ are obtained by maximizing the following log likelihood, with respect to $\mathcal{P}_n$

$$L(D_n; \mathcal{P}_n) = \sum_{i,k} \ln p(\boldsymbol{\Theta}_{ik}; \mathcal{P}_n, \pi_0). \tag{8}$$

We omit the TF-bin index from π_0 for brevity, although π_0 is computed bin-wise. The mixture model given by (6) leads to a summation inside the logarithm in (8), which does not allow for a closed form maximization. Instead, maximizing the expectation of the complete data likelihood under the posterior distribution of the unobservable z, represents a significantly easier problem [15]. This expectation, also known as the Q-function, is given for our model by

$$\sum_{i,k} \Big\{ p(z_{ik} = 0 \,|\, \boldsymbol{\Theta}_{ik}; \widehat{\mathcal{P}}_{n-1}) \left[\ln\, (1 - \pi_0) + \ln\, \mathcal{U}(\boldsymbol{\Theta}_{ik}) \right] + \sum_{z \neq 0} p(z_{ik} = z \,|\, \boldsymbol{\Theta}_{ik}; \widehat{\mathcal{P}}_{n-1}) \left[\ln \pi_z \pi_0\, p(\boldsymbol{\Theta}_{ik} \,|\, z_{ik} = z; \mathcal{P}_n) \right] \Big\}, \tag{9}$$

where the posterior distribution of z_{ik} is computed with respect to the old parameters $\widehat{\mathcal{P}}_{n-1}$. Evaluating (9), and maximizing it with respect to $\mathcal{P}_n$ represents an iteration of the EM algorithm, guaranteeing that $L(D_n; \mathcal{P}_n) > L(D_n; \mathcal{P}_{n-1})$ [15]. Setting the derivatives of (9) with respect to $\mathcal{P}_n$ to zero, the standard M-step for a (GMM) is obtained, as the terms due to $z = 0$ do not depend on $\mathcal{P}_n$. Introducing $P_z(n) = \sum_{i,k} p(z_{ik} = z \,|\, \boldsymbol{\Theta}_{ik})$, the new parameters are computed as

$$\pi_z = \frac{P_z(n)}{LK}, \quad \boldsymbol{\mu}_z = \frac{\sum_{i,k} p(z_{ik} = z \,|\, \boldsymbol{\Theta}_{ik})\, \boldsymbol{\Theta}_{ik}}{P_z(n)}, \quad \boldsymbol{\Sigma}_z = \frac{\sum_{i,k} p(z_{ik} = z \,|\, \boldsymbol{\Theta}_{ik})\, (\boldsymbol{\Theta}_{ik} - \boldsymbol{\mu}_z)(\boldsymbol{\Theta}_{ik} - \boldsymbol{\mu}_z)^{\mathrm{T}}}{P_z(n)}. \tag{10}$$

To compute the posteriors $p(z_{ik} = z \,|\, \boldsymbol{\Theta}_{ik})$, we express them as follows

$$p(z_{ik} = z \,|\, \boldsymbol{\Theta}_{ik}) = p(z_{ik} = z \,|\, \boldsymbol{\Theta}_{ik}, z \neq 0)\, \pi_0 \;+\; p(z_{ik} = z \,|\, \boldsymbol{\Theta}_{ik}, z = 0)\, (1 - \pi_0). \tag{11}$$

Next, noting that for $z_{ik} \neq 0$ the second term equals zero, the posterior for $z_{ik} \neq 0$ can be written by applying the Bayes theorem to the first term, i.e.,

$$p(z_{ik} = z \,|\, \boldsymbol{\Theta}_{ik}) = \pi_0 \cdot \frac{p(\boldsymbol{\Theta}_{ik} \,|\, z_{ik} = z)\, \pi_z\, \pi_0}{\sum_{z' \neq 0} p(\boldsymbol{\Theta}_{ik} \,|\, z_{ik} = z')\, \pi_{z'}\, \pi_0}. \tag{12}$$

Hence, by virtue of the proposed model in Sect. 2, the speech presence uncertainty is inherently considered when computing the model parameters. The EM iteration given by (9)–(12) can be efficiently implemented using sufficient statistics for GMMs. The reader is referred to our work in [11] where the EM algorithm was implemented in this manner.

