MULTIPLE SOURCE TRACKING BY SEQUENTIAL POSTERIOR KERNEL DENSITY ESTIMATION THROUGH GSCT

Alessio Brutti, Francesco Nesta

Fondazione Bruno Kessler-Irst
via Sommarive 18, 38123 Trento, Italy
{brutti|nesta}@fbk.eu

ABSTRACT

This paper introduces a novel framework for tracking the TDOAs of multiple sources whose acoustic activities overlap in time. Assuming the number of sources to be known, multiple disjoint particle filters estimate the posterior kernel density of the propagation parameters. An approximated instantaneous kernel density is provided through the Generalized State Coherence Transform, which improves the source interference rejection across the dimensions using a frequency-normalized non-linearity. Results obtained from an experimental evaluation on synthetic data show that the proposed framework enables localization and tracking of bidimensional TDOAs for 7 competitive sources recorded under reverberant conditions and with only 3 microphones.

1. INTRODUCTION

Accurate estimation of the time difference characterizing the arrivals of acoustic waves at two microphones is a crucial step in many speech related applications [1]. Recently, scenarios where several audio sources are active at the same time have received an increasing interest from the scientific community calling for effective solutions for tracking of multiple targets, e.g. audio-video conferencing, home automation, etc.. In the context of the Blind Source Separation (BSS) effective methods for localization of multiple sources have been proposed as the Generalized State Coherence Transform (GSCT) [2], which was also shown to be a robust solver for the permutation problem of the frequency-domain BSS [3]. Observation vectors, describing the acoustic propagation of the sources, are generated from the demixing matrices estimated by applying a complex-valued Independent Component Analysis (ICA) to different frequency bins. By a non-linear integration of the multidimensional phase coherence of the observed vectors, the GSCT generates multimodal and multivariate likelihoods where the maxima correspond to the TDOA vectors that best fit the observed propagation parameters of the sources. With a proper choice of the non-linearity the GSCT approximates the kernel density estimation of the multidimensional time-delay propagation [4]. However, as the dimensionality increases an exhaustive search for the maxima in the multivariate likelihoods becomes inefficient and computationally infeasible. Furthermore, the GSCT is able to give an instantaneous picture of the acoustic propagation but cannot explicitly deal with time-varying mixing conditions.

Sequential Bayesian methods and Particle Filters (PF) represent an effective solution to the acoustic tracking task by evaluating the posterior probability density function (PDF) of the target state based on all available measurements [5, 6]. Approaches combining BSS and PF have been recently presented in [7] and [8]. Following this direction, we propose a general framework for multiple source tracking which combines the robustness of the GSCT with the flexibility of PF approaches. Although the multimodal nature of GSCT suggests the adoption of multidimensional state spaces, sources are tracked in a disjoint fashion so that curse of dimensionality is avoided. A kernel-like measurement exclusion approach is adopted to avoid the collapse of different filters on the same target. The framework operates in the TDOA domain in order to define a source tracking method which is as general as possible. In fact, if the microphone geometry is known the source spatial locations can be derived from the estimated TDOAs. On the other hand the TDOA estimates can be used to approximate the mixing parameters of the sources and opportunely drive BSS algorithms. Note that in this work we assume that the number of active sources is known and focus instead on accuracy and robustness of tracking.

This paper is organized as follows. Section 2 describes the GSCT technique while Section 3 introduces the sequential Bayesian tracking and the adopted PF implementation. Section 4 describes the experimental set up and the obtained results. Finally, Section 5 concludes the paper with final remarks and possible future research directions.

2. GENERALIZED STATE COHERENCE TRANSFORM AND FREQUENCY-DOMAIN BSS

Let us assume that N sources are recorded by an array of M elements and indicate with h_{mn} the generic impulse response between the *m-th* microphone and *n-th* source. The sampled signal acquired by the *m-th* microphone is the result of the combination of the filtered versions of all the signals emitted by the sources:

$$y_m(l) = \sum_{n=0}^{N-1} h_{mn}(l) * s_n(l) + \eta_m(l) \tag{1}$$

where $\eta_m(l)$ is the environmental noise, l denotes sample index and $*$ indicates convolution. By taking the Discrete Fourier Transform (DFT) of the impulse responses we define the acoustic propagation observation vector of the *n-th* source at the *k-th* frequency bin as:

