14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

TRACKING MULTIPLE ACOUSTIC SOURCES USING PARTICLE FILTERING

F. Antonacci, D. Riva, D. Saiu, A. Sarti, M. Tagliasacchi, S. Tubaro

Dipartimento di Elettronica e Informazione - Politecnico di Milano - Italy

ABSTRACT

In this paper we deal with the problem of localizing and tracking multiple acoustic sources by means of microphones pairs. We assume that the propagation takes place in a reverberating environment such as an office room. The problem is tackled by combining two well known techniques. First, for each pair of microphones, source de-mixing is carried out using the TRINICON algorithm. TRINICON exploits the fact that the original sources are statistically independent in order to estimate appropriate de-mixing filters. The impulse responses of such filters exhibit peaks related to the TDOA (Time Difference of Arrival) of each microphones pair. In the second step, such observations are combined using a particle filter with a dynamic model representing the positions and the velocities of the sources. Simulations demonstrate that the proposed system enables to accurately tracking moving acoustic sources in reverberating environments ($\pm 10cm$ in a $5m \times 5m$ room with $T_{60} \leq 0.450s$).

1. INTRODUCTION

The problem of tracking wideband acoustic sources in reverberating environments is relevant in several applications, including seismology, sonar and speech. In this paper we specifically address the case of speech signals. Localizing and tracking multiple speakers talking in the same room can be used, for example, to automatically steer camera sensors in video-conferencing applications.

In the literature, several works address the problem of localizing one acoustic source. In [1] a tutorial review of TDOAs (Time Differences of Arrival) estimation algorithms is presented, with particular emphasis to the case of multi-path propagation typical of reverberating environments. Also in [2], localization techniques based on TDOAs measurements are compared with each other. The LCLS (Linear Correction Least Squares) [3] algorithm outperforms the other approaches in simulations consisting in only one still source. The problem of tracking is further addressed in [4], where TDOA estimates obtained either with a GCC (Generalized Cross Correlation) or AEDA (Adaptive Eigenvalue Decomposition Algorithm) are combined with a particle filter [5] in order to account for the source dynamics.

The proposed solution builds on [4], extending it in order to take into account multiple moving sources in reverberating environments. Unlike [4], a more sophisticated pre-processing stage is needed in order to (partially) separate the sources. In the proposed system, we use the TRINICON algorithm [6], that was shown to perform well in mildly reverberating environments. TRINICON builds upon the assumption that sources are statistically independent and the signals received by the microphones can be modeled as a convolutive mix of the original sources. In a similar way as conventional ICA (Independent Component Analysis) algorithms, TRINICON estimates the de-mixing filters by maximizing the non-gaussianity of the output. An estimate of the TDOAs can be obtained by analyzing the extrema of the de-mixing filters. In [7], a preliminar study shows how the TRINICON algorithm can be used to perform source localization and tracking. Nevertheless, one pair of microphones is used, therefore estimating only the DOAs (Directions Of Arrival) of the two sources.

The work presented was developed within VISNET, a Network of Excellence (http://www.visnet-noe.org), funded by the European Commission

