MULTI TARGET ACOUSTIC SOURCE TRACKING WITH AN UNKNOWN AND TIME VARYING NUMBER OF TARGETS

Maurice Fallon and *Simon Godsill*

Signal Processing Laboratory
University of Cambridge
Trumpington Street, Cambridge
CB2 1PZ, UK
Email: mff25@cam.ac.uk

ABSTRACT

Particle Filter-based Acoustic Source Tracking algorithms track (online 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 previously introduced multi-target (MTT) methodology to allow for an unknown and time-varying number of speakers. Finally examples show typical tracking performance in a number of different scenarios with simultaneously active speech sources.

Index Terms— Monte Carlo methods, Microphones, Acoustic tracking, Filtering, Speech processing

1. 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 very recently from tracking single-source recordings in synthetic environments [1], to tracking in real and challenging environments [2], and recently to tracking multi-source recordings [3]. However these algorithms typically assume that the source(s) are active from the start of the algorithm and run to its end without any major silent pauses - which is obviously an over-idealisation.

Previously we introduced a methodology for multi-target tracking of acoustic sources. The method avoided data association by use of the track-before-detect paradigm, [4], and tracked multiple sources simultaneously. Again this technique assumed knowledge of the number of sources in the surveillance region as well as their initial positions. In the following an entirely probabilistic strategy is proposed which identifies newly active sources, keeps track of them and removes them when they become inactive.

Note that the Steered Beamformer Function (SBF) is used to isolate localisation information from each frame of microphone array audio, as previously used in [2, 3].

2. EXISTENCE GRID

An important part of our particle filtering algorithm is an effective proposal mechanism for initiating new targets and deleting existing ones. An approach which does not include such a carefully designed data-dependent proposal mechanism is likely to suffer from poor exploration of the variable dimension target space. To achieve this goal we adopt an existence grid approach, based quite closely upon [5], but with likelihood functions carefully designed for our acoustic localisation framework. This existence grid is a low resolution grid overlayed on the surveillance region and updated at each iteration to reflect our belief in the existence of target(s) in each of the cells of the grid.

Evaluating the SBF function using a low band of frequencies, in this case $\Omega \in [100, 400]$Hz, reduces the peaked nature of the underlying surface, as discussed in [6]. As a result a low resolution grid, with J cells with cell dimensions in the order of 60-120cm across, can provide a coarse estimate of regional activity for the current frame of audio. Using the Bayesian update framework discussed by in [5], this estimate can be combined with previous data to give a posterior estimate of activity in each cell. It is important to note that because of these two design choices the computational draw of this module is very small, especially when compared with the ensuing particle filter.

While details of the updating procedure can be found in [5], it is necessary to design likelihood functions for each existence cell, given it contains at least one target, $p(z_j|o_j = 1)$, or no targets, $p(z_j|o_j = 0)$.

Having first used a normal CDF to map the SBF values onto the range $[0, 1]$ (similar to that used in [3]), the likelihood

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

functions for cell j will then be as follows:

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

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 pdfs in the interval $[0,1]$. z_j is the (CDF-transformed) low frequency steered response power evaluated at the centre of cell j. Variance and noise floor constants used herein are as follows, based on careful tuning to real datasets: Active Source: $\sigma_1 = 0.02$ and $q_1 = 7$; Inactive Source: $\sigma_0 = 0.4$ and $q_0 = 40$. Note the large difference between the variances used - which illustrates that an *active source measurement* is deemed to be much more informative than an *inactive source measurement*.

The whole procedure produces, at each time frame k and for each cell j, a probability g_j for activity of targets. These values, in association with the configuration of active targets within particles at the previous time frame, are used to propose target initiations and deletions within the particle filter, which is now described.

3. TRACKING FRAMEWORK

The tracking system will utilise a variable-dimension particle filter to keep track of the time-varying number of sources present in the room. The strategy is similar to the framework of [7], combined with an activity grid-based target proposal method similar to [5]. The number of targets, T_k, within each individual particle may vary in the range $\{0,\ldots,T_{\text{max}}\}$, representing the number of speakers deemed to be active at any given time k. T_{max} is the maximum number of simultaneously active speakers and is chosen to be 3 in our experiments. An individual particle state-space, containing T_k targets at time k, is defined as follows

$$\mathcal{A}_k = (\alpha_k^1,\ldots,\alpha_k^{T_k},T_k) \quad (2)$$

with an associated particle weighting w_k. Each target, α_k^t, will contain position and velocity components in the $\mathcal{X}$ and $\mathcal{Y}$-dimensions, $\alpha_k^t = (x_k^t, y_k^t, \dot{x}_k^t, \dot{y}_k^t)$. The aim of particle filtering is to update the posterior probability density for the entire vector given in (2) using information drawn from the current measurement set, $\mathbf{Z}_k$.

3.1. Data Model

Each active target within the state-space system will be modelled to evolve according to a nonlinear state transition equation based on the Langevin dynamical model which has been used previously in this field, see [1, 2, 3]. This will allow us to form the dynamic model in for a target α_k^t which has been active in frames $k-1$ and k. In terms of probability densities we have then that $p(\alpha_k^t|\alpha_{k-1}^t) = \mathcal{N}(\alpha_k^t;\mathbf{f}(\alpha_{k-1}^t),\sigma_{\mathbf{e}}^2)$ where the formulation is as in Eq. (8) of [6].

Within our framework we propose also to model the random appearance ('birth') and disappearance ('death') of speakers. For simplicity we will assume at most one target may appear or disappear at each time step:

$$T_k = T_{k-1} + \epsilon_k \quad (3)$$

and will do so with a prior probability distribution

$$p(T_k|T_{k-1}) = \begin{cases} \Pr(\epsilon_k=-1) = h_d \\ \Pr(\epsilon_k=0) = 1-h_b-h_d \\ \Pr(\epsilon_k=1) = h_b \end{cases} \quad (4)$$

where h_b and h_d are probabilities of incrementing and decrementing the number of targets, respectively. In general h_b and h_d will be set equal, except when T_{k-1} is equal to 0 or T_{max}.

The prior state distribution of new target births, $p_0(\alpha_k^t)$, 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. To maintain the generality of our approach 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, $p_0(x_k^t, y_k^t) = \mathcal{U}_{\mathcal{S}}(x_k^t, y_k^t)$, where $\mathcal{S}$ will be the volume of the entire surveillance region. Secondly, the prior distribution of the velocity components $p_0(\dot{x}_k^t, \dot{y}_k^t)$ will be initiated normally around zero velocity to give

$$p_0(\alpha_k^t) = p_0(x_k^t, y_k^t) \times p_0(\dot{x}_k^t, \dot{y}_k^t) \quad (5)$$

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

$$p(\mathcal{A}_k|\mathcal{A}_{k-1}) = p_\alpha(\alpha_k^{1:T}|\alpha_{k-1}^{1:T}, T_k, T_{k-1}) p_T(T_k|T_{k-1}) \quad (6)$$

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

$$p_\alpha(\alpha_k^{1:T_k}|\alpha_{k-1}^{1:T_k}, T_k, T_{k-1}) = \begin{cases} \prod_{t=1,t\neq t'}^{T_{k-1}} p(\alpha_k^t|\alpha_{k-1}^t) & \text{if } T_k = T_{k-1}-1 \\ \prod_{t=1}^{T_{k-1}} p(\alpha_k^t|\alpha_{k-1}^t) & \text{if } T_k = T_{k-1} \\ p_0(\alpha_k^{T_k}) \times \prod_{t=1}^{T_{k-1}} p(\alpha_k^t|\alpha_{k-1}^t) & \text{if } T_k = T_{k-1}+1 \end{cases} \quad (7)$$

and t' is the target removed at time k.

3.2. Sequential Monte Carlo Methods

As mentioned above our goal is to estimate the joint posterior distribution of the target states recursively, and we adopt the standard two step Bayesian update rule. However for many models of interest the evaluation of the integral and update steps is intractable. As a result Sequential Monte Carlo methods have been proposed to approximate the recursion for such complex measurement or dynamical models. The basic idea is that a complex probability distribution can be represented

as a set of weighted Monte Carlo importance samples, see [8] for a recent survey.

The problem at hand has many state variables and moreover has a time-varying number of speakers encoded into it. Hence, instead of sampling from the dynamical model alone, as would be done in the standard bootstrap versions of particle filtering, we will instead sample the ith particle for the new state vector from an appropriately selected proposal function

$$\begin{aligned}\mathcal{A}_k^{(i)} &\sim q(\mathcal{A}_k|\mathcal{A}_{k-1}^{(i)}, \mathbf{Z}_{1:k}) \\ &\sim q_\alpha(\alpha_k^{(:)}|\alpha_{k-1}^{(:)(i)}, T_k^{(i)}, T_{k-1}^{(i)}, \mathbf{Z}_{i:k}) q_T(T_k|T_{k-1}^{(i)}, \mathbf{Z}_{i:k})\end{aligned} \quad (8)$$

where $q_\alpha(\cdot)$ and $q_T(\cdot)$ 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 (see again [8] for details).

According to (8), we first propose the new target number in time-frame k by first removing unsupported targets and then adding targets to newly active regions of the existence grid as follows.

1. Removal of targets: Using the existence cell probabilities evaluated in Section 2, $g_{1:J}$, a set of relative probabilities for the removal of a target are evaluated, using Eq. (52) of [5], $\nu_1, \ldots, \nu_{T_{k-1}}$. The sum of these terms, representing the overall probability of any one of the targets being removed, is $\mathcal{V}_k$. This, along with a constant probability for the removal of no target, are used to draw a decision of whether a target is removed or not

$$q(T_{k|k-1}|T_{k-1}^{(i)}) = \begin{cases} \Pr(\epsilon_{k|k-1} = -1) = \bar{\mathcal{V}}_k \\ \Pr(\epsilon_{k|k-1} = 0) = 1 - \bar{\mathcal{V}}_k \end{cases}$$

Should the removal of a target be decided upon, a random draw from the set of target removal probabilities is made, the associated target is removed and the intermediatory target number is decremented as follows $T_{k|k-1}^{(i)} = T_{k-1}^{(i)} - 1$. Otherwise no action is made.

2. Initiation of new targets: In a similar manner to the above a set of relative probabilities for the addition of a new target are evaluated, $\kappa_1, \ldots, \kappa_J$ and the sum of the these probabilities is $\mathcal{K}_k$. A decision is then made

$$q(T_k^{(i)}|T_{k|k-1}^{(i)}) = \begin{cases} \Pr(\epsilon_k = 0) = \bar{\mathcal{K}}_k \\ \Pr(\epsilon_k = 1) = 1 - \bar{\mathcal{K}}_k \end{cases}$$

where $(1 - \bar{\mathcal{K}}_k)$ is the (normalised) probability of adding no new targets. Should a new target addition be decided upon, a random draw is made to chose a cell in which it should be initiated.

Having selected the cell, the target position is initialised using a weighted combination of a uniform distribution within the physical region of cell, S_j, and a normal distribution centred on the weighted mean of any particle states currently existing in that cell, the idea being that some particles may have detected the correct object position in an earlier time frame,

$$\begin{aligned}\alpha_k^{t(i)} &\sim q_0(\alpha_k^t|\mathbf{Z}_{1:t}) \\ &\sim \beta\mathcal{N}(\alpha_k^t; \bar{\alpha}_k^{(j)}, \bar{\sigma}_k^{2(j)}) + (1-\beta)\mathcal{U}_{\mathcal{S}_j}(\alpha_k^t)\end{aligned} \quad (9)$$

3. Updating of persistent target positions: Finally the states of targets persisting from time-step $k-1$ are propagated using the Langevin dynamical model, $\alpha_k^{t(i)} \sim q(\alpha_k^t|\alpha_{k-1}^{t(i)}, \mathbf{Z}_k)$.

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 by be birthed and another removed and finally no change in the target set may occur from the previous time-step.

3.3. Importance Weights

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

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{l(\mathbf{Z}_k|\mathcal{A}_k^{(i)}) p(\mathcal{A}_k^{(i)}|\mathcal{A}_{k-1}^{(i)})}{q(\mathcal{A}_k^{(i)}|\mathcal{A}_{k-1}^{(i)}, \mathbf{Z}_{1:k})}. \quad (10)$$

where the likelihood term is determined up to a constant of proportionality by using a likelihood ratio calculation, as in [4, 3]. The SBF grid with a 10cm density integrated over the frequency range of 200-6000Hz. The formulation as a likelihood ratio implies that we only need evaluate this function at the grid cells that contain targets, and the computation needs only be made once and stored for each grid cell, and not for each particle containing a target within that cell. This SBF surface is computed separately from the low resolution SBF required for the activity grid detector in Section 2 and the likelihood ratio is calculated as

$$l(\mathbf{Z}_k|\mathcal{A}_k^{(i)}) = \prod_{t=1}^{t=T_k} \exp(\frac{2z_t - 1}{2\sigma_N^2}) \quad (11)$$

where the target is located in cell t and z_t is derived from the steered response power of the SBF steered to the centre of that cell in the same way as (1). This likelihood ratio is the same as was used in [3]. Note that this ratio is a special case of the likelihoods of the form found in 1 with $\sigma_0 = \sigma_1 = \sigma_N$ and $q_1 = q_0 = 0$.

Finally it should be noted that because of the temporal discontinuity of speech, for multiple acoustic source tracking it is necessary to trade off the better tracking accuracy of a dominant source against improved tracking stability of weaker, less active sources. This trade-off involves careful choice of the likelihood parameters and judicious use of resampling strategy parameters.

4. EXPERIMENTS

To test the algorithm, a set of recordings were made in a typical office room with twelve microphones spaced around a

roughly 5m x 5m space and illustrated in 2. The setup and other details were identical to that used in [3]. 500 particles were used in iterations which allowed for in realtime operation in MATLAB on a typical PC.

Figure 1 depicts tracking performance in the $\mathcal{X}$ and $\mathcal{Y}$-dimensions for two alternating speakers taking part in a conversation. The duration of the sample is 20 seconds. The location of each source during **active** speech is indicated by a solid line. Overlayed is the results of a typical run of the algorithm. Every twentieth estimate of the estimated active source location is given while the variance of the estimates is indicated by error bars. The algorithm can be seen to correctly identify and track the active source and to quickly switch between the two speakers.

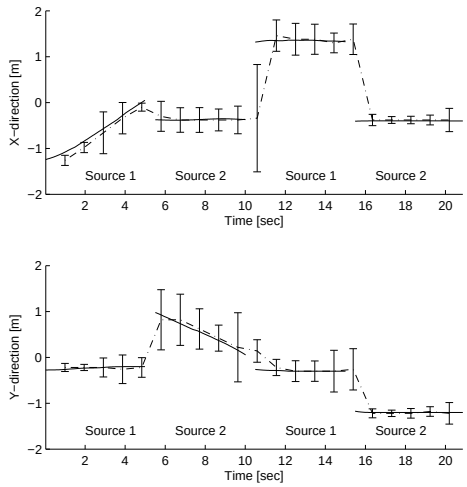

Fig. 1. Example of tracking two sources in conversation. Note that at 10 seconds the error bars indicate high uncertainty in the silent gap between the speakers.

Figure 2 illustrates the tracking of two sources alternating between activity and inactivity which includes segments in which both sources are simultaneously speaking. The upper plot illustrates tracking performance in both X and Y dimensions with source position estimates indicated by crosses. The lower plot illustrates the number of sources estimated to be active (again indicated by crosses) compared to the number that actually were. The algorithm is seen to preform both of the tasks successfully.

5. CONCLUSION

A probabilistic algorithm for the detection and tracking of an unknown and time varying number of speaker has been proposed and demonstrated. While there exists considerable scope for further optimisation of the algorithm, the results illustrate an ability to track more than one source simultaneously and in real-time. The main limitation of the algorithm is a maximum number of simultaneous active sources, could perhaps be improved by notch filtering of dominant speakers.

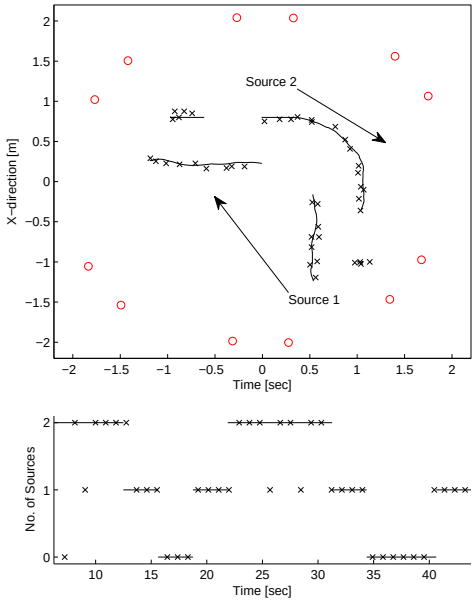

Fig. 2. Tracking two simultaneously active sources.

6. REFERENCES

[1] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *Proc. ICASSP 2001*, 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 Trans. on Speech and Audio Processing*, 2003.

[3] M. Fallon and S. Godsill, "Multi target acoustic source tracking using track before detect," in *Proceedings of WASPAA 2007*, Oct. 2007.

[4] D.J. Salmond and H. Birch, "A particle filter for track-before-detect," in *American Control Conference, 2001. Proceedings of the*, 2001, vol. 5.

[5] M. Morelande, C. Kreucher, and K. Kastella, "A Bayesian approach to multiple target detection and tracking," *IEEE Trans. on Signal Proc.*, 2007.

[6] 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.

[7] W. Ng, J. Li, S. Godsill, and S.K. Pang, "Multitarget initiation, tracking and termination using bayesian monte carlo methods," *The Computer Journal*, , no. 6, 2007.

[8] O. Cappé, S.J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential monte carlo," *IEEE Proceedings*, , no. 5, 2007.