Given the estimated parameter set $\mathcal{P}_n$ at each frame n, the sequence $\left[\boldsymbol{\mu}_{z,1}, \boldsymbol{\mu}_{z,2}, \ldots, \boldsymbol{\mu}_{z,n} \right]$ represents the estimated track for the s-th speaker, whereas the posterior probabilities can be used to design TF masks or spatial filters for BSS [7, 9, 11].

4 Robust Counting and Tracking of Sources

The EM iteration, described in Sect. 3, is coupled with an outlier control and a source counting mechanism. The outlier control is based on the likelihood of the incoming data under the current parameter estimates and implicitly imposes smooth

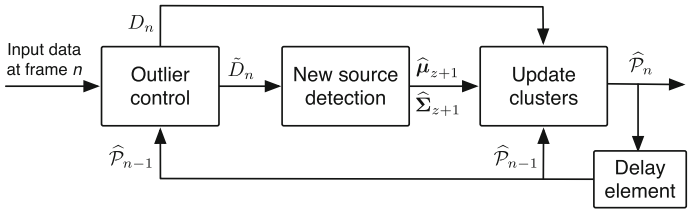

Fig. 1 Diagram of the proposed system. The parameters $\widehat{\mathcal{P}}_{n-1}$ from the previous frame are used to remove outliers and to run an EM iteration at the current frame n

speaker tracks, as described next. A block diagram of the proposed online clustering framework is illustrated in Fig. 1.

4.1 Outlier Control and Smoothness of Estimated Tracks

Noise and reverberation result in outliers which often exceed the number of reliable data points. If there are erroneous SPP estimates in the incoming data, the ML-based criterion in Sect. 3 forces the parameter estimates to fit noisy data that do not accurately represent the speaker locations. Without a motion model or track smoothness constraint, even moderate amount of outliers lead to significant tracking errors. We propose a data-driven approach for trimming the set D_n to contain only positions $\boldsymbol{\Theta}_{ik}$ that belong to a specified confidence region of at least one of the clusters. Given the confidence probability p, the data used in the EM step at frame n needs to satisfy the following, for at least one $z \in [1, S]$

$$(\boldsymbol{\Theta}_{ik} - \boldsymbol{\mu}_{z,n-1})^{\mathrm{T}} \boldsymbol{\Sigma}_{z,n-1}^{-1} (\boldsymbol{\Theta}_{ik} - \boldsymbol{\mu}_{z,n-1}) \leq \Psi_p. \tag{13}$$

As the quadratic form on the left hand side in (13) follows a Chi-squared distribution with two degrees of freedom, Ψ_p and p are related as $p = 1 - e^{-\frac{\Psi_p}{2}}$ [16]. The data trimming step implicitly imposes smoothness of the speaker tracks and assists the detection of new speakers, as described in Sect. 4.2.

Note that the points with low SPP, and hence a low impact on $\widehat{\mathcal{P}}_n$, can be removed from D_n without evaluating (13). In this manner, the storage and computation complexity are notably reduced, as due to speech sparsity the number of low SPP points is significant. Thresholding based on voice activity detection is often a required step in acoustic source clustering and tracking [6, 11, 12]. In this work, due to the incorporated SPP and outlier control, this step is optional and is done only to reduce the computational cost.

4.2 Removing and Adding Speakers

Removing speakers. Inspired by the sparse EM variant [17], where at frame n only the Gaussian components with large responsibility for observing the current data D_n are updated, we propose to first update the mixture coefficients π_z using the data D_n, update the mean and covariance according to (10) only for the components z for which $\pi_z > \pi_{\text{thr}}$, and freeze the means and covariances otherwise. The frozen parameters indicate inactive sources that are removed if their parameters are frozen longer than a pre-defined number of frames L_{frz}. The value L_{frz} is often referred to as *time-to-live* in tracking literature. In this work, L_{frz} was fixed to 60 frames, corresponding to 1.9 s. Alternatively, L_{frz} can be computed online based on the predicted travelled distance within a silent period. If a speaker travels a large distance without acoustic activity, the track can not be recovered when acoustic activity is resumed. This behavior is due to the smoothness requirement implicitly imposed by (13). Instead, the speaker is removed after L_{frz} frames, and promptly re-detected as a new speaker when activity is resumed.

