IEEE International Conference on
Acoustics, Speech, and Signal Processing (ICASSP)
Orlando, Florida, May 2002

FAST BAYESIAN ACOUSTIC LOCALIZATION

Stanley T. Birchfield and Daniel Kahn Gillmor

Quindi Corporation, 480 S. California Ave., Palo Alto, California 94306
{birchfield, dkg}@quindi.com

ABSTRACT

We derive a probabilistic formulation, based upon Bayes' rule, for the acoustic localization problem. The resulting formula is shown to be closely related to the energy of a conventionally beamformed signal. We then present a close approximation to both which is much faster to compute — by two orders of magnitude with our experimental setup. The fast algorithm is essentially a generalization of approaches based upon time delay estimates (TDE's), by applying the principle of least commitment. Experiments on real signals demonstrate accurate localization in noisy, reverberant environments (less than 3 dB SNR) several times faster than real time.

1. INTRODUCTION

Two common ways of determining the location of an acoustic source using an array of microphones are beamforming [1, 2] and time-delay estimation [3]. In beamforming, the signals from the microphones are time-shifted and summed, and the energy of the reconstructed signal indicates the likelihood of the source being the location corresponding to the time delays used. Although beamforming can give good results, the computation is generally too expensive to allow the likelihoods to be computed at all possible locations. As a result, sophisticated search strategies are employed which are inevitably unable to ensure that the global maximum has been found.

Traditionally, time-delay estimation has been a two-step process. First, the signals from pairs of microphones are correlated, and the peak of each correlation vector is taken as an estimate of the time delay between the microphones in the pair. Then these estimates are combined, using any of several techniques, to determine the location of the acoustic source. Although such methods are fast, they generally perform poorly in highly reverberant environments.

In [4], the *principle of least commitment* was applied to acoustic localization, resulting in a fast, robust algorithm for a compact microphone array. Like time-delay estimation, the method involves correlating pairs of microphone signals, but instead of taking the peak of each correlation vector, all the correlation values from all the vectors are accumulated in a common coordinate system.

This paper extends that work by deriving, using Bayes' rule, a probabilistic way of combining the information from all the microphones. Surprisingly, the resulting formula is shown to be very similar to that of conventional beamforming. We then offer a computationally efficient alternative, which in many scenarios can be expected to yield results indistinguishable from either the probabilistic or beamforming approaches. Because it does not make the plane-wave assumption inherent in [4], the algorithm presented here is much simpler and is also applicable to non-compact microphone arrays.

2. BAYESIAN DERIVATION

Suppose we have N microphones and a source signal $s(t)$ propagating through a generic free space with noise. The signal acquired by the ith microphone, $i = 1, \ldots, N$ can be modeled as

$$x_i(t) = g_i(t) * s(t - \tau_i) + \xi_i(t),$$

where τ_i is the propagation time $\xi_i(t)$ is additive noise, and g_i is the acoustic impulse response of the channel between the source and the ith microphone [5].

Let each value of $x_i(t)$ be treated as an estimator for τ_i and for $s(t - \tau_i)$. Using Bayes' rule, the *a posterior* probability that the source location is $\mathbf{q}$ is

$$\mathcal{P} = P(\mathbf{q}, s|x_1, \ldots, x_N) = \frac{P(x_1, \ldots, x_N|\mathbf{q}, s)P(\mathbf{q}, s)}{P(x_1, \ldots, x_N)}.$$

Ignoring the denominator, which is just a normalization constant, and assuming that the prior $P(\mathbf{q}, s)$ is uniform, this reduces to the maximum likelihood: $P(x_1, \ldots, x_N|\mathbf{q}, s)$.

Let us assume that the sound source is audible and in a fixed location during the time interval $[t_0 - W, t_0 + W]$, where $2W$ is a window size. For simplicity, let us also assume that $\xi_i(t)$ is independent zero-mean white Gaussian noise with variance σ_i^2, and let us ignore reverberation by setting $g_i(t)$ is the Dirac delta function. In that case, every

value of the microphone signals are considered as independent measurements, leading to

$$\begin{aligned}\mathcal{P}' &= P(x_1, \ldots, x_N|\mathbf{q}, s) = \prod_{i=1}^{N} P(x_i|\mathbf{q}, s) \\ &= \prod_{i=1}^{N} e^{-\int_{t_0-W}^{t_0+W} \frac{[x_i(t+\tau_i)-s(t)]^2}{2\sigma_i^2}\, dt},\end{aligned}$$

where $\tau_i = \| \mathbf{q} - \mathbf{m}_i \| / c$, if $\mathbf{m}_i$ is the location of microphone i and c is the speed of sound in the medium.