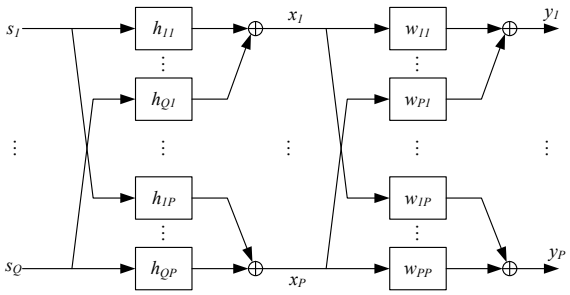

Figure 1: Linear MIMO model for BSS

In this paper we enhance the localization and tracking capability of the algorithm described in [7] by formulating the problem in the state-space. The state is represented by the positions and velocities of the moving sources. In order to infer information about the state, we have access to the observed TDOAs estimated using TRINICON. Since the functional form relating the observation to the state is non-linear, Kalman filtering is not the suitable solution. In order to apply Kalman filtering to non-linear problems, one must recur to linearization about the estimate of the current state. This technique, known as Extended Kalman Filtering, is effective when state probability density function is mono-modal. In our case, underlying statistics are generally multi-modal due to the fact that we are addressing the case of multiple moving sources. Therefore we apply a particle filter that was shown to perform well for the case of non-linear observation models and non-Gaussian statistics [5]. The main idea behind particle filtering is to approximate the posterior PDF of state given the observations by flooding the state-space with particles. For the problem at hand, particles represent potential positions and velocities of the sources and each is characterized by a weight, which measures the likelihood with respect to the observations. We will show how particle filtering can be effectively applied to solve the localization and tracking problem.

The rest of this paper is organized as as follows: in Section 2 we briefly summarize the TRINICON algorithm. In Section 3 we illustrate how the output of the TRINICON algorithm can be used to estimate TDOAs. In Section 4 the proposed algorithm based on particle filtering is presented. Finally, Section 5 and Section 6 illustrate the simulation setup and the experimental results.

2. OVERVIEW OF THE TRINICON ALGORITHM

In order to properly model room reverberations, a convolutive mixing model is generally suitable, as it represents the signal received by each of the P microphones as the sum of delayed and filtered versions of the sources:

$$x_p(n) = \sum_{q=1}^{Q} \sum_{k=0}^{M-1} h_{qp}(k) s_q(n-k)\,, \qquad (1)$$

where Q is the number of active acoustic sources and $h_{qp}(k)$, $k = 0,\ldots,M-1$ denotes the coefficients of the finite impulse response (FIR) filter model from the q-th source to the p-th sensor. In the following, it is assumed that the number of source signals

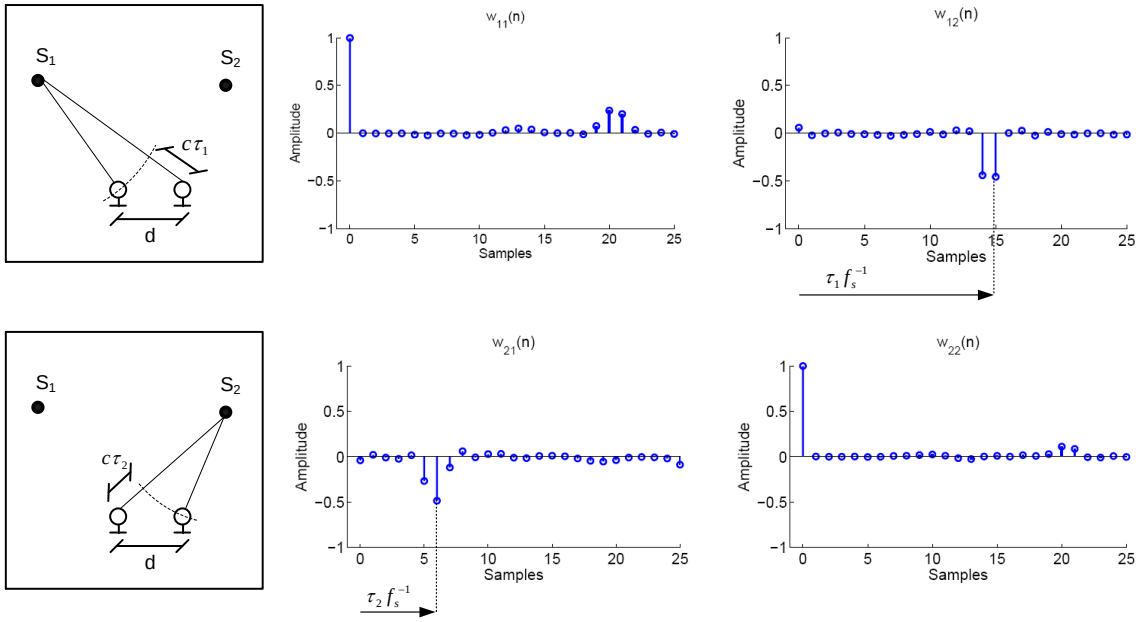

Figure 2: Sources configuration and corresponding de-mixing filters.