Adding speakers. Let us denote by $\widetilde{D}_n$ the set of all the points that were removed from D_n by (13), and for which the SPP satisfies $\pi_0 > p_{sp}$. Assuming that there is no new speaker at frame n, the cardinality $|\widetilde{D}_n|$ is low as all points with high SPP are modeled by the current GMM and remain in the set D_n. On the contrary, if a new speaker appears, a cluster of points is present in $\widetilde{D}_n$. To verify the existence of a new cluster we first take the maximum $\theta_{a,\max}$ of the DOA histogram for each array a, computed using the TF bins in $\widetilde{D}_n$. Consider a set $\widetilde{D}'_n \subset \widetilde{D}_n$ such that for all TF bins (i,k) corresponding to the points in $\widetilde{D}'_n$, the following holds

$$\widetilde{D}'_n = \{\boldsymbol{\Theta}_{ik} \,|\, |\theta_{a,\max} - \theta_{a,ik}| < \Delta_\theta, \forall a\}. \tag{14}$$

The threshold Δ_θ is chosen such that the intersection region satisfying (14) is sufficiently small, so that a large cardinality $|\widetilde{D}'_n|$ indicates, with high probability, a new source activity in that region. Denote by $\boldsymbol{\mu}_{z+1}$ the mean of the data points in $\widetilde{D}'_n$. To ensure that the newly detected source does not model an already existing one we impose a limit on the minimum distance between two detected sources, leading to the last condition for adding a source

$$||\boldsymbol{\mu}_{z+1} - \boldsymbol{\mu}_s||_2 > d_{\min}, \forall s \wedge |\widetilde{D}'_n| \geq \xi \tag{15}$$

where ξ is the minimum number of points required to declare a new speaker. If (15) is satisfied, a new Gaussian is initialized with mean $\boldsymbol{\mu}_{z+1}$ and a scaled identity covariance, and the standard EM iteration is executed with the new model. Note that although only one new speaker per frame can be detected, there is no constraint on the number of new speakers that appear in the same frame. Due to the outlier-based speaker detection, both speakers will be detected, however with at least one frame delay between the two detections.

In addition, the following check is continuously exectuted $||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||_2 < d_{\text{mrg}}$ so that clusters of speaker i and j are merged in the case of crossing tracks. After the speakers move away from each other, the two separate clusters and the corresponding trajectories are recovered, as shown in the results. Although in certain applications, it might be desirable to maintain separate speaker tracks regardless of crossing, the proposed system in this work was developed in view of the signal extraction applications using spatial filters, where speakers with small inter-speaker distances cannot be separated by the spatial filter as separate sound sources. Therefore, we choose to merge the the tracks of crossing trajectories.

5 Performance Evaluation

The proposed algorithm was evaluated in a simulated $6 \times 5 \times 3$ m room. Clean speech signals were convolved with room impulse responses for moving sources using the software in [19]. Diffuse babble noise [18] and uncorrelated sensor noise were added to the speech signals. The STFT frame length was 64 ms, with 50 % overlap, at a sampling rate of 16 kHz. Three uniform circular arrays of diameter 2.9 cm and three omnidirectional microphones per array were employed. All relevant processing parameters are summarized in Table 1. Scenarios with different reverberation times T_{60}, noise levels, number of sources, and motion patterns were examined. The system is tested in dynamic situations with appearance of new sources, speech pauses, and sources with crossing trajectories.

Experiment 1. In this experiment, we tested the tracker for a fixed number of moving speakers. We started the algorithm with an unknown number of speakers and once all were detected, the RMSE between the true and the estimated tracks was computed. The results in Table 2 are averaged over time frames, over speakers, and over three scenarios with different motion patterns. Two, three, and four simultaneously active speakers were tracked for signal-to-noise ratios of 11 and 21 dB and speeds of 0.23 and 0.34 m/s. It can be observed that while the system is robust to noise and reverberation, accuracy decreases for faster speaker movement. The sensitivity to speed can be controlled by the number of frames L that constitute

Table 1 Parameters used in the implementation

L	Δ_θ	p_{sp}	p	π_{thr}	L_{frz}	ξ	d_{min}	d_{mrg}
30	10°	0.85	0.95	0.07	60	30	0.5	0.3