$$\begin{aligned} \bar{\mathbf{r}}_{nk} &= [\bar{r}_{nk}^1, \cdots, \bar{r}_{nk}^P] \\ \bar{r}_{nk}^p &= \frac{\{\mathbf{H}(k)\}_{a_p n}}{\{\mathbf{H}(k)\}_{b_p n}} \Big/ \left| \frac{\{\mathbf{H}(k)\}_{a_p n}}{\{\mathbf{H}(k)\}_{b_p n}} \right| \end{aligned} \tag{2}$$

where $\mathbf{H}(k)$ is the $M \times N$ mixing matrix corresponding to the transfer function between sources and microphones, a_p and b_p indicate the microphone indexes of a generic *p-th* microphone pair and P is the total number of used microphone pairs. The ratios in eq. 2 can be modeled as:

$$\bar{r}_{nk}^p = e^{-j2\pi f_k \Delta_n^p(k)} \tag{3}$$

where f_k is the real frequency corresponding to the *k-th* frequency bin and $\Delta_n^p(k)$ is the time-delay of the *n-th* source related to the microphone pair p at the *k-th* bin, which is frequency invariant in anechoic environments. If $N = M$ an estimation of $\mathbf{H}(k)$ can be obtained through a complex-valued ICA algorithm, independently applied to the time-series of each frequency bin, resulting from the Short-Time Fourier Transform (STFT) of the recorded mixtures [9]. Taking the inverse of each ICA demixing matrix $\mathbf{W}(k)$, a scaled and permuted estimation of $\mathbf{H}(k)$ is obtained as $\mathbf{W}^{-1}(k) = \overline{\mathbf{H}}(k)\mathbf{\Pi}(k)\mathbf{\Lambda}(k)$, where $\mathbf{\Pi}(k)$ and $\mathbf{\Lambda}(k)$ are a generic permutation and scaling matrices and $\overline{\mathbf{H}}$ is the underlying mixing matrix estimated by the ICA algorithm. Thanks to the scaling invariance, eq. 2 can be computed substituting $\mathbf{W}^{-1}(k)$ in place of $\mathbf{H}(k)$. Note that however, the permutation ambiguity is not solved. In fact the $\bar{\mathbf{r}}_{nk}$ computed through $\mathbf{W}^{-1}(k)$ does not necessarily represent the acoustic propagation of the same source for each frequency bin.

In [2] it was shown that even if $N > M$ (i.e. the separation problem is underdetermined) it may be assumed that the dominance of the sources is sparse in the time-frequency domain. As a consequence, different ICA adaptations may be run at each frequency bin k and several time-segments to obtain a set of observation vectors which gives a full description of the acoustic propagation of all the sources. In this paper we focus only on the frequency sparseness and consider the GSCT obtained using observations derived from a single time-segment. As explained in the next session, temporal redundancy will be exploited through the use of particle filtering.

Once $\overline{\mathbf{r}}_{nk}$ have been estimated, a likelihood of multidimensional TDOAs can be obtained through the GSCT as:

$$\mathrm{GSCT}(\mathbf{T}) = \sum_k \sum_n g[D(\overline{\mathbf{r}}_{nk},\, \mathbf{c}(k,\mathbf{T}))] \tag{4}$$

where $g[\cdot]$ is a generic non-linear decreasing monotonic function, $D(\cdot,\cdot)$ is a distance metric and $\mathbf{c}(k,\mathbf{T})$ is the ideal propagation model defined as:

$$\mathbf{c}(k,\mathbf{T}) = [e^{-j2\pi f_k \tau^1}, \cdots, e^{-j2\pi f_k \tau^P}] \tag{5}$$

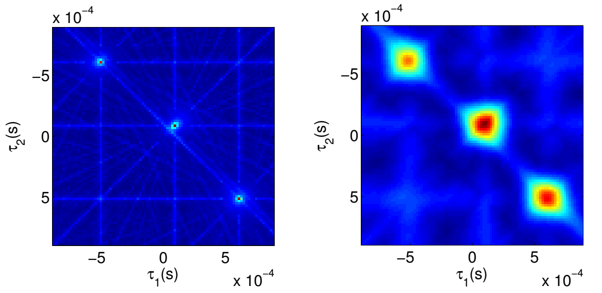
where $\mathbf{T} = [\tau^1, \cdots, \tau^P]$ is the vector of time-delay parameters. According to eq. 4, GSCT has the shape of a kernel estimator and the non-linearity $g[\cdot]$ may be optimized by a proper statistical model for $\overline{\mathbf{r}}_{nk}$. Assuming that the reverberation is diffuse and that the sound propagating over the direct-path prevails, the ratios in eq. 2 may be considered a transformed sample of normally distributed time-delays, centered on the true TDOAs. Since the ratios describe the acoustic propagation of multiple sources, the underlying distribution of the corresponding time-delays is expected to be multimodal. If the distribution is approximated with a mixture of spherical Gaussians of same variance, up to a scaling factor, a kernel density estimation is obtained as:

$$f(\mathbf{T}) = \sum_k \sum_n e^{-\frac{||\mathbf{T}-\mathbf{T}_{nk}||^2}{2h^2}} \tag{6}$$

where h is the kernel bandwidth and $\mathbf{T}_{nk}$ is the time-delay observation vector corresponding to $\overline{\mathbf{r}}_{nk}$. Because of the phase wrapping, $\mathbf{T}_{nk}$ can not be unambiguously determined by $\overline{\mathbf{r}}_{nk}$. A way to circumvent such ambiguity is to consider all the possible TDOA vectors $\tilde{\mathbf{T}}_{nk}$ according to the inverse transformation of eq. 2, realizing then a mixture of wrapped Gaussian distributions [10]. However, the computation of the resulting likelihood becomes infeasible for $P > 1$ since the amount of the possible TDOA vectors $\tilde{\mathbf{T}}_{nk}$ dramatically increases with the number of phase wrappings (which depends on the frequency and microphone spacing) [4]. It can be shown [4] that $f(\mathbf{T})$ is efficiently approximated in correspondence of the modes of the distribution through the GSCT, when $D(\cdot,\cdot)$ is the Euclidean distance and $g[\cdot]$ is a frequency-dependent Gaussian kernel:

$$f(\mathbf{T}) \simeq \mathrm{GSCT}(\mathbf{T}) = \sum_k \sum_n \frac{1}{2\pi f_k} e^{-\frac{||\mathbf{c}(k,\mathbf{T})-\overline{\mathbf{r}}_{nk}||^2}{(2\pi f_k)^2 \cdot 2h^2}}\,. \tag{7}$$

One of the advantages introduced by eq. 7 compared with the original frequency-independent non linearity proposed in [2] (i.e. $g[x] = 1 - \tanh[\alpha \cdot x]$ in eq. 4) is that the resulting likelihoods have a smoother representation in the time-delay domain. Figure 1 provides a comparison between examples of bidimensional likelihoods obtained for a real-world case of three sources and three microphones whose spacing is 0.3 meters and for moderate reverberation (i.e. $T_{60} = 300ms$). It can be noted that the likelihood computed with the original non-linearity in [2] has a very sparse and sharp representation with three main maxima at points corresponding to the time-delays of each source. On the other hand the GSCT computed with eq. 7 has a much smoother representation and with less artifacts. This is a desired characteristic for the likelihood function in order to be successfully combined with a particle filtering approach. It is known that likelihoods very sharp and with many local maxima do not marry well with a Bayesian framework.



(a) GSCT computed as in [2] (b) GSCT computed as in eq. 7

Figure 1: Comparison between bidimensional GSCT likelihoods for $N = 3$, $M = 3$ and $P = 2$.

In general the PDF of the time-delays may be approximated by a sparse mixture of Gaussians, i.e. the Gaussian components related to the propagation of different sources slightly overlap to each other. According to this assumption if there is no spatial aliasing, with an appropriate choice of the bandwidth the kernel is able to isolate the contribute of state observations related to different sources. The resulting likelihood is then expected to be equivalent to the sum of the single distributions defined by each Gaussian. If the bandwidth is overestimated the resolution of the kernel might be not sufficiently high to isolate observation vectors related to different sources. Specifically, the choice of the bandwidth becomes crucial when $P > 1$ since the likelihood may be enhanced at point corresponding to *ghost* locations, i.e. the cross points between the modes of the density marginalised in each dimension.

The theoretical reasons for which the likelihood is enhanced in these points lies in the interpretation of the statistical model considered for the underlying time-delays [4][11]. If there is no spatial aliasing the time-delay propagation of each source is modeled by a single Gaussian, whose variance depends on the reverberation and the mean on the source location with respect to the microphones. In these conditions the degree of sparseness of the Gaussian components depends on the spatial diversity of the source location and on the reverberation. In the presence of spatial aliasing, due to the wrapping uncertainty, the time-delay density related to the propagation of each individual source is equivalently represented by a mixture of wrapped Gaussians [10] and eq. 7 would approximates the wrapped kernel density. Consequently, the resulting variance of each individual time-delay density is larger than the *aliasing-free* case and the sparseness assumption, which is on the basis of the successfulness of the kernel, does not hold anymore. In this case the bandwidth needs to be reduced, theoretically till the limit case where the kernel becomes an impulse centered in $\mathbf{T}$. However, due to the low data density, the smaller is the kernel the noisier is the resulting likelihood.

