European Signal Processing Conference (EUSIPCO)
Vienna, Austria, September 2004

A UNIFYING FRAMEWORK FOR ACOUSTIC LOCALIZATION

Stanley T. Birchfield

Dept. of Electrical and Computer Engineering
Clemson University, Clemson, SC 29634
stb@clemson.edu

ABSTRACT

Recent advances in acoustic localization have combined the advantages of the traditional methods of beamforming and time-delay estimation, leading to techniques that are both accurate and fast. We present a unifying framework that reveals the relationships between beamforming, time-delay estimation, Bayesian formulation, hemisphere sampling, and accumulated correlation. We then experimentally compare the algorithms, on both compact and distributed microphone arrays, showing that the recent technique of accumulated correlation, although much less computationally expensive, exhibits performance comparable to that of beamforming.

1. INTRODUCTION

Traditionally, two common methods for determining the location of a sound source have been beamforming [14, 6, 8] and time-delay estimation (TDE) [4, 5, 13, 11]. Beamforming generally produces better results because it takes all the information into account before making a decision, while TDE techniques are more computationally efficient because they rely upon the fast operation of cross-correlation.

Recently an algorithm called *accumulated correlation* was proposed to combine the advantages of these two approaches [2]. Like beamforming, it is accurate because it takes all the information into account before making a decision. Like TDE, it is computationally efficient because it uses cross-correlation as its basic operation. Accumulated correlation builds upon the successful *hemisphere sampling* algorithm [1] that was shown to be more accurate than traditional TDE methods without significantly more computation. Unlike the latter, though, accumulated correlation makes no plane-wave assumption, thus making it applicable to both compact and distributed microphone arrays.

In this paper we place accumulated correlation into a unified framework that highlights the relationships between it and the classical techniques of beamforming and time-delay estimation, as well as the Bayesian formulation and hemipshere sampling. We also provide an experimental analysis of the algorithm, answering some of the questions that remain from [2]. For example, we show that the algorithm not only is applicable to distributed microphone arrays *in theory*, but that it also works well on such arrays *in practice.* We demonstrate on real audio data that the algorithm performs as well as beamforming on both compact and distributed microphone arrays.

2. ALGORITHMS

The algorithms under consideration in this paper fall into two categories: traditional techniques and recent advances.

2.1 Traditional techniques

Traditional delay-and-sum beamforming computes the likelihood that a sound source is at location $\mathbf{q}$ by measuring the energy of the reconstruted signal at that location:

$$\mathcal{L}_{beam}(\mathbf{q}) = \int_{t_0-\frac{W}{2}}^{t_0+\frac{W}{2}} \left[\sum_{i=1}^{N} x_i(t+\tau_{i,\mathbf{q}}) \right]^2 dt,$$

where N is the number of microphones, x_i is the signal received by the ith microphone, $\tau_{i,\mathbf{q}}$ is the travel time for sound to reach microphone i from location $\mathbf{q}$, and t_0 and W are the center and width of the integration window, respectively [6, 7, 8, 14]. In [2] it was shown that this equation can be expressed as the sum of two terms: $\mathcal{L}_{beam}(\mathbf{q}) = 2V_C(\mathbf{q}) + V_E(\mathbf{q})$, where V_C measures the pairwise similarity between the received signals, and V_E is the combined energy in all the signals:

$$\begin{aligned} V_C(\mathbf{q}) &= \sum_{i=1}^{N} \sum_{j=i+1}^{N} \int_{t_0-\frac{W}{2}}^{t_0+\frac{W}{2}} x_i(t+\tau_{i,\mathbf{q}}) x_j(t+\tau_{j,\mathbf{q}})\, dt \\ V_E(\mathbf{q}) &= \sum_{i=1}^{N} \int_{t_0-\frac{W}{2}}^{t_0+\frac{W}{2}} x_i^2(t+\tau_{i,\mathbf{q}})\, dt \end{aligned}$$

Delay-and-sum beamforming is similar to a Bayesian formulation: $\mathcal{L}_{Bayes}(\mathbf{q}) = \frac{1}{N}\left[V_C(\mathbf{q}) + (1-N)V_E(\mathbf{q})\right]$ [2]. In fact, in the context of maximizing the likelihood, the two formulations are identical except for how they weight the energy term. When the signals are stationary this energy term will have no effect on the likelihoods and the Bayesian and beamforming equations become identical.

