PARTICLE FILTER BEAMFORMING FOR ACOUSTIC SOURCE LOCALIZATION IN A REVERBERANT ENVIRONMENT

Darren B. Ward[*]

Department of Electrical and Electronic Engineering
Imperial College of Science, Technology and Medicine
London SW7 2BT, U.K.

Robert C. Williamson[†]

Telecommunications Engineering, RSISE
The Australian National University
Canberra ACT 0200, Australia

ABSTRACT

Traditional acoustic source localization uses a two-step procedure requiring intermediate time-delay estimates from pairs of microphones. An alternative single-step approach is proposed in this paper in which particle filtering is used to estimate the source location through steered beamforming. This scheme is especially attractive in speech enhancement applications, where the localization estimates are typically used to steer a beamformer at a later stage. Simulation results show that the algorithm is robust to reverberation, and is able to accurately follow the source trajectory.

1. INTRODUCTION

The ability to localize an acoustic source is critical for the correct performance of speech enhancement algorithms using microphone arrays. Acoustic source localization is also finding increasing use for automatic camera-steering in videoconferencing applications.

Traditionally, source localization has been a two-stage procedure. First, time-delay estimates (TDEs) are computed for different pairs of microphones, and then the individual TDEs are combined to estimate the source location.

A conceptually simple approach to source localization is beamforming, where the source location is estimated by calculating the steered output power of a beamformer over a set of candidate locations. This has the advantage that if one wants to use the array for speech enhancement (and not just localization), the beamformer output has already been computed. However, there are significant disadvantages to the scheme that have prevented it from being used for source localization: (i) it suffers from poor resolution; (ii) searching over all possible source locations is computationally expensive; and (iii) in reverberant environments, there may be spurious peaks in the effective beampattern.

Recently a new framework for TDE source localization was proposed by Vermaak and Blake [1], based on a Sequential Monte Carlo (or *particle filtering*) approach to state-space estimation [2]. This is a two-stage approach in which TDEs are first calculated using generalized cross-correlation (or any other TDE scheme), and then a likelihood model is used to determine the source location based on the obtained TDEs. By including multi-hypothesis testing, this method has the distinct advantage that it can cope with spurious peaks in the cross-correlation function caused by reverberation. Furthermore, because of the state-space estimation framework, there is no need to explicitly combine disparate TDEs to find a candidate source location. An alternative idea that also avoids the need for triangularization from TDEs has recently been proposed in [3].

In this paper our aim is to use beamformer-based source localization within the particle filtering framework. By using particle filters, we eliminate the need for a comprehensive search over the source location space. The proposed scheme has the advantage over conventional TDE-based schemes that it does not require intermediate calculation of TDEs from microphone pairs.

2. MICROPHONE ARRAY MODEL

Consider an array of M microphones used in a reverberant room. For a single source located within the room, let the signal received at the mth microphone at time k be:

$$x_m(k) = u_m(k) \star s(k) + n_m(k), \quad m = 1, \ldots, M, \quad (1)$$

where $s(k)$ is the source signal, $u_m(k)$ is the room impulse response (RIR) between the source and the mth microphone, $n_m(k)$ is additive noise (assumed uncorrelated with the signal $s(k)$, and also uncorrelated from microphone to microphone), and $\star$ denotes convolution. The RIR can be written as

$$u_m(k) = a_m(k) + r_m(k) \quad (2)$$

where $a_m(k)$ is the component of the RIR due to the direct-path, and $r_m(k)$ is the component of the RIR due to rever-

[*] This work was performed while the first author was a visitor at the Australian National University.

[†] This work was partially supported by the Australian Research Council.

beration. The Fourier transform of the direct-path RIR is given by

$$A_m(\omega) = \frac{1}{||\ell_{\rm s} - \ell_m||} \, e^{-j\omega c^{-1}||\ell_{\rm s}-\ell_m||} \tag{3}$$

where $\ell_{\rm s}$ is the source location vector, ℓ_m is the microphone location vector, and Euclidean distance is denoted by $||\cdot||$.

