IEEE International Conference on
Acoustics, Speech, and Signal Processing (ICASSP)
Salt Lake City, Utah, May 2001

ACOUSTIC SOURCE DIRECTION BY HEMISPHERE SAMPLING

Stanley T. Birchfield and Daniel Kahn Gillmor

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

ABSTRACT

A method for estimating the direction to a sound source, using a compact array of microphones, is presented. For each pair of microphones, the signals are prefiltered and correlated. Rather than taking the peak of the correlation vectors as estimates for the time delay between the microphones, all the correlation vectors are accumulated in a common coordinate system, namely a unit hemisphere centered on the microphone array. The maximum cell in the hemisphere then indicates the azimuthal and elevation angles to the source. Unlike previous techniques, this algorithm is applicable to arbitrary microphone configurations, handles more than two microphone pairs, and has no blind spots. Experiments demonstrate significantly increased robustness to noise, compared with previous techniques.

1. INTRODUCTION

Determining the location of an acoustic source, or sound source, using an array of microphones is generally a two-step process. First, the relative times of arrival for the sound to reach the different microphones are computed. Secondly, these time-delay estimates (TDE's) are used to determine the sound source location.

The first step is a well-studied problem. Several techniques have been proposed over the years to estimate the time delays in the presence of noise and reverberation (the accumulated sum of echos), including various prefilters [4, 6, 9], and eigenvalue decomposition [3].

Depending upon the microphone array geometry, the second step can have several interpretations. For systems with just two microphones, the angle corresponding to the TDE yields the single angle along the axis connecting the microphones [3, 9]. With four microphones arranged in a compact square, one can intersect the cones corresponding to the two opposite pairs to find the azimuthal and elevation angles indicating the direction to the sound source [1, 2]. Finally, with more microphones spaced far apart relative to the distance to the sound source, the actual 3D position of the sound source can be found by minimizing a nonlinear function, once an initial guess is given [1, 10].

In this paper, we address the problem of determining the direction to a sound source from a compact array of microphones. Instead of using the peak of the correlation vector for each pair of microphones to determine the direction, our algorithm maps the entire correlation vector from each pair into a common coordinate system, namely a sampled unit hemisphere centered on the microphone array. The accumulated sum of these vectors is then used to determine the direction. The algorithm thus follows the *principle of least commitment* [7] by delaying decisions as long as possible.

This method has several advantages over previous techniques. First, unlike [1, 2, 8], it is applicable to arbitrary microphone configurations within a compact space. Second, it does not have any blind spots, unlike the cone intersection method, as we shall see in Section 5.1. Finally, because it avoids making decisions prematurely, and because it handles more than two microphone pairs, it achieves additional robustness to noise and reverberation.

2. TIME-DELAY ESTIMATION

For a source signal $s(n)$ propagating through a generic free space with noise, the signal acquired by the ith microphone can be modeled as a convolution [3]:

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

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

To estimate the time delay $\tau = \tau_i - \tau_j$ between two microphones i and j, one can compute the generalized cross correlation (GCC) between the two signals [4, 6]:

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

where $X_i(f)$ and $X_j(f)$ are the Fourier transforms of $x_i(n)$ and $x_j(n)$, respectively. If the filter $\Psi(f)$ is 1, then the technique reduces to standard cross correlation between the two signals. More commonly, $\Psi(f)$ is a prewhitening filter (to flatten the magnitude of the power spectrum), which typically improves results in the presence of reverberation [4, 6]. Intuitively, notice that a flat power spectrum corresponds to

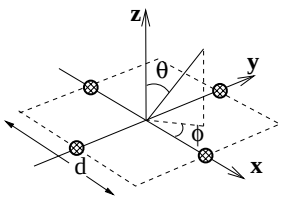


Fig. 1. The four-microphone array configuration.

a delta autocorrelation function, for which correlation works perfectly.

We use the filter $\Psi(f) = 1/(|X_i(f)X_j^*(f)|)$, commonly known as PHAT or CSP [4, 6], which normalizes the crosspower spectrum $X_i(f)X_j^*(f)$ to remove all magnitude information, leaving only the phase.