Since we don't have access to the original source s, we use the maximum likelihood estimator (MLE):

$$\hat{s}(t) = \frac{1}{N} \sum_{i=1}^{N} x_i(t + \tau_i).$$

Substituing $\hat{s}$ for s, assuming all the σ_i's are equal, and taking the logarithm, we get

$$\log \mathcal{P}' = -\sum_{i=1}^{N} \int_{t_0-W}^{t_0+W} [x_i(t + \tau_i) - \hat{s}(t)]^2\, dt,$$

which reduces, after some algebraic manipulation, to

$$\mathcal{L}_{bayes}(\mathbf{q}) = \log \mathcal{P}' = \frac{2}{N} V_C - \frac{N-1}{N} V_E, \qquad (1)$$

where

$$V_C = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \int_{t_0-W}^{t_0+W} x_i(t + \tau_i) x_j(t + \tau_j)\, dt$$

is the sum of samples taken at various offsets from the cross-correlations of each pair of microphone signals, and

$$V_E = \sum_{i=1}^{N} \int_{t_0-W}^{t_0+W} x_i^2(t + \tau_i)\, dt$$

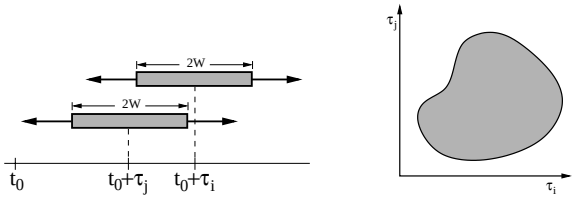
is the combined energy in all the signals.

2.1. Comparison with beamforming

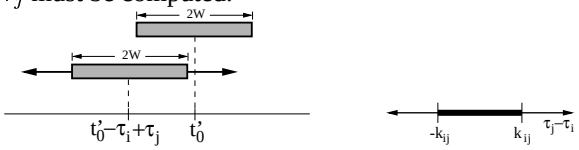
Conventional beamforming [6] computes the energy of the reconstructed signal:

$$\begin{aligned}\mathcal{L}_{beam}(\mathbf{q}) &= \int_{t_0-W}^{t_0+W} \left[\sum_{i=1}^{N} x_i(t + \tau_i) \right]^2 dt \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{t_0-W}^{t_0+W} x_i(t + \tau_i) x_j(t + \tau_j)\, dt \\ &= 2V_C + V_E \qquad (2)\end{aligned}$$

Notice the similarity between Eqs. (1) and (2). Both methods seek to maximize the cross-correlation between microphone signals, but the way they treat energy is different. In one case the total energy is minimized, while in the other it is maximized.



With V_C, both windows move. All combinations of τ_i and τ_j must be computed.



With V_C', one window remains fixed. Only the values along a line bounded by $|\tau_j - \tau_i| \leq k_{ij}$, where $k_{ij} = \lfloor \frac{r d_{ij}}{c} \rfloor$, are computed.

Fig. 1. Why V_C' is more efficient than V_C.

3. A FASTER METHOD

Computing V_C for every possible location $\mathbf{q}$ is expensive. As a result, researchers have proposed various techniques for efficiently searching the space. Invariably, however, these methods are prone to finding local maxima rather than the global maximum. In this section, we show how to efficiently compute a close approximation to V_C.

First notice that, by a change of variables, V_C can be expressed as

$$V_C = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \int_{t_0+\tau_i-W}^{t_0+\tau_i+W} x_i(t) x_j(t - \tau_i + \tau_j)\, dt.$$

A tremendous gain in efficiency can be realized by instead computing

$$V_C' = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \int_{t_0'-W}^{t_0'+W} x_i(t) x_j(t - \tau_i + \tau_j)\, dt,$$

where t_0' is a constant.

Figure 1 illustrates the difference between these two formulas. With V_C, the window over which the integral is computed varies for both microphone signals under consideration. As a result, the integral must be computed for every possible combination of τ_i and τ_j corresponding to the hypothesized locations under consideration. The question that V_C is helping to answer is, "Where was the source that generated sound at time t_0?"

In contrast, V_C' ignores the source generation time t_0 and instead asks, "Where was the source that generated sound that was *heard* by microphone i at time t_0'?" One of the microphone signal windows is held constant while the other varies relative to it. As a result, the integral is computed

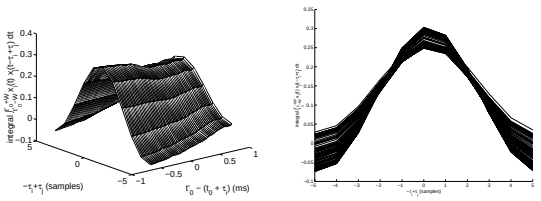


Fig. 2. Even with a large microphone array, V_C' is a good approximation to V_C because the value of the integral is largely independent of t_0'.

for a one-dimensional set of values $\tau_j - \tau_i$, which is simply the cross-correlation between the two signals. Moreover, because $|\tau_j - \tau_i| \leq d_{ij}/c$, where $d_{ij} = \| \mathbf{m}_i - \mathbf{m}_j \|$ is the distance between the microphones, the range of values is determined solely by the microphone array geometry (and the sampling rate, in discrete processing), independent of the number of hypothesized locations.