Assume that the data received at each microphone is collected over a frame of K samples, and denote the data at the mth microphone for frame t as

$$\mathbf{x}_{t,m} = [x_{t,m}(0) \cdots x_{t,m}(K-1)], \tag{4}$$

where $x_{t,m}(k) = x_m(K(t-1)+k)$. Stack the microphone frames to form the array frame matrix

$$\mathbf{x}_t = \begin{bmatrix} \mathbf{x}_{t,1} \\ \vdots \\ \mathbf{x}_{t,M} \end{bmatrix}, \tag{5}$$

which represents the data received at the array during time frame t.

3. PARTICLE FILTER BEAMFORMING

3.1. Localization through steered beamforming

Using a beamformer for source localization is a conceptually simple idea. The aim is to scan the beamformer over a set of candidate source locations, and then choose the source location as that which gives the maximum beamformer output power. In the frequency domain, the output of a beamformer steered to a location ℓ is

$$Y(\ell, \omega) = \frac{1}{M} \sum_{m=1}^{M} H_m(\ell,\omega)\; X_m(\omega), \tag{6}$$

where $X_m(\omega)$ is the Fourier transform of $x_m(k)$, and $H_m(\ell,\omega)$ is the beamforming weight for the mth microphone that steers the beamformer towards the desired source location ℓ. Of the many candidate beamformers, the simplest is the delay-sum beamformer (DSB) in which the weights are given by

$$H_m(\ell,\omega) = e^{-j\omega c^{-1}(d_{\max} - ||\ell - \ell_m||)}, \tag{7}$$

where c is the speed of wave propagation, $d_{\max}$ is the maximum possible distance between any two points within the room, and ℓ_m is the location of the mth microphone.

Let the frequency-averaged output power of the beamformer be

$$|\bar{Y}(\ell)|^2 = \sum_k W(\omega_k)\, |Y(\ell,\omega_k)|^2, \tag{8}$$

where $W(\omega_k)$ is an arbitrary frequency weighting function. If the DSB was steered toward the true source location, then one would expect the beamformer output power to be large. Thus, one could (potentially) find the source location as

$$\hat{\ell} = \arg\max_{\ell} |\bar{Y}(\ell)|^2. \tag{9}$$

Performing a search over candidate source locations is, however, computationally burdensome. Moreover, in a reverberant environment the beamformer output power has several spurious maxima, and the true source location may not always be the global maximum (this phenomena is also apparent in the TDE problem).

In the remainder of this paper we formulate the beamforming source localizer within a state-space framework, thereby making it into a viable acoustic localization scheme.

3.2. State-space estimation with particle filters

The source localization problem can be formulated in a state-space estimation framework by associating the source location at time t with an unobserved state vector $\boldsymbol{\alpha}_t$. The problem is then to learn about the state given measurements of the microphone array signals. Let the source state at time t be

$$\boldsymbol{\alpha}_t = \left[x_{\rm s}, y_{\rm s}, z_{\rm s}, \dot{x}_{\rm s}, \dot{y}_{\rm s}, \dot{z}_{\rm s}\right]^T, \tag{10}$$

where $[x_{\rm s}, y_{\rm s}, z_{\rm s}]^T$ is the source location in Cartesian coordinates, and $[\dot{x}_{\rm s}, \dot{y}_{\rm s}, \dot{z}_{\rm s}]^T$ is the source velocity.

