TRACKING A VARYING NUMBER OF SPEAKERS USING PARTICLE FILTERING

Angela Quinlan and Futoshi Asano

AIST Information Technology Research Institute,
Tsukuba, Ibaraki 305-8568,
JAPAN

ABSTRACT

The extension of particle filtering techniques to the multiple speaker case is difficult as two distinct problems must now be addressed. Firstly, the active speakers must be identified and their locations estimated, requiring the use of multi-dimensional likelihoods, and then each speaker must be correctly associated with his corresponding location. In this paper we propose a multi-speaker tracking algorithm in which the number of active speakers is determined by estimating the profile of the noise-plus-reverberation covariance matrix eigenvalues. The multi-dimensional likelihoods are then decoupled using the Expectation Maximization (EM) algorithm. The tracking accuracy is improved by the inclusion of a pause detection step and estimation of the noise-plus-interference covariance matrix. The results show the benefits of the proposed methods under difficult tracking situations.

Index Terms— Particle filtering algorithms, Source number estimation, Noise-plus-reverberation covariance matrix estimation, Multiple source tracking, Microphone arrays.

1. INTRODUCTION

The ability to track the locations of a varying number of speakers in the presence of background noise and reverberation is of great interest due to the vast number of potential applications. Particle filtering offers a robust method of tracking moving sources by recursively updating the location estimates using a two-step process of prediction and filtering.

While various particle filtering methods have been applied to the problem of tracking a single speaker e.g. [1, 2], the extension of these techniques to the case of multiple speakers is not straightforward. This is because in the situation of multiple speakers two distinct problems have to be solved, the estimation of the locations, involving multi-dimensional likelihoods, and also the association of each estimate with the correct source track. Furthermore, as one or more of the speakers may not be speaking it is also necessary to estimate which speakers are "active" at a given time.

Recently a method for tracking multiple sources using audio signals only was proposed in [3]. In this case the computational complexity due to the multiple sources is reduced by exploiting the signal separation characteristics of the Expectation Maximization (EM) algorithm to estimate the particle filter weights. This method was then extended in [4] in order to avoid confusion of the source tracks in situations where one of the sources is inactive for a significant lengths of time.

However, in more difficult tracking situations when there are more than two sources speaking intermittantly the methods proposed in [3] and [4] can no longer accurately track the speakers. This failure can be partly attributed to the fact that in this situation it is no longer sufficient to randomly assign the speech activity status. Instead a more accurate method of distinguishing the active sources is needed.

In this paper we address this problem and apply the resulting method to a difficult tracking scenario using live recordings of multiple moving speakers. Firstly the number of active sources is estimated directly from the received data using an extension of the method proposed in [5]. This method is based on predicting the profile of the noise-plus-reverberation covariance matrix eigenvalues and is particularly well suited to situations where a small number of data samples is available, as is the case when tracking moving sources. The inclusion of a pause detection step also allows us to continuously update the estimate of the background noise-plus-reverberation matrix.

2. PROBLEM FORMULATION

We consider the model of an array of M microphones located in a sound field generated by N_a active sources, which are assumed to be non-coherent.

Then, taking the short-term Fourier transform of the signals received by the microphones at time t, we obtain the following data model:

$$\mathbf{y}(\omega,t) = \mathbf{A}(\omega,t)\,\mathbf{s}(\omega,t) + \mathbf{n}(\omega,t), \tag{1}$$

where ω is the frequency under consideration. In what follows we omit the frequency index for the sake of simplifying the notation. $\mathbf{A}(\omega,T)$ is the matrix of the L direct path transfer function vectors:

$$\mathbf{A} = [\mathbf{a}(\theta_1),\ldots,\mathbf{a}(\theta_L)], \tag{2}$$

with $\theta_l, l = 1,\ldots,N_a$ representing the 2D directions of the N_a sources. $\mathbf{s}(t) = [S_1(t),\ldots,S_{N_a}(t)]^T$ is the source spectrum vector, and $\mathbf{n}(t) = [N_1(t),\ldots,N_M(t)]$ is the background noise spectrum vector. The signal and noise covariance matrices are defined respectively as:

$$\mathbf{R_{ss}} = E\left[\mathbf{s}(t)\,\mathbf{s}^H(t)\right],$$
$$\mathbf{R_{nn}} = E\left[\mathbf{n}(t)\,\mathbf{n}^H(t)\right],$$

where the superscript H denotes the conjugate transpose of the matrix.

