A SEQUENTIAL MONTE CARLO APPROACH FOR TRACKING OF OVERLAPPING ACOUSTIC SOURCES

Alessio Brutti, Maurizio Omologo, Piergiorgio Svaizer

Fondazione Bruno Kessler - CIT - IRST
via Sommarive 18, 38050, Trento, Italy
phone: + (39) 0461314529, fax: + (39) 0461314591, email: brutti@fbk.eu
web: shine.fbk.eu

ABSTRACT

This paper describes a novel approach based on the sequential Monte Carlo method for tracking of multiple sources overlapping in time. The algorithm is designed for multi-microphone applications and manipulates GCC-PHAT measurements in order to obtain robust likelihoods for all active sources. An experimental analysis was conducted on real data acquired with different microphone configurations. Results show the effectiveness of the proposed algorithm.

1. INTRODUCTION

Sound source localization in reverberant and noisy environments has been investigated for decades. The problem consists in finding the positions of active sources given the measurements provided by a set of acoustic sensors [2]. Besides background diffused noise and localized noise sources, the main issue for a localization algorithm is reverberation which is caused by wave reflections that typically occur in enclosures [11]. In general the source localization problem in multimicrophone scenarios is tackled by evaluating and combining GCC-PHAT functions [10], also known as CSP [14], at a set of microphone pairs.

In recent years, sequential Bayesian methods and Sequential Monte Carlo (SMC) simulations have become very common and have proved to provide an efficient and robust solution to the localization problem [17, 18]. Bayesian methods offer a general probabilistic framework by considering the posterior density function of the target state based on all available measurements [8]. Monte Carlo methods approximate the posterior [1] and are also known as Particle Filter (PF) techniques. One of the main reasons for the popularity of PF approaches is that they can be used without assuming linearity or Gaussianity of the problem, which hardly ever holds in real scenarios. However, particular expedients and proper tuning must be adopted to obtain effective solutions in real applications where, due to the sparseness of speech, the acoustic information may be lacking for long periods.

Monte Carlo methods have been also successfully applied to the multisource scenario by extending in a straightforward manner the algorithms used for single source tracking. The idea is that either multimodal or multidimensional likelihoods, for multitarget states, can be obtained from acoustic observations when more sources are simultaneously active. In this paper we describe a novel approach to multiple source tracking specifically designed to handle conditions where strong time overlapping between sources occurs. The method derives from the post-processing technique presented in [4].

This work was partially supported by the EU under the STREP Project DICIT (FP6 IST-034624). Further details can be found at http://dicit.fbk.eu.

2. TRACKING FRAMEWORK

Sequential Bayesian methods are an alternative to exhaustive maximization/minimization of cost functions. In such state-space trackers we consider a state variable at time k, including information on the position and the speed of a target in a Cartesian coordinate system:

$$\mathbf{S}_k = [x_k\ y_k\ z_k\ \dot{x}_k\ \dot{y}_k\ \dot{z}_k]^T$$

The acoustic sensors provide at each time instant k a set of measurements, which we refer to as $\mathbf{O}_k$, associated to the current state. In a probabilistic framework, the state of the target is computed using the posterior density function (PDF) conditioned to all the available measurements, from time 1 to time k: $p(\mathbf{S}_k|\mathbf{O}_{1:k})$, where $\mathbf{O}_{1:k} = \{\mathbf{O}_1, \ldots, \mathbf{O}_k\}$. For instance the PDF mean can be used as target state estimation.

