Sound Source Localization in Reverberant Environments using an Outlier Elimination Algorithm *

Ea-Ee Jan †and James Flanagan

CAIP Center, Rutgers University, Piscataway, New Jersey 08855
e-mail: ejan@watson.ibm.com and jlf@caip.rutgers.edu

I. Abstract

Differences in arrival times of acoustic waves at multiple sensors permit the computation of source location. The computation depends upon delay estimation between sensor pairs. In severe acoustic environments, the estimates are degraded by reverberation and interfering noise, and some estimates are poor, constituting outlier. This report describes a computational method for outlier elimination to improve the accuracy of source location.

II. Introduction

Microphone arrays effectively capture desired sound and mitigate interfering noise and reverberation in multipath environments [2, 3, 4]. Correct steering of these systems requires knowledge of the source location. Also, in teleconferencing applications, source finding systems assist participants in localizing and visualizing the talker. Video cameras can also be "slaved" to the location finder.

A simple means for locating the source is merely to search the space for the direction of maximum energy. A more favored approach for source localization employs a Time Delay Estimator (TDE). Source location can then be estimated from the angles of signals arriving from pairs of sensors, using triangulation procedures [1, 5, 7].

A reverberant environment degrades the time delay estimate. Coincident arrival of reflected signals can sometimes make reflected signals stronger than the direct path signal, leading to a source location estimate which may correspond to an image. An algorithm to eliminate incorrect TDE's is desired to improve the accuracy of source localization. This report describes an outlier detector using a projection method [4]. The outlier technique is applied to a source localization system based on the Cross-Power Spectrum Phase Time Delay Estimator [7]. Performance of the system is evaluated in experimental rooms under three values of reverberation characterized as slight, moderate and high.

* This research is supported by NSF contracts MIP-9121541 and MIP-9314625.

†Dr. Jan is now with IBM Human Language Technologies Group, T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

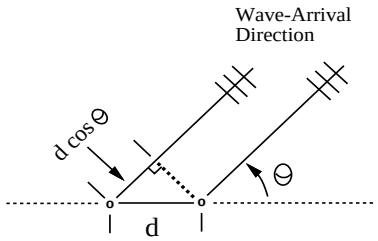

Figure 1: Time delay between a microphone pair with respect to an arriving signal with incident angel of θ.

III. Time delay estimation in a multipath environment

In a multipath environment, the output of a transducer, $x(t)$, associated with the source signal, $s(t)$, can be expressed as:

$$x(t) = \sum_{i=0}^{k} \alpha_i s(t - \tau_i) + n(t) \qquad (1)$$

where k is the number of reflections considered, α_i is the amplitude of the i^{th} reflection, τ_i is the corresponding path delay, and $n(t)$ is ambient noise. The direct path is associated with $i = 0$.

But, in order to calculate τ_0, Eq. (1) cannot be usefully employed without *a priori* information about $s(t)$. Under the favorable condition where the direct path wavefront dominates the reflections, and two transducers are close enough, e.g., 10-20 cm, the output of those transducers, $x_1(t)$ and $x_2(t)$, will be similar except for a small delay equal to the difference in arrival time of the direct path signals . This is expressed as the approximation

$$x_2(t) = x_1(t - \tau) + n(t). \qquad (2)$$

Note that this is also an assumption for the delay-and-sum beamformer. Therefore, ignoring the effects of noise, the delay can be estimated from the inverse Fourier trans-

form of the normalized cross-power spectrum [7]:

$$f(t) = \mathcal{F}^{-1} \frac{X_1(w)\, X_2(w)^*}{|X_1(w)|\ |X_2(w)|}, \qquad (3)$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform, and $*$ denotes complex conjugate. The delay τ is the time index associated with the peak value of $f(t)$. This approach is based on the use of phase difference information only to perform an interchannel time delay estimation. This delay estimator is computationally convenient and appears more resistant to noise and reverberation than approaches based on cross-correlation or adaptive filtering.

The estimated time delay between a pair of microphones defines a hyperboloid surface upon which the source could be located. If a two-dimensional problem is encountered, and if the source is at least 1 or 2 meters from the transducer, a plane wave approximation is sufficient. From Fig. (1), the incident angle of the arriving signal for a given delay τ is $\theta = \cos^{-1} \frac{\tau \cdot c}{d}$, where c is the speed of sound. The line approximately representing the arriving signal passing through the center of a microphone pair can be formulated.