2.1. Particle Filtering Algorithm

Using the framework of Bayesian hidden state sequence estimation, the particle filtering algorithm estimates the locations (θ_l) of moving targets by combining the information received from the observations with any available prior knowledge of the source transition model.

The hidden variable vector is defined as [6]:

$$\chi(t) = \left[N_a(t), \chi_1(t), ..., \chi_{N_a(t)}(t)\right], \quad (3)$$

where $N_a(t)$ is the number of active sources, and the required parameters of the ith target are defined as $\chi_l(t) = (\theta_l, s_l(t))$. θ_l defines the location as in (2), and $s_l(t)$ is a Boolean variable which denotes whether the ith source is switched on/off. The observation variables, $\mathbf{Z}(t)$, are composed of the audio signals $z(t)$ received by the microphone array.

The posterior probability distribution $P\left(\chi_{1|T}|\mathbf{Z}_{1|T}\right)$ specifies the likelihood of each possible $\chi_{1|T}$ given the observations $\mathbf{Z}(t)$. The estimated hidden variable vector, $\widehat{\chi}_{1|T}$, should then be selected so as to maximize this distribution.

Unfortunately, the distribution $P\left(\chi_{1|T}|\mathbf{Z}_{1|T}\right)$ is not available. However, under certain non-restrictive assumptions, the required distribution can instead be approximated in accordance with Bayes' theorem using the measurement likelihood $P(\mathbf{Z}(t)|\chi(t))$, and the state transition probability, $P(\chi(t), \chi(t-1))$[6]:

$$P\left(\chi_{1|T}|\mathbf{Z}_{1|T}\right) \propto \prod_{l=1}^{N_a} P(\mathbf{Z}(t)|\chi(t)) P(\chi(t), \chi(t-1)), \quad (4)$$

As no closed-form solution exists for (4) we approximate this distribution at a number of discrete points - or particles. Then, according to the central limit theorem as the number of particles increases towards infinity this approximation approaches the true posterior density.

The basic particle filtering framework can then be applied as a two-step *prediction* and *filtering* process. The prediction step consists of propagating the particles according to the motion model. In the method presented here we use a random walk motion model.

In the filtering step the propagated particles are weighted according to the measurement likelihood corresponding to this particle location. The particles are then re-sampled according to these importance weights.

The final estimate of the source locations can then be found by taking the mean of the re-sampled particles.

$$\widehat{\chi} = \frac{1}{N_p} \sum_{i=1}^{N_p} \chi^{\mathbf{i}}, \quad (5)$$

where N_p is the total number of particles and $\chi^{\mathbf{i}}$ is the parameter vector associated with the ith particle.

3. ESTIMATING THE NUMBER OF ACTIVE SOURCES

From equation (3) it can be seen that in the case of a varying number of sources it is necessary to estimate $N_a(t)$ the number of active sources as well as their respective locations. There are two main approaches to this problem. In the first case the particle filter can be applied to the joint problem of estimating and tracking the sources present. However this approach leads to high computational complexity. Therefore in this paper we instead use the alternative approach of firstly estimating the number of sources present and then using the particle filter to perform the tracking.

The MDL [7] and AIC [8] criteria are traditionally used for source number estimation. However, both these approaches are based on an assumption of white noise and are known to consistently overestimate the number of sources present when reverberation is present. In what follows we use the method proposed in [9] extended to cover reverberant environments as detailed in [5]. This method is based on the eigenvalue decomposition of the sample covariance matrix of the received signal:

$$\mathbf{R}_{yy}(t) = \frac{1}{N} \sum_{n=1}^{N} y(t) y^H(t), \quad (6)$$

where N is the number of data frames the covariance matrix is averaged over. The number of eigenvalues corresponding to the signal subspace, the so-called signal eigenvalues, is equal to the number of sources present. Consequently the source number estimation problem then becomes one of distinguishing between the signal and noise eigenvalues.

Under the proposed scheme the smallest observed eigenvalue is assumed to be a noise eigenvalue, corresponding to a noise subspace dimension of $P = 1$. Then letting $P = P + 1$ for each subsequent step until $P = M - 1$, the predicted profile of the noise-only eigenvalues is found recursively according to an exponential profile as detailed in [9].

In situations where at least one source is present the predicted noise-eigenvalues are then corrected to account for the presence of a reverberant tail [5]. The relative differences between the predicted noise eigenvalues, $\left[\hat{\lambda}_m, \ldots \hat{\lambda}_M\right]$ and $[\lambda_m, \ldots \lambda_M]$ the observed eigenvalues, are found from:

$$r_m = \frac{\lambda_m - \hat{\lambda}_m}{\hat{\lambda}_m}, \qquad m = 1, \ldots, M-1. \quad (7)$$

r_m is then compared to a threshold value η_m in order to determine if and when a break from the noise-only profile has occurred.

4. ESTIMATION OF MEASUREMENT LIKELIHOOD USING EXPECTATION MAXIMIZATION

When tracking N_T sources the measurement likelihood distribution is an N_T-dimensional distribution and accordingly the computational complexity grows exponentially as the number of sources increases. A solution to this complexity problem proposed in [3] is the use of the Expectation Maximization (EM) algorithm. The main feature of the EM algorithm is that it decouples the N_T-dimensional likelihood distribution into N_T 1-dimensional distributions which can be calculated in parallel.This decoupling of the sources is achieved by decomposing the observed microphone signals into 'complete data' vectors which correspond to the signal due to each source:

$$\mathbf{y}(t) = \sum_{l=1}^{N_a} \mathbf{x}_l(t) = \mathbf{H}\mathbf{x}(t), \quad (8)$$

where

$$\mathbf{x}_l(t) = \mathbf{a}(\theta_l) S_l(t) + n_l(t)$$

$$\mathbf{x}(t) = \left[\mathbf{x}_1^T(t), \ldots, \mathbf{x}_{Na}^T(t)\right];$$

$$\mathbf{H} = [\mathbf{I}, \ldots, \mathbf{I}];$$

and the matrix $\mathbf{I}$ denotes the identity matrix. $\mathbf{n}_l(t)$ an arbitrary decomposition of the noise vector, which must satisfy $\mathbf{n}(t) = \sum_{l=1}^{N_a} \mathbf{n}_l(t)$ and $\mathbf{R}_{nl} = E[\mathbf{n}_l(t) \mathbf{n}_l(t)]$.

The likelihood of the complete data is then given by:

$$L_{xl}(\theta_l, \gamma_l|\mathbf{X}_l) = \Psi_{xl} \exp\left(-\frac{1}{2} tr\left[\mathbf{C}_{xl}\mathbf{K}_{xl}^{-1}\right]\right), \quad (9)$$

where:

$$\Psi_{xl} = (2\pi)^{-MN} \left[\det \mathbf{K}_{xl}\right]^{-N/2};$$

$$\mathbf{K}_{xl} = \gamma_l \mathbf{a}\left(\theta_l\right) \mathbf{a}^H\left(\theta_l\right) + \mathbf{R}_{nl}, \tag{10}$$

and the sample covariance matrix of the complete data $\mathbf{X}_l$ is given by:

$$\mathbf{C}_{xl} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_l\left(n\right) \mathbf{x}_l^H\left(n\right). \tag{11}$$

As the complete data is not known $\mathbf{C}_{xl}$ cannot be found directly and must instead be estimated using the following equations as in the Expectation step of the EM algorithm:

$$\mathbf{C}_{xl} = E\left[\mathbf{C}_{xl} | \mathbf{C}_y; \widehat{\mathbf{K}}_y\right]$$

$$= \widehat{\mathbf{K}}_{xl} - \widehat{\mathbf{K}}_{xl}\left(\widehat{\mathbf{K}}_y\right)^{-1} \widehat{\mathbf{K}}_{xl} + \widehat{\mathbf{K}}_{xl}\left(\widehat{\mathbf{K}}_y\right)^{-1} \mathbf{C}_y \left(\widehat{\mathbf{K}}_y\right)^{-1} \widehat{\mathbf{K}}_{xl}, \tag{12}$$

with:

$$\widehat{\mathbf{K}}_y = \sum_{l=1}^{N_a} \widehat{\mathbf{K}}_{xl} \tag{13}$$

$$\widehat{\mathbf{K}}_{xl} = \widehat{\gamma}_l \mathbf{a}\left(\widehat{\theta}_l\right) \mathbf{a}^H\left(\widehat{\theta}_l\right) + \widehat{\mathbf{R}}_{nl}. \tag{14}$$

Now, the importance weight for the particle filtering expression in (4) is calculated using $\mathbf{C}_{xl}$ as defined in (12) and $\widehat{\mathbf{K}}_{xl}^{-1}$, where $\widehat{\mathbf{K}}_{xl}$ is defined in (14).

$$L\left(\mathbf{y}_{t|t+N} | \chi\left(\mathbf{t}\right)\right) := \exp\left(-\frac{1}{2} tr\left[\mathbf{C}_{xl}^p \widehat{\mathbf{K}}_{xl}^{-1}\right]\right). \tag{15}$$

As this expression defines the likelihood at an individual frequency, the overall measurement likelihood is then given by:

$$P\left(\mathbf{z}_{t|t+N} | \chi\left(\mathbf{t}\right)\right) = \prod_{\omega} L\left(\mathbf{y}_{t|t+N}\left(\omega\right) | \chi\left(\mathbf{t}\right)\right). \tag{16}$$

5. NOISE-PLUS-REVERBERATION COVARIANCE ESTIMATION

In order to accurately model the received data the noise covariance matrix estimate $\widehat{\mathbf{R}}_{\mathbf{nn}}$ should take into account both background noise and reverberation. This estimate must also be continuously updated to reflect the non-stationarity of the background noise. The significant improvement in the tracking results when $\widehat{\mathbf{R}}_{\mathbf{nn}}$ is continuously estimated from the received data instead of assuming stationary white noise were previously demonstrated in [4].

In order to estimate the noise-plus-covariance matrix we firstly apply a pause detection step on a frequency-by-frequency basis using the noise characterization method proposed in [10]. In this step a noise threshold is applied to each frequency subband in order to determine which subbands contain signal components. The noise threshold η is calculated as:

$$\eta\left(\omega, k\right) = \beta E\left(\omega, k-1\right); \tag{17}$$

where k is the block index, (with N frames in a block). $E\left(\omega, k-1\right)$ is the energy of the previous noise estimate at the given frequency ω, and β is a constant value lying between 1.5 and 2.5.

Then, if:

$$E\left(\omega, k\right) > \eta\left(\omega, k\right) \tag{18}$$

the frequency value ω is determined to contain signal components, and is included in (16) in order to find the measurement likelihood.

Meanwhile, if the frequency component is determined to contain no signal component, the noise-plus-reverberation covariance matrix estimate for this frequency can then be found. The resulting covariance estimate is then smoothed over time:

$$\mathbf{R}_{nn}\left(\omega\right) = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{R}_{nn}\left(q, \omega\right) \tag{19}$$

where Q is the number of previous values used and is selected to match the statistics of the background noise.

6. DATA ASSOCIATION

The problem of data association - i.e. association of each location estimate with its corresponding source, arises when or more of the sources is inactive. In this paper we make a decision on which source or sources are active by comparing the measurement likelihoods for each source.

$$P_l\left(\mathbf{z}_{t|t+N} | \chi\left(\mathbf{t}\right)\right) = \prod_{i=1}^{N_p} P_l\left(\mathbf{z}_{t|t+N} | \chi^{\mathbf{i}}\left(\mathbf{t}\right)\right) \tag{20}$$

The active sources are then estimated to be the sources resulting in the N_a highest values of $P_l\left(\mathbf{z}_{t|t+N} | \chi\left(\mathbf{t}\right)\right)$. Only the active sources are then used to calculate the measurement likelihood, as discussed in section 2.1.

The locations of the inactive sources are estimated from the propagation model. By continuing to estimate the location of the inactive sources in this manner the correct location can be estimated as soon as the source becomes active once more, however the location estimates of the inactive sources cannot be expected to be very accurate. For this reason we distinguish between active and total errors in the experimental results as seen in section 7.

7. EXPERIMENTAL SETUP AND RESULTS

The proposed method was tested using recordings taken in a medium sized meeting room with a reverberation time of $500ms$. Four people then moved around the room, while speaking intermittently.

The speech was recorded using a uniform circular array of 8 microphones which was placed at ceiling height, and the distance between the microphone array and the speakers was sufficient to ensure far-field conditions. The recorded signals were divided into frames of length $32ms$, with an averaging interval of $N = 9$, or approximately $0.1s$.

The true trajectory of the speakers was found using the Zone Positioning System ZPS-3D by Furukawa Co.,Ltd. and is depicted by the dashed lines in fig 1. Using the Zone Positioning System a badge is pinned on the chest of each of the speakers and the location of the badge is then tracked.

In fig 1 the layout for the experiments is shown. From the Root Mean Square Error (RMSE) values shown in tables 1 it can be seen that the estimation of the background noise from the data improves the performance compared to the case where white background noise is assumed. From table 2 it can be seen that the inclusion of the pause detection and active source number estimation steps also lead to significant reduction in the RMSE values.

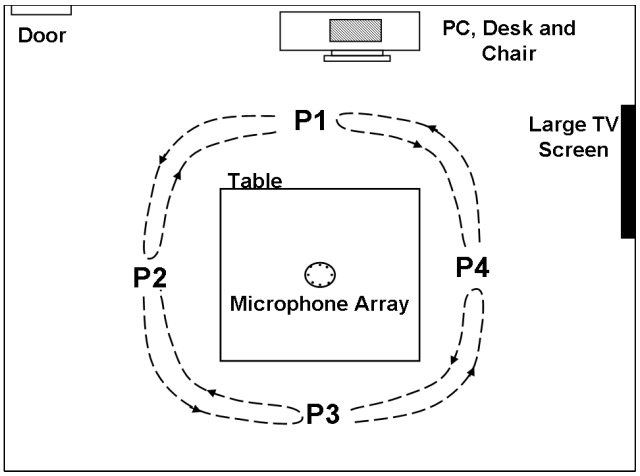

Fig. 1. Experimental layout. The four people are denoted P1, P2, P3, P4, and the dashed line traces their movement. The microphone array is set at ceiling height.

	White Noise RMSE	Estimated Noise RMSE
Source 1	1.49 m	0.67 m
Source 2	1.37 m	0.71 m
Source 3	1.50 m	1.04 m
Source 3	1.29 m	1.12 m
Average Over 4 Sources	**1.41 m**	**0.89 m**

Table 1. Root Mean Square Error (RMSE) values when all the sources are assumed to be active at all times.

8. CONCLUSION

This paper proposes a particle filtering scheme for tracking multiple speakers based on the approach proposed in [3]. This method was extended to include a pre-tracking active source number estimation step which is robust to the presence of reverberation. The results show that the proposed method can successfully track multiple sources even in difficult scenarios.

9. ACKNOWLEDGMENT

Angela Quinlan would like to acknowledge the support of a Japanese Society for the Promotion of Science (JSPS) postdoctoral fellowship. This research was partly supported by JSPS Kakenhi(A), no.18200007.

10. REFERENCES

[1] J. Vermaak and A. Blake, "Nonlinear Filtering for Speaker Tracking in Noisy and Reverberant Environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Salt Lake City, UT, 2001.

[2] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Montreal, Quebec, 2004.

[3] M. Kawamoto, F. Asano, H. Asoh, and K. Yamamoto, "Particle Filtering Algorithms for Tracking Multiple Sound Sources Using Microphone Arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Honolulu, Hawaii, 2007.

[4] A. Quinlan, M. Kawamoto, F. Asano, H. Asoh, and K. Yamamoto, "Tracking a Varying Number of Sound Sources using Particle Filtering," in *IASTED Conference on Signal and Image Processing SIP 2007*, Honolulu, Hawaii, 2007.

[5] A. Quinlan and F. Asano, "Detection of Overlapping Speech in Meeting Recordings Using the Modified Exponential Fitting Test," in *Proc. 15th European Signal Processing Conference (EUSIPCO)*, Poznan, Poland, 2007.

[6] A. Doucet, N. de Freitas, and E. N. Gordon, in *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[7] J. Rissanen, "Modelling by Shortest Data Description Length," *Automatica*, vol. 14, pp. 465–471, 1978.

[8] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contr.*, vol. 19, no. 6, pp. 716–723, 1974.

[9] A. Quinlan, J.-P. Barbot, P. Larzabal, and M. Haardt, "Model Order Selection for Short Data: An Exponential Fitting Test (EFT)," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, p. Article ID 71953, 2007.

[10] H. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Detroit, MI, 1995.

	White Noise RMSE		Estimated Noise RMSE	
Error	**Total**	**Active**	**Total**	**Active**
Source 1	1.30 m	0.99 m	0.57 m	0.43 m
Source 2	1.34 m	1.11 m	0.63 m	0.47 m
Source 3	1.48 m	1.02 m	0.58 m	0.39 m
Source 3	1.15 m	0.69 m	0.61 m	0.47 m
Average Over 4 Sources	**1.32 m**	**0.95 m**	**0.60 m**	**0.44 m**

Table 2. Root Mean Square Error (RMSE) values for the case where the active sources are estimated. Total Error is the error for the entire tracking time, Active Error is the error over the time each source was determined to be active.