Since an analytical closed-form solution is non tractable unless the problem has Gaussian and linear properties, an iterative approach is adopted. Let us assume that $p(\mathbf{S}_{k-1}|\mathbf{O}_{1:k-1})$ is known at time $k-1$. The solution is obtained by iterating the following equation set [8, 1, 17, 18]:

$$\begin{aligned} p(\mathbf{S}_k|\mathbf{O}_{1:k-1}) &= \int p(\mathbf{S}_k|\mathbf{S}_{k-1})\, p(\mathbf{S}_{k-1}|\mathbf{O}_{1:k-1})\, d\mathbf{S}_{k-1} \\ p(\mathbf{S}_k|\mathbf{O}_{1:k}) &\propto p(\mathbf{O}_k|\mathbf{S}_k)\, p(\mathbf{S}_k|\mathbf{O}_{1:k-1}) \end{aligned} \quad (1)$$

It includes a *prediction step* where $p(\mathbf{S}_k|\mathbf{S}_{k-1})$ models the transition between one state to the next one (motion model) and an *update step* where $p(\mathbf{O}_k|\mathbf{S}_k)$ is called measurement likelihood.

Given the above described framework, SMC is an approximation method that represents the PDF at time k through N weights $w_k^{(n)}$ $n = 1, \ldots, N$ associated to a set of samples of the state space (particles) $\mathbf{S}_k^{(n)}$. Starting from a set of pairs $\left(\mathbf{S}_{k-1}^{(1:N)}, w_{k-1}^{(1:N)}\right)$ the PDF at time $k-1$ is represented as:

$$p(\mathbf{S}_{k-1}|\mathbf{O}_{1:k-1}) \sim \sum_{n=1}^{N} w_{k-1}^{(n)} \delta\left(\mathbf{S}_{k-1} - \mathbf{S}_{k-1}^{(n)}\right) \quad (2)$$

where $\delta(\cdot)$ is the Dirac impulse. The iterative prediction-update process introduced above allows to compute a new set of weighted particles $\left(\mathbf{S}_k^{(1:N)}, w_k^{(1:N)}\right)$, approximating the

new PDF, through which the target state can be estimated for instance as weighted average:

$$\hat{\mathbf{S}}_k = \sum_{n=1}^{N} w_k^{(n)} \mathbf{S}_k^{(n)}. \quad (3)$$

One of the approaches that best fits the acoustic tracking problem is based on the Sampling Importance Re-sampling (SIR) method described by the following steps [8]:

i A set of particles $\tilde{\mathbf{S}}_k^{(1:N)}$ is computed by propagating with a proper motion model the previous set $\mathbf{S}_{k-1}^{(1:N)}$.

ii A weight $w_k^{(n)}$ is associated to each new particle according to the measurement likelihood. Weights are normalized to sum up to 1.

$$w_k^{(n)} = \frac{w_{k-1}^{(n)} p\left(\mathbf{O}_k | \tilde{\mathbf{S}}_k^{(n)}\right)}{\sum_{n=1}^{N} w_{k-1}^{(n)} p\left(\mathbf{O}_k | \tilde{\mathbf{S}}_k^{(n)}\right)} \quad (4)$$

iii The target state is estimated using equation (3).

iv Given the PDF approximated by $\left(\tilde{\mathbf{S}}_k^{(1:N)}, w_k^{(1:N)}\right)$, as in (2), a new set of particles is obtained by resampling.

2.1 Measurement Likelihood

Traditional PF implementations compute the measurement likelihood either using the output of a steered beamformer [13] or considering the distance between the hypothesized time differences of arrivals and the estimated ones at several microphone pairs [17]. In this paper we instead adopt an approach derived from acoustic map analysis such as the Global Coherence Field (GCF) [5], also known as SRP-PHAT [2]. In traditional source localization methods based on GCF, GCC-PHATs are combined in order to compute maps of enclosures, representing the plausibility that a source is active in a given point. Localization is then carried out by maximizing the resulting map. Therefore, given a grid of points $\mathbf{s}$ defined over the area of interest and M microphone pairs, a GCF acoustic map is computed for each point as follows [5]:

$$\mathrm{GCF}(k, \mathbf{s}) = \frac{1}{M} \sum_{m=0}^{M-1} C_m\left(k, \psi_m(\mathbf{s})\right) \quad (5)$$

where $C_m(k,\tau)$ is the GCC-PHAT function computed at pair m and time instant k, while $\psi_m(\mathbf{s})$ is the geometrically computed time difference of arrival at pair m if the source is in $\mathbf{s}$. In our implementation, the measurement likelihood for each particle is computed as:

$$p\left(\mathbf{O}_k | \tilde{\mathbf{S}}_k^{(n)}\right) = \mathrm{GCF}\left(k, \tilde{\mathbf{S}}_k^{(n)}\right) \quad (6)$$