An alternative way to tackle this problem is to improve eq. 7 in order to explicitly take into account the interference problem that occurs if the bandwidth is overestimated [11]. An effective improvement is obtained modifying eq. 7 as follows:

$$\tilde{f}(\mathbf{T}) = \sum_k \max_n \left[\frac{1}{2\pi f_k} e^{-\frac{D(\overline{\mathbf{r}}_{nk},\mathbf{c}(k,\mathbf{T}))^2}{(2\pi f_k)^2 \cdot 2h^2}} \right] \tag{8}$$

In our recent contribution we proposed $D(\cdot,\cdot)$ to be a generic reshaped γ-norm in order to obtain a better reduction of the source interference in the multidimensional space, consequently reducing the likelihood at the ghost locations [11]. In this work we choose $D(\cdot,\cdot)$ to be the Chebyshev norm, i.e. $||\mathbf{c}(k,\mathbf{T}) - \overline{\mathbf{r}}_{nk}||_\infty$, which is sufficient to give stable performance for the considered microphone spacings. A more detailed analysis of the meaning of eq. 8 and of its behavior in different conditions is provided in [11].

With the above kernel definitions the GSCT likelihood becomes an approximated kernel density and can be directly used in a Bayesian statistical framework for source localization and tracking. Moreover, the source locations can be estimated from the full likelihood without modeling the joint conditional PDF because we assume that sources dominate different directions of propagation, and hence the corresponding time-delays are distributed in different hyperspheres of the whole multidimensional TDOA space.

3. SEQUENTIAL BAYESIAN TRACKING

In a generative Bayesian tracking framework the target states $\mathbf{x}_t$ at time t, in our case the set of TDOAs, are obtained by evaluating the PDF $p(\mathbf{x}_t|\mathbf{z}_{1:t})$, where $\mathbf{z}_{1:t} = \{\mathbf{z}_1, \dots, \mathbf{z}_t\}$ is the set of all the available observations up to time t. Assuming that $p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})$ is known and that the target states evolve as a 1-st order Markov chain, the PDF can be iteratively computed by adopting the two following prediction-update steps [6, 12]:

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) &= \int p(\mathbf{x}_t|\mathbf{x}_{t-1})\, p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})\, d\mathbf{x}_{t-1} \\ p(\mathbf{x}_t|\mathbf{z}_{1:t}) &\propto p(\mathbf{z}_t|\mathbf{x}_t)\, p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) \end{aligned} \tag{9}$$

The above recursion is completely described by the dynamic model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, that describes the state evolution over time, the observation likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$ and the initial distribution $p(\mathbf{x}_0)$.
When assumptions on Gaussianity and linearity of the problem do not hold a closed-form derivation of the involved distributions is not feasible. A solution is offered by PF methods that represent the PDF at time t through I weights $w_t^{(i)}$ $i = 1, \dots, I$ associated to a set of samples of the state space (particles) $x_t^{(i)}$. Among several available implementations, whose main differences lie in the adopted importance distribution and resampling strategy, Sampling Importance Resampling (SIR) has been shown to marry well with tracking of multiple acoustic sources [13, 14].

3.1 PF Implementation

In the proposed tracking framework, the target state is defined in the TDOA domain $\mathbf{x}_t = \left[\tau^1, \cdots, \tau^P\right]^T$ and particle weights are obtained evaluating the GSCT function of eq. 8.
When N targets are being tracked the above Bayesian framework can be adopted by considering the joint state that combines each single target state $\mathbf{x}_t = \left[\mathbf{x}_t^1, \cdots, \mathbf{x}_t^N\right]$. Unfortunately, joint tracking suffers from the so called "curse of dimensionality": the number of particles needed as well as the computational costs grow exponentially with the number of targets. This problem can be circumvented tracking the sources in a disjoint fashion, i.e. instantiating a filter for each target. This can be done if the PDF is separable:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \prod_n p(\mathbf{x}_t^n|\mathbf{z}_{1:t-1}) \tag{10}$$