When M pairs of sensors are employed, a linear system with M simultaneous equations is formulated. This over-determined linear system can be simply expressed as $\mathbf{A}\,\mathbf{x} = \mathbf{b}$. The intersection of M lines from M pairs of microphones is the prospective answer. However, due to noise and numerical errors, more than one intersection point is found. For this over-determined system, a Least Mean Square (LMS) error solution is usually applied.

IV. Outlier detector using the projection method

Due to severe multipath distortion, and the non-stationary properties of speech signals, the TDE may give an incorrect answer. In addition, the direct wavefront may be weaker than a coincidence of reflections, also inducing a wrong estimation. An outlier detector is necessary to discard the incorrect TDE's and improve performance in the over-determined system.

The upper left graph of Fig. (2) illustrates an experimental measurement in which 2 out of 6 TDE's are incorrect. Six bearing lines represent the six TDE's. A total of $6 \times 5/2 = 15$ intersections can be found. However, a total of 5 + 4=9 intersection points are generated because of two erroneous lines. (5 from the first line, and 4 from the second one.) From these data, the standard LMS solution yields an unreliable answer.

It is more practical to get rid of the incorrect bearing estimates (line equations) instead of removing the incorrect intersection points. The incorrect bearing line can be detected by projection method described as follows.

The matrix form of the over-determined system, $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, can be reformulated by simultaneous equations. (N equations with M unknowns)

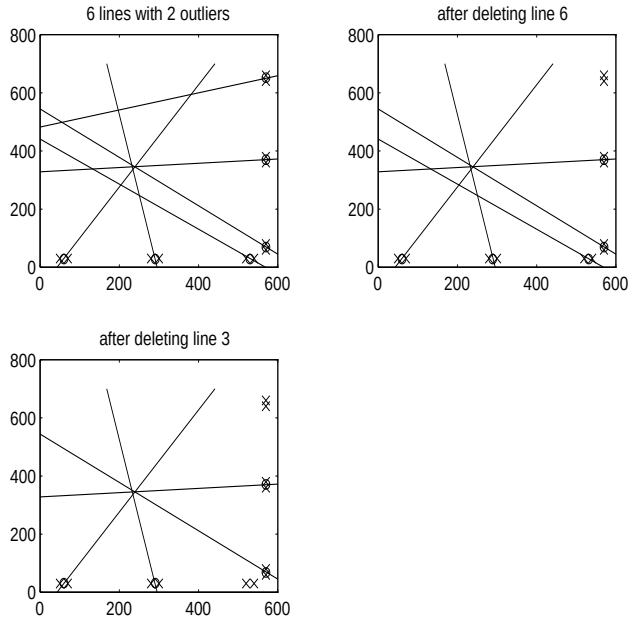

Figure 2: An example of iterations of outlier detection and elimination. The system is composed of 6 lines, two of them incorrect.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1M}x_M &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2M}x_M &= y_2 \\ \vdots & \vdots \\ a_{N1}x_1 + a_{N2}x_2 + \cdots + a_{NM}x_M &= y_N \end{aligned} \qquad (4)$$

where N > M.

Starting from an initial guess $\mathbf{x}^0$, the algorithm iterates over all the equations by projecting the solution on the hyperplanes represented by each individual equation. At step $i+1$, the projected solution is

$$\mathbf{x}^{i+1} = \mathbf{x}^i + \frac{e_i}{|\mathbf{a}_p|^2}\mathbf{a}_p \qquad (5)$$

where $\mathbf{a}_i$ is the i^{th} row of matrix $\mathbf{A}$. At the i^{th} iteration, the p^{th} row is utilized, where $p = i \bmod \mathrm{N}$, thus the equations cycle over. The error at the i^{th} iteration is:

$$e_i = y_p - \mathbf{a}_p \mathbf{x}^i \qquad (6)$$

The iterations are terminated when either the error (L1 or L2 norm of vector $\mathbf{e}$) is within a pre-determined threshold, or the number of iterations is over the pre-set maximum number of iterations. Theoretically, this method asymptotically converges to the Least Mean Square solution [8, 6] without computing the inverse matrix of $(\mathbf{A}^T\mathbf{A})$. This is efficient when the number of simultaneous