Let $\mathbf{x}_{1:t}$ denote the concatenation of all the array frame matrices (5) up to frame t. Estimating the state $\boldsymbol{\alpha}_t$ would be straightforward if one could directly calculate the conditional density $p(\boldsymbol{\alpha}_t|\mathbf{x}_{1:t})$, since then one would have a direct measure of how likely a particular state was based on the measured microphone signals. Unfortunately, in practice this posterior filtering density is unavailable. It can be calculated, however, from [4]

$$p(\boldsymbol{\alpha}_t|\mathbf{x}_{1:t}) \propto p(\mathbf{x}_t|\boldsymbol{\alpha}_t)\, p(\boldsymbol{\alpha}_t|\mathbf{x}_{1:t-1}), \tag{11}$$

where $p(\mathbf{x}_t|\boldsymbol{\alpha}_t)$ is the likelihood (or measurement density). The prediction density $p(\boldsymbol{\alpha}_t|\mathbf{x}_{1:t-1})$ is given by [4]

$$p(\boldsymbol{\alpha}_t|\mathbf{x}_{1:t-1}) = \int p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1})\, p(\boldsymbol{\alpha}_{t-1}|\mathbf{x}_{1:t-1})\; d\boldsymbol{\alpha}_{t-1} \tag{12}$$

where $p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1})$ is the state transition density, and $p(\boldsymbol{\alpha}_{t-1}|\mathbf{x}_{1:t-1})$ is the prior filtering density. Although no closed-form solution exists for (11) and (12), these recursions can be approximated through Monte Carlo simulation of a set of particles (representing the source state) having associated discrete probability masses.

3.3. Proposed algorithm

In order to apply particle filters to the beamforming source localization problem, there are two requirements. First, Form an initial set of particles $\{\boldsymbol{\alpha}_0^{(i)}, i = 1 : N\}$ and give them uniform weights $w_0^{(i)} = 1/N, i = 1 : N$. Then, as each new frame of data is received:

1. Resample the particles from the previous frame $\{\boldsymbol{\alpha}_{t-1}^{(i)}\}$ according to their weights $\{w_{t-1}^{(i)}\}$ to form the resampled set of particles $\{\tilde{\boldsymbol{\alpha}}_{t-1}^{(i)}, i = 1 : N\}$
2. Predict the new set of particles $\{\boldsymbol{\alpha}_t^{(i)}\}$ by propagating the resampled set $\{\tilde{\boldsymbol{\alpha}}_{t-1}^{(i)}\}$ according to the source propagation model
3. Calculate the FFT of each microphone frame, and denote it by $X_m(\omega_k)$
4. Calculate the DSB weights in each frequency bin for each particle

$$H_m(\ell_{\boldsymbol{\alpha}}^{(i)}, \omega_k) = e^{-j\omega_k c^{-1}(d_{\max} - \|\ell_{\boldsymbol{\alpha}}^{(i)} - \ell_m\|)}$$

   where $\ell_{\boldsymbol{\alpha}}^{(i)}$ is the location of the ith particle, $\boldsymbol{\alpha}_t^{(i)}$

5. Calculate the frequency-averaged output of the beamformer steered to each particle

$$\bar{Y}(\ell_{\boldsymbol{\alpha}}^{(i)}) = \frac{1}{K}\sum_{k=1}^{K} \sqrt{W(\omega_k)} \times \frac{1}{M}\sum_{m=1}^{M} H_m(\ell_{\boldsymbol{\alpha}}^{(i)}, \omega_k)\, X_m(\omega_k)$$

6. Weight the new particles according to the likelihood function

$$w_t^{(i)} = p(\mathbf{x}_t|\boldsymbol{\alpha}_t^{(i)}) = \phi(|\bar{Y}(\ell_{\boldsymbol{\alpha}}^{(i)})|^2)$$

   and normalize so that $\sum_i w_t^{(i)} = 1$

7. Estimate the current source location as the weighted sum of the particle locations

$$E\{\ell_t\} = \sum_{i=1}^{N} w_t^{(i)} \ell_{\boldsymbol{\alpha}}^{(i)}$$

8. Store the particles and their respective weights $\{\boldsymbol{\alpha}_t^{(i)}, w_t^{(i)}, i = 1 : N\}$

Fig. 1. Algorithm for particle filter beamforming.

a state transition density $p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1})$, or equivalently, a model of how the states propagate is required. We will use the Langevin propagation model described in [1].