so that posteriors are approximated by independent sets of weighted samples $\left(x_{t-1}^{(ni)}, w_{t-1}^{(ni)}\right)$. Under this assumption, the state evolution of each target is modeled, independently, as a Brownian movement by adding Gaussian noise to the previous particle positions:

$$\mathbf{x}_t^n \sim \mathcal{N}\left(\mathbf{x}_{t-1}^n, \sigma_x^n\right)$$

As for the observation likelihood, thanks to the kernel approximation of GSCT, it can in turn be assumed to be locally separable, allowing a sequential independent update of the weights of each filter. However, since filters evolve independently, a data exclusion process is necessary to encourage filters to track all the targets. Each particle weight is thus computed using a mask whose aim is to remove the already tracked modes in the likelihood:

$$\begin{aligned} w_t^{(ni)} &= p\left(\mathbf{z}_t|x_t^{(ni)}\right) \\ &= \psi\left(\tilde{f}\left(x_t^{(ni)}\right)\right) \cdot \Upsilon(x_t^{(ni)}, \overline{\mathbf{x}}_t) \end{aligned} \tag{11}$$

where $\psi(\cdot)$ is a generic contrast function (e.g. $\psi(x) = x^3$) and $\overline{\mathbf{x}}_t = \left[\overline{\mathbf{x}}_t^1, \dots, \overline{\mathbf{x}}_t^N\right]$ is the maximum a posteriori estimation of the combined state space:

$$\begin{aligned} \hat{i}_n &= \arg\max_i w^{(ni)} && (12) \\ \overline{\mathbf{x}}_t^n &= x_t^{n\hat{i}_n} && (13) \end{aligned}$$

while $\Upsilon(\cdot)$ is defined as:

$$\Upsilon(x_t^{(ni)}, \overline{\mathbf{x}}_t) = \begin{cases} 0 & \text{if } \min_{\nu \neq n} |x_t^{(ni)} - \overline{\mathbf{x}}_t^{\nu}| < h; \\ 1 & \text{otherwise.} \end{cases} \tag{14}$$

The mask $\Upsilon(\cdot,\cdot)$ implicitly solves the data association problem using a minimum distance criterion. The function $\psi(\cdot)$ in eq. 11 is adopted to increase the concavity of the GSCT function around the modes and speed up the particle convergence toward the target. Note that the likelihood of the particles $\tilde{f}\left(x_t^{(ni)}\right)$ is dynamically normalized in the range between 0 and 1. Finally, once the particle weights have been normalized in order to sum up to the unity, the target state estimation $\hat{\mathbf{x}}_t^n$ is obtained as the expectation of the PDF approximated by the *n-th* set of weighted particles:

$$\hat{\mathbf{x}}_t^n = \int \mathbf{x}_t^n p(\mathbf{x}_t^n|\mathbf{z}_{1:t}) d\mathbf{x}_t^n = \sum_i w_t^{(ni)} x_t^{(ni)} \tag{15}$$

Note that $\tilde{f}\left(x_t^{(ni)}\right)$ is the GSCT function computed as in eq. 8 using the observation vectors $\overline{\mathbf{r}}_{nk}$ estimated by ICA at the current time t. In our implementation t represents an index for discrete time instants since, as explained in section 4, the estimation of the observation vectors is performed applying a batch off-line ICA implementation over sliding segments.

It is worth underlining that even though the propagation model should ensure spatio-temporal continuity, trajectory swaps may occur between the filters, in particular when the speech activities of some sources interrupt for some frames. The association ambiguity between targets and filters is known as "external permutations" in the BSS community. Although this problem is beyond the goal of this paper, trajectory swaps can be mitigated by finding, at each iteration, the permutation that maximizes the continuity between contiguous estimated state vectors.

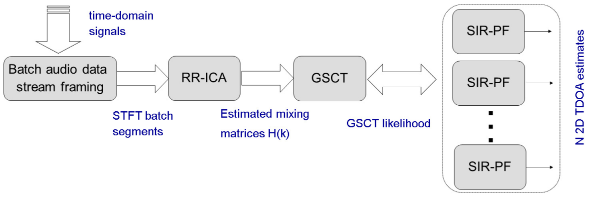


Figure 2: Overall block diagram of the proposed tracking framework.