A classic alternative to beamforming is time-delay estimation (TDE), which involves two steps. First, the cross-correlation is computed between each microphone pair:

$$R_{ij}(\tau) = \int_{t'_0-\frac{W}{2}}^{t'_0+\frac{W}{2}} x_i(t) x_j(t-\tau)\, dt, \qquad (1)$$

where t'_0 is the approximate time at which the sound was heard as opposed to the time t_0 at which the sound was generated [2]. By applying R_{ij} to a range of discrete values, a cross-correlation vector $\mathbf{v}_{ij}$ of length $2\lfloor \frac{dr}{c} \rfloor + 1$

is generated, where d is the distance between the two microphones, r is the sampling rate, and c is the speed of sound. Each element of $\mathbf{v}$ indicates the likelihood that the sound source is located near a half-hyperboloid centered at the midpoint between the two microphones, with its axis of symmetry the line connecting the two microphones.

The second step of TDE methods estimates the location of the sound source using the peaks of the cross-correlation vectors. For this step a number of approaches have been proposed [4, 5, 11, 13].

2.2 Recent advances

A few years ago, a technique for compact microphone arrays called *hemisphere sampling* was introduced [1]. Like TDE, the first step of hemisphere sampling is to compute the cross-correlations between pairs of microphone signals using Eq. (1). In the second step, however, instead of using just the peaks of the cross-correlation vectors, the direction to the sound source is determined by mapping *all* the elements of the vectors onto a hemisphere surrounding the microphone array. The peak of the entire hemisphere then indicates the sound source direction. By accumulating all the values in a common coordinate system, this method follows the principle of least commitment because it delays the decision as long as possible, resulting in more robust behavior. The increased accuracy of hemisphere sampling over the linear intersection variant of TDE [4] was demonstrated in [1].

More recently, this concept of accumulating correlation vectors by mapping them to a common coordinate system was extended to handle arbitrary microphone array geometries, in a technique called *accumulated correlation* [2]. The resulting computation is surprisingly simple:

$$\mathcal{L}_{corr}(\mathbf{q}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} R_{ij}(\tau_{j,\mathbf{q}} - \tau_{i,\mathbf{q}}), \tag{2}$$

where R_{ij} again comes from Eq. (1). Like TDE and hemisphere sampling, accumulated correlation is a two-step process: first the cross-correlation vector is computed for each microphone pair, then all the elements of the vectors are mapped onto a common coordinate system to yield a likelihood for each candidate location.

2.3 Discussion

Not only is there a natural connection between TDE and accumulated correlation due to their shared use of cross-correlation as the basic operation, but there is also a connection between accumulated correlation and delay-and-sum beamforming. It can be shown [2] that $\mathcal{L}_{corr}(\mathbf{q}) = V'_C$, where V'_C is identical to V_C except for a change of the integration limits (by substituting $t'_0 - \tau_i$ for t_0 in the equation for V_C above).

In comparing V_C and V'_C, a key parameter is the *maximum relative discrepancy* $\tau_{\max} = \max_{i\in\{1,\ldots,N\},\mathbf{q}\in\mathcal{Q}} |\bar{\tau} - \tau_{i,\mathbf{q}}|/W$, where $\bar{\tau}$ is the average time delay between any microphone and any candidate sound source location and $\mathcal{Q}$ is the set of such locations. If this maximum relative discrepancy is small then V'_C and V_C will be approximately equal, but as it increases the two alternatives diverge.

Because it is based on the fast operation of cross-correlation, V'_C in many common scenarios is more computationally efficient than V_C, sometimes by several orders of magnitude (see Section 4). In fact, exhaustively searching the space using V'_C can often be performed in real time. Thus, when the maximum relative discrepancy is small, V'_C obviates the need for complicated search strategies [8] or multi-hypothesis methods [12, 14] that are often needed to make beamforming practical. Moreover, since the search is exhaustive, temporal smoothing of either V_C or V'_C reduces the effects of spurious global peaks.

2.4 Frequency-domain formulations

It is well-known that cross-correlation in the time domain is equivalent to multiplication in the frequency domain:

$$R_{ij}(\tau) = \mathcal{F}^{-1}\{X_i(f)X_j^*(f)\}, \tag{3}$$

where

$$X_i(f) = \mathcal{F}\{x_i(t)\} = \int_{t'_0-\frac{W}{2}}^{t'_0+\frac{W}{2}} x_i(t)e^{-j2\pi ft}\,dt$$

is the Fourier transform of $x_i(t)$, and * denotes the complex conjugate. Notice that the window is centered at t'_0, as in Eq. (1). Combining Eqs. (3) and (2) yields a frequency-domain formulation of accumulated correlation, or V'_C.

