JOINT ACOUSTIC SOURCE LOCATION AND ORIENTATION ESTIMATION USING SEQUENTIAL MONTE CARLO

Maurice Fallon, Simon Godsill*
Signal Processing Group
University of Cambridge, UK
{mff25|sjg30}@cam.ac.uk

Andrew Blake
Microsoft Research
Cambridge, UK
ablake@microsoft.com

ABSTRACT

Standard acoustic source localization algorithms attempt to estimate the instantaneous location of a source based only on current data from a microphone sensor array. This is done regardless of previous location estimates. However more recent Sequential Monte Carlo based approaches have instead posed the problem using an evolving state-space framework. In this paper we take this approach further by exploiting the directionality of human speech sources. This allows us to estimate the orientation of the source within the room. Finally combining previous source localization methods with this work we outline how both parameters - location and orientation - may be estimated jointly. Examples are given of performance in a typically reverberant real office environment for both a stationary and a moving source.

1. INTRODUCTION

Localization and tracking of sources is a topic of great importance in many fields from seismology to sonar. Location and orientation estimation and tracking in an audio environment, and more specifically speech, will be the focus of this paper. The use of these estimates has become important in several applications including automatic camera or microphone steering. The setup envisaged in this work is of several microphones spatially distributed around a typically reverberant room such as an office or living room. Traditionally localization algorithms have been based upon time delay estimation (TDE) on signals from multiple microphone pairs followed by often complex triangulation of the resultant estimates [1].

While sufficient in low noise and mildly reverberant conditions these methods break down in even moderately reverberation conditions - giving false peaks which lead to incorrect location estimates. As a result recent attention has turned to Sequential Monte Carlo (SMC) methods [2], commonly known as particle filters, which provide a simple yet adaptable framework for parameter estimation and tracking. Advantages of these methods include robust operation in adverse noise conditions and it allows for an elegant integration of the non-linear relationship between the TDE values and source position. Also SMC forms an easily extendable framework - which is illustrated by the orientation extension introduced in this paper.

The remainder of this paper details this extension as follows: Section 2 illustrates the effect that speech directivity has on signal correlation and how this information can be used to estimate orientation. Section 3 details the generic time delay framework as used in previous literature. Section 4 incorporates these estimates into a particle filtering framework from which a novel likelihood function for orientation is then formed in Section 5. Finally in Section 6 experimental results are detailed. Conclusions and discussion follow in Section 7.

* This work was supported by Microsoft Research. Preliminary results available at http://www-sigproc.eng.cam.ac.uk/~mff25/research.html

2. SPEECH DIRECTIONALITY AND THE GCC

Speech Directionality, as experimentally explored by [3], is the effect of non-uniform radiation of sound from the mouth and its absorption by the head and torso. For human speakers this effect causes as much as 15dB front/back attenuation differential at high frequencies, which is illustrated in Figure 1.

For clustered sensor arrays using the far-field assumption, such as those used in coherent signal processing, directionality is negligible. However when one chooses to instead use a sparse distributed sensor network the effect of directionality becomes very relevant. Indeed the setups envisaged in [4, 5] have the source surrounded by the sensors.

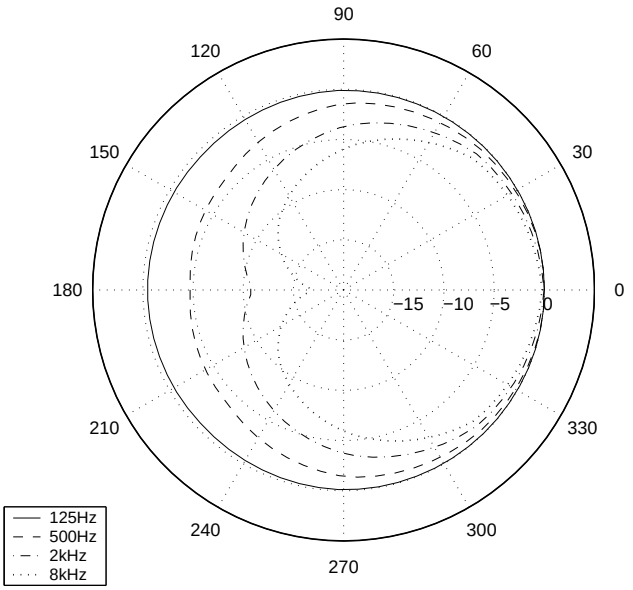

Figure 1: *Human speaker directional response (dB) obtained by least squares fitting of data from [3]. Graph taken from [6].*