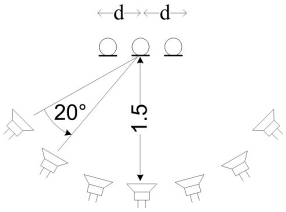


Figure 3: The experimental setup. 3 microphones capture the signals emitted by up to 7 sources positioned along a circle whose radius is 1.5 meters. The angular distance between sources is 20°.

4. EXPERIMENTAL EVALUATION

The proposed approach is evaluated on a synthetic data set generated with the image method [15] in a 5×8m room with reverberation time 0.5s. Very challenging scenarios are taken into account

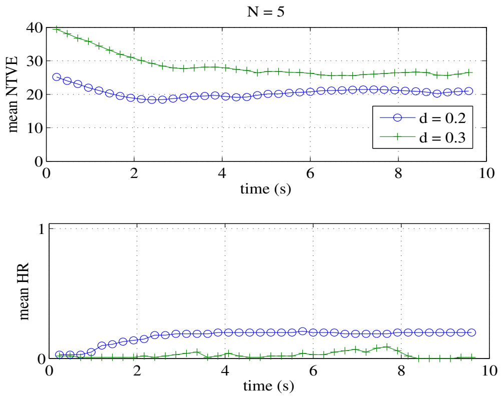


Figure 4: Mean NTVE and mean HR when $N = 5$ and the inter-microphone distance d is 0.2 and 0.3 meters. The GSCT is computed as in [2].

where only three microphones are used to acquire the speech signals emitted by several omnidirectional sources whose acoustic activities overlap in time. For each scenario several runs are performed to account for the stochastic nature of the tracking framework.

4.1 Algorithm description

Figure 2 sketches a block diagram of the data flow in the proposed tracking scheme. Time-domain signals, sampled at $f_s = 16$kHz are transformed by an STFT analysis with windows of 1024 points shifted of 256 samples. The Recursively Regularized implementation of ICA (RR-ICA) [9] is independently applied to sliding blocks of 30 STFT frames, shifted of 15 frames. Note, no link is defined between ICA adaptations of different blocks, so that a batch off-line ICA implementation is realized. Denoting the maximum allowable TDOA as $\tau_{\max} = \frac{c}{d}$, where c and d are the sound speed and the microphone spacing respectively, the bandwidth h is fixed to $h = \frac{\tau_{\max}}{10}$.
As far as the particle filter is concerned, 700 particles are allocated for each filter while $\sigma_x^n = \frac{\tau_{\max}}{20}$ in the dynamical model. In order to ensure that filters keep monitoring the whole state space, the 20% of particles with lowest weigths are propagated with a higher speed ($\sigma_x^n = \frac{\tau_{\max}}{3}$) with the intent to alleviate the effects of potential local maxima in the likelihood. Filters are initially instantiated randomly in the state space (i.e. $p(\mathbf{x}_0^n)$ is uniform).

4.2 Metrics

Performance is measured in terms of Normalized TDOA Vector Error (NTVE) which is the euclidean distance between the hypothesized target position $\hat{\mathbf{x}}_t^n$ and the reference one $\tilde{\mathbf{x}}_t^n$, normalized with respect to the maximum distance $2\tau_{\max}\sqrt{P}$:

$$\varepsilon_t^n = 100 \cdot \frac{\|\hat{\mathbf{x}}_t^n - \tilde{\mathbf{x}}_t^n\|}{2\tau_{\max}\sqrt{P}}. \tag{16}$$

Each estimation is further labeled as correct if the localization error is below 5%. This allows the introduction of the *Hit Rate* (HR) as the ratio between the number of correctly tracked targets over the number of targets. We further introduce the *mean NTVE* and *mean HR* which are the average over all the experimental runs of NTVE and HR respectively.
Since the addressed localization task does not foresee any source identification, estimated and reference positions are associated on a minimum-distance maximum-hit criterion: among all the possible permutations, the one with lowest average localization error and highest hit rate is selected.

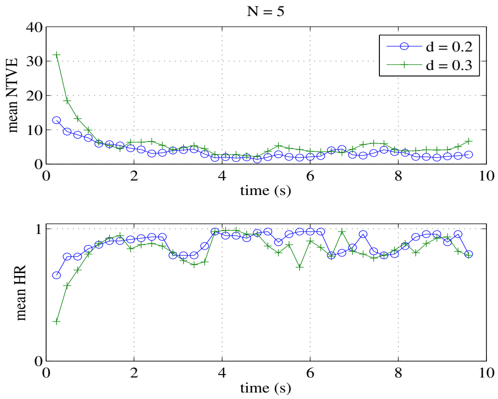


Figure 5: Mean NTVE and mean HR when $N = 5$ and the inter-microphone distance d is 0.2 and 0.3 meters. The GSCT is computed as in eq. 8.