Table 2 Average root mean squared error in meters between the true speaker location and the mean of the respective Gaussian

SNR (dB)	speed (m/s)	$T_{60} = 200$ ms			$T_{60} = 400$ ms		
		two	three	four	two	three	four
21	0.23	0.16	0.14	0.12	0.18	0.15	0.22
21	0.34	0.23	0.17	0.17	0.24	0.22	0.25
11	0.23	0.16	0.12	0.19	0.20	0.16	0.24
11	0.34	0.25	0.23	0.24	0.26	0.26	0.26

the short-term data. Smaller L allows for faster adaptation suited for higher speeds, however leads to quickly varying parameter estimates and less smooth tracks. Experiments showed that $L \in [25, 35]$ offers a good tradeoff between responsiveness and stability.

Experiment 2. We tested the tracker in several challenging scenarios which often occur in practice. In Fig. 2, we illustrate a four-speaker example with $T_{60} = 200$ ms and velocity 0.34 m/s for each speaker. Snapshot II shows a detection of a new speaker, where even in multitalk scenarios, a new speaker is detected in less than 0.5 s after the first activity. Snapshots I and III show stabilized clusters of three and four speakers, respectively. Finally, snapshot IV is taken at the frame where the third speaker is discarded, and only two Gaussians track the remaining two speakers.

Experiment 3. In the last experiment, the tracker was tested in a triple-talk scenario when the trajectories of two speakers cross. The tracking result is visualized in Fig. 3, where the x and y coordinates of the true and estimated tracks are plotted across time. The crossing happens around second 5 on the time axis, where it is visible that for a short period the two speakers are represented by a single Gaussian component. When speakers split, around second 7, the tracker promptly detects a new speaker, assigns a new Gaussian distribution to it, and continues to track the three speakers.

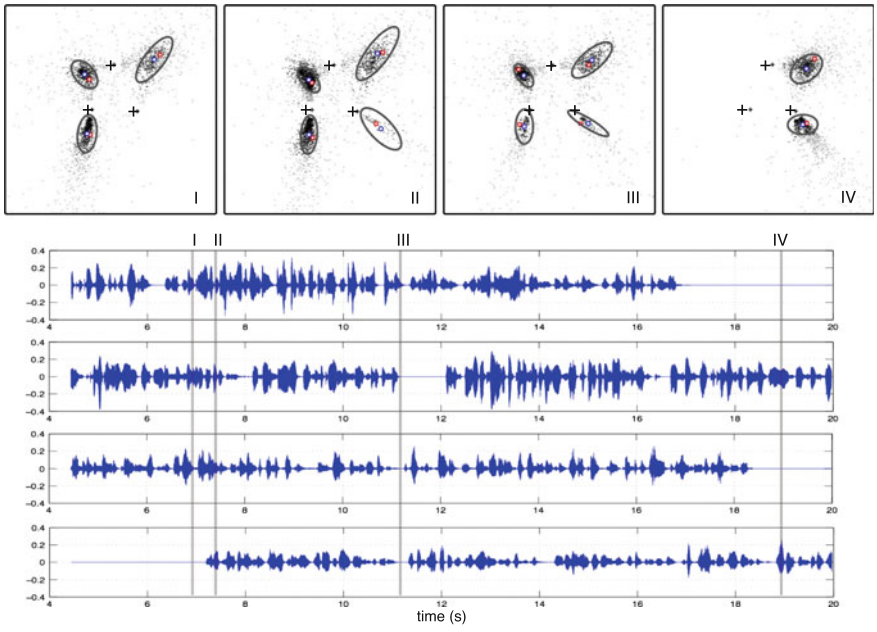

Fig. 2 Snapshots of the tracking at different times and the signals of each speaker. The number at the *right bottom corner* of each snapshot relates to the signal segment as indicated by the markers. SNR 21 dB, T_{60} = 200 ms, velocity 0.34 m/s. The *black dots* denote the points in D_n that are used in the current EM iteration, whereas the *gray dots* denote the discarded points $\widetilde{D}_n$. The *plus signs* denote the microphone arrays