Notice that $\psi_m(\cdot)$ depends only on the first three components of the particle state, i.e. it does not depend on speed. Other types of acoustic maps are suitable to be adopted as likelihood measurements. For instance the Oriented Global Coherence Field [3] allows one to deduce also information about the orientation of the source and is more effective when microphones are distributed along the perimeter of a room. In order to use such a likelihood in a PF implementation, an extra dimension must be added to the state space which accounts for the orientation of the source [7]. A further alternative is represented by the use of multiplication rather than summation in equation (5) [15].

2.2 PDF approximation

SIR works properly when information for likelihood computation is available. But it may fail when the source is silent or the observations $\mathbf{O}_k$ are not reliable due to background noise or reverberation. Even if the method can deal with short time pauses thanks to the motion model, small changes in the propagation model may affect the overall performance considerably. Moreover, during the above mentioned lacks of information a different speaker may take turn. The PDF approximation presented in equation (2) is hence modified in order to address the above mentioned issues. First of all we consider two hypotheses, based on GCF information:

- $\mathscr{H}_0$: the target is silent;
- $\mathscr{H}_1$: the target is active;

In $\mathscr{H}_1$, the original SIR algorithm is used. Conversely, in $\mathscr{H}_0$ a uniform distribution over the particles is adopted, i.e. all weights are forced to the same value to approximate the PDF. In practice particles only propagate according to the motion model and no resampling occurs. Therefore the PDF is approximated as follows:

$$\begin{aligned} p(\mathbf{S}_k | \mathbf{O}_{1:k}) &= \eta_k \sum_{n=1}^{N} w_k^{(n)} \delta\left(\mathbf{S}_k - \mathbf{S}_k^{(n)}\right) \\ &+ (1-\eta_k) \frac{1}{N} \sum_{n=1}^{N} \delta\left(\mathbf{S}_k - \mathbf{S}_k^{(n)}\right) \end{aligned} \quad (7)$$

where η_k is a flag indicating whether the observations at time k are generated by the target. If we denote as w'_k the maximum non-normalized weight at time k and introduce a threshold ϑ_R, η_k is defined as:

$$\eta_k = \begin{cases} 1 & \text{if } w'_k \geq \vartheta_R \\ 0 & \text{otherwise} \end{cases}$$

In a similar way one may use an energy-based function for η_k, as in [13], or adopt a fuzzy approach. In order to guarantee quick reaction to speaker turn taking and ensure a fast convergence when the user moves while being silent, the motion model is not used when hypothesis $\mathscr{H}_0$ lasts for more than n_D consecutive frames and a Gaussian distribution is used instead:

$$\begin{aligned} p(\mathbf{S}_k | \mathbf{O}_{1:k}) &= (1-\rho_k)\eta_k \sum_{n=1}^{N} w_k^{(n)} \delta\left(\mathbf{S}_k - \mathbf{S}_k^{(n)}\right) \\ &+ (1-\rho_k)(1-\eta_k) \frac{1}{N} \sum_{n=1}^{N} \delta\left(\mathbf{S}_k - \mathbf{S}_k^{(n)}\right) \\ &+ \rho_k \mathscr{N}(\mu_k, \sigma) \end{aligned} \quad (8)$$

where $\mu_k = \arg\max\{\mathrm{GCF}(\mathbf{s},\mathbf{s})\}$ and σ is a parameter of the algorithm. When $\rho_k = 1$ both measurements and motion model are not used since the particle filter information is assumed to be out-of-date and not reliable. At a given time k, ρ_k is a function of n_D and of k_R that is the last time index at which $\eta_k = 1$.

$$\rho_k = \lfloor \frac{k - k_R}{n_D} \rfloor \quad (9)$$

2.3 Motion Model

Assuming that particle trajectories along different coordinates are independent each other, the propagation step of the PF process is handled by adopting a variation of the Langevin model [13, 17]. In our implementation an adaptive motion model controlled by the value of $(k-k_R)$ is implemented. In practice, if $\mathscr{H}_0$ holds particles are encouraged to explore a larger area until, after some steps, they are spread over the whole state space. On the contrary, when speech activity is detected, particles follow the dynamics estimated by the model.