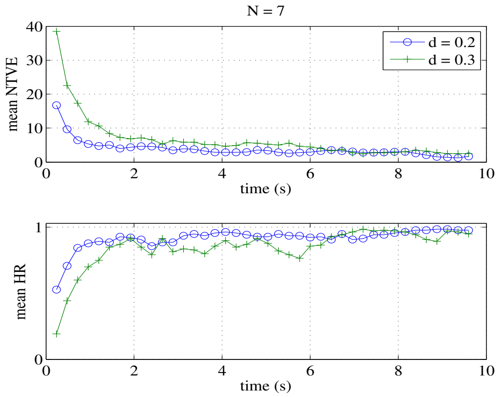


Figure 6: Mean NTVE and mean HR when $N = 7$ and the inter-microphone distance d is 0.2 and 0.3 meters. The GSCT is computed as in eq. 8.

4.3 Static Sources

We consider a first scenario where 5 and 7 static sources are positioned on a circle at 1.5m from the microphones, as depicted in Figure 3. The length of the speech signals is 10 seconds. Figure 4 shows the envelopes of mean NTVE and mean HR for 5 sources considering two different microphone distances (i.e $d = 0.2$ and $d = 0.3$) when the GSCT is computed with the original non-linearity in [2]. As expected, the PF cannot converge correctly due to the sparse and sharp representation of the likelihoods. Figure 5 shows the same envelopes but obtained computing the GSCT as defined in eq. 8. Note that the proposed method converges very quickly to all the target positions and keep track of them ensuring an average error that is around 5% of the maximum error. It is worth noting that the system convergence time and stability are slightly worse when $d = 0.3$. Large microphone distances allow higher spatial resolution and coverage. Unfortunately, as the spatial aliasing is increased, the resulting GSCT modes become sharper, which makes them harder to detect. Moreover the likelihood in correspondence of *ghost* locations could result enhanced, affecting the stability of the system. In order to tackle this problem future experiments will be carried out exploiting the appealing behavior of "Super-Chebishev" metrics, as proposed in [11]. Figure 6 reports the results obtained when 7 sources are active, computing the GSCT as in eq. 8. The plots confirm that the proposed method can accommodate 7 active sources ensuring not only very good spatial detection but also providing very high accuracy and quick convergence. Finally, Figure 7 reports the average performance for 7 sources when $d = 0.2$ and

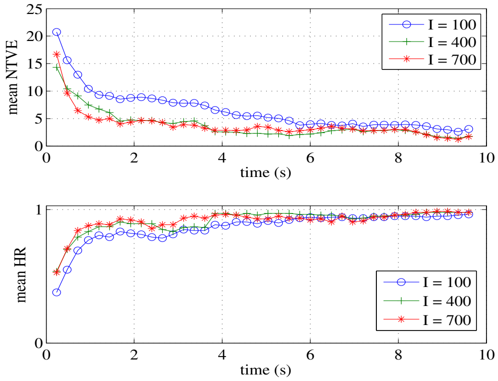


Figure 7: Mean NTVE and mean HR computed for different numbers of particles, when $N = 7$ and the inter-microphone distance d is 0.2. The GSCT is computed as in eq. 8.

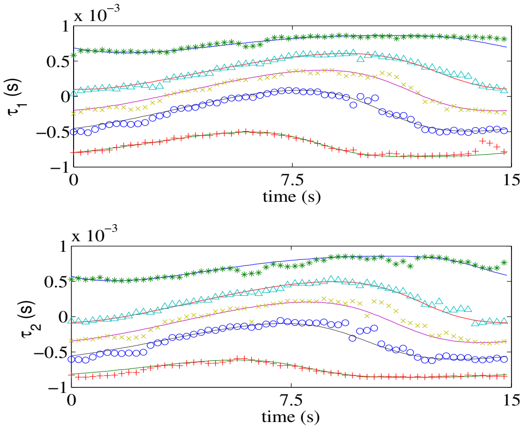


Figure 8: Reference (continuous line) and estimated (points) trajectories on each dimension for 5 sources moving in circles.

different particle numbers are adopted. It is clear how increasing the number of hypotheses (i.e. number of particles) tends to speed up the convergence process. Conversely, the accuracy seems not to take benefit from a large number of particles.