How good of an approximation is V_C' to V_C? Figure 2 displays the value of the integral of a typical speech signal as a function of t_0' over 14 ms, which corresponds to the maximum time-shift $\max |t_0' - t_0 - \tau_i|$ for a 5m $\times$ 5m room with microphones around the perimeter. Notice that the value of the integral changes by just 25%, but more importantly, the ordering of the hypotheses is preserved. That is, the peak is always the correct $\tau_j - \tau_i$ independent of the value of t_0'. Therefore, when the speech utterance is long relative to the maximum time-shift, we can expect the relative values computed with V_C' to be very similar to the relative values computed with V_C, even with large microphone arrays.

Notice that V_C' involves essentially the same computation as the first step of TDE-based locators. Instead of taking the peak, however, all the correlation vectors are combined to improve robustness to noise, similar to that done in [1, 7]. Because we are able to compute the a reasonably dense sampling of the entire probability density function in well under real time, there is no advantage in using the particle filtering techniques of [7].

Let us now compare the computational complexity of the two alternatives. (We ignore the computational cost of time-shifting the input signals, which often requires interpolation.) In both formulations, the integral is evaluated $2WQ\binom{N}{2}$ times, where Q is the number of hypothesized locations $\mathbf{q}$ being evaluated. Evaluating the integral requires $2Wr$ multiplies, where r is the sampling rate. Thus, the total number of multiplies to compute V_C is $2WrQ\binom{N}{2}$. For V_C', evaluating the integral is simply a lookup, but the precomputation requires $2WrK\binom{N}{2}$ multiples, where $K \leq 2\lfloor \frac{rd_{max}}{c} \rfloor + 1$ and d_{max} is the maximum distance between any two microphones. Thus V_C' requires at most $Q\binom{N}{2} + 2WrK\binom{N}{2}$ multiplies. The ratio of the number of multiplies for V_C' to V_C is given by $\frac{1}{2Wr} + \frac{K}{Q}$. For typical scenarios, the second term dominates. We have a compact array of four microphones spaced 15 cm apart. We have chosen a sample domain with 2500 sample locations, which easily gives us as much resolution as the sampling rate of our input signals. With these numbers, V_C' requires just over 1% of the multiplies as V_C. Thus, V_C', in this scenario, is faster by two orders of magnitude. Of course, with a large microphone array and a small number of locations, the computational advantage will decrease, as long as every pair of microphones is cross-correlated.

Because we have been able to achieve excellent results without V_E, our algorithm relies solely upon V_C'. A summary of the three methods is shown in the following table.

$$\begin{aligned} \mathcal{L}_{bayes}(\mathbf{q}) &= \frac{2}{N}V_C - \frac{N-1}{N}V_E \\ \mathcal{L}_{beam}(\mathbf{q}) &= 2V_C + V_E \\ \mathcal{L}_{corr}(\mathbf{q}) &= V_C' \end{aligned}$$

4. HANDLING MULTIPLE SOUND SOURCES

By computing likelihoods for all possible locations $\mathbf{q}$, the methods presented in the previous sections are capable of finding mulitple simultaneous sound sources. To see this, let τ_{ki} represent the travel time between source s_k to microphone i, and let C_k represent the correlation sample inside the integral due to a single source s_k:

$$\begin{aligned} C_k &= x_i(t)x_j(t - \tau_i + \tau_j) \\ &= s_k(t - \tau_{ki})s_k(t - \tau_i + \tau_j - \tau_{kj}). \end{aligned}$$

It is straightforward to show that the correlation sample, C_{12}, due to two simultaneous sound sources s_1 and s_2, is

$$\begin{aligned} C_{12} &= x_i(t)x_j(t - \tau_i + \tau_j) \\ &= [s_1(t - \tau_{1i}) + s_2(t - \tau_{2i})] \cdot \\ &\quad [s_1(t - \tau_i + \tau_j - \tau_{1j}) + s_2(t - \tau_i + \tau_j - \tau_{2j})] \\ &= C_1 + C_2 + \xi', \end{aligned}$$

where

$$\begin{aligned} \xi' &= s_1(t - \tau_{1i})s_2(t - \tau_i + \tau_j - \tau_{2j}) \\ &\quad + s_2(t - \tau_{2i})s_1(t - \tau_i + \tau_j - \tau_{1j}). \end{aligned}$$

If s_1 and s_2 are uncorrelated, then the expected value of the cross terms in ξ' are zero, which means that $E[\xi'] = 0$. Thus, ξ' can be treated as an additive zero-mean noise source. The value computed at any location is approximately the sum of the values that would have been computed with either source alone. Thus, we can detect multiple simultaneous speakers by looking for multiple local maxima in the set of possible locations.

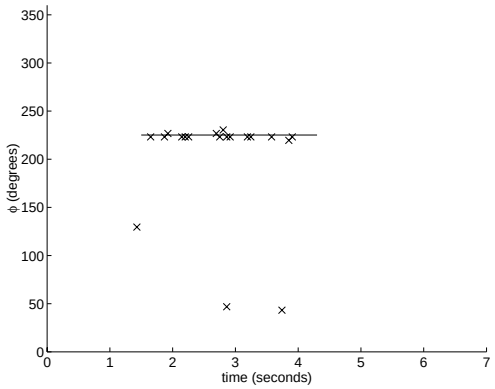


Fig. 3. Results with one speaker. The line indicates ground truth.

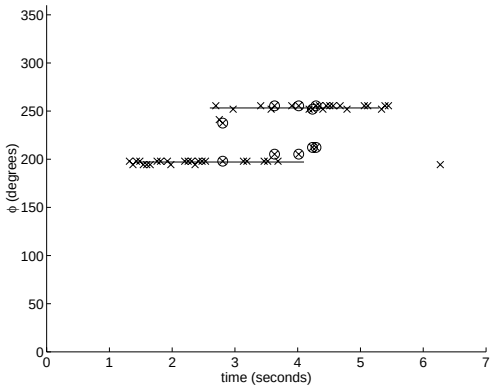


Fig. 4. Results with two simultaneous speakers, one starting later than the other. The lines indicate ground truth. The circles indicate audio frames containing multiple peaks.

5. EXPERIMENTAL RESULTS

For our experiments, an array of four microphones, arranged in a square with a diagonal of 15 cm, was placed on a table in a conference room. A recording of a male voice counting from one to ten was played through a speaker on the table. The hypothesized locations consisted of a sampling over the surface of a hemisphere (1m radius) centered on the center of the array. This surface was sampled at 100 longitudes and 25 latitudes. $\mathcal{L}_{corr}$ was measured at each location $\mathbf{q}$ for each 55 ms window of audio sampled at 44.1 kHz. PHAT prefiltering was applied, as well as bandpass filtering between 3 and 4 kHz. The entire algorithm takes less than 7 ms to process a 55 ms frame of audio, on a 450 MHz Pentium III.

Figure 3 shows the results of a single speaker at a distance of 2 m, with SNR ranging from 0.2 to 2.8 dB. An x is placed at every local maxima above a threshold. The line indicates ground truth. During the time of the recording, the median error in pan angle (ϕ) was 1.97 degrees.

Shown in Figure 4 are the results of playing the recording through two speakers 1.1 m from the array, starting the second one when the first was approximately halfway finished. The SNR ranged from 4.8 to 6.3 dB. The algorithm is able to simultaneously detect multiple speakers.

6. CONCLUSION

We have used Bayes' rule to derive an optimal algorithm for performing acoustic localization from an array of microphones and showed that it is very similar to conventional beamforming. We then presented a new algorithm which closely approximates the Bayesian formulation yet is orders of magnitude faster than conventional beamforming localization techniques for compact microphone arrays. The new algorithm also more closely approximates the Bayesian formulation because of its decreased reliance on the energy term. The algorithm was shown to produce excellent results for a single speaker and for multiple speakers.

7. REFERENCES

[1] H. F. Silverman and S. E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone array data," *Computer Speech and Language*, vol. 6, no. 2, pp. 129–152, 1992.

[2] R. Duraiswami, D. Zotkin, and L. Davis, "Active speech source localization by a dual coarse-to-fine search," in *Proc. of the IEEE ICASSP*, 2001.

[3] M. S. Brandstein and H. F. Silverman, "Practical methodology for speech source localization with microphone arrays," *Computer Speech and Language*, vol. 11, no. 2, pp. 91–126, 1997.

[4] S. T. Birchfield and D. K. Gillmor, "Acoustic source direction by hemisphere sampling," in *Proc. of the IEEE ICASSP*, 2001.

[5] Y. Huang, J. Benesty, and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for realtime acoustic source localization system," in *Proc. of the IEEE ICASSP*, pp. 937–940, 1999.

[6] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, Michael S. Brandstein and D. B. Ward, Eds. Springer Verlag, 2001.

[7] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.