3. OVERLAPPING SOURCE TRACKING

In this section we present our approach for performing tracking of sources overlapping in time. Without loss of generalization, we limit our analysis to two simultaneous sources. The proposed algorithm can be extended to deal with a larger number of sources in a straghtfoward way, even though generally GCC-PHAT fails to perform when more than 3 sources overlap.

Under ideal conditions,when two sources are active GCC-PHAT shows two evident peaks at the time lags corresponding to the two real time differences of arrivals. Unfortunately, in real conditions the two peaks hardly ever appear at the same time since one of the two tends to absorb the overall coherence. This phenomenon is due to different spectral contents and to different nature of signals produced by the sources as well as to distance from microphones and source orientation. Therefore, it is more likely to observe two alternating peaks and in this case short-term spatio-temporal clustering may be a convenient approach [6, 12]. However this method fails when constructive interferences between sources occur generating fake peaks or when one of the sources is dominant in the long term preventing the information associated to the second one from appearing. In [4] a method for highlighting weak peaks in GCF maps is presented which attempts to de-emphasize the dominant peak at GCC-PHAT level. The idea presented here is to embed that technique in a PF framework in order to guarantee availability of measurements for all sources.

3.1 Proposed Approach

For tracking of two simultaneous sources we use two parallel filters instead of extending the dimension of the state by adding coordinates and speed of the second target [16]. The reason is purely implementative as with this solution the GCC-PHAT de-emphasis can be easily embedded. The two filters are identified by two separate and independent populations of particles $\mathbf{S}_k^{(1:N,i)}, i=0,1$ whose estimation output is defined as $\hat{\mathbf{S}}_k^{(i)}$. The proposed approach needs a multimodal PDF that in general is not available for the reasons explained above. Hence, the weight of each particle is computed by applying beforehand the GCC-PHAT de-emphasis process presented in [4] in order to remove the contribution of the competitive target.

Let us consider the target state estimation $\hat{\mathbf{S}}_{k-1}^{(1)}$ obtained from the set $\mathbf{S}_{k-1}^{(1:N,1)}$. Before computing the weights of each $\tilde{\mathbf{S}}_k^{(n,0)}$ the theoretical time delay $\psi_m(\hat{\mathbf{S}}_{k-1}^{(1)})$ is used to modify all the GCC-PHAT functions by means of a notch mask as follows:

$$C_m^{(0)}(k,\tau) = \phi\left(\tau, \psi_m\left(\hat{\mathbf{S}}_{k-1}^{(1)}\right)\right) \cdot C_m(k,\tau). \qquad (10)$$

Weights are then computed using the new function $C_m^{(0)}(k,\tau)$ in (5). The same process is used to remove contributions associated to $\hat{\mathbf{S}}_{k-1}^{(0)}$ before computing weights of each $\tilde{\mathbf{S}}_k^{(n,1)}$:

$$C_m^{(1)}(k,\tau) = \phi\left(\tau, \psi_m\left(\hat{\mathbf{S}}_{k-1}^{(0)}\right)\right) \cdot C_m(k,\tau) \qquad (11)$$

In our implementation the function $\phi(\cdot)$ was chosen to belong to the following class:

$$\phi(\tau,\nu) = \alpha\left[1 - e^{-\frac{|\tau-\nu|^p}{b}}\right] \qquad (12)$$