3. MICROPHONE ARRAY GEOMETRY

Although the technique of this paper is applicable to arbitrary microphone configurations, to facilitate the discussion consider the specific array shown in Figure 1, with four microphones placed at the midpoints of the sides of a square in the xy plane, centered at the origin. The length of each side is d. Let c be the speed of sound and r be the sampling rate.

By applying R_{ij} to a range of discrete values, we generate a correlation vector **v** for each pair of microphones i and j. Each element v_k, $k = -\lfloor \frac{dr}{c} \rfloor, \ldots, \lfloor \frac{dr}{c} \rfloor$, 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. At distances sufficiently far from the microphones (approximately $2d$ from the center), the half-hyperboloid is well approximated by the asymptotical cone having angle

$$\alpha_k = \cos^{-1}\left(\frac{ck}{dr}\right) \tag{1}$$

with respect to the axis of symmetry. As shown in Figure 2, a given pair of microphones has poor resolution near its axis, but together two perpendicularly-oriented pairs cover the full 360 degrees with acceptable resolution.

4. MAPPING TO THE UNIT HEMISPHERE

4.1. Coincident pairs of microphones

Because the midpoints of the two opposite pairs of microphones coincide, all the asymptotical cones have vertices at this intersection point and can therefore be mapped to a common hemispherical coordinate system centered at that point, without knowing the distance to the sound source. Using a hemisphere assumes that all sound sources are in the half-space defined by $z > 0$.

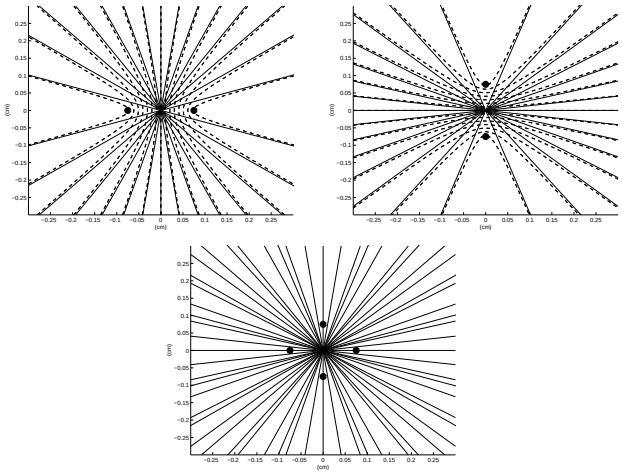


Fig. 2. The asymptotes (i.e., intersection of asymptotical cones with the xy plane) of the two opposite microphone pairs separately and together. The dashed curves are the hyperbolas.

Let us define h_p to be a function defined on the unit hemisphere such that $h_p(\theta, \phi)$ indicates the likelihood that the sound source is located in the (θ, ϕ) direction, given the correlation vector from microphone pair p. As shown in Figure 1, the angles are those of a spherical coordinate system, so that θ is the angle with respect to the z axis, and ϕ is the angle, in the xy plane, with respect to the x axis.

Let l be the line connecting the two microphones, and let γ be the angle between l and the x axis. For the opposing pairs, then, $\gamma = 0$ and $\gamma = \frac{\pi}{2}$. To determine $h_p(\theta, \phi)$, we first compute the angle between l and the ray designated by (θ, ϕ):

$$\alpha = \cos^{-1}(\sin\theta\cos(\phi - \gamma)). \tag{2}$$

Since every asymptotical cone intersects the hemisphere along a semicircle parallel to the z axis, as shown in Figure 3a, we linearly interpolate along the surface of the hemisphere between the two cones nearest α:

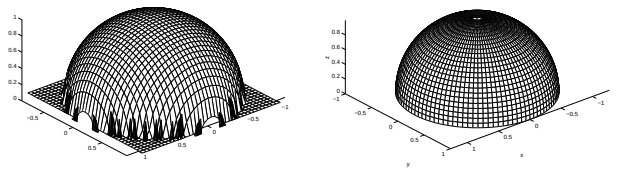
$$h_p(\theta, \phi) = \frac{(\alpha_{k+1} - \alpha)v_k + (\alpha - \alpha_k)v_{k+1}}{\alpha_{k+1} - \alpha_k}, \tag{3}$$

where k is obtained by inverting Eq. (1):

$$k = \left\lfloor \frac{dr}{c}\cos\alpha \right\rfloor.$$

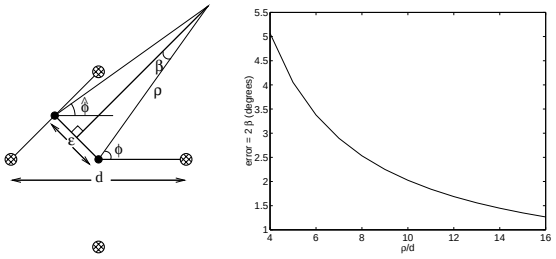
4.2. Non-coincident pairs of microphones

Further redundancy can be achieved by matching not only the two pairs of opposite microphones, but also the four pairs of adjacent ones. Because the midpoints of the adjacent microphone pairs are not coincident with those of the



(a) (b)

Fig. 3. (a) The intersections of the cones of the two opposite pairs with the hemisphere. The black regions near the equator are the blind spots of the cone intersection method, as described in Section 5.1. (b) The division of the unit hemisphere into equally spaced latitudes and longitudes.



(a) (b)

Fig. 4. (a) Top view of four microphones arranged in a square. The sound source is at a distance ρ from the center of the microphone configuration. (b) The maximum error versus ρ/d.

opposite pairs, however, the results cannot be mapped to the unit hemisphere without first estimating the distance $\hat{\rho}$ to the sound source. The point $(\theta', \phi', \hat{\rho})$ in the off-center coordinate system is converted to $(\theta, \phi, \tilde{\rho})$ in the common coordinate system, and $\tilde{\rho}$ is ignored. Then Eqs. (2) and (3) are used to compute $h_p(\theta, \phi)$, with $\gamma = \pm\frac{\pi}{4}$ or $\pm\frac{3\pi}{4}$.

Let us examine the amount of error introduced when the estimated distance $\hat{\rho}$ is different from the true distance ρ. From Figure 4a, we see that, in the worst case ($\hat{\rho} = \infty$), the azimuthal error is bounded by

$$\phi - \hat{\phi} = 2\beta = 2\sin^{-1}\left(\frac{\epsilon}{2\rho}\right) = 2\sin^{-1}\left(\frac{d}{\rho(4\sqrt{2})}\right).$$

Figure 4b plots the error versus the ratio ρ/d. Notice that, if the sound source is at least $4d$ from the array, the error is less than 5.1^0, which is close to the resolution of the localization technique under ideal conditions. With a better distance estimate, the error becomes even smaller. Therefore, the additional robustness from these pairs should outweigh any possible error due to the non-coincidence of the centers.

With coordinate transformations similar to those described above, any microphone pair can be mapped to a common hemisphere. This method can thus accommodate arbitrary microphone array configurations in which the entire array is compact with respect to the distance to the sound source.

4.3. Combining all the pairs

For each pair p, the function h_p is computed at a discrete set of points, spaced at equal latitudes and longitudes around the hemisphere, as shown in Figure 3b. The final result is obtained by summing the sampled functions for the P microphone pairs [5]:

$$h(\theta, \phi) = \sum_{p=1}^{P} h_p(\theta, \phi).$$

The direction to the sound source is then given by the point with the maximum value:

$$(\theta, \phi) = \arg\max_{\theta,\phi} h(\theta, \phi).$$

5. EXPERIMENTAL RESULTS

In this section we compare three algorithms: the intersection of cones (IC) [1], the sampled hemisphere with two opposing microphone pairs (SH2), and the sampled hemisphere with all six pairs (SH6). We used the microphone layout of Figure 1, with $d = 15$ cm and $r = 44.1$ kHz. The hemisphere was divided into 25 latitudes and 100 longitudes, as shown in Figure 3b. For SH6, we used $\hat{\rho} = 1$ m.