The other requirement is a likelihood function of the microphone data, i.e., $p(\mathbf{x}_t|\boldsymbol{\alpha}_t)$. The pseudo-likelihood model[1] that we propose is

$$p(\mathbf{x}_t|\boldsymbol{\alpha}_t) = \phi(|\bar{Y}(\ell)|^2), \tag{13}$$

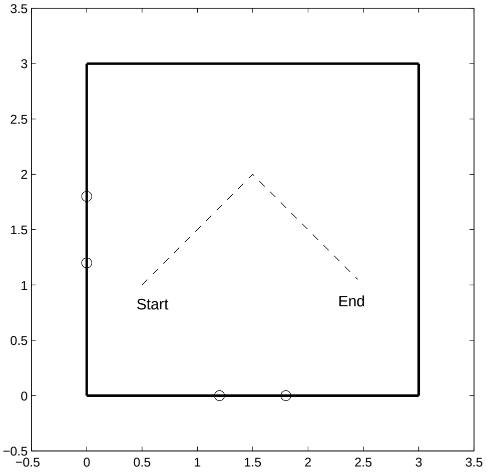

Fig. 2. Plan view of room used for simulations. Microphone locations are denoted by the circles. The dashed line indicates the source trajectory.

where $|\bar{Y}(\ell_{\boldsymbol{\alpha}})|^2$ is the output power of a DSB steered to the source location $\ell_{\boldsymbol{\alpha}}$ (where $\boldsymbol{\alpha}$ is the source state), and $\phi(\cdot)$ is a real-valued function that we will call the *likelihood shaping function* (LSF). The LSF is a non-linear function that modifies the beampattern to narrow the main beam and reduce the level of the sidelobes, thus making the resulting likelihood function more amenable to recursive estimation. We have found that good performance is achieved with a LSF of

$$\phi(x) = x^i, \tag{14}$$

where typically $i = 2, 3$, or 4.

The proposed algorithm is summarized in Fig. 1.

4. RESULTS

To demonstrate the proposed algorithm we now present results of a simulated acoustic localization problem. The acoustic environment was a $3 \times 3 \times 2.5$ meter room with a reverberation time of 200 msec. Four omni-directional microphones were located as shown in Fig. 2 (denoted by the circles), which is a plan view of the room at a height of 1.5 m. The source trajectory (which is also at a height of 1.5 m) is denoted by the dashed line.[2] The impulse responses between the source and the microphones were simulated using the image method [5]. The source signal was the speech utterance *"Draw every outer line first, then fill in the interior"*, taken from the TIMIT database. Uncorrelated white noise resulting in a SNR of 30 dB was added to each microphone signal. In calculating the frequency-averaged

[1] We call this a pseudo-likelihood, since it is not a true pdf.

[2] The source localization problem we consider here is only two-dimensional, since we assume the height of the source is known.

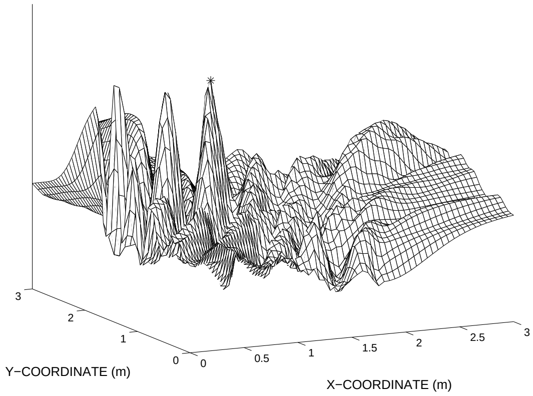

Fig. 3. Beampattern of microphone array at $t = 0.7$ s. The source is located at $(x, y) = (0.85, 1.35)$ denoted by the asterisk.