equals the number of sensors ($Q = P$). The goal of BSS (Blind Source Separation) is to find a corresponding de-mixing system, as illustrated in Figure 1, where the output signals $y_q(n), q = 1,\ldots,P$ are described by:

$$y_q(n) = \sum_{p=1}^{P} \sum_{k=0}^{L-1} w_{pq}(k) x_p(n-k)\,. \qquad (2)$$

Recently, the problem of BSS for the case of multiple acoustic sources has been addressed in [8], where an iterative algorithm is used to minimize the inter-channel statistical dependency. This algorithm, originally based only on second order statistics, has been extended by the TRINICON framework [6][8]. Following the same guidelines as ICA, TRINICON efficiently exploits the nongaussianity of the sources to improve source separation. The fundamental idea is that the sources are statistically independent and that separation is achieved when the joint inter-channel PDF of the separated signals can be factored out in the product of the PDFs of each channel.

3. SOURCE LOCALIZATION USING TRINICON

The TRINICON algorithm has been successfully used as a preprocessing stage to perform localization of multiple acoustic sources [7]. For the case of $P = Q = 2$ (two sources and two microphones), it is shown that the TDOAs can be estimated from the de-mixing filters w_{pq} as follows:

$$\hat{\tau}_1 = (\arg\max_n |w_{12}(n)| - \arg\max_n |w_{22}(n)|) f_s^{-1}\,, \qquad (3)$$

$$\hat{\tau}_2 = (\arg\max_n |w_{11}(n)| - \arg\max_n |w_{21}(n)|) f_s^{-1}\,, \qquad (4)$$

where f_s denotes the sampling frequency. Knowledge of TDOAs for a single microphones pair allows us to determine a source locus position consistent with the observed TDOAs. Such a locus position is an hyperbola, but it can be confused as a straight line (DOA) when the distance from the microphones pair is much larger than the distance between microphones. In Figure 2 a specific source configuration and the corresponding estimated de-mixing filters are plotted. The locations of the sources can be estimated by triangulating DOAs obtained by two or more microphones pairs.

When only two microphones pairs are used, we need some a priori information in order to achieve a correct localization. This is due to the fact that the TRINICON algorithm suffers from the permutation problem typical of ICA approaches. In fact, one cannot assign a TDOA to a specific source. When TRINICON is run independently on two separate microphones pairs, four TDOAs are estimated, but we do not know which ones belong to the first or to the second source. An illustrative example is depicted in Figure 3: correct source locations are represented with a black dot, while false locations (falling inside the test room) are represented with a crossed circle. For the case of two microphones pairs two DOAs are estimated, therefore obtaining four intersection points by triangulation. When three or more microphones pairs are available, sources can be localized. In this situation, correct and false sources locations are the intersection of, respectively, three and two DOAs.

Equations (3) and (4) show that the information contained in the de-mixing filters is only partially exploited to determine the TDOAs. In other words, only the positions of the global maxima/minima of the filters are needed to achieve source localization, whereas the complete de-mixing filters are used to properly separate the sources.