When formulating beamforming in the frequency domain [8, 14], careful attention must be paid to the limits of the integral to avoid confusing V_C with V'_C. SRP-PHAT, for example, where the SRP stands for "steered response power [beamformer]," is actually a frequency-domain formulation of accumulated correlation with a PHAT prefilter and an energy term [7]. The derivation of SRP-PHAT assumes that the maximum relative discrepancy is small, thus enabling V_C to be approximated by V'_C.

2.5 Prefilters

To reduce the effect of reverberation, a common practice is to apply a prefilter to the signals such as the phase transform (PHAT): $X_i(f)/|X_i(f)|$ [7, 10, 9]. By dividing the signal by its magnitude, PHAT treats all frequencies the same. PHAT can be applied either to time-domain or frequency-domain formulations.

3. UNIFYING FRAMEWORK

All of the approaches above can be written as follows:

$$\mathcal{L}(\mathbf{q}) = \mathcal{G}\left(\int_{\mathcal{T}(i,\mathbf{q})-\frac{W}{2}}^{\mathcal{T}(i,\mathbf{q})+\frac{W}{2}} x_{ij,\mathbf{q}}(t)\,dt\right) + \alpha V_E,$$

where $x_{ij,\mathbf{q}}(t) = x_i(t)x_j(t - \tau_{i,\mathbf{q}} + \tau_{j,\mathbf{q}})$. $\mathcal{G}(\cdot)$ is the function that combines the results from the different microphone pairs, such as $\mathcal{G}_\Sigma = \sum_{i=1}^N \sum_{j=i+1}^N$, $\mathcal{G}_{\Sigma\Sigma} = \mathcal{G}_\Sigma + \sum_{j=1}^N \sum_{i=j+1}^N$, $\mathcal{G}_{LI}$ for linear intersection [4], $\mathcal{G}_{HS}$ for hemisphere sampling [1], and so on. The function