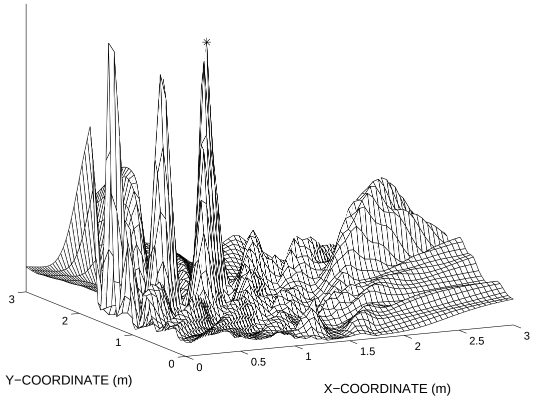

Fig. 4. Pseudo-likelihood function at $t = 0.7$ s. The source is located at $(x, y) = (0.85, 1.35)$ denoted by the asterisk.

beamformer output power (8), we used a frequency weighting function that was unity between 300 and 3000 Hz and zero elsewhere. The sampling frequency was 8 kHz.

As described above, the beampattern of the microphone array will exhibit several peaks due to reverberation (and spatial aliasing effects at high frequencies). Figure 3 shows an example of the steered beamformer output power. We used the LSF (14) with $i = 3$. As seen in Fig. 4, this has the desired effect of reducing the level of the smaller peaks (corresponding to the sidelobes) and concentrating the distribution around the modes.

The source localization algorithm was run with 50 particles, whose positions were uniformly randomly initialized with zero initial velocity. The microphone signals were processed in frames of 64 msec with 50% overlap. Results are shown in Fig. 5. After the first few frames, the particles are able to lock onto the source and track its location over time.

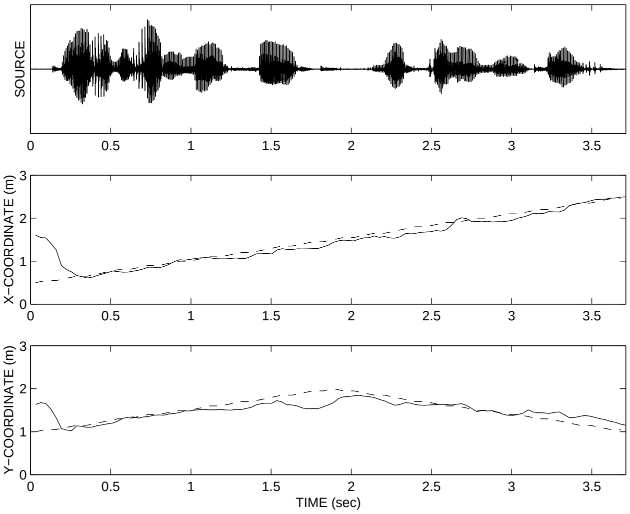

Fig. 5. Results of proposed algorithm. The top plot shows the source speech signal. The bottom two plots show the estimated (solid) and true (dashed) source trajectory in the x and y directions as a function of time.

5. CONCLUSIONS

Localization using a steered beamformer is a conceptually simple idea, although calculating the beampattern over candidate source locations is computationally expensive. In this paper we have used a particle filtering framework to develop a beamformer-based source localizer that is computationally undemanding. This scheme has the advantage that it does not require intermediate calculation of time-delay estimates, and is very attractive for speech enhancement applications where a beamformer output must be calculated anyway. Simulations show that it also provides good results in reverberant environments.

6. REFERENCES

[1] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. ICASSP-01*, Salt Lake City, UT, USA, May 2001.

[2] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, Berlin, Germany, 2001.

[3] S.M. Griebel and M.S. Brandstein, "Microphone array source localization using realizable delay vectors," in *Proc. WASPAA-01*, New Paltz, NY, USA, Oct. 2001.

[4] N.J. Gordon, D.J. Salmond, and A.F.M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proc. F, Commun., Radar & Signal Process.*, vol. 140, no. 2, pp. 107–113, Apr. 1993.

[5] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.