4. SOURCE TRACKING USING PARTICLE FILTERING

In this section we describe the proposed algorithm that combines the observations obtained with TRINICON with a state-space model capturing the source dynamics. We consider the case of two sources and $M \geq 3$ microphones pairs. According to the notation in [4], we introduce a localization function $f_p^m(\tau)$ ($p = 1,2$ is the source index and $m = 1,\ldots M$ is the index of a microphones pair). Exploiting the de-mixing filters $w_{pq}^m(n)$ estimated using TRINICON at the m-th microphones pair, we define the following localization functions, one for each source:

$$f_1^m(\tau) = (|w_{12}^m(n - arg\max_n(|w_{22}^m(n)|)|) f_s^{-1}\,, \qquad (5)$$

$$f_2^m(\tau) = (|w_{21}^m(n - arg\max_n(|w_{11}^m(n)|)|) f_s^{-1}\,. \qquad (6)$$

The locations of the peaks of the localization functions in (5) and (6) give us an estimation of the TDOAs. Due to reverberations, localization functions may present spurious peaks, thus generating

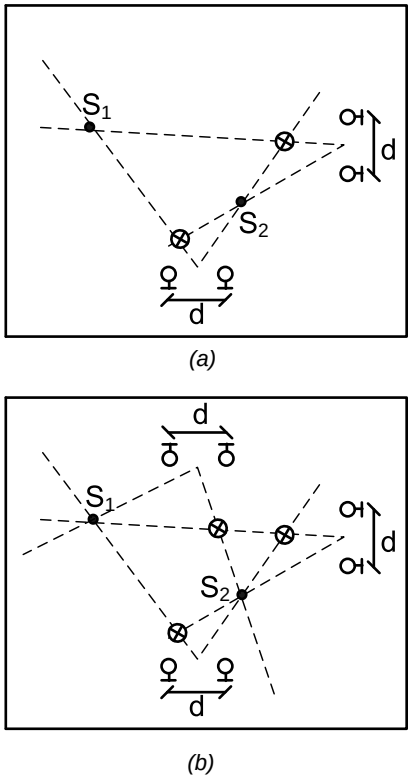

Figure 3: Sources configurations and estimated DOAs. a) Two microphones pairs. b) Three microphones pairs.

outliers. We employ a particle filtering approach in order to filter out spurious peaks by exploiting source dynamics. The underlying idea is that peaks corresponding to the sources, differently from spurious peaks, follow a dynamic model through time.

In our formulation, the state information associated with each particle at time t is described by the $\alpha(t)$ vector, describing the source position and velocity:

$$\alpha(t) = [X(t), Y(t), \dot{X}(t), \dot{Y}(t)]\,. \tag{7}$$

We notice that at this stage, we do not distinguish between the two sources, i.e. we do not assign to the particles a label indicating the source it belongs to. This will be done in a later stage. The update equation of the dynamic system is

$$\alpha(t) = T(\alpha(t-1), \mathbf{n_1}(t))\,, \tag{8}$$

where $\mathbf{n_1}(t)$ is a noise term. At time t, a new measurement $\tau(t)$ becomes available. It is related to the unobserved state $\alpha(t)$ through the equation

$$\tau(t) = S(\alpha(t), \mathbf{n_2}(t))\,, \tag{9}$$

where $\tau(t) = [\tau_1^1, \tau_2^1, \ldots, \tau_1^M, \tau_2^M]^T$ and $\mathbf{n_2}(t)$ is the measurement noise term. For the problem at hand, τ_p^m is related to the state-space model by

$$\begin{aligned}\tau_p^m(t) =& (\sqrt{|X(t)-X_1^m|^2 + |Y(t)-Y_1^m|^2} + \\ &- \sqrt{|X(t)-X_2^m|^2 + |Y(t)-Y_2^m|^2})c + n_p^m(t)\,,\end{aligned} \tag{10}$$