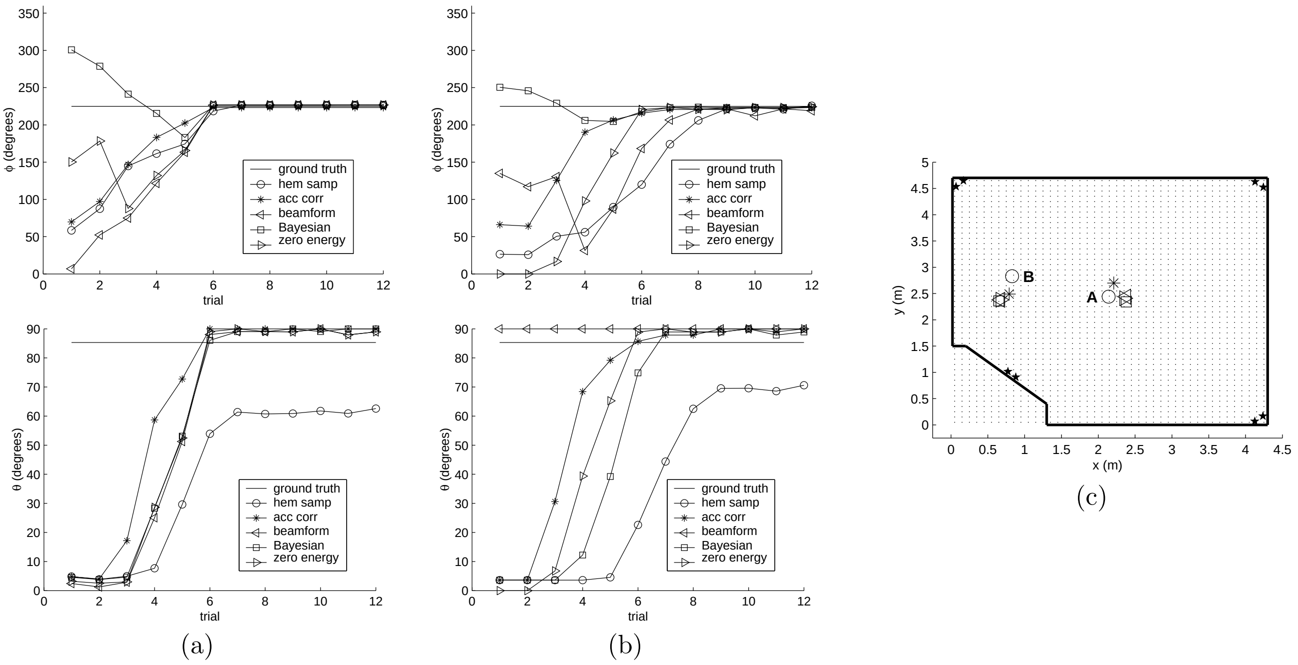

Figure 1: Results of algorithms on compact array (a) without PHAT and (b) with PHAT. The top row shows the pan angle (ϕ), while the bottom row shows the tilt angle (θ). (c) Top-down view of the room used in the experiments (4.3m × 4.7m × 2.6m), along with the microphones (indicated by stars) of the distributed array. For the distributed experiments the dots show the candidate locations considered by the algorithms, the circles show the two actual sound source locations (A and B), and the average computed locations of the algorithms are shown using the same symbols as the other figures.

$\mathcal{T}(i, \mathbf{q})$ indicates the center of the window and is generally one of two values: $\mathcal{T}_G = t_0 + \tau_{i,\mathbf{q}}$ (used by V_C), or $\mathcal{T}_H = t'_0$ (used by V'_C).

The algorithms, then, differ by their choice of the functions $\mathcal{G}$ and $\mathcal{T}$ and parameter α, as summarized in the following table:

method	$\mathcal{G}$	$\mathcal{T}$	α
Bayesian [2]	$\mathcal{G}_\Sigma$	$\mathcal{T}_G$	$(1-N)/2$
beamforming [14, 6, 8]	$\mathcal{G}_\Sigma$	$\mathcal{T}_G$	$1/2$
zero energy	$\mathcal{G}_\Sigma$	$\mathcal{T}_G$	0
accumulated correlation [2]	$\mathcal{G}_\Sigma$	$\mathcal{T}_H$	0
SRP-PHAT [7]	$\mathcal{G}_{\Sigma\Sigma}$	$\mathcal{T}_H$	1
hemisphere sampling [1]	$\mathcal{G}_{HS}$	$\mathcal{T}_H$	0
linear intersection [4]	$\mathcal{G}_{LI}$	$\mathcal{T}_H$	0

From this table we see that acoustic localization algorithms can generally be divided into two categories: those using $\mathcal{T}_G$ and those using $\mathcal{T}_H$. The latter comprises all TDE methods [4, 5, 13, 11] that, due to space constraints, are omitted from the table. The role of accumulated correlation is clear: Like the TDE methods it uses $\mathcal{T}_H$ and is therefore fast, and like beamforming it uses $\mathcal{G}_\Sigma$, making it robust. Keep in mind that, for any algorithm using $\mathcal{T}_G$, $\mathcal{G}_\Sigma$ can be replaced by $\mathcal{G}_{\Sigma\Sigma}$ if α is multiplied by two. Thus, SRP-PHAT is identical to beamforming except for $\mathcal{T}$.[1]

4. EXPERIMENTS

In this section we compare the performance of the algorithms on compact and distributed arrays of microphones in a real environment. All experiments were conducted in the conference room shown in Figure 1(c) using a sampling rate of 44.1 kHz. Reverberation time of the room was approximately 200 ms, results were smoothed with a temporal half-life of 250 ms, and ground truth was estimated by hand using a tape measure.

4.1 Compact array

In the first experiment, a compact array of four omnidirectional microphones were arranged in a square with a distance of 15cm between opposing microphones. The maximum relative discrepancy was 1.3%. The array was placed horizontally on a large table in the center of the room, while a computer speaker played a recording of a male voice counting from one to ten repeatedly.

Twelve trials were captured by increasing the volume on the speaker in each succeeding trial, thereby increasing the signal-to-noise ratio (SNR). For each trial the algorithms of accumulated correlation, hemisphere sampling, beamforming, Bayesian, and zero energy (which is simply V_C) were used to compute the pan and tilt angles from the microphone array to the sound source. The angles, averaged over all non-overlapping 55-ms windows for each trial, are shown in Figure 1(a) and (b) both with and without the PHAT prefilter. For V'_C PHAT was computed on the cross-correlation signal, while for

[1] Another difference is the PHAT prefilter, but this is a separate computation that can be applied to any of the algorithms in the table.