Parameters b and p determine the sharpness of the notch, while α is a normalization factor that guarantees that:

$$\sum_{\tau=-\tau_{max}}^{\tau_{max}} \phi(\tau,\nu)\cdot C_m(\tau) = \sum_{\tau=-\tau_{max}}^{\tau_{max}} C_m(\tau) \qquad (13)$$

where τ_{max} is the maximum allowed time delay. As shown in Figure 1, in practice a null is set around the lag that corresponds to the time difference of arrival of the source to remove. Small values of b and p correspond to very selective functions. On the other hand, large values of the parameters enlarge the notch. In general a trade-off is needed to obtain functions that ensure the removal of one of the peaks without affecting the other one. When applied to acoustic map based localization, de-emphasis proved to be an efficient tool for locating two simultaneous sources [4].

Notice that if only one source is present, one of the filters "captures" the target and the second one either randomly distributes its particles as if there is no acoustic activity or tracks ghosts if any is present. Conversely, if de-emphasis were not performed both filters would track the first available source and would neglect the other one.

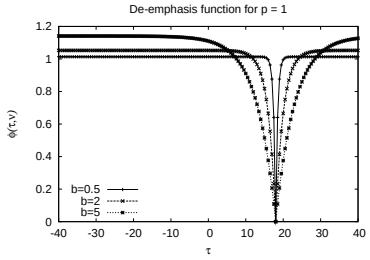

Figure 1: Example of functions $\phi(\cdot)$ for different values of b when $\nu = 18$ and $p = 1$. The normalization factor α guarantees that $\phi(\cdot)$ sums up to 1.

4. EXPERIMENTAL ANALYSIS

The proposed algorithm is evaluated on real data acquired with two different sensor settings: the first one makes up a Distributed Microphone Network (DMN), while the second one consists of a linear array. In order to simulate overlapping sources, a human speaker is recorded while pronouncing a phonetically rich sentences at 5 different positions

and orientations. Each sentence is about 10 second long. Recorded signals from each single source session are then summed up assuming a linear sound propagation model and source independence, which are both acceptable hypotheses in the given context. A set of 10 position combinations was considered for evaluation in both the settings under analysis. Experiments are restricted to bi-dimensional horizontal localization due to limited vertical coverage of the given microphone configurations and due to constant speaker height. The sampling rate is 44.1 kHz in the DMN case and 48 kHz in the linear array setting. The analysis window is 2^{14} samples long with a 25% overlap. The number of particles is set to 900 and $n_D = 5$ while the threshold ϑ_R depends on the experimental set up in use. Those values have been defined in a heuristic way in order to optimize performance.
The capabilities of our localization algorithm are measured in terms of average RMS error computed on both sources. Let us define as $\mathbf{p}_k^{(i)}$ the actual position of the i-th source ($i = 0, 1$) at time frame k. The localization error in estimating the position of the i-th source is expressed as the Euclidean distance between the estimated and the actual position:

$$e_k^{(i)} = \|\hat{\mathbf{S}}_k^{(i)} - \mathbf{p}_k^{(i)}\| \tag{14}$$

the average RMS error is computed as follows:

$$\overline{\text{rms}} = \sqrt{\frac{\sum_{k=1}^{N_t}\left[\left(\left(e_k^{(0)}\right)^2 + \left(e_k^{(1)}\right)^2\right)/2\right]}{N_t}} \tag{15}$$

where N_t is the overall number of processed frames. Since no source label is available in the localization estimation, estimates are associated to sources on a minimum distance criterion in an exclusive way.
For comparison we consider an upper bound defined as the performance of a PF implementation for single source tracking based on GCF likelihood when sources do not overlap. The outputs are then combined in order to generate the upper bound localization performance for the double-talk case.
As further metric, we consider also the tracking rate, defined as the percentage of localization frames for which the original PDF is used (i.e. $\eta_k(1-\rho_k) = 1$).