This effect causes reduced correlation between signals recorded at sensors positioned at different angles relative to the mouth. There

are two different reasons for this. Firstly the attenuation effect varies greatly with frequency which leads to non-uniform absorbtion of the sound by the source's head. Secondly as the relative angle increases the energy in the direct path portion of the signal falls in comparison to that of the multi-path portion caused by reflections - in effect increasing the reverberation portion of the recorded signal. While this effect is very difficult to quantify because of the numerous factors, it is illustrated in Figure 2. In Figure 3 we go on to show that with an increasing relative angle between the source's orientation and the microphone pair there is a corresponding fall in signal cross-correlation. In other words microphones near to front of a speaker's head record the source signal with greater correlation. Using this effect we will now introduce a model which utilizes this effect to estimate speaker orientation.

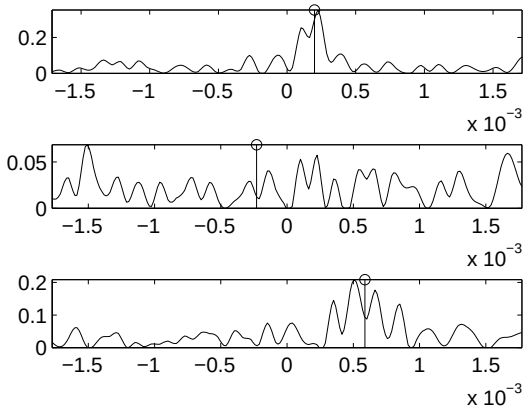

Figure 2: *Effect of orientation on signal correlation: GCC functions formed from the same source signal recorded at microphone pairs situated at* 0° *(top),* 180° *(middle) and* 90° *(bottom) relative to source orientation. The true source delay is indicated with a vertical line, the x-axis is relative delay between the recorded signals (in seconds) while the y-axis is the cross-correlation magnitude.*

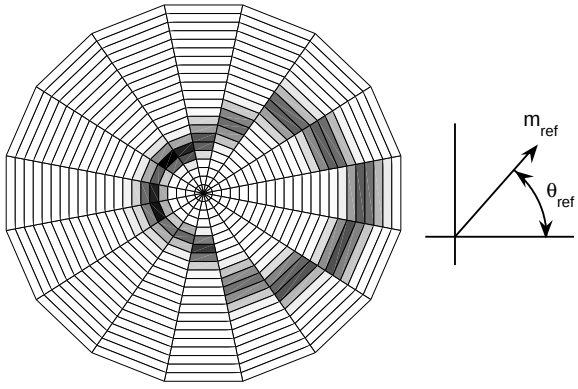

Figure 3: *Plot of the distribution of GCC peak magnitude,* m_{rel}*, versus relative angle between source heading and microphone pair,* θ_{rel}*, illustrating that there is greater signal correlation when the relative angle is small. The source signal was white noise and the environment was a typical office room.*

3. MEASUREMENT MODEL

We will concern ourselves with the problem of tracking the location and orientation of a single speaker in the XY-plane using TDOA measurements taken from a set of N_m spatially distributed microphone pairs. While the paper concentrates on the 2-D case it should be stated that any solution may be quite easily extended to full 3-D tracking. The assumed environment is a typically noisy and reverberant room and the microphones are assumed to be omnidirectional.

The core mathematical framework used in this paper follows that introduced by [7]. The received signal at sensor m is related to the transmitted source as follows

$$x_m(t) = h_m(t) * s(t) + e_m(t) \tag{1}$$

where $s(t)$ is the source signal, $h_m(t)$ is the impulse response and $e_m(t)$ is additive noise. In theory having estimated the impulse response it would then be possible to evaluate the source position. Unfortunately in challenging reverberant conditions estimating a rapidly changing time varying impulse response is a very difficult task. While there do exist some methods which attempt to do this, such as adaptive eigenvalue decomposition (AED, [8]), it has yet to be demonstrated in challenging multi-path environments. As a simplification [7, 9] and others ignore the multi-path portion of the signal and instead assume a simple direct path model only. The signal model then becomes

$$x_m(t) = \alpha_m s(t - \tau_m) + e_m(t) \tag{2}$$

where signal received at a particular microphone is simply a delayed and attenuated version of the source signal. The amplitude parameter, α_m, and the delay parameter τ_m, are then a simple function of the distance between source and sensor, d_m, as follows

$$\alpha_m \propto \frac{1}{d_m} \qquad \tau_m = \frac{d_m}{c} \tag{3}$$

where c is the speed of sound. Because of speech's highly variable nature the amplitude parameter, α_m, will not be considered for estimation. Meanwhile various techniques to estimate time delay parameter, τ_m, will be discussed in the following section.