where (X_1^m, Y_1^m) and (X_2^m, Y_2^m) are the Cartesian coordinates of the m-th microphones pair. We can collect all measurements up to time t in the vector $\tau_{1:t} = [\tau(1), ..., \tau(t)]$. We want to estimate the posterior probability density function $p(\alpha(t)|\tau_{1:t})$. No closed-form solution exists for it except for the case where $S(\cdot, \mathbf{n}_2(t))$ and $T(\cdot, \mathbf{n}_1(t))$ are linear and the noise terms $\mathbf{n}_1(t)$ and $\mathbf{n}_2(t)$ are Gaussian. For the latter case, Kalman filtering provides the optimal solution. Equation (10) shows that the function $S(\cdot, \mathbf{n}_2(t))$ is strongly non-linear. This problem can be solved through a linearization of equation (10) around the current estimate of the state-space variables. This technique is normally referred to as Extended Kalman Filter (EKF). EKF linearization is suitable only when $p(\alpha(t)|\tau_{1:t})$ is unimodal. In the case of multi-modal posterior PDF a non-trivial initialization phase is needed. Particle filtering solves this problem in a different way, representing the posterior PDF $p(\alpha_t|y_{1:t})$ through samples (particles). Particle filtering assigns to each particle a weight that is proportional to the likelihood of the observed measurements.

In the following paragraphs we summarize the main steps of proposed algorithm based on particle filtering.

1. *Initialization*: The state space is flooded with uniformly distributed N_s particles $\alpha^i(0)$, $i = 1, \ldots, N_s$. Every particle is assigned a weight $w^i(0)$ equal to $\frac{1}{N_s}$
2. *Dynamic model evolution*: Each particle is shifted according to the following dynamic model (for details see [4]:

$$\dot{X}(t) = a_x \dot{X}(t-1) + b_x F_x(t), \tag{11}$$

$$X(t) = X(t-1) + \Delta T \dot{X}(t), \tag{12}$$

$$a_x = exp(-\beta_x \Delta T), \tag{13}$$

$$b_x = v_x \sqrt{1 - a_x^2}\,, \tag{14}$$

   where $F_x(t)$ is an i.i.d sequence characterized by a Gaussian distribution. The dynamic model described by equations (11)-(14) is extensively used in literature under the name of Langevin model. One problem related to the use of the Langevin dynamic model is the fact that particles can move beyond the perimeter of the room in which the system is installed. In our implementation out of boundaries particles are removed and replaced with new particles in a random location inside the room.
3. *Weight assignment*: At each time instant t, a new weight $w^i(t)$ is assigned to each particle $\alpha^i(t)$depending on the likelihood of the particle given the observed measurements. The weighting operation starts from TDOA estimation provided by the M separators which form our localization system. Every pair of microphones provides K TDOA candidates $\hat{\tau}^{m,k}$, corresponding to the positions of the $K/2$ maximum values of each localization function. The likelihood function can be written as

$$F_m(\alpha(t)) = \sum_{k=1}^{K} q_k \mathcal{N}(\tau_{\alpha(t)}; \hat{\tau}^{m,k}, \sigma^2) + q_0\,, \tag{15}$$

   where $\mathcal{N}(x; \mu, \sigma^2)$ is the probability of extracting x from a Gaussian distribution having mean μ and variance σ^2. $\tau_{\alpha(t)}$ denotes the TDOA value corresponding to the particle $\alpha(t)$. The likelihood function is the sum of K Gaussian PDFs weighted by $q_k = (1-q_0)K^{-1}$.
   Equation (15) gives the likelihood of each particle $\alpha^i(t)$ with respect to the observations obtained from the m-th microphones pair. The overall particle weight can be evaluated by combining likelihood values of each of the M pairs as follows

$$w^i(t) = F(\alpha^i(t)) = \prod_{m=1}^{M} F_m(\alpha^i(t))\,. \tag{16}$$

   Finally, particle weights are normalized in order to respect the following condition:

$$\sum_{i=1}^{N_s} w^i(t) = 1\,. \tag{17}$$

4. *Source label assignment*: The iterative application of steps 2. and 3. causes the particles to cluster around the positions occupied by the two sources. This allows us to perform a clustering

algorithm in the state-space in order to assign to each particle a label corresponding to the source it belongs to. In our system, we perform a k-means clustering algorithm.