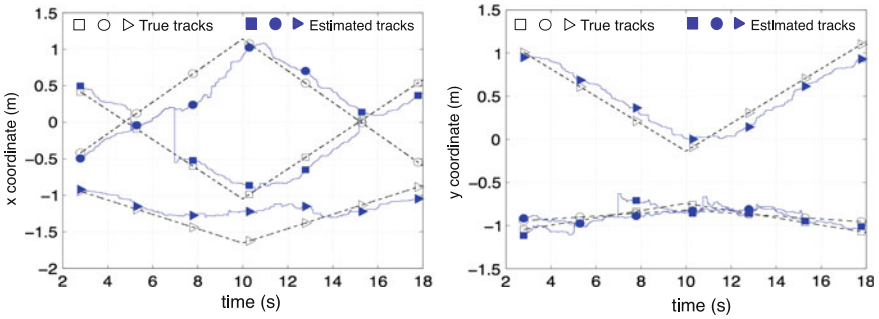

Fig. 3 Visualizing the speaker tracks for a scenario with crossing trajectories. SNR = 21 dB, T_{60} = 200 ms, velocity 0.34 m/s

6 Conclusions

A maximum likelihood framework for multi-speaker detection and tracking was proposed, based on clustering of narrowband position estimates. The position estimates are obtained by using at least two distributed arrays. Speech presence uncertainty and outlier control are incorporated in a unified manner, resulting in a system that is robust to noise and reverberation, estimates the number of speakers online, and allows for track recovery even in situations where the sources have crossing trajectories. As a by-product of the clustering-based tracking, each TF bin is classified to the dominant source providing means to design data-dependent spatial filters for blind source separation.

References

1. Fallon FC, Godsill JS (2012) Acoustic source localization and tracking of a time-varying number of speakers. IEEE Trans Audio Speech Lang Process 20:1409–1415
2. Gehrig T, Klee U, McDonough J, Ikbal S, Wölfel M, Fügen C (2006) Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters. Interspeech
3. Yilmaz O, Rickard S (2004) Blind separation of speech mixture via time-frequency masking. IEEE Trans Signal Process 52:1830–1847
4. Mandel M, Ellis D, Jebara M (2006) An EM algorithm for localizing multiple sound sources in reveberant environtments Proceedings of Neural Information Processing System
5. Schwartz O, Gannot S (2014) Speaker tracking using recursive EM algorithms. IEEE Trans Audio Speech Lang Process 22:392–402
6. Loesch B, Yang B (2008) Source number estimation and clustering for underdetermined blind source separation. Proceedings of international workshop on acoustic signal enhancement
7. Madhu N, Martin R (2011) A versatile framework for speaker separation using a model-based speaker localization approach. IEEE Trans Audio Speech Lang Process 19:1900–1912
8. Taseska M, Habets EAP (2014) Informed spatial filtering with distributed arrays. IEEE Trans Audio Speech Lang Process 22:1195–1207

9. Souden M, Kinoshita K, Delcroix M, Nakatani T (2014) Location feature integration for clustering-based speech separation in distributed microphone arrays. IEEE Trans Audio Speech Lang Process 22:354–367
10. Plinge A, Fink GA (2014) Multi-speaker tracking using multiple distributed microphone arrays. Proceedings of IEEE international conference on acoustics, speech and signal processing
11. Taseska M, Habets EAP (2013) An online EM algorithm for source extraction using distributed microphone arrays. Proceedings of European signal processing conference
12. Lehmann EA, Johansson AM (2007) Particle filter with integrated voice activity detection for acoustic source tracking. EURASIP J Appl Signal Process
13. Araki S, Sawada H, Mukai SMR (2006) DOA estimation for multiple sparse sources with normalized observation vector clustering. Proceedings of IEEE international conference on acoustics, speech and signal processing
14. Taseska M, Habets EAP (2012) MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator. Proceedings of international workshop acoustic signal enhancement
15. Demptster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Statist Soc 39:1–38
16. Bar-Shalom Y (2001) Estimation with applications to tracking and navigation. Wiley & Sons
17. Neal RM, Hinton GE (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. Learning in Graphical Models, Kluwer Academic Publishers, p 355–368
18. Habets EAP, Gannot S (2007) MATLAB implementation for: generating sensor signals in isotropic noise fields. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/noise-generators
19. Habets EAP Available: http://www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator