MULTI TARGET ACOUSTIC SOURCE TRACKING USING TRACK BEFORE DETECT

Maurice Fallon[*] *and Simon Godsill*

Signal Processing and Communications Laboratory
Cambridge University Engineering Dept
Trumpington Street, Cambridge, CB2 1PZ, UK.
mff25,sjg30@cam.ac.uk

ABSTRACT

Particle Filter-based Source Localisation algorithms attempt to track 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 introduces a multi-target methodology for acoustic source tracking. The methodology is based upon the Track Before Detect (TBD) framework. The algorithm also implicitly evaluates the source activity using a variable appended to the state vector. Examples of typical tracking performance are given using a set of real speech recordings with two sources active simultaneously.

1. INTRODUCTION

Localisation and tracking of speech sources has become an increasingly active area of research. This straightforward problem is complicated by the existence of background noise and reverberation. Furthermore speech by its nature is highly non-stationary - alternating between periods of high activity during a sentence and silence. While algorithms have been presented to track a single source, [1], extensions to the multi target environment have had limited success [2].

A novel algorithm for single source tracking will be presented using the Track Before Detect (TBD) methodology within a particle filtering framework. Using TBD allows a significant proportion of the computation associated with the evaluation of the likelihood function to be avoided. An extension of the framework to tracking multiple sources simultaneously is then illustrated in Section 4. By using the TBD methodology this algorithm avoids the need to associate the measurements to a particular source - an issue of great complication in Multi Target Tracking (MTT) [3].

2. ACOUSTIC SOURCE TRACKING

This paper concerns itself with the problem of tracking the location of moving speech source(s) in the $\mathcal{XY}$-plane. We consider N_m microphones in a typically noisy and reverberant room using the basic framework initially introduced by [4].

Assuming a batch of synchronised data of L samples from each sensor is available at time k, $\mathbf{X}_k = [\mathbf{x}_1(k), \ldots, \mathbf{x}_{N_m}(k)]^T$, we will use a *localisation* function to make a transformation between the audio frame data and a location estimate - $\hat{l} = \mathbf{f}(\mathbf{X_k})$.

Localisation measurement models are divided into two groups - those that provide indirect measurements from each microphone or microphone pair (such as Generalised Cross Correlation) which are then combined to give an overall location estimate and those that use the entire microphone data frame to make a single estimate such as the Steered Beamformer (SBF). We will focus on using the SBF as the measurement function for this paper.

2.1. Measurement Function: Steered Beamformer

The Steered Beamformer (SBF) value is a measure of how likely a full audio frame originated from a specific location. For continuity we will maintain the same notation used in [5] (Equation 2). The SBF steered to the physical location $l = [x, y]$ will be denoted $\mathcal{S}(l)$. It is noted that computation of the SBF represents a large proportion of the computational effort for any particle filter which utilises it.

2.2. Observations from real data

A number of observations are detailed here regarding the performance of the SBF - based on real recorded audio. Space limitations do not permit a technical discussion however. Firstly the width of SBF peaks are determined by the frequencies used to calculate the SBF. Because speech's maximum frequency is about 4000Hz the overall peak will typically have a width of about 10cm (above the noise floor) as discussed in [1].

As previously observed, [5], speech is a highly non-stationary signal meaning that successive frames may give clear distinct SBF peaks while others may be useless. We tackle this problem with an activity detector in Section 3.2.3. Finally when two sources are active simultaneously, the more active source is typically dominant in the resultant SBF trace. Clear SBF peaks from each source are generally observed in the momentary silent gaps between the other source's words and sentences.

3. TRACKING FRAMEWORK

3.1. Bayesian Filtering and SMC

We will define the source state vector at time k to be

$$\alpha_k \triangleq (x_k, \dot{x}_k, y_k, \dot{y}_k, \lambda_k) \tag{1}$$

where x_k and $\dot{x}_k$ are position and velocity of the source, respectively, in the $\mathcal{X}$-direction and similarly for the $\mathcal{Y}$-direction. The parameter λ_k, a source activity indicator, will be introduced in Section 3.2.3. Solution of the tracking problem will require the estimation of the source position portion of this vector - (x_k, y_k) at each time step using the Chapman-Kolmogorov equations:

$$\begin{aligned} p(\alpha_k|\mathbf{Z}_{1:k-1}) &= \int p(\alpha_k|\alpha_{k-1})p(\alpha_{k-1}|\mathbf{Z}_{1:k-1})d\alpha_{k-1} \\ p(\alpha_k|\mathbf{Z}_{1:k}) &\propto p(\mathbf{Z}_k|\alpha_k)p(\alpha_k|\mathbf{Z}_{1:k-1}). \end{aligned} \tag{2}$$

[*] This work was by Microsoft Research through the European PhD Scholarship Programme in collaboration with Professor Andrew Blake.

This non-linear and non-Gaussian problem has no closed form solution. An alternative approach is Sequential Monte Carlo (SMC) which attempts to carry out the above integrations on a large set of weighted discrete samples, also known as particles, which can then be used to estimate the posterior density. A general overview of the principles and background to SMC, or *particle filtering* as it is generally known, can be found in [6]. In the following sections the various components that are required to solve this problem will be introduced.

Source Dynamical Model: We will use the Langevin dynamical model introduced by Vermaak [4] and retained by Ward et al. [1]. The parameter values chosen will be $\beta = 6$Hz and $\bar{v} = 0.6$m/sec. In Section 4.2.1 the tracking algorithm is modified to track more than one source using a repulsive force as a modification to this dynamical model.

3.2. Track Before Detect

Classical approaches to tracking typically require an initial step to first extract a small number of position measurements from the raw sensor output (such as raw radar scans) using sensor signal processing. However this step usually requires a thresholding process which can lead to a loss of information. Also to calculate this function at a sufficient density of points so as guarantee the observation of the source peak, using the full frequency range of interest, is computationally prohibitive as noted by [1, 7]. As mentioned in Section 2.2 the SBF function peak widths are related to the range of frequencies used to calculate the SBF. It is suggested that a grid resolution of 10cm is sufficient to observe the majority of peaks.

In [7] the authors illustrated a method which made use of a hybrid particle filter using both grid points and free moving particles to track a moving source. However there appears to be no obvious way to extend this method to MTT as there isn't a measurement set which can be assigned to either source (as the measurement function is only evaluated at the actual particle locations). An alternative approach is now introduced from the TBD literature [8].

3.2.1. *TBD Likelihood*

A Bayesian TBD particle filter provides an approximation to the target state directly from the pixel array data. It is assumed that at each time step k, a pixel grid of IJ resolution cells is read simultaneously and that an individual pixel (i, j) has an intensity of $z_{ij}(k)$.

Readers are directed to Salmond and Birch, [8], in which the TBD framework is built up. Briefly the background noise is modelled as a zero mean Gaussian with variance of σ_N^2 for all pixels (i, j) - $p_N(z_{ij}|x, y) = \mathcal{N}(z_{ij}; 0, \sigma_N^2)$. If however the source is located within the grid pixel the pixel likelihood will be $p_{S+N}(z_{ij}|x, y)$ $= \mathcal{N}(z_{ij}; I, \sigma_N^2)$ where I is the intensity due to the source. The resultant likelihood ratio will then be

$$l(z_{ij}|x, y) = \frac{p_{S+N}(z_{ij}|x, y)}{p_N(z_{ij}|x, y)} = \exp\left(\frac{-I(I - 2z_{ij})}{2\sigma_N^2}\right). \quad (3)$$

To apply this method to the Acoustic Source Tracking (AST) problem, it is first necessary to recognise that the initial assumption of TBD - that only the single SBF pixel in which the source is located is influenced by the source's speech - is not true. The SBF is a continuous function and can be evaluated at any continuous location. However the shape of the SBF is defined by the frequencies used to calculate it and hence a grid of density of 10cm is sufficient to observe all promising peaks.

Equation 3 requires that the SBF values be normally distributed with known mean and variance statistics. As a result a non-linear mapping is necessary to adjust the highly varying SBF values to such a distribution.

3.2.2. *Magnitude Mapping*

As noted in Section 3.2.1 the measurement function is based on a likelihood function calculated for a set of pixels rather than a continuous function. As such the measurement related to a particular particle, positioned at a continuous location (x_k, y_k), is that of the point at the centre of the *pixel grid* in which it lies.

From a study of the SBF function, it was noted that for a particular environment and experimental setup the SBF will exhibit a typical distribution of noise values when no speech is active. Meanwhile SBF measurements from the correct source location will typically be above this distribution when the source is active. By tuning the likelihood function to give low likelihoods for SBF values within the noise distribution and larger likelihoods for large values it is possible to isolate the useful measurements without a strict thresholding. To do this we apply a nonlinear mapping to the SBF values as follows:

$$z = \Phi(\mathcal{S}; \bar{\mathcal{S}}, \sigma_{\mathcal{S}}^2) \quad (4)$$

where Φ is a normal cumulative distribution function with mean $\bar{\mathcal{S}}$ and variance $\sigma_{\mathcal{S}}^2$ applied to an SBF intensity of $\mathcal{S}$. These parameters are calibrated in advance or online so that its mean, $\bar{\mathcal{S}}$, lies between the mean of the noise distribution and magnitude of typical source peaks[1].

The measurements are now mapped onto the range $\Psi \in (0, 1)$ and we will now set the intensity value, I, to unity. As the measurement range is now truncated it is necessary to introduce a truncation constant to normalise the range. Truncation of the normal distribution $N(\mathcal{S}(l), 0, \sigma_{\mathcal{S}}^2)$ at the limits of $(0, 1)$ will require identical truncation constants as follows $c_N = c_{S+N} =$ $2\left(\text{erf}\left((\sqrt{2}\sigma_{\mathcal{S}}^2)^{-1}\right)\right)^{-1}$. As the likelihood ratio for a pixel is the ratio of these two likelihoods (see [8]), it is unaltered.

The final likelihood ratio can be stated as follows

$$l(z_{ij}|x, y) = \begin{cases} \exp\left[\frac{2z_{ij}-1}{2\sigma_N^2}\right] & \text{for } |i\Delta - x| < \Delta/2 \\ & \text{and } |j\Delta - y| < \Delta/2 \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

3.2.3. *Activity Indicator Variable*

As Lehmann and Johansson discussed, [5], the temporally discontinuous nature of speech must be recognised to allow for a complete AST system. However instead of measuring source activity indirectly using a Voice Activity Detector (VAD), we propose to detect activity directly from the SBF function itself. It will also allow us to track the activity of multiple sources simultaneously - something that would not be possible with a speech energy activity detector.

Firstly we add an activity indicator variable, $\lambda_k \in \{0, 1\}$, to the state vector. This variable will change according to a Markovian birth/death process with pre-determined probabilities as suggested in [8]. We choose the probability of birth, $P_B = 0.3$, and the probability of death, $P_D = 0.1$. The probability of activity of the source will simply be the proportion of active particles.

[1] The choice of a CDF is *not* intrinsic to this algorithm. A similarly shaped function would be sufficient.

Inactive particles drift via the dynamical model with the likelihood ratio set to unity. The final likelihood weighting function will become

$$l(z_{ij}|x,y) = q(\alpha) \propto \begin{cases} \exp\left[\frac{2z_{ij}-1}{2\sigma_N^2}\right] & \text{for } \lambda = 1 \\ & \text{and } |i\Delta - x| < \Delta/2 \\ & \text{and } |j\Delta - y| < \Delta/2 \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

4. MULTI TARGET TRACKING USING TBD

Multi-target TBD is a relatively new extension of the TBD methodology, [9]. According to the TBD methodology it is assumed to be only possible for a source to influence pixels in which it is located or a region surrounding the true location if smearing has occurred due to the sensor. Hence as suggested by [9], we will consider the sources to behave independently when widely separated. Tracking in this scenario will be identical to the single source case in Section 3.2.1. Alternatively when sources are closely spaced a joint likelihood will be considered. The transition between these two states is explained in Section 4.2.2. Two sources[2] cannot separate or coalesce of course. For this reason we will introduce a source-to-source repulsive effect to preclude this behaviour.

4.1. Disjoint tracking of Multiple Sources

Consider the case of two sources that have a large separation. The sources may be considered to be independent as in the single target scenario. A state vector for the source s at time frame k is

$$\alpha_{l,k}^s = (x_k^s, \dot{x}_k^s, y_k^s, \dot{y}_k^s, \lambda_k^s) \tag{7}$$

with an associated weighting w_k^s. The generic dynamical model will again be used as the transition prior.

Because the TBD method uses only pixels co-located with this source to generate a likelihood function, we can again use the single source methodology. As a result, the likelihood ratio for source i will be identical to the single source in Equation 6 and the particle weights $w_k^* \propto q_k$.

4.2. Joint Tracking of Multiple Sources

Now instead consider a joint state vector for two sources at time k:

$$\begin{aligned} \alpha_k &= (\alpha_k^1, \alpha_k^2) & (8) \\ \alpha_k^1 &= (x_1, y_1, \dot{x}_1, \dot{y}_1, \lambda_k^1) & (9) \\ \alpha_k^2 &= (x_2, y_2, \dot{x}_2, \dot{y}_2, \lambda_k^2) & (10) \end{aligned}$$

with a single associated weighting w_k. As in the case of joint source tracking, the individual sources will be propagated according to the dynamical model. However we will modify the model to disallow two speech sources to coalesce.

4.2.1. Source Repulsion Mechanism

Consider two source particles; the distance between each will be $d_{12} = \|(x_1, y_1), (x_2, y_2)|$ and the angle between them will simply be $\theta_{12} = \angle\{(x_1, y_1), (x_2, y_2)\}$, as illustrated in Figure 1. We shall propose that beyond a certain distance, $d_{12} > d_{\text{rep}}$, the sources are neither attracted to one another nor repulsed - simply moving independently with the usual Langevin motion model from Section 3. However when sources become closer than this, $d_{12} \leq d_{\text{rep}}$, a repulsive effect will force them apart. This force is modelled as an accelerating force applied in the opposite direction of θ_{12} - much like a pair of polar equal magnets. A simple squared function works satisfactorily

$$F_{\text{rep}}(\alpha_k) = \begin{cases} a_{\text{rep}}(d_{12} - d_{\text{rep}})^2 & \text{if } d_{12} \leq d_{\text{rep}} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where a_{rep} and d_{rep} are constants chosen empirically to give reasonable behaviour, as illustrated in Figure 1. For the first source, this force is then projected into the $\mathcal{X}$ and $\mathcal{Y}$ direction to give $F_{\text{rep,x}}^1(\alpha_k)$ and $F_{\text{rep,y}}^1(\alpha_k)$. Meanwhile the force applied to the second source (for the $\mathcal{X}$-direction component) is the equal opposite force $F_{\text{rep,x}}^2(\alpha_k) = -F_{\text{rep,x}}^1(\alpha_k)$, with a similar force for the $\mathcal{Y}$-direction. See Figure 1 for a graphical illustration of these forces. These projected forced are added to the original dynamical model as follows (in this case for the $\mathcal{X}$-coordinate of source s):

$$\begin{aligned} \dot{x}_k^s &= a_x \dot{x}_{k-1}^s + b_x F_x + F_{\text{rep},x}^s(\alpha_k) & (12) \\ x_k^s &= x_{k-1}^s + dT \dot{x}_k^s. & (13) \end{aligned}$$

Finally the likelihood ratio for each source position, $\tilde{q}(\alpha_k^s)$, is again based on the SBF image pixel as in Equation 6. Assuming pixel-to-pixel independence and that the sources are not co-positional, the product of the two likelihood ratios is used to give an overall likelihood ratio $q(\alpha) = \prod_{s=1}^{N_s} \tilde{q}(\alpha_k^s)$.

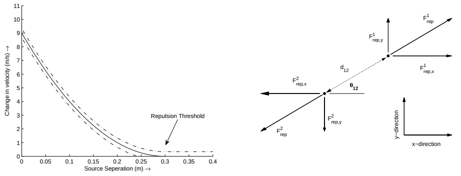

Figure 1: Illustration of repulsion effect: as source separation, d_{12}, falls below the threshold, d_{rep}, the force becomes more and more significant. The decomposition of the resultant force is shown on the right.

4.2.2. Transition between states

The decision to transition between the joint and disjoint particle filters is based on the MMSE estimate of the source particles and their variances. While other estimators, such as KL Divergence, might have been tried this method has proven to be sufficient in practice. The decision is as follows

$$I_c = \begin{cases} 1 & \text{if } d_{12,\text{MMSE}} \leq d_{\text{thres}} \\ & \text{or } (d_{12,\text{MMSE}} - \sigma_1 - \sigma_2) \leq \sigma_{\text{thres}} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where I_c is the state decision indicator and σ_1 and σ_2 are the variances of the two particle cluster positions.

5. AUDIO EXPERIMENTS

The recording environment was a typical office room, measuring roughly 7m x 7m. Twelve microphones were set up around the centre of the room. The positions of the microphones are illustrated as circles in Figure 3. Accurate ground-truth locations for the source and the microphones was provided via a commercial camera-based motion capture system. The source used was a computer loudspeaker transmitting typical conversational speech.

Single Target: Figure 2 illustrates the performance for a small portion of single source tracking. The filter estimates $\mathcal{X}$ and $\mathcal{Y}$ positions quite correctly. Note how the uncertainty of estimates

[2] This paper will concern itself only with a two source scenario. The extension to three or more sources is straightforward and will be published in future publications.

Source	$\bar{\epsilon}$ (m)	MSTD (m)	TLP (%)
	Example 1 - 32sec of audio		
1	0.110	0.076	2
2	0.114	0.118	10

Table 1: Illustrative Results for the SBF TBD particle filter tracking two sources for the examples in Section 3. Number of particles used was 1000. Average algorithm runtime was 92.14sec.

grow during sections of speaker inactivity i.e. silence. The activity detector (top-right plot) measures this activity directly from the particle filter output. As a result it switches between source activity and inactivity more regularly than a simple Voice Activity Detector.

Space limitations preclude a thorough comparison of the various single target tracking algorithms such as the GCC-based particle filter, [4], and various SBF-based particle filters, [1]. However experiments carried out by the authors have shown that the proposed algorithm gives similar performance for mean error, $\bar{\epsilon}$, and the percentage of tracks which fail completely (TLP, as explained in [1]) with typically a lower mean standard deviation of the particle cluster (MSTD). The caveat being that the number of particles used for the TBD version was 1000 compared to 100 for the other methods giving increased stability. It is anticipated that the SBF-TBD will comfortably run realtime on a typical modern computer with thousands of particles.

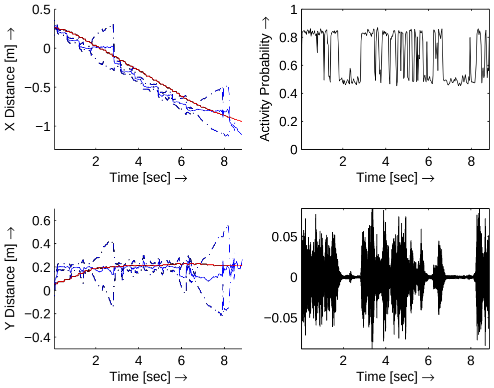

Figure 2: Example of single target tracking. Tracking performance in the X and Y directions is shown in left side figures respectively. The correct path is shown in red, the estimate is blue and variance bars are in dotted blue. Top right figure shows the evolution of the activity variable and the bottom right is the speech signal.

Performance of MTT-TBD Algorithm: Figure 3 shows a sample path for two active sources. Results are then presented in Table 1. Source 1 is a female voice and Source 2 a male voice. Individual audio samples were recorded separately and then linearly mixed before the MTT algorithm was run.

The results show that tracking of two sources speaking simultaneously is possible and that performance is only somewhat degraded when compared to the single source case - this despite the fact that dual source recordings will have a much lower proportion of useful peak measurements due to cross-signal interference. Finally the computation time is increased by only a factor of two by adding a second source.

6. FUTURE WORK AND CONCLUSIONS

A multi target TBD algorithm has been proposed which tracks two simultaneously active speech sources. This approach allows for a vast increase in the number of particles used and in algorithm stability without an increase in the computational effort. Performance for two source examples was seen to be similar to single source ASL algorithms. Further testing with more challenging data sets is necessary to evaluate the method's full performance capability.

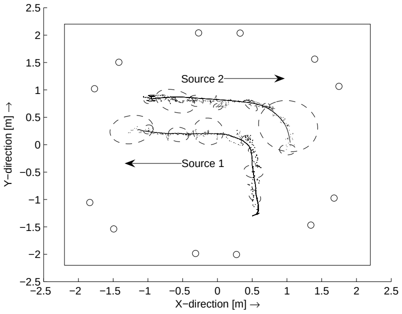

Figure 3: Example showing two sources moving in a room, which was used to test the performance of the algorithm. An example of the tracking performance is overlayed on each plot. Uncertainty ellipses are shown every 100 frames.

Future Work: A consequence of the TBD algorithm is that likelihoods are calculated for every pixel which contains a particle. As the cluster size increases during silence, computation will increase as more pixel values are evaluated. This can be controlled by halting a particle track after extended silence. Furthermore an algorithm module which allows for the initiation and removal of source tracks has yet to be proposed.

7. REFERENCES

[1] 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*, vol. 11, no. 6, pp. 826–836, Nov. 2003.

[2] 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, Sept. 2006.

[3] Y. Bar-Shalom and W. D. Blair, *Multitarget-Multisensor Tracking: Applications and Advances*. Artech House, 2004, vol. 3.

[4] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *Proc. ICASSP 2001*, vol. 5, pp. 3021–3024, 2001.

[5] 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*, vol. 2007, 2007, Article ID 50870, 11 pages.

[6] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo methods in practice*. Springer-Verlag, 2000.

[7] 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.

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

[9] C. Kreucher, M. Morelande, K. Kastella, and A. Hero, "Particle filtering for multitarget detection and tracking," in *Aerospace, 2005 IEEE Conference*, March 2005.