5. *Source localization*: The estimated source locations and velocities correspond to the centroids of the clusters as defined by

$$\mathbf{c}_p = \frac{\sum_{i=1}^{N_p} w_p^i(t)\alpha_p^i(t)}{\sum_{i=1}^{N_l} w_p^i(t)}, \qquad (18)$$

where N_p denotes the number of particles contained in every cluster p and α_p^i is the i-th particle belonging to the p-th cluster.

We have shown in Section 3 that triangulation of DOA is effective only when more than two pairs of microphones are used. The same is true using a particle filtering approach. In fact, we notice that equation (15) combines together observed TDOAs of both sources, since it is not possible to assign a TDOA to a specific one. For this reason, when two microphones pairs are used ($M = 2$) the likelihood function $F(\cdot)$ is multi-modal and it is characterized by four peaks in the same positions as the intersections of the DOAs. As before, increasing the number of microphones pairs solves this ambiguity problem and only two peaks survive in the likelihood function.

5. SIMULATION SETUP

In this section we illustrate the simulation setup used in our tests. First, we describe how we simulate the movement of the sources. Then, we illustrate the room geometry, the positions of the microphones and the reflection coefficients used in our experiments.

5.1 Source Movement Simulation

The original source signals are male speech segments sampled at $f_s = 16kHz$. Such signals are convolved with the impulse responses that characterize the propagation from the sources to the microphones in the reverberating environments. Impulse responses are synthesized using a room simulation tool based on a fast beam tracing algorithm (for details see [9]) every 1 s. In order to obtain a finer time granularity (0.125 s), at intermediate time instants impulse responses are estimated using an interpolation technique. With this framework, we can efficiently simulate source trajectories in complex environments.

5.2 Room geometry

We conduct our simulations in a $5m \times 5m \times 2.7m$ test room. The reflection coefficient is varied in the range 0.2-0.9. Table 5.2 reports the relationship between the reflection coefficient and the reverberation time T_{60}.

Simulations are carried out on a specific trajectory roughly at the center of the test room as illustrated in Figure 4. The same figure also shows microphones located in the proximity of the side walls of the room. They are organized in four independent pairs. Adjacent microphones are $40cm$ apart. In order to cope with the ambiguity introduced by sources permutation, in our experiments we used $M = 4$ microphones pairs.

Table 1: Relationship between reflection coefficient (ρ) and reverberation time (T_{60}) in the test $5m \times 5m \times 2.7m$ room

ρ	T_{60}
0.2	0.11 s
0.3	0.13 s
0.4	0.19 s
0.5	0.24 s
0.6	0.32 s
0.7	0.45 s
0.8	0.57 s
0.9	0.61 s

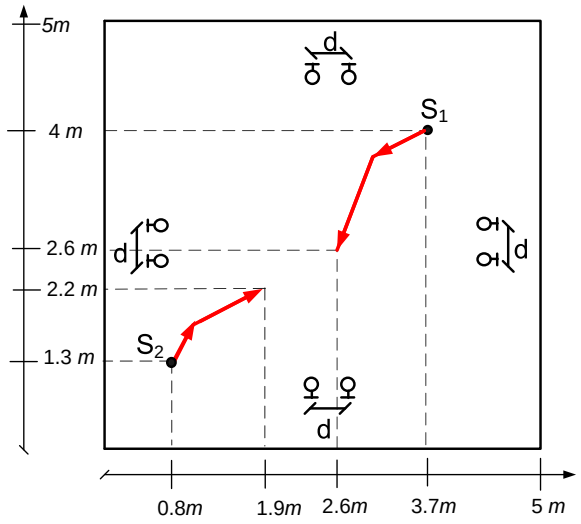

Figure 4: Environment, microphone locations and sources trajectories used in our tests.