	— location A —		— location B —	
Algorithm	μ_x	μ_y	μ_x	μ_y
acc corr	7	26	4	34
beamforming	24	0	16	45
Bayesian	24	10	18	47
zero energy	21	3	16	45

Table 1: Absolute error in x and y (cm) for the distributed array experiments.

V_C it was computed on the individual signals. A bandpass filter of 3 to 4 kHz was used throughout [6, 14]. On a 550 MHz PIII, the computing time per 55-ms window was 6 ms for accumulated correlation and 3969 ms for beamforming.

As the SNR increased, the algorithms generally performed better, as one would expect. Somewhat surprisingly, better results were obtained without PHAT. This decrease in performance was perhaps due to the fact that PHAT accentuates those parts of the signal with low SNR [9, 3]. Without PHAT, all algorithms performed well after the SNR reached 2 dB (trial 6), the only exception being the hemisphere sampling algorithm which was never able to estimate the tilt accurately because of its plane-wave assumption.

4.2 Distributed array

In the next experiment eight microphones were arranged in four pairs, one pair per corner of the room, as shown in Figure 1(c). The maximum relative discrepancy was 8.5%. A computer speaker played the same audio file as before at two separate locations, one near the center of the room at $(2.1, 2.4, 1.5)$m and one off-center at $(0.8, 2.8, 1.7)$m, with average SNR of 2.3 dB and 2.6 dB, respectively. The bandpass filter was used, but PHAT was not. The computing time per 100-ms window was 146 ms for accumulated correlation and 7452 ms for beamforming.

The results obtained by averaging the locations of each non-overlapping 100-ms window in both x and y are shown in Figure 1(c) and the absolute errors of the locations in Table 1. The errors are on the order of 0 to 47 cm, while the difference in error between algorithms along any given axis is between 2 and 26 cm. Although accumulated correlation exhibits the least error overall, the differences between the algorithms are too small relative to the accuracy of ground truth (the width of the speaker itself was 9 cm) to declare a clear winner. All algorithms perform comparably.

5. CONCLUSION

In this paper we have provided a unifying framework for acoustic localization algorithms including beamforming, TDE, hemisphere sampling, the Bayesian formulation, and accumulated correlation. The algorithms were experimentally compared on both a compact and a distributed array of microphones in a real environment. Although accumulated correlation requires orders of magnitude less computation, it performs comparably to the accurate but expensive method of beamforming.

REFERENCES

[1] S. T. Birchfield and D. K. Gillmor. Acoustic source direction by hemisphere sampling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[2] S. T. Birchfield and D. K. Gillmor. Fast Bayesian acoustic localization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[3] M. S. Brandstein. Time-delay estimation of reverberated speech exploiting harmonic structure. *Journal of the Acoustical Society of America*, 105(5):2914–2919, 1999.

[4] M. S. Brandstein, J. E. Adcock, and H. F. Silverman. A closed-form method for finding source locations from microphone-array time-delay estimates. In *ICASSP*, volume 5, pages 3019–3022, 1995.

[5] M. S. Brandstein and H. F. Silverman. Practical methodology for speech source localization with microphone arrays. *Computer Speech and Language*, 11(2):91–126, 1997.

[6] N. Checka, K. Wilson, V. Rangarajan, and T. Darrell. A probabilistic framework for multi-modal multi-person tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[7] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein. Robust localization in reverberant rooms. In M. S. Brandstein and D. B. Ward, editors, *Microphone Arrays*. Springer Verlag, 2001.

[8] R. Duraiswami, D. Zotkin, and L. Davis. Active speech source localization by a dual coarse-to-fine search. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[9] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, Aug. 1976.

[10] M. Omologo and P. Svaizer. Use of the crosspower-spectrum phase in acoustic event location. *IEEE Transactions on Speech and Audio Processing*, 5(3), 1997.

[11] P. Svaizer, M. Matassoni, and M. Omologo. Acoustic source location in a three-dimensional space using crosspower spectrum phase. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 231–234, 1997.

[12] J. Vermaak and A. Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *ICASSP*, 2001.

[13] H. Wang and P. Chu. Voice source localization for automatic camera pointing system in videoconferencing. In *ICASSP*, pages 187–190, 1997.

[14] D. B. Ward and R. C. Williamson. Particle filter beamforming for acoustic source localization in a reverberant environment. In *ICASSP*, 2002.