4.1 Distributed Microphone Network

As a first study case we consider a DMN as the one adopted in the CHIL project[1] where a set of microphones is distributed all around a room, typically grouped in small arrays. A DMN consisting of 7 arrays with 3 microphones in a row is used for recordings in a room whose dimensions are $5.9 \times 4.8 \times 5$ m. The environmental noise is low but the reverberation time is quite challenging and is equal to 0.7 s. The positions of the 7 arrays are shown in Figure 2 where they are labeled with "T". Since we do not combine microphones belonging to different arrays, the resulting number of used pairs is $M = 21$. Figure 2 shows also the positions where the sources were recorded and, by means of arrows, their orientation.

Figure 3 reports performance for different configuration of the de-emphasis function when the speech activity threshold ϑ_R is 0.6. The tracking rate is always between 87% and 90% of the overall processed frames. Notice how the proposed method permits to almost reach the upper bound when specific parameters of the de-emphasis function are selected. Best performance is obtained when $p = 1$ and $b = 3.5$ corresponding to $\overline{\text{rms}} = 152.9$ mm, with an upper bound of 129.2 mm.

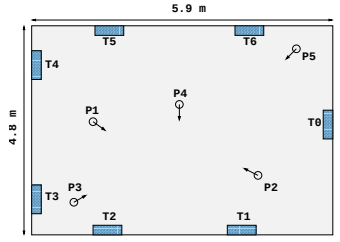

Figure 2: Microphone array (T) and source (P) positions in the DMN setting. Arrows indicate the orientation of the speaker. Arrays are represented by boxes.

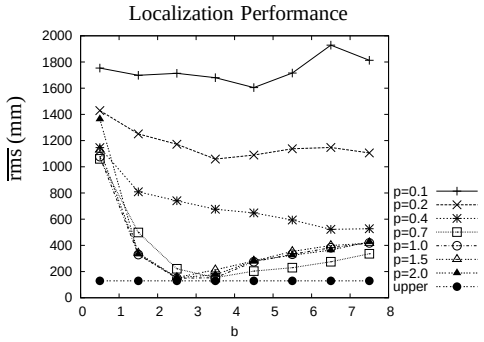

Figure 3: Performance of a multispeaker tracker in a DMN set up for different parameters of the de-emphasis function. Full circles indicates the upper bound.

It is worth noting that accurate tuning is needed in the given scenario. Using very small values of p and/or b does not work because the notch would miss the GCC-PHAT peaks to remove. Conversely, very wide functions corresponding to large values of b seem to be too invasive and affect also some useful information. Good parameters for the proposed method are $p > 0.7$ and $b \in [2.5, 4]$.

4.2 Linear Array

In a second experiment a linear array consisting of 7 microphones placed at 32 cm distance was used, resulting in $M = 6$ pairs of adiacent microphones. It is part of the more complex harmonic array adopted in the DICIT project[2]. The positions of the array and of the 5 sources are described in Figure 4 which shows also the orientations of the speakers. The recordings were carried out in a $3.5 \times 5 \times 4$ m room whose reverberation time is about 0.35 s.

Although the distance estimation is more prone to errors when using a linear array, localization results obtained in this configuration are very satisfactory as reported in Figure 5 and the best performance is very close to the upper bound. In these conditions the best value for ϑ_R is 0.45 which results in a tracking rate of about 90%. The best performance is

[1] For further details see: http://chil.server.de

[2] For further details see: http://dicit.fbk.eu

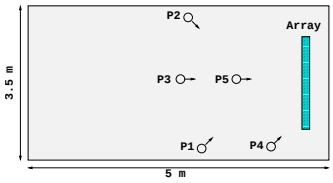

Figure 4: Scheme of the source positions under investigation. The bar on the right represents the linear array.

$\overline{\text{rms}} = 144$ mm when $p = 1.5$ and $b = 5.5$ against an upper bound equal to 102.7 mm. Notice that in this setting, larger values of p are needed (1.5, 2) because it is not possible to take advantage from a more effective spatial distribution of the microphones in space as in the DMN case.