In order to quantitatively assess the localization efficiency, we measure the localization error as the Euclidean distance between the estimated position $[\hat{X}_p(t,T_{60}),\hat{Y}_p(t,T_{60})]$ of the p-th source at time t for a given reverberation time T_{60} and the ground truth $[X_p(t),Y_p(t)]$. The overall performance index is the average of the localization errors of the two sources.

6. EXPERIMENTAL RESULTS

We have carried out simulations in order to test the localization and tracking efficiency of the proposed algorithm based on particle filtering, both for static and moving sources. In the following, we also compare the performance obtained using particle filtering with respect to using simple triangulation of DOAs.

6.1 Sources localization experiments

First, we analyze the efficiency of the proposed algorithm for the case of static sources. In this test sources are in fixed locations ($[X_1 = 1.0m, Y_1 = 1.5m, X_2 = 3.25m, Y_2 = 3.5m]$). Figure 5 shows the localization error as a function of reverberation time. We notice that the localization efficiency tends to decrease as the the reverberation time increases. This is due to the fact that spurious peaks might be present in the de-mixing filters estimated by TRINICON. Figure 5 shows a significant performance improvement provided by the proposed algorithm based on particle filtering with respect to DOAs triangulation. The enhanced performance can be justified by the fact that the proposed algorithm takes into account K peaks in the localization function, thus achieving a sort of fractional sample TDOA resolution. Moreover, the particle filtering approach is less sensitive to outliers. This latter fact explains why the gap becomes larger for mildly reverberating environments ($0.3s < T_{60} < 0.6s$), where peaks in the de-mixing filters tend to be less pronounced.

6.2 Sources tracking experiments

In the case of moving sources, localization accuracy still experiences a noticeable improvement using particle filtering. Figure 6 plots the ground truth and estimated trajectories, respectively with a continuous line and circles. Small localization problems are experienced when sources change directions, due to the memory embedded in state-space equation (particles tend to preserve their momentum), but a few observations are sufficient to recover a good estimation.

Figure 7 depicts the overall localization error as a function of the reverberation time. As for the static case, we can notice that

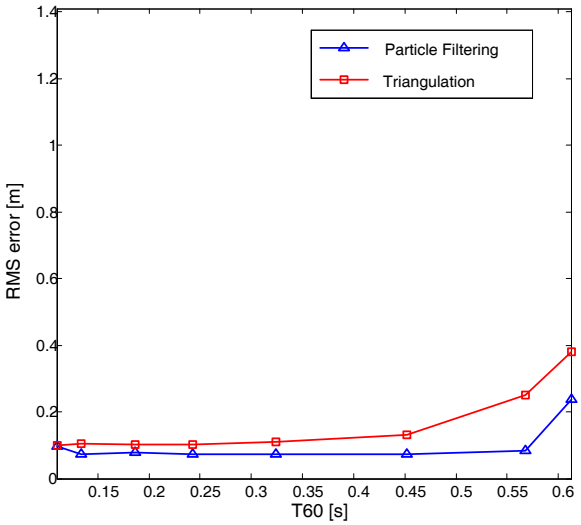

Figure 5: Localization performances with static sources using particle filtering (triangles) and DOAs triangulation (squares).

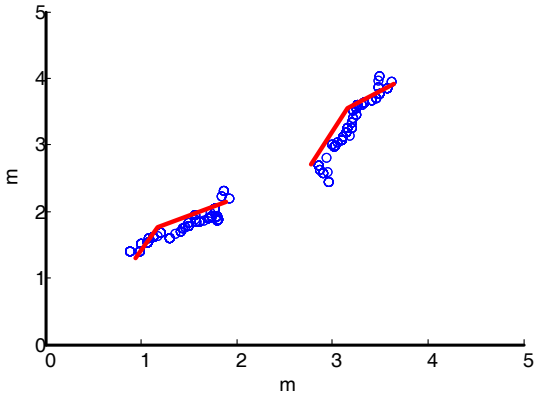

Figure 6: An example of source tracking using particle filtering. Correct and estimated trajecories are drawn, respectively, with continous line and circles

particle filtering sensibly increases localization capabilities with respect to DOAs triangulation.