3.1. Time Delay Estimation Model

A vast body of literature on the subject of time delay estimation exists. For multichannel delay estimation, a number of techniques attempt to estimate the set of delays over all microphones which best fit the recorded data. This approach is used for the SRP-PHAT beamforming localizer, which is essentially a multichannel cross-correlation [1]. This is expanded upon and used as the measurement model for the importance sampling particle filter in [4]. This method assumes that by correlating the data from all microphones we can make our estimate in some way more robust than a set of pairwise delay estimates alone. However since the previously mentioned directivity effect causes non-adjacent recordings to be relatively uncorrelated, it may not be particularly useful in practice for distributed microphones.

As a result we will concentrate on the most basic of the time delay estimation methods - pairwise cross-correlation using the phase transformed generalized cross-correlation (PHAT-GCC) as introduced by [10]. Taking synchronized frames of data with length L samples at time frame k from sensor m, that is

$$\mathbf{x}_m(k) = [x_m(kL), x_m(kL+1), \ldots, x_m(kL+L-1)], \tag{4}$$

we form a N_m by L matrix as follows

$$\mathbf{X}_k = \begin{bmatrix} \mathbf{x}_1(k) \\ \vdots \\ \mathbf{x}_{N_m}(k) \end{bmatrix}. \quad (5)$$

A generalized TDE function makes a transformation between this raw data vector and a set of time delays, represented as $\mathbf{T}_k = \mathbf{f}(\mathbf{X}_k)$. For the GCC this transformation may be made as follows: First pairing up the sensors into N_p pairs, we evaluate the signal Fourier transforms, where for example $X_{p,2}(\omega)$ represents the Fourier transform of the audio recorded by second sensor in pair p. Next we form the GCC for pair p as follows

$$R_p(\tau) = \frac{1}{2\pi} \int \Psi(\omega) S_{x_{p,1}x_{p,2}}(\omega) e^{j\omega\tau} d\omega \quad (6)$$

where $S_{x_{p,1}x_{p,2}}(\omega) = E[X_{p,1}(\omega)X^*_{p,2}(\omega)]$ is the signal cross power spectrum. $E[\cdot]$ and $(\cdot)^*$ stand for the expectation and complex conjugate operators respectively. The weighting function will be the commonly used phase transform

$$\Psi_p(\omega) = \frac{1}{\|X_{p,1}(\omega)X^*_{p,2}(\omega)\|}. \quad (7)$$

This time delay estimator is used because of its robust behavior in the face of changing speech source signal characteristics although it is noted that performance is much reduced in strong reverberation [11]. Having formed the GCC function the corresponding set of promising candidate time delay estimates, $T_k^{(p)} = (\hat{\tau}_1^{(p)}, \ldots, \hat{\tau}_{N_c}^{(p)})$, are then isolated. These are simply the delays corresponding to the largest N_c peaks in the GCC function.

The entire set of delay measurements for a particular time frame k is then $\mathbf{T}_k = (T_k^{(1)}, \ldots, T_k^{(N_p)})$. The expected delay, at microphone pair p, for a given source position, $l = (x, y)$, is then given by

$$\tau^{(p)} = c^{-1} \left(\|l - l_{p,1}\| - \|l - l_{p,2}\| \right). \quad (8)$$

Note that the TDE measurements are limited by the separation of the intra-pair microphones; where the positions of the microphones in pair p are

$$\begin{aligned} l_{p,1} &= (x_{p,1}, y_{p,1}) \\ l_{p,2} &= (x_{p,2}, y_{p,2}), \end{aligned} \quad (9)$$

The maximum possible time delay is $\tau_{\max}^{(p)} = c^{-1}\|l_{p,1} - l_{p,2}\|$.