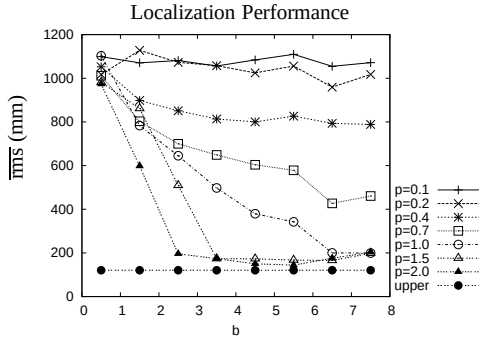

Figure 5: Performance of a multispeaker tracker in a linear array set up for different parameters of de-emphasis function.

5. CONCLUSIONS AND FUTURE WORK

This paper presented a sequential Bayesian approach for tracking of two sources overlapping in time. Since in these conditions likelihoods do not show a clear multimodal nature, the proposed method attempts to highlight weak modes to ensure good tracking. Experiments on real data acquired with different microphone configurations show the very good capabilities of the described method which allows one to obtain performance close to the considered upper bound.
Including birth and death processes through Random Finite Sets [9] should permit to deal with tracking of a varying number of sources in a more efficient way by allocating and deallocating filters. Therefore a future activity will be towards this direction. However, the current implementation does not affect tracking in case a single source is active and there is no need of knowing the exact number of sources but just the maximum number.
Finally, a deep comparison between different likelihoods derived from acoustic map analysis, as for instance Oriented Global Coherence Field, is needed in an attempt to improve the tracking performance of the proposed method.

REFERENCES

[1] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line nonlinear/non-gaussian bayesian tracking. *IEEE Trans. on Signal Processing*, 50(2), February 2002.

[2] M. Brandstein and D. Ward, editors. *Microphone Arrays*. Springer, 2001.

[3] A. Brutti, M. Omologo, and P. Svaizer. Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays. In *Proc. of Interspeech*, Lisbon, 2005.

[4] A. Brutti, M. Omologo, and P. Svaizer. Localization of multiple speakers based on a two step acoustic map analysis. In *Proc. of IEEE ICASSP*, Las Vegas, 2008.

[5] R. DeMori. *Spoken Dialogue with Computers*. Academic Press, London, 1998. Chapter 2.

[6] E. D. DiClaudio, R. Parisi, and G. Orlandi. Multisource localization in reverberant environments by ROOT-MUSIC and clustering. In *Proc. of IEEE ICASSP*, Istanbul, 2000.

[7] M. Fallon, S. Godsill, and A. Blake. Joint acoustic source localization and orientation estimation using sequential monte carlo. In *Proc. of 9th Int. Conference on Digital Audio Effects*, Montreal, 2006.

[8] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proc. of Radar and Signal Processing*, 140(2), April 1993.

[9] J. Goutsias, R. Mahler, and H. Nguyen. *Random Sets Theory and Applications*. Springer-Verlag New York, 1997.

[10] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustic, Speech and Signal Processing*, 24(4), 1976.

[11] H. Kuttruff. *Room Acoustics*. Elsevier Applied Science, 1991.

[12] G. Lathoud and J. Odobez. Short-term spatio-temporal clustering applied to multiple moving speakers. *IEEE Trans. on Audio, Speech and Language Processing*, 15(5), July 2007.

[13] E. A. Lehmann and A. M. Johansson. Particle filter with integrated voice activity detection for acoustic source tracking. *Eurasip Journal on Advances in Signal Processing*, 2007.

[14] M. Omologo and P. Svaizer. Acoustic event localization using a crosspower-spectrum based technique. In *Proc. of IEEE ICASSP*, Adelaide, 1994.

[15] P. Pertilä, T. Korhonen, and A. Visa. Measurement combination for acoustic source localization in room environment. *Eurasip Journal on Audio, Speech and Music Processing*, 2008.

[16] J. Valin, F. Michaud, and J. Rouat. Robust 3D localization and tracking of sound sources using beamforming and particle filtering. In *Proc. of IEEE ICASSP*, Tolouse, 2006.

[17] J. Vermaak and A. Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *Proc. of IEEE ICASSP*, Salt Lake City, 2001.

[18] 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*, 11(6), November 2003.