4.4 Moving Sources

Since the PF methods have been devised specifically for source tracking, a further experiment is conducted where 5 sources move along circles with radius 0.5m, completing a full loop. Circle centers are the source positions in Figure 3. The length of the speech signals is approximately 15 seconds. Figure 8 reports the estimated source trajectories against the references in the TDOA domain and shows that the adopted dynamic model is suitable for both static as well as moving sources[1].

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a novel framework for tracking multiple acoustic sources with a limited amount of sensors. Our method consists in a disjoint particle filter implementation whose likelihood is obtained from the GSCT function. Experimental results show that the proposed method is capable of locating up to 7 sources in a reverberant environment using only 3 microphones. Experiments show also that the proposed method can effectively track 5 moving sources in the same reverberant set up.

It is worth noting that the GSCT does not require any particular geometry for the microphone array. On condition that ICA is able to correctly estimate the mixing parameters, multiple distributed microphone pairs can be used in order to increase both resolution and coverage in the spatial domain [16]. Thus, future works will concern possible extensions to distributed microphone array contexts for both source separation, source position and orientation estimation [17].

Concerning trajectory swaps, future investigation will focus on robust spectral features for measurement-target association.

A further open issue is related to the detection of the number of sources which is not addressed in this paper. Solutions based on the so called Track-Before-Detect paradigm already exist and will be taken into account for future developments [13, 14].

[1] A video clip showing a real-time system implementing the described method is available at http://shine.fbk.eu/research/demoGSCT.

REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

[2] F. Nesta and M. Omologo, “Generalized state coherence transform for multidimensional localization of multiple sources,” in *WASPAA*, 2009.

[3] B. Loesch, Nesta F., and B. Yang, “On the robustness of the multidimensional state coherence transform for solving the permutation problem of frequency-domain ICA,” in *ICASSP*, 2010.

[4] F. Nesta and M. Omologo, “Approximated kernel density estimation for multiple TDOA detection,” in *ICASSP*, 2011.

[5] N.J. Gordon, D.J. Salmond, and A.F.M Smith, “Novel approach to nonlinear/non-gaussian bayesian state estimation,” *IEE Proc. of Radar and Signal Processing*, vol. 140, no. 2, 1993.

[6] M. S. Arulampalam et al., “A tutorial on particle filters for on-line nonlinear/non-gaussian bayesian tracking,” *IEEE Trans. on Signal Processing*, vol. 50, no. 2, February 2002.

[7] F. Antonacci et al., “Tracking multiple acoustic sources using particle filtering,” in *EUSIPCO*, 2006.

[8] P. Teng, A. Lombard, and W. Kellermann, “Disambiguation in multidimensional tracking of multiple acoustic sources using a gaussian likelihood criterion,” in *ICASSP*, 2010, pp. 145–148.

[9] F. Nesta, P. Svaizer, and M. Omologo, “Convolutive bss of short mixtures by ica recursively regularized across frequencies,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 624 –639, march 2011.

[10] P. Smaragdis and P. Boufounos, “Learning source trajectories using wrapped-phase hidden markov models,” in *IWAENC*, 2005, pp. 114–117.

[11] F. Nesta and A. Brutti, “Self-clustering non-euclidean kernels for improving the estimation of multidimensional TDOA of multiple sources,” in *HSCMA*, 2011.

[12] D. Ward, E. Lehmann, and R. Williamson, “Particle filtering algorithms for tracking an acoustic source in a reverberant environment,” *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, November 2003.

[13] M. Fallon and S. Godsill, “Multi target acoustic source tracking using track before detect,” in *WASPAA*, 2007, pp. 102 –105.

[14] P. Pertilä and M.S. Hämäläinen, “A track before detect approach for sequential bayesian tracking of multiple speech sources,” in *ICASSP*, 2010, pp. 4974 –4977.

[15] E. Lehmann and A. Johansson, “Prediction of energy decay in room impulse responses simulated with an image-source model,” *Journal of the Acoustical Society of America*, vol. 124(1), pp. 269–277, July 2008.

[16] F. Nesta and M. Omologo, “Cooperative wiener-ica for source localization and separation by distributed microphone arrays,” in *ICASSP*, 2010.

[17] A. Brutti, M. Omologo, and P. Svaizer, “Speaker localization based on Oriented Global Coherence Field,” in *Interspeech*, 2006, pp. 2606–2609.