Meanwhile as indicated by Figure 3 the **magnitude** of the GCC peaks are indicative of the relative angle between the source heading and the microphone pair direction. This information will be used to estimate orientation. For the set of delay measurements identified above the matching GCC function magnitudes are identified as follows

$$\hat{m}_c = R_p(\hat{\tau}_c^{(p)}) - m_{\text{th}} \quad (10)$$

where m_{th} is an experimentally determined constant to correct for the GCC function noise floor and obviously candidate peaks are chosen to have magnitudes greater than the noise floor.

In the same way that the delay measurement vector was constructed above the magnitude measurement vector for microphone pair p is $M_k^{(p)} = (\hat{m}_1^{(p)}, \ldots, \hat{m}_{N_c}^{(p)})$. The set of magnitude measurements for a particular time frame k will then become $\mathbf{M}_k = (M_k^{(1)}, \ldots, M_k^{(N_p)})$. Finally the overall combined measurement vector is then a combination of the above, $\mathbf{D}_k = (\mathbf{T}_k, \mathbf{M}_k)$.

How we decide upon which observations to consider is open to some discussion - for example thresholding of the GCC function is the simplest method. It is also useful to limit the number of candidate peaks to a reasonable maximum using a constant $N_{c,max}$. Either way the set of candidate peaks should contain all of those likely to be due to a speech source.

4. PARTICLE FILTER FRAMEWORK

In this section a general particle filter framework for localization is introduced. We will use a framework adapted from that initially proposed by [7] and also used by [5]. The source state vector at time k will be defined as

$$\alpha_k \triangleq (\alpha_{l,k}, \alpha_{o,k})$$

where the source state vector is made up of two different sections: one each for location and orientation respectively. These are in turn defined as

$$\begin{aligned} \alpha_{l,k} &\triangleq (x_k, \dot{x}_k, y_k, \dot{y}_k) \\ \alpha_{o,k} &\triangleq (\theta_k, \dot{\theta}_k) \end{aligned} \quad (11)$$

where, for example, $(x_k, \dot{x}_k)$ are source position and velocity, in the $\mathcal{X}$-dimension. Estimation of this state vector will jointly provide us with source position (x_k, y_k) and orientation (θ_k). The time delay, $\tau_k^{(p)}$, that one would expect for this source state vector, for microphone pair p, is found by substituting the location parameter pair, (x_k, y_k), into Equation 8. The set of resultant source time delays is then $\mathbf{T}_{\alpha,k} = (\tau_{\alpha,k}^{(1)}, \ldots, \tau_{\alpha,k}^{(N_p)})$.

Source movement in the $\mathcal{X}$, $\mathcal{Y}$ and θ dimensions is assumed to be independent and can be decoupled as a result. The state dynamics will be modeled by a first-order Markov process specified by its initial state and state transition distributions which are of the form $p(\alpha_0)$ and $p(\alpha_k|\alpha_{k-1})$ respectively, whose specifics can be found in [7].

The tracking problem itself involves recursive estimation of the posterior filtering distribution, $p(\alpha_k|\mathbf{D}_{1:k})$, using Bayes' Theorem as follows

$$\begin{aligned} p(\alpha_k|\mathbf{D}_{1:k-1}) &= \int p(\alpha_k|\alpha_{k-1}) p(\alpha_{k-1}|\mathbf{D}_{1:k-1}) d\alpha_{k-1} \\ p(\alpha_k|\mathbf{D}_{1:k}) &\propto p(\mathbf{D}_k|\alpha_k) p(\alpha_k|\mathbf{D}_{1:k-1}). \end{aligned} \quad (12)$$

The first step is the *prediction step* which will use the combined dynamics model, $p(\alpha_k|\alpha_{k-1})$, introduced in the succeeding sections to propagate the previous posterior, $p(\alpha_{k-1}|\mathbf{D}_{1:k-1})$, so as to give an estimate of the predictive distribution $p(\alpha_k|\mathbf{D}_{1:k-1})$. The second step is the *update step*, where the likelihood, $p(\mathbf{D}_k|\alpha_k)$, is combined with the predictive distribution to obtain the posterior filtering distribution at time k.

This problem is by nature non-linear/multi-modal and as such no closed-form filtering framework exits. However, Sequential Monte Carlo (SMC) methods provide accurate yet simple and computationally efficient estimation strategies for this framework. Essentially SMC, also known as particle filtering, involves a Monte Carlo implementation of the recursions in Eq. (12) by using a large set of discrete samples, or particles, with associated weights.