5.1. Blind spots

For some locations on the hemisphere, the two cones have no intersection, and IC is therefore unable to compute a solution. These "blind spots," as we call them, in theory cover 8.4% of the hemisphere — using the above values for d and r — and all occur near the equator ($\theta > 75^o$), as shown in Figure 3a. In practice, however, we have found the blind spots to have a much larger influence than the figure indicates, appearing when $45^o < \theta \leq 90^o$.

5.2. Experiment

In the experiment, two individuals conversed candidly in a typcial office, with an array of four cheap omnidirectional microphones in the center of the room on a table. The room was approximately 7 by 3 meters, with brick and sheetrock walls, and background noise consisting mostly of computer fans. The people were located at approximately $(-2.8, 0, 0.5)$ and $(0.6, 1.9, 0.5)$ meters, i.e., $(\theta, \phi) = (76^o, 180^o)$ and $(80^o, 72^o)$, respectively.

The audio segments containing speech by either speaker were manually labeled. These segments were then divided

	SH6	SH2	IC
blind spots	0	0	61.0%
ϕ error	14.5%	80.5%	96.9%
θ error	18.1%	64.8%	84.7%
σ_ϕ	14.4^o	39.7^o	111.2^o
σ_θ	13.7^o	16.1^o	18.4^o

Person 1

	SH6	SH2	IC
blind spots	0	0	44.9%
ϕ error	38.9%	54.5%	77.3%
θ error	81.5%	74.3%	71.3%
σ_ϕ	34.7^o	91.9^o	95.0^o
σ_θ	16.1^o	19.1^o	21.1^o

Person 2

Fig. 5. Comparison of the three algorithms in a real, noisy environment. Blind spots were excluded for the θ and ϕ calculations of the last column.

into non-overlapping 50 ms frames, and the algorithms were run on each frame independently.

Results of the three algorithms are shown in Figure 5. The first row displays the percentage of frames for which IC could compute no answer, due to the null intersection of cones (blind spots). The next two rows show the percentage of frames for which the reading was at least ten degrees[1] away from the measured ground truth, and the final two rows show the standard deviation in the two angles.

We learn several things from these numbers. First, IC can compute no answer for a significant fraction of the frames. Second, in nearly every category SH2 significantly outperforms IC, and SH6 greatly outperforms SH2. The only exception to this is the θ error of Person 2, where IC performs slightly better than the others. Finally, the accuracy of ϕ is in general significantly better than the accuracy in θ. This is perhaps due to additional reflections from the table on which the microphones sit.

6. CONCLUSION

In this paper a method for determining the direction to a sound source using a compact array of microphones has been described. The technique has no blind spots and can handle arbitrary microphone configurations, unlike previous techniques. Moreover, increased robustness to noise was demonstrated on real audio signals in a real environment. This algorithm could be used by itself or as the first stage of a system to determine the three-dimensional coordinates of the source. Further work could include investigating the accuracy of the method in the presence of multiple simultaneous sound sources.

[1] Similar relative results are achieved with other thresholds.

7. REFERENCES

[1] Michael S. Brandstein, John E. Adcock, and Harvey F. Silverman, "A closed-form method for finding source locations from microphone-array time-delay estimates," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1995, vol. 5, pp. 3019–3022.

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

[3] Yiteng Huang, Jacob Benesty, and Gary W. Elko, "Adaptive eigenvalue decomposition algorithm for real-time acoustic source localization system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999, vol. 2, pp. 937–940.

[4] Charles H. Knapp and G. Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[5] Masatoshi Okutomi and Takeo Kanade, "A multiple-baseline stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 4, pp. 353–363, Apr. 1993.

[6] Maurizio Omologo and Piergiorgio Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, 1997.

[7] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Englewood Cliffs, NJ: Prentice Hall, 1995.

[8] 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.

[9] Alex Stéphenne and Benoît Champagne, "Cepstral prefiltering for time delay estimation in reverberant environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1995, vol. 5, pp. 3055–3058.

[10] Piergiorgio Svaizer, Marco Matassoni, and Maurizio 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*, 1997, vol. 1, pp. 231–234.