Acoustic Source Localisation and Tracking of a Time-Varying Number of Speakers

Maurice F. Fallon, *Member, IEEE,* and Simon Godsill, *Member, IEEE*

***Abstract*—Particle Filter-based Acoustic Source Tracking algorithms track (on-line and in real-time) the position of a sound source - a person speaking in a room - based on the current data from a distributed microphone array as well as all previous data up to that point. This paper develops a multi-target tracking (MTT) methodology to allow for an unknown and time-varying number of speakers in a fully probabilistic manner and in doing so does not resort to independent modules for new target proposal or target number estimation as in previous works. The approach uses the concept of an existence grid to propose possible regions of activity before tracking is carried out with a variable dimension particle filter — which also explicitly supports the concept of a null particle, containing no target states, when no speakers are active. Examples demonstrate typical tracking performance in a number of different scenarios with simultaneously active speech sources.**

***Index Terms*—Tracking Filters, Sequential Estimation, Particle Filtering, Acoustic Source Location, Multi-target Tracking.**

I. Introduction

THE application of particle filtering to speech source localisation and tracking (AST) is an increasingly active area of research. A seemingly simple problem at the outset, AST is complicated by the existence of noise sources, reverberation, other speech sources and - possibly most challenging of all - the non-stationarity of speech. The field has developed from tracking single-source recordings in synthetic environments [1], to real and challenging environments [2]. Approaches have however assumed that a single source is active from the start of the algorithm to its end without any major silent pauses — clearly an over-idealisation.

Previously we introduced a methodology for multi-target tracking of acoustic sources [3] which avoided data association by using the track-before-detect (TBD) paradigm [4] and tracked multiple sources simultaneously. Again this technique assumed knowledge of the number of sources as well as their initial positions. In the following, a **fully probabilistic** algorithm is proposed which identifies newly active sources, keeps track of them and removes them when they become inactive. Likely targets are proposed using an existence grid [5] before being accurately tracked using TBD [3]. In particular the approach does not resort to a seperate modules for target proposal or removal.

Before detailing this framework we will briefly discuss other approaches suggested to solve this problem (Sec II). Sec III will present in detail the particle filter tracking algorithm we will use. Particles will be proposed using an existence grid, detailed in Sec IV. Finally, in Sec V typical performance with audio data recorded using a 12-element microphone array will be illustrated. Note that the Steered Beamformer Function (SBF) is used to isolate localisation information from each frame of audio, as previously used in [2], [3].

Maurice Fallon is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Email: mfallon@mit.edu

Simon Godsill is with the Signal Processing and Communications Laboratory, Cambridge University Engineering Dept, Trumpington Street, Cambridge, CB2 1PZ, UK. Email: sjg30@cam.ac.uk

Tracking examples can be viewed at http://people.csail.mit.edu/mfallon

This work was funded by Microsoft Research through the European PhD Scholarship Programme.

Manuscript received January 19, 2011; revised March 11, 2011.

II. Tracking target activity

Some ad-hoc approaches have attempted to deterministically identify active speech targets based around heuristic decisions and to then perform tracking. Sturim *et al* [6] first proposed a Kalman-filter tracking solution of this form. Similar methods using particle filter-based tracking followed [7]. Alternatively, Lehmann and Williamson [8] introduced an algorithm which allows for switching between conversational sources (i.e non-simultaneously active sources) when one of the speakers stops speaking. The system does not, however, attempt to determine if a source is actually active — with particle states spreading out across the room when the speaker was inactive. As a result, recovering from silences was dependent on using a large number of particles. It did not probabilistically support particles with no active sources. Application of random finite set (RFS) theory to this field has also been proposed [9]. That approach used a Generalized Cross Correlation (GCC) measurement function, however the GCC is not well suited to AST because when targets cross in front of a microphone pair the pairwise measurement-to-source assignment is ambigous. For the SBF inter-pair correlations (ignored by the GCC) resolve this ambiguity.

Recent research such as [10] have retained the TBD approach while adding an 'initialization filter' to propose new speakers separate to the tracking filter. Meanwhile [11] carried out explicit data association of all possible source locations before carrying out radial position tracking on a robot.

Meanwhile, within the general field of (military) tracking a number of methodologies have been introduced to keep track of the number of active sources in a more principled manner such as the Independent Partition and Joint Probabilistic Data Association filters. Acoustic targets are, however, discontinuous with a dramatically varying SNR. Our proposal instead maintains a single joint target state updated using observations drawn from acoustic data. Before describing our target proposal mechanism (in Sec IV) and the likelihood function, we will first outline the higher level tracking algorithm with which the proposed targets will be tracked.

III. Tracking Framework

In this section a variable dimension particle filter is proposed in which each particle represents a single un-partitioned estimate of the underlying state space. It will keep track of the time-varying number of sources present in the room, including the possibility that no targets are active at all, which is a novel contribution of the paper.

The number of targets, S_k, within each individual state vector may vary in the range $\{0, \ldots, S_{\text{max}}\}$, representing the number of speakers deemed to be active at any given time k. S_{max} is the maximum number of speakers and is chosen to be 3 in our experiments, although in principle the methods extend to more 'crowded' environments as well. See Sec. VI for more discussion. An individual state vector, containing S_k targets at time k, is defined as follows

$$\mathcal{A}_k = (\alpha_k^1, \ldots, \alpha_k^{S_k}, S_k) \tag{1}$$

Each target, α_k^s, contains position and velocity components in the $\mathcal{X}$

and $\mathcal{Y}$-dimensions, as follows

$$\alpha_k^s = (x_k^s, y_k^s, \dot{x}_k^s, \dot{y}_k^s) \tag{2}$$

The aim of the particle filter is to update the posterior probability density for the entire vector (Eq 1) using information from the measurements, $\mathbf{Z}_k$. These measurements are defined on a dense but discrete grid of J cells covering the tracking space of interest, and computed using a specially optimised steered beamformer (SBF) directed at each cell's centre.

A. Prior Speaker Models

Within our framework we propose to model the random appearance ('birth') and disappearance ('death') of speakers. These prior models are unrelated to the audio data and attempt to capture the behaviour of speaker and/or the environment in question. For example this could support information about typical sentence length and the duration of inter-syllable silences. It could also be tuned to short sentences (e.g. a speaker controlling a television) versus longer sentences (e.g. lecturing scenario). We have endeavored to use only a weak prior model so as to maintain generality.

a) Prior target number model: This is carried out as two consecutive transition processes. First, the removal process will provide a prior model of how the number of targets is likely to change given the possibility of removing a target

$$S_{k|k-1} = S_{k-1} + \epsilon_{k|k-1} \tag{3}$$

to give an intermediate estimate of the target number, $S_{k|k-1}$. It will do so with a prior removal probability distribution

$$p(S_{k|k-1}|S_{k-1}) = \begin{cases} \Pr(\epsilon_{k|k-1} = -1) = h_d \\ \Pr(\epsilon_{k|k-1} = 0) = 1 - h_d \end{cases} \tag{4}$$

where $h_d = 0.05$ is the probability of decrementing the number of targets - that is the prior probability that a speaker will stop speaking at that specific time. This means that we expect a speaker to stop speaking every few seconds. The target for removal, say s', is chosen randomly with probability $1/S_{k-1}$.

Then the addition of a new target will be carried out in a similar way as follows

$$S_k = S_{k|k-1} + \epsilon_k \tag{5}$$

to give the final target number estimate S_k. It will do so with a prior addition probability distribution

$$p(S_k|S_{k|k-1}) = \begin{cases} \Pr(\epsilon_k = 0) = 1 - h_b \\ \Pr(\epsilon_k = 1) = h_b \end{cases} \tag{6}$$

where h_b is the probability of incrementing the number of targets — that is the prior probability of a speaker (spontaneously) starting to speak. h_b and h_d are set equal, except when $S_{k-1} = 0$ where we will set $h_d = 0$, or when $S_{k-1} = S_{\max}$ where we will set $h_b = 0$.

The overall prior probability distribution for the number of targets is then simply the product of the individual probabilities

$$p(s'|S_k, S_{k|k-1}, S_{k-1}) = p(S_{k|k-1}|S_{k-1})p(s')p(S_k|S_{k|k-1}) \tag{7}$$

and then s' and $S_{k|k-1}$ may be discarded moving to time $k+1$.

b) Prior Distribution of new target positions: Secondly, the prior state distribution of new target births, $p_0(\alpha_k^s)$, may be chosen to reflect areas of the room in which new speakers are more likely to appear — such as near the doorways of a room. No such information will be used at this stage and the prior distribution of the location parameters will be set to be uniform across the cell, $p_0(x_k^s, y_k^s) = \mathcal{U}_{\mathcal{T}}(x_k^s, y_k^s)$, where $\mathcal{T}$ will be the area of the entire surveillance region. Furthermore, the prior distribution of the velocity components $p_0(\dot{x}_k^s, \dot{y}_k^s)$ will be initiated as a Gaussian around zero velocity to give

$$p_0(\alpha_k^s) = p_0(x_k^s, y_k^s) \times p_0(\dot{x}_k^s, \dot{y}_k^s) \tag{8}$$

Thus the overall prior distribution of the full state vector, $\mathcal{A}_k$, can be stated as follows

$$p(\mathcal{A}_k|\mathcal{A}_{k-1}) = p_\alpha(\alpha_k^{1:S_k}|\alpha_{k-1}^{1:S_{k-1}}, S_k, S_{k-1}, s')p_S(S_k, s'|S_{k-1}) \tag{9}$$

where the portion of the prior related to the target positions can be broken down as follows

$$p_\alpha(\alpha_k^{1:S_k}|\alpha_{k-1}^{1:S_{k-1}}, S_k, S_{k-1}, s') = \begin{cases} \prod_{s=1, s\neq s'}^{S_{k-1}} p(\alpha_k^s|\alpha_{k-1}^s) & \text{if } \epsilon_{k|k-1} = -1, \epsilon_k = 0 \\ \prod_{s=1}^{S_{k-1}} p(\alpha_k^s|\alpha_{k-1}^s) & \text{if } \epsilon_{k|k-1} = 0, \epsilon_k = 0 \\ p_0(\alpha_k^{S_k}) \times \prod_{s=1, s\neq s'}^{S_{k-1}} p(\alpha_k^s|\alpha_{k-1}^s) & \text{if } \epsilon_{k|k-1} = -1, \epsilon_k = 1 \\ p_0(\alpha_k^{S_k}) \times \prod_{s=1}^{S_{k-1}} p(\alpha_k^s|\alpha_{k-1}^s) & \text{if } \epsilon_{k|k-1} = 0, \epsilon_k = 1 \end{cases} \tag{10}$$

where s' is the target removed at time k (if any). Note that when a target is added, the new target is added at the position S_k in the vector $\mathcal{A}_k$, whereas deletion can occur randomly to any target from α_{k-1}, randomly chosen with probability $1/S_{k-1}$.

B. Sequential Monte Carlo Methods

Our goal is to estimate the joint posterior distribution of the target states recursively, and we adopt the standard two step Bayesian update rule. As the evaluation of the integral and update steps is often intractable, sequential Monte Carlo methods are used to approximate the recursion for such complex measurement or dynamical models. The idea is that a complex probability distribution can be represented as a set of weighted Monte Carlo importance samples.

The problem at hand has many state variables and a time-varying number of speakers. Hence, instead of sampling from the dynamical model alone (as in bootstrap filtering), [12], we will instead sample the pth particle for the new state vector from a data–dependent proposal function

$$\begin{aligned} \mathcal{A}_k^{(p)} &\sim q(\mathcal{A}_k|\mathcal{A}_{k-1}^{(p)}, \mathbf{Z}_{1:k}) \\ &\sim q_\alpha(\alpha_k^{1:k}|\alpha_{k-1}^{(1:S_k)(p)}, S_k^{(p)}, S_{k-1}^{(1:S_k)(p)}, \mathbf{Z}_{1:k}) q_S(S_k, S_{k|k_1}|S_{k-1}^{(p)}, \mathbf{Z}_{1:k}) \end{aligned} \tag{11}$$

where $q_\alpha(\cdot)$ and $q_S(\cdot)$ are importance sampling functions for the position/velocity and target number states respectively, and an appropriate correction is then made for the bias introduced in the importance weighting step. According to (Eq 11), we first propose the new target number in time-frame k by removing unsupported targets and then add targets to newly active regions of an existence grid (as defined in Sec IV) as follows.

1. Removal of targets: A decision on whether to remove a target from a particle is randomly made according to probability $\bar{\kappa}_0$:

$$q(S_{k|k-1}|S_{k-1}^{(p)}) = \begin{cases} \Pr(\epsilon_{k|k-1} = -1) = 1 - \bar{\kappa}_0 \\ \Pr(\epsilon_{k|k-1} = 0) = \bar{\kappa}_0 \end{cases} \tag{12}$$

Should a removal be decided upon, a random draw from the set of properly normalised removal probabilities ($\bar{\kappa}_{1:S_{k-1}}$) is then made to choose a target s' for removal. The evaluation of these removal probabilities is explained in Sec IV-E. Having removed a target, the intermediary target number is decremented:

$$S_{k|k-1}^{(p)} = S_{k-1}^{(p)} - 1 \tag{13}$$

Otherwise no action is taken.

2. Initiation of new targets: In a similar manner to the above an addition decision is then made as follows

$$q(S_k^{(p)}|S_{k|k-1}^{(p)}) = \begin{cases} \Pr(\epsilon_k = 0) = \bar{\nu}_0 \\ \Pr(\epsilon_k = 1) = 1 - \bar{\nu}_0 \end{cases} \tag{14}$$

Should a new addition be decided upon, a random draw is made to choose a cell for the new target using the set of normalised addition probabilities ($\bar{\kappa}_{1:S_J}$), again see Sec IV-E for details.

Having selected the cell, the target position is initialised using a weighted combination of a uniform distribution within the physical region of cell, $\mathcal{T}_j$, and a normal distribution centred on the weighted mean of any particle states currently existing in that cell, $\bar{\alpha}_{k-1}^{(j)}$ and with variance equal to $\bar{\sigma}_{k-1}^{2(j)}$, the idea being that some particles may have detected the correct object position in an earlier time frame,

$$\begin{aligned} \alpha_k^{s(p)} &\sim q_0(\alpha_k^s|\mathbf{Z}_{1:k}) \\ &\sim \beta\mathcal{N}(\alpha_k^s;\bar{\alpha}_{k-1}^{(j)},\bar{\sigma}_{k-1}^{2(j)}) + (1-\beta)\mathcal{U}_{\mathcal{T}_j}(\alpha_k^s) \end{aligned} \quad (15)$$

The parameter β is set to be 0.7 in what follows.

3. Updating of persistent target positions: Finally the states of targets persisting from time-step $k-1$ to k are propagated using the Langevin dynamical model which has been used previously in this field, see [2], [8], $\alpha_k^{s(p)} \sim q(\alpha_k^s|\alpha_{k-1}^{s(p)},\mathbf{Z}_k)$. While more advanced dynamic models could have been used, this model showed adequate performance in practice. Parameter values used were similar to those mentioned in these works. The use of a more accurate model (for example learned from typical user motions) could possibly reduce the required number of particles.

In this way four distinct events can occur: one target may be birthed to a particle, one may be removed from a particle, a target may be birthed and another removed and, finally, no change in the particle's target set may occur from the previous time-step except dynamical propagation.

C. Importance Weights

Having determined the particle set for the current iteration, the importance weights will be updated using

$$w_k^{(p)} \propto w_{k-1}^{(p)} \frac{l(\mathbf{Z}_k|\mathcal{A}_k^{(p)})p(\mathcal{A}_k^{(p)}|\mathcal{A}_{k-1}^{(p)})}{q(\mathcal{A}_k^{(p)}|\mathcal{A}_{k-1}^{(p)},\mathbf{Z}_{1:k})} \quad (16)$$

where the likelihood term $l(\mathbf{Z}_k|\mathcal{A}_k^{(p)})$ is determined up to a constant of proportionality by using a likelihood ratio calculation, as in [4], [3]. The formulation as a likelihood ratio implies that we only need evaluate this function at the grid cells that contain targets, and the computation need only be made once for each grid cell (see [3] for more details), and used by each particle containing a target within that cell. The likelihood ratio is calculated as

$$l(\mathbf{Z}_k|\mathcal{A}_k^{(p)}) = \prod_{s=1}^{S_k} l(z_{i_s}|\alpha_k^{s(p)}) \quad (17)$$

where $l(z_{i_s},\alpha_k^{s(p)})$ is the individual target likelihood ratio for target s located in cell i_s.

As already implied, data Z_k are obtained on a discrete grid of spatial cells, covering the tracking region of interest. The measurement value, z_{i_s}, is derived from the steered response power of the SBF steered to the centre of that cell, evaluated using

$$\mathcal{S}(r) = \int_\Omega \left| \sum_{m=1}^{N_m} X_m(\omega)W_m(\omega)\mathrm{e}^{j\omega T_m(r)/c} \right|^2 d\omega \quad (18)$$

at the cell centre, $r_j = (x_j, y_j)$ where $X_m(\Omega)$ is the Fourier transform of a frame of audio recorded at microphone m and ω is the set of integration frequencies. The total number of microphones is denoted as N_m. The measured quantity $\mathcal{S}(r)$ is known as the Steered Response Power (SRP). The exponential term is used to correct for the time-of-flight between the speech source and each sensor, $\tau_m = T_m(r)/c$, where the distance between the steering location and the known microphone position m is $T_m(r) = \|r - r_m\|$. The SBF cells have a 10cm spacing and are integrated over the frequency range $\Omega \in [200, 6000]$Hz, in practice implemented as a summation over DFT bins.

To better define the measurement function a nonlinear mapping of the SRP values, $z(r) = \Phi(\mathcal{S}(r))$, will redistribute measurements on the range $z(r) \in \{0, 1\}$ using a normal CDF

$$z_r = \Phi(\mathcal{S}(r_j); \bar{\mathcal{S}}, \sigma_{\mathcal{S}}^2) \quad (19)$$

where the mean and variance of the distribution will be set to be $\bar{\mathcal{S}} = 5500$ and $\sigma_{\mathcal{S}} = 500$. These parameters are chosen after careful review of the data — such that typical measurements recorded in noise will be at the lower end of this range, while measurements for active sources will be at the upper end of the range.

The likelihood ratio is calculated as a ratio of the signal-plus-noise likelihood function and the noise only likelihood function using the following expressions

$$\begin{aligned} l(z_{i_s}|\alpha_k^{s(p)}) &= \frac{p_{S+N}(z_{i_s}|\alpha_k^{s(p)})}{p_N(z_{i_s}|\alpha_k^{s(p)})} \\ p_{S+N}(z_{i_s}|\alpha_k^{s(p)}) &= c_1(\mathcal{N}(z_{i_s};1,\sigma_1) + q_1) \\ p_N(z_{i_s}|\alpha_k^{s(p)}) &= c_0(\mathcal{N}(z_{i_s};0,\sigma_0) + q_0) \end{aligned} \quad (20)$$

where the likelihood functions are both normal distributions with variances $\sigma_0 = 0.5$ and $\sigma_1 = 0.01$ and the constants were typically set to zero $q_1 = 25$ and $q_0 = 20$ so as to support some heavy tailed behaviour. Finally c_0 and c_1 are normalisation constants. Using likelihood ratio is important as it allows us to avoid evaluating this (continuous) function across the entire region, which would be required for proper normalisation of the underlying likelihood function, as outlined in [4]. To our knowledge this issue was not dealt with in previous works.

IV. Existence Grid

An important part of our approach is an effective proposal mechanism for initiating new targets and deleting old ones. Without a carefully designed data-dependent proposal mechanism the algorithm is likely to suffer from poor exploration of the variable dimension target space. To achieve this we adopt an existence grid approach, based closely upon [5], but with likelihood functions carefully designed for acoustic localisation. This existence grid is a low resolution grid reflecting our belief in the existence of target(s) in each of the grid cells and is used to evaluate the removal and addition probabilities mentioned in Sec III-B.

A. Design Choices

As mentioned previously, the Steered Beamformer function (SBF, see Eq 18) provides an indirect measure of how likely it is that a speech sample originated at a particular location. The function allows for two free design parameters (a) the frequency range used for the integration (affecting the precision of the location estimates) and (b) the set of locations evaluated (affecting the spatial extent of the evaluated surface).

Evaluating the entire surface using the full range of frequencies is, of course, impractical and hence a compromise is necessary. Evaluating the SBF function using a low band of frequencies (in our case we have chosen $\Omega \in [100, 400]$ Hz) reduces the peaked nature of the underlying surface as it is limited by the signal wavelengths. Fig 1 illustrates the SBF evaluated using two different frequency ranges. Thus a low resolution grid with cell dimensions in the order of 60-120cm across, can provide a coarse estimate of speaker activity. Using the Bayesian update framework discussed by Moreland *et al*

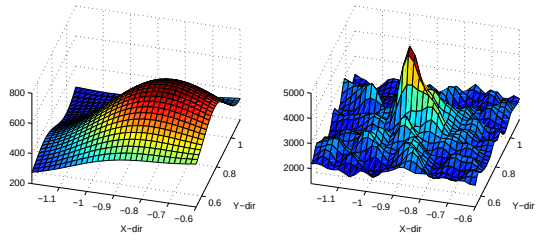

Fig. 1. *Comparison between SBF functions for a 12 microphone audio frame. Left: the SBF surface for the frequency range 100-400Hz. Right: the same surface for 200-6000Hz. The true source position was* $[-0.92, 0.80]$. *Only a* ***low-density*** *version of lower frequency surface will be evaluated for the existence grid in Sec IV.*

	(a)	(b)	(c)
Frequencies used	200-6000	100-400	2000-6000
3dB peak width	5cm	50cm	5cm
Grid cell size	N-A	80cm	∼5cm
No. Frequencies	742	38	742
Total evaluations	∼100	100	∼6400
Relative Computation	∼742×100	38×100	∼742×6400
Used in this algorithm	Yes	Yes	No

TABLE I
Computational comparison between (a) Likelihood-SBF, (b) Existence-SBF and (c) Likelihood-SBF Grid

[5], the obtained values can be combined with previous data to give a posterior estimate of cell activity — the existence grid.

Finally, because of the two design choices mentioned above the computational draw of this module is very small — especially when compared with the ensuing particle filter. Table I provides a comparison between the evaluation of this surface, the final particle filter likelihoods and also a full density SBF grid.

B. Desirable traits of the Existence Grid

In determining source activity, we will choose to place higher weight on quickly finding newly active sources than on quickly removing sources that have become inactive. Operating at 30Hz, activity will be recognised within an existence cell after just a couple of cycles. Conversely when the source becomes inactive or leaves a grid cell, the existence cell value will die away gradually over the course of a second — returning to the background level.

C. Evaluating the Grid

First the SBF function (as in Eq 18), will be evaluated for a grid of J cells each of size, $(\triangle x, \triangle y)$, spread across the surveillance region. The cells will be numbered $j = 1, \ldots, J$. Again the SRP values will then be transformed onto a range $[0, 1]$ using the CDF mentioned in Sec. III-C. In this case the mean and variance chosen were $\bar{\mathcal{S}} = 450$ and $\sigma_{\mathcal{S}} = 50$ - as only the low band of frequencies are involved in the SBF integration. Using this measurement grid as an input, we will now update an existence grid across the surveillance region.

Moreland et al, [5], presented a two step Bayesian update rule for the measure of the evidence that a source exists in a particular cell which is defined as g_j^{t-1} for cell j at time $t-1$ (Eqs 48 and 49 from their paper). First, the previous estimate is updated using our prior information of how we expect the source's activity to evolve:

$$g_j^{t|t-1} = g_j^{t-1}(1 - \zeta_j^t) + (1 - g_j^{t-1})\epsilon_j^t \tag{21}$$

where $g_j^{t|t-1}$ is the updated measure. Subsequently we update the existence values using information drawn from the current measurements

$$\begin{aligned} g_j^t &= \frac{g_j^{t|t-1} p(z^t|o_j^t = 1)}{p(z^t)} \\ &= \frac{g_j^{t|t-1} p(z^t|o_j^t = 1)}{g_j^{t|t-1} p(z^t|o_j^t = 1) + (1 - g_j^{t|t-1}) p(z^t|o_j^t = 0)} \end{aligned} \tag{22}$$

which gives us the existence measure for the current iteration. Note that g_j^t is bounded within the range $[0, 1]$. o_j^t is a binary label of either inactivity ($o_j^t = 0$) or activity ($o_j^t = 1$). A large value means a high chance of the existence of a target in that cell but it is not a probability, per se. To implement this update rule we need to formulate the likelihood functions $p(z_j|o_j = 1)$ and $p(z_j)|o_j = 0)$.

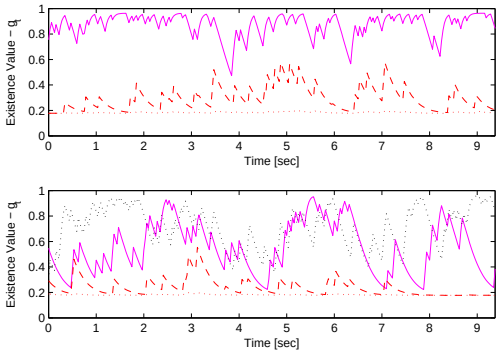

Fig. 2. *Upper: Evolution of existence grid values for a single source. Lower: Two moving sources speaking simultaneously. The Magenta solid line and dotted black is the evolution of the existence grid values of the source cells, while dashed red is the maximum existence grid value of empty cells. See Sec IV-D.*

D. Existence Grid Likelihood Functions

The likelihood functions we shall propose will simply be as follows, for cell j:

$$\begin{aligned} p(z_j|o_j^k = 1) &= c_1(\mathcal{N}(z_j; 1, \sigma_1) + q_1),\ 0 < z_j < 1 \\ p(z_j|o_j^k = 0) &= c_0(\mathcal{N}(z_j; 0, \sigma_0) + q_0),\ 0 < z_j < 1 \end{aligned} \tag{23}$$

where q_1 and q_0 allow some heavy-tailed behaviour in both active and inactive cases. c_0 and c_1 are the normalising constants necessary to normalise the probability density functions in the interval $[0, 1]$. z_j is the (CDF-transformed) low frequency steered response power evaluated at the centre of cell j. After optimisation the following parameters were used: active source $\sigma_1 = 0.018$ and $q_1 = 15$; inactive source $\sigma_0 = 0.54$ and $q_0 = 25$.

The likelihood functions highly reward measurements deemed to have originated from the source (modified SRP value ~ 1) but only very mildly weight against less informative clutter measurements (modified SRP value $\sim$ 0). Fig 2 illustrates the evolution of the existence function for two recorded samples: while the grid provides an indication of source activity, it does so in a somewhat unreliable manner. This issue is returned to in Sec V.

The entire procedure produces, at each time frame k and for each cell j, a measure g_j of the activity of target(s) within that cell. These values are used, with the active targets from the previous time frame, to propose target initiations and deletions within the particle filter, which is now described.

E. Target Proposal Mechanism

Having evaluated the existence grid values, next we find probabilities for adding a new target or removing an existing target which will

be used to propose the particle set. To do so we will use a slightly modified approach from [5], given here in shortened form.

c) Addition Probabilities:: Consider a particle at time $k-1$ made up of s^{k-1} targets, $\mathcal{A} = (\alpha_1, \ldots, \alpha_{s^{k-1}})$, located in cells $(l_1, \ldots, l_{s^{k-1}})$. The relative probability of adding a new target to a specific cell, j, will be the product of the probability of a target existing in that cell with each of the probabilities of a target *not* existing in each of the other (vacant) cells,

$$\nu_j = \frac{g_j^k}{1-g_j^k} \prod_{i \in A^k} (1 - g_i^k) \tag{24}$$

and the probability of no target being added will be as follows

$$\nu_0 = \prod_{i \in A^t} (1 - g_i^t) \tag{25}$$

The normalised set of addition probabilities, $\mathcal{V} = \{\nu_0, \ldots, \nu_J\}$, are denoted where $\sum_{j=0:J} \bar{\nu}_j = 1$. $\bar{\nu}_0$ is the probability of no target being added while the sum of the remaining values represents the probability of any of the targets being added.

d) Removal Probabilities:: The set of relative probabilities of a target *not existing* in the cell l_s^k given a target combination is found using Eq 53 from [5]:

$$\tau_s^k = \beta_s^k \frac{1 - g_{l_s^k}^k}{1/s^{k-1} \sum_{r=1}^{s^{k-1}} (1 - g_{l_r^k})}. \tag{26}$$

where β_s^k is prior target occupancy constant which can be used to reflect regions which are more or less likely to be occupied.

This set of values correspond to the existence grid values, g_j^k, and are used to form the removal probabilities, $\kappa_{0:S_{k-1}}$, and their normalised versions, $\bar{\kappa}_{0:S_{k-1}}$, which correspond to those given in the previous section for addition. While the method allows only a single target to become active or become inactive at a particular time frame, this will not effect our envisaged application.

Thus this provides us with a logical framework for proposing particle sets to reflect the underlying activity of the different regions of the surveillance space which can in turn be used to update the posterior distribution of the number of targets and their positions.

An alternative approach would be to propose completely new targets into newly active existence grid cells which are maintained in a nursery before being transferred to the mature target set when reinforced by subsequent audio frames. This would make for an interesting comparison with our proposed method.

Due to the temporal discontinuity of speech, one must trade off the better tracking accuracy of a dominant source against improved tracking stability of weaker, less active sources. A careful choice of likelihood and resampling parameters is required. For example, when two sources are active one can typically expect only 45% of frames to give accurate location estimates while a similar proportion will contain contain clutter measurements. Because of this it is important to ensure that re-sampling occurs infrequently to avoid degeneracy.

V. Experiments

So as to test the algorithm, a set of recordings were made in a typical office room with 12 microphones spaced around the roughly 5m x 5m space illustrated in Fig 5. The setup and other details were identical to that used in [3]. The number of particles was fixed at 500 which allowed for realtime operation in MATLAB on a typical PC (1.20GHz, Dual Core, 2GB RAM). Note that given the size of the room available for these experiments the mean velocity of the speakers is quite low - corresponding to a slow walking pace. For higher walking speeds it is anticipated that more particles would be required to support diverse dynamics as well as possibly lower frequency of accurate measurements.

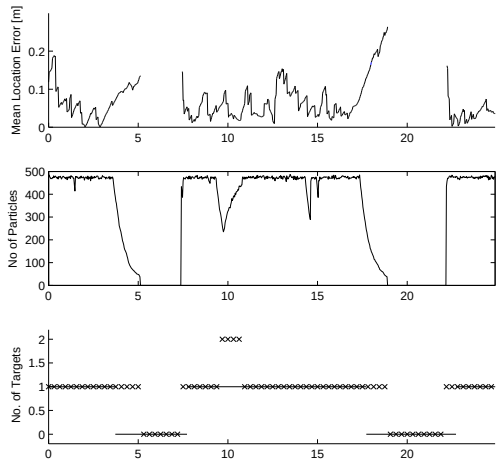

Fig. 3. *Tracking of a single intermittent speech source moving in the path illustrated in Fig 5. Top figure: position error; centre figure: the number of active particles; bottom figure: mean number of estimated sources (crosses) versus the true number (line). See Sec V-A for details.*

A. Tracking Examples

Intermittent Single Speaker: Fig 3 shows the performance of the algorithm for a single intermittent speech source moving in the path indicated by Source 2 in Fig 5 over the course of 25 seconds. The system correctly identified source activity and inactivity. As intended the source activity was more quickly determined than inactivity. Position error is typically below 0.1m, although the average error is effected by periods in which the source becomes silent but before the tracking particle set disappeared.

Two Conversing Speakers: Fig 4 depicts tracking performance in the $\mathcal{X}$ and $\mathcal{Y}$-dimensions for two alternating speakers taking part in a 20 second conversation. The location of each source during **active** speech is indicated by a red dashed line, while the algorithm's tracking performances is indicated by a solid blue line. Variance of the estimate is indicated by error bars. The algorithm is seen to correctly identify and track the active source and to quickly switch between the speakers.

Two Overlapping Speakers: Fig 5 illustrates the tracking of two sources alternating between activity and inactivity including when both sources are simultaneously speaking. The upper plot illustrates tracking performance while the lower plot illustrates the number of sources estimated to be active. As mentioned previously the existence grid gives a coarse indication of regional activity. Having proposed particles in these broad locations, the more accurate particle filter then tracks the source location precisely. The algorithm is seen to preform tracking of both of the sources successfully — both when they were active and where they were inactive.

B. Monte Carlo Simulations

Finally we will present the results of a series of Monte Carlo simulations to present performance in a quantitative manner. As mentioned previously, the proposed algorithm is unique in determining the presence, activity and continuity of speech sources in an entirely probabilistic manner[1] and as such no method which can be compared with the proposed algorithm. Hence the results presented give only an illustrative example of the algorithm's performance.

Tracking of one intermittent source is examined here using the first sample from Sec V-A which follows the path of Source 2 in Fig 5

[1]Although the approach taken by Ma et al. [9] has the capacity, if re-implemented using the SBF measurement framework, to behave similarly.

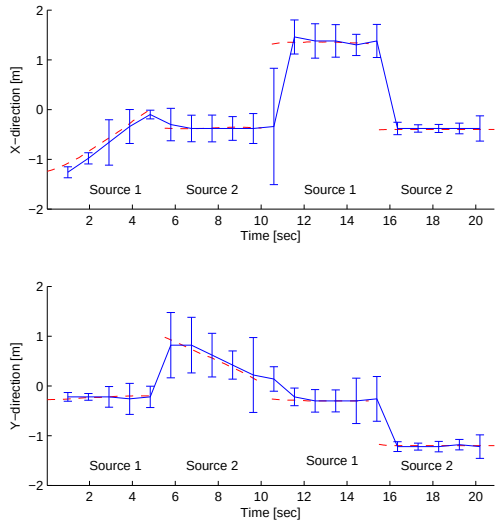

Fig. 4. *Tracking two sources in conversation using the algorithm presented in Sec III. Solid lines show the estimated position while dotted is the ground truth. Note how at 10 seconds the error bars indicate high uncertainty in the silent gap between the speakers before continuing accurate tracking.*

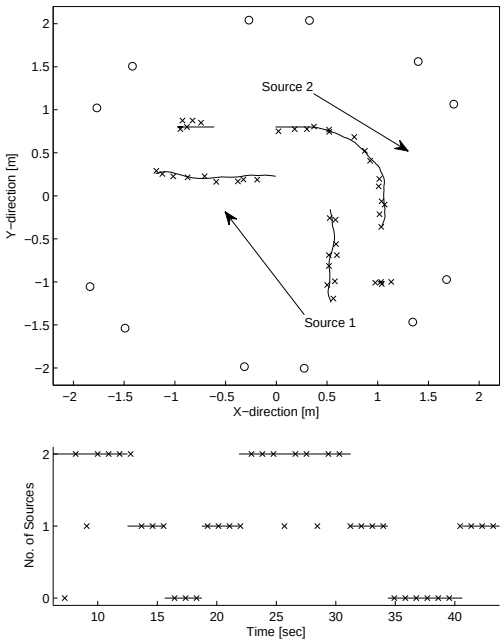

Fig. 5. *Top figure: tracking of two intermittent and overlapping speakers (see Sec III). Lower figure: the estimate of the number of sources deemed active (crosses) compared to the number that actually were active (line).*

was used (before the linear addition of Source 1). This consisted of three periods of silence followed by three periods of speech activity and was 67 seconds in duration. The algorithm was run 50 times and the results were averaged. The following metrics are presented to illustrate performance:

1) Mean Location Error (of frames when the source is active): 0.055m. This error is of a similar size of a person's mouth and within the margin of error of the ground truth system.
2) Percentage of (active) frames that 70% of particle weight lies within 0.2m of the source position: 98.9%. Illustrating that tracking is stable — although it must be acknowledged that the test sample, while realistic, was not very challenging.
3) Mean error of the estimated number of targets: 0.394 targets **more** are estimated. Our implementation of the system overestimates the number of targets so as to avoid missing a target which is important in a number of envisaged applications.
4) Mean time taken for particle weight within 0.2m of the source position to rise to 70% (when becoming active): 0.28 seconds. As envisaged in Sec IV, quickly detecting new sources.
5) Mean time taken for particle weight within 0.2m of the source position to fall to 30% (when becoming inactive): 0.87 seconds. Inactive sources are removed more slowly.

VI. Conclusions

A **fully probabilistic** entirely integrated algorithm for the detection and tracking of an unknown and time varying number of speakers has been proposed and demonstrated with real audio recordings. While there exists scope for further optimisation of the algorithm, the results illustrates the ability of the system to track more than one source simultaneously in real-time in a computationally efficient manner. In particular the algorithm does not rely on external modules to propose target or to keep track of targets. Additionally, this system supports null particles, explicitly containing no target states when none are supported by the audio data, which is a unique yet probabilistically correct approach.

Improvement of the stability of the existence grid mechanism is still possible. Currently the existence grid is implemented using a grid of non-overlapping cells which can lead to instability when a target moves from one cell to the next. An alternative system utilising two interleaved mesh grids could possibly remove this instability while requiring only a small increase in computing power.

A limitation of the algorithm (and AST in general) is the maximum number of active sources (about 3). This is due to the shared acoustic channel which results in a reduced frequency of observations with an increasing number of speakers and the breakdown of the tracking algorithm. This restriction could be improved with notch filtering or binary masking of dominant speakers to expose the weaker speaker.

References

[1] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 3021–3024, 2001.
[2] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, Nov. 2003.
[3] M. Fallon and S. Godsill, "Acoustic source localisation and tracking using Track Before Detect," *IEEE Transactions on Speech and Audio Processing*, vol. 18, no. 6, pp. 1228–1242, 2010.
[4] D. J. Salmond and H. Birch, "A particle filter for track-before-detect," in *Proceedings of the American Control Conference*, vol. 5, 2001, pp. 3755–3760.
[5] M. Morelande, C. Kreucher, and K. Kastella, "A Bayesian approach to multiple target detection and tracking," *IEEE Transactions on Signal Processing*, vol. 55, pp. 1589–1604, May 2007.
[6] D. E. Sturim, M. S. Brandstein, and H. F. Silverman, "Tracking multiple talkers using microphone array measurements," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 1997, pp. 371–374.
[7] K. Nakadai, H. Nakajima, M. Murase, S. Kaijiri, K. Yamada, T. Nakamura, Y. Hasegawa, H. G. Okuno, and H. Tsujino, "Robust tracking of multiple sound sources by spatial integration of room and robot microphone arrays," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, May 2006.
[8] E. A. Lehmann and R. C. Williamson, "Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments," *EURASIP Journal on Applied Signal Processing*, 2006.

[9] W.-K. Ma, B.-N. Vo, S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using tdoa measurements: A random finite set approach," *IEEE Transactions on Signal Processing*, vol. 54, pp. 3291–3304, Sep. 2006.
[10] P. Pertilä and M. Hämäläinen, "A track before detect approach for sequential Bayesian tracking of multiple speech sources," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010.
[11] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, pp. 216–228, 2007.
[12] N. Gordon, D. Salmond, and A. F. M. Smith, "Novel approach to nonlinear and non-Gaussian Bayesian state estimation," *IEE Proceedings (F)*, vol. 140, pp. 107–113, 1993.