As a final remark, we have to point out that the information provided as output by the particle filtering algorithm is much richer than the cluster centroids used as source position estimates in our tests. In fact, particle filtering estimates the full posterior PDF given the TDOAs observations, which is only partially described by the cluster centroids.

7. CONCLUSIONS

In this paper we present an algorithm to perform localization and tracking of multiple acoustic sources in reverberating environments. Our approach uses TRINICON as a preprocessing step, in order to determine the TDOAs for each pair of microphones. We show that at least three microphones pairs are needed when TDOAs are separately estimated for each pair. Our current research activities are focused on improving the tracking efficiency especially at high reverberation time by adaptively tuning the parameters of the likelihood function. Moreover, we are extending the work for the case of $P > 2$ and $Q > 2$.

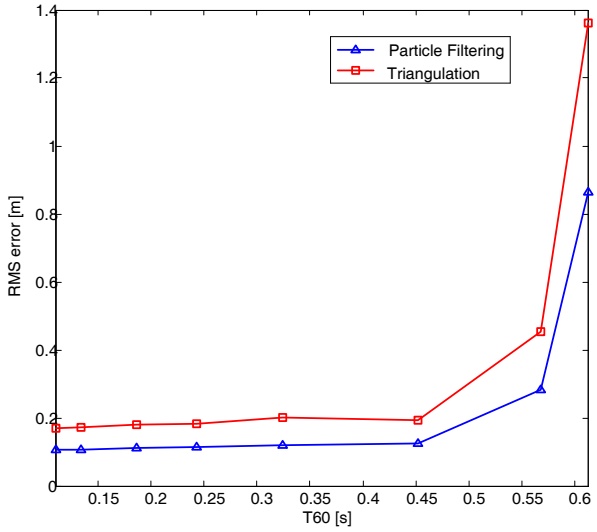

Figure 7: Localization performances with dynamic sources using particle filtering (triangles) and using triangulation (squares)

REFERENCES

[1] Y. Huang and J. Benesty, Time Delay Estimation, in Audio Signal Processing for the Next-Generation Multimedia Communication Systems, J. Benesty and Y. Huang, Eds., Springer, New York, 2003.

[2] Y. Huang and J. Benesty, G. W. Elko Source Localization, in Audio Signal Processing for the Next-Generation Multimedia Communication Systems, J. Benesty and Y. Huang, Eds., Springer, New York, 2003.

[3] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, Real-time passive source localization: an unbiased linear-correction least-squares approach, IEEE Trans. Speech Audio Processing, vol. 9, no. 8, pp. 943956, Nov. 2001

[4] D.B. Ward, E.A. Lehmann, R.C. Williamson, "Particle Filtering Algorithms for Tracking an Acoustic Source in a Reverberant Environment", *Speech and Audio Processing*, *IEEE Transactions on* Volume 11, Issue 6, Page(s):826 - 836, Nov. 2003

[5] M.S.Arulampalam, S.Maskell, N.Gordon and T.Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian bayesian tracking," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 174-188, Feb. 2002

[6] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A Versatile Framework for Multichannel Blind Signal Processing," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp 889−92, vol. 3, Montreal, Canada, May 2004

[7] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, W. Kellermann "Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp 97-100, vol. 3, Philadelphia, PA, Mar 2005

[8] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: A unified treatment", in J. Benesty and Y. Huang, editors, Audio Signal Processing for Next-Generation Multimedia Communication Systems. Kluwer Academic Publishers, Boston, Feb. 2004.

[9] M. Foco, P. Polotti, A. Sarti, S. Tubaro, "Sound Spatialization Based on Fast Beam Tracing in the Dual Space", Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-03), London, Great Britain, Sep. 2003