Having propagated the particles in the first step (prediction) and reweighting them with the likelihood function in the second

step (update) the particles are then resampled according to the new weights to have uniform weight distribution. Simple bootstrap filtering [2] and stratified resampling are used here. Other more elaborate schemes could easily be incorporated into our framework.

We will now go into some detail to explain the dynamics models used for orientation and location and finally the likelihood functions, including the novel orientation likelihood function.

4.1. Localization Movement Model

The motion, in the $\mathcal{X}$-coordinate, will be as described by Equation 1 in [7]. The parameter values in that paper are retained, viz $\beta_{\mathcal{X}} = 10$Hz and $\bar{v}_{\mathcal{X}} = 1$m/sec. Movement in the $\mathcal{Y}$-coordinate will be independent to this but will use an identical framework. Note that a more accurate and complex model could be trained on a set of representative movements but this has not been done here so as to maintain generality and simplicity.

4.2. Orientation Movement Model

For the orientation movement model we will again use the model from Equation 1 [7] where the orientation portion of the state, at time k, is defined as $\alpha_{o,k} \triangleq [\theta_k, \dot{\theta}_k]$ and the parameters $\beta_{\mathcal{X}}$ and $\bar{v}_{\mathcal{X}}$ are replaced by $\beta_\theta = 10$Hz and $\bar{v}_\theta = 0.1$rad/sec. Note that as in all the following orientation calculations, particle orientations are limited to the range $[-\pi : \pi]$ with wrap-around applied to all particles propagated to positions outside this range.

5. MEASUREMENT LIKELIHOOD MODEL

The next step is to outline the likelihood function for joint orientation and location tracking. Movement in each domain (i.e. orientation or location) is assumed to be independent and as a result is it possible to create separate likelihood functions for each. The overall likelihood function is simply the product of the individual likelihood functions

$$\begin{aligned} p(\mathbf{D}_k|\alpha_k) &= p(\mathbf{D}_k|\alpha_{l,k}, \alpha_{o,k}) \\ &= p(\mathbf{T}_k|\alpha_{l,k})p(\mathbf{M}_k|\alpha_k). \end{aligned} \quad (13)$$

Note that to track speaker location alone, as per [7], it is simply a case of setting $p(\mathbf{M}_k|\alpha_k) \propto 1$ and removing all orientation parameters from the state vector. While conversely orientation-only tracking, with a fixed and known source location, is possible by setting $p(\mathbf{T}_k|\alpha_{l,k}) \propto 1$ and removing the location parameters. This is further discussed in Section 5.2 and experimentally illustrated in the Section 6.

Note: For the remainder of this section the time index k has been suppressed for ease of reading.

5.1. Localization Likelihood Functions

The localization likelihood function will follow a similar form to that introduced in [7]. This framework allows two hypotheses: either one of measurements is due to the source and the rest are due to clutter, $\mathcal{H}_0$, or alternatively all of the measurements are due to clutter, $\mathcal{H}_1$.

Hypothesis $\mathcal{H}_0$: For microphone pair p, if a particular candidate delay measurement, $\hat{\tau}_c^{(p)}$, is due to the true source we will represent the resultant likelihood function using a normal distribution as follows

$$p(\hat{\tau}_c^{(p)}|\alpha_l, \mathcal{H}_0) = c_{\alpha_l}\mathcal{N}(\hat{\tau}_c^{(p)}; \hat{\tau}_\alpha, \sigma_l^2) \quad \text{for } |\hat{\tau}_c^{(p)}| \leq \tau_{\max}^{(p)} \quad (14)$$

where c_{α_l} is a normalizing constant due to the limited admissible delay region. The TDE is assumed to be corrupted by Gaussian observation noise with variance σ_l^2. This assumption is further discussed in [7]. The normalizing constant is obtained using the Gaussian error function.

Hypothesis $\mathcal{H}_1$: The likelihood of one of the measurements being associated with clutter is given by a uniform distribution within the admissible interval

$$p(\hat{\tau}_c^{(p)}|\alpha_l, \mathcal{H}_1) = \mathcal{U}_{\mathcal{D}}(\hat{\tau}_c^{(p)}) \quad (15)$$

These measurements are summed together to give a final likelihood function for microphone pair p:

$$p(T^{(p)}|\alpha_l) = \sum_{c=0}^{N_c} q_c p(T^{(p)}|\alpha_l, \mathcal{H}_c) \quad (16)$$

where q_c are the prior hypothesis probabilities which are commonly held to be equal but may also be chosen to reflect confidence in the measurements. Finally the overall likelihood function is the product of the likelihood functions for each microphone pair

$$p(\mathbf{T}|\alpha_l) = \prod_{p=1}^{N_p} p(T^{(p)}|\alpha_l) \quad (17)$$

5.2. Proposed Orientation Likelihood Functions

In this section a novel algorithm is proposed which estimates an orientation likelihood function, $p(\mathbf{M}|\alpha_o)$, for the particle set at the current time. (A number of different algorithms were implemented however space restrictions inhibit their discussion here). It uses information from the measurement magnitude vector, $\mathbf{M}$, as well as particle locations and orientations to create a representative likelihood function for each microphone pair, which are then combined to give an overall likelihood function.

As mentioned above if the source is stationary then orientation-only tracking may be implemented if the source position, $(\mathbf{x}, \mathbf{y})$, is known *a priori*. Examples of each type of tracking are shown in Section 6.

The hypotheses of the measurements being due to clutter or the true source are similar to that used in the previous section and are calculated for each microphone pair. Note that for this algorithm the maximum number of candidate pairs is limited, $N_{c,max} \leq 1$, therefore the individual microphone pair measurement vectors contain only a single element i.e. $M^{(p)} = \hat{m}^p$.

Hypothesis $\mathcal{H}_0$: Consider microphone pair p, with microphones positions defined as $l_{p,1}$ and $l_{p,2}$. The midpoint of the microphone pair is then simply

$$l_{p,\text{mid}} = \frac{l_{p,1} + l_{p,2}}{2}. \quad (18)$$

Now consider a particle with position $l = (x, y)$; its position relative to the midpoint of microphone pair p is given by

$$r^{(p)} = l - l_{p,\text{mid}}. \quad (19)$$

Using this vector the likelihood function takes the form of a normal distribution as follows:

$$p(\hat{m}^{(p)}|\alpha, \mathcal{H}_0) \propto \mathcal{N}(\hat{\theta}^{(p)}; \theta, \sigma_{\hat{\theta}}^{(p)}). \quad (20)$$

where mean and variance are determined using the relevant magnitude measurement data, $\hat{m}^{(p)}$, as

$$\hat{\theta}^{(p)} = \angle(r^{(p)}), \quad \sigma_{\hat{\theta}}^{(p)} = \frac{a}{\|\hat{m}^{(p)} r^{(p)}\|} \tag{21}$$

where a is a constant experimentally determined using the data in Figure 3. As such the mean of the likelihood function is the relative angle between source and microphone pair while the standard deviation is inversely proportional to the measurement magnitude vector, $\hat{m}^{(p)}$, and the distance between the source and the microphone pair.

Hypothesis $\mathcal{H}_1$: The clutter hypothesis for this algorithm is again a uniform distribution but must be formed for each microphone pair p:

$$p(M^{(p)}|\alpha, \mathcal{H}_1) = \mathcal{U}_{\mathcal{D}}(M^{(p)}). \tag{22}$$

From this the likelihood function, due the measurements from a particular microphone pair, is as follows

$$p(M^{(p)}|\alpha) = q_T p(M^{(p)}|\alpha, \mathcal{H}_0) + (1 - q_T) p(M^{(p)}|\alpha, \mathcal{H}_1) \tag{23}$$

where q_T is again the prior probability that the source is present in microphone pair p. Finally the complete likelihood function is product of the individual microphone pair likelihood functions

$$p(\mathbf{M}|\alpha) = \prod_{p=1}^{N_p} p(M^{(p)}|\alpha). \tag{24}$$

This algorithm has the advantage that the overall angle estimate need not be limited to within the angular distribution of the microphones. In other words it can produce an estimate in the full range $[-\pi : \pi]$ independent of microphone positioning.

6. REAL AUDIO EXPERIMENTS

6.1. Experimental Setup

A recording environment was a typical room measuring roughly 5.5m x 6.1m x 2.8m containing typical office furniture. (Note that none of the experiments were carried out on simulated data because of the ease of gathering real data). A set of 12 microphones, arranged into pairs, were set up at the same horizontal height - 1.2m. The intra-pairing spacing was uniformly 60cm. Accurate ground-truth location and orientation of the source and the microphones was provided via a commercial motion capture system. Accuracy of this system is estimated to be sub-millimeter [12]. The source used was a computer loudspeaker transmitting typical conversational speech. The samples were taken from digital recordings of BBC radio presenters. The duration of the audio samples varied from 20 seconds to several minutes. The recorded signals were band-passed to remove interfering components outside of the 200-6000Hz range.

The following parameters were used in the particle filter algorithm: Audio Sampling Rate, 16kHz; Number of particles, $N_p = 150$; Resampling Threshold, $N_{th} = 0.5$; Frame Length, $L = 512$ samples; Frame Overlap percentage, 50%; Update Rate, $\triangle T$ = 31.25Hz; a, 0.2; q_T, 0.2 and σ_l, $6e^{-5}$ (for the joint tracker only). While the number of particles used is much more than that used by [5] it is still sufficiently small to allow for real time performance and as such no optimization to reduce the number has been attempted.

6.2. Typical Tracking Examples

Two examples of typical recordings are discussed in this section. Both of these examples are typical tracking results using the algorithm proposed in Section 5.2. The first is an example of orientation-only tracking of a stationary source as seen in Figure 4. The particles are seen to clearly track the source's changing orientation.

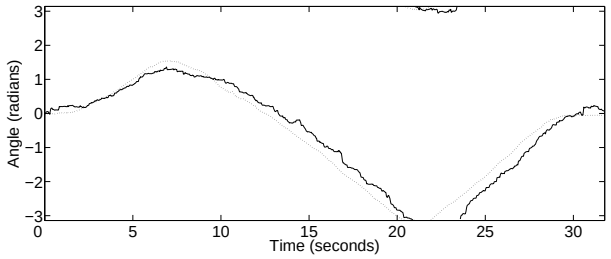

Figure 4: *Example of orientation tracking performance (solid line) of a real source which is turning but stationary at the center of an office room. The dotted line represents the ground truth orientation.*

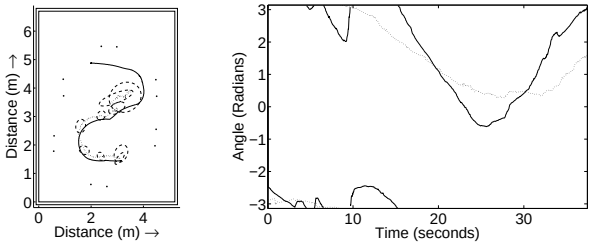

Figure 5: *Example of joint location (left) and orientation (right) tracking performance (solid line) of a real moving and turning source (dotted line). Particle position and tracking performance is indicated by the loops of equal variance (dashed lines, top). The microphone positions and the room boundary are also shown.*

The second example illustrated is of the source simultaneously moving and turning around the room. Figure 5 shows the results for joint tracking. The particle filter comfortably tracks the movement and orientation for the first two-thirds of the signal - as expected. The remainder of the test illustrates the performance limitations of bearings only tracking - and why speaker directivity and must be considered at a deeper level. At this source location and orientation only the microphone pair at the top-right of the room can provide bearings estimates - causing reduced performance radially relative to this microphone pair as the GCC functions from the other microphone pairs show very little correlation because of the directivity of the source.

6.3. Comparative Results

During this final section the source also turns quickly anti-clockwise causing difficulty for the orientation tracker. The particle filter does regain the correct track after a period of inaccuracy. The cause of this inaccuracy is twofold: firstly the impulse responses will be non-stationary across the data frames - leading to poor orientation estimates. Secondly the source orientation is changing rapidly - which the tracking system does not respond to because its behavior is outside of the typical movement envisaged by the movement model. Both of these problems and possible solutions are discussed in the Conclusions section.

Path	1	2	3	4	5
Location MSE	0.104	0.061	0.013	0.024	-
Location MSTD	0.051	0.070	0.06	0.074	-
Orientation MSE	0.018	0.098	0.311	0.113	0.145
Orientation MSTD	0.073	0.227	0.115	0.377	0.300

Table 1: *Results for experimental tracking for each recording. The units of the performance measures are as follows: m^2, m, rad^2, and rad.*

To evaluate performance the proposed algorithm was tested on 5 different audio recordings. The paths the speaker took in four of these recordings are shown in Figure 6. The source movement was approximately constant at typical walking speeds. Orientation was maintained in the direction of movement. The fifth recording is of a stationary source located at the co-ordinates (2.7m,3.9m) turning on its axis as shown in Figure 4. Because of the nature of the microphone pair spacings and the different paths and source signals used performance is expected to vary from recording to recording. Each algorithm was run 50 times and the statistics were then averaged to give the results in Table 1. The statistics evaluated are mean square error (MSE) and mean standard deviation (MSTD) of the particle cloud as suggested by [5]. The first statistic give an indication of tracking performance while the second is an estimate of tracking stability.

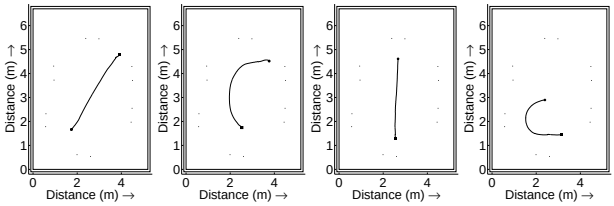

Figure 6: *Source paths used to evaluate performance statistics. Each path runs from the square marker to the circle marker and lasts approximately 20-30 seconds, the room boundary and microphone positions are also shown. The paths are numbered 1-4 from left to right.*

The results in Table 1 illustrate that the algorithm can robustly estimate both orientation and location jointly. As expected Path 1 has better orientation estimation performance as it does not change orientation while it moves across the room. Paths 2-5 all show similar orientation estimation performance. Paths 2 and 3 illustrate good tracking in the face of changes in both orientation and location.

7. CONCLUSIONS

In this paper the problem of orientation and location estimation of a moving speech source has been proposed and discussed. We introduced an algorithm to estimate and track speaker orientation while maintaining the same framework as previous localization literature. This allows us to simultaneously estimate both parameters jointly.

Results of real audio experiments were promising; however future work is still necessary. For example it is necessary to correct for the bias caused by an uneven distribution of microphones around the source. The likelihood function is by no means perfect and as such a good deal of experimentation is required to achieve optimal performance. In a situation where only one microphone pair gave active measurements the performance was predictably poorer. A solution to this problem may be possible by considering the effect of speaker directivity within the localization estimator.

8. REFERENCES

[1] J. DiBiase, H. Silverman, and M. Branstein, *Robust Localization in Reverberant Rooms in Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer, 2001, ch. 8, pp. 157–180.

[2] N. Gordon, D. Salmond, and A. F. M. Smith, "Novel approach to nonlinear and non-Gaussian Bayesian state estimation," *Proc. IEE-F*, vol. 140, pp. 107–113, 1993.

[3] H. K. Dunn and D. W. Farnsworth, "Exploration of pressure field around the human head during speech," *J. Acoust. Soc. Am.*, vol. 10, pp. 184–199, Jan. 1939.

[4] E. A. Lehmann and R. C. Williamson, "Particle filtering design using importance sampling for acoustic source localisation and tracking in reverberant environments," *EURASIP J. on Applied Sig. Proc.*, vol. 2006, Special issue on Advances in Multi-Microphone Speech Proc., article ID 17021, 9 pages, Jan. 2006.

[5] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech and Audio Proc.*, vol. 11, no. 6, pp. 826–836, Nov. 2003.

[6] T. Betlehem and R. C. Williamson, "Acoustic beamforming exploiting directionality of human speech sources," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc. (ICASSP'03),* Hong Kong, China, april 2003, pp. 365–368.

[7] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, vol. 5, pp. 3021–3024, 2001.

[8] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP J. on Applied Sig. Proc.*, vol. 2006, Special issue on Advances in Multi-Microphone Speech Proc., article ID 26503, 19 pages, 2006.

[9] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. Speech and Audio Proc.*, vol. 12, no. 5, pp. 499–508, Sept. 2004.

[10] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 24, pp. 320–327, Aug. 1976.

[11] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech and Audio Proc.*, vol. 4, no. 2, pp. 148–152, Mar. 1996.

[12] PhaseSpace Inc., "PhaseSpace motion capture," Retrieved June 29th, 2006, [Online] http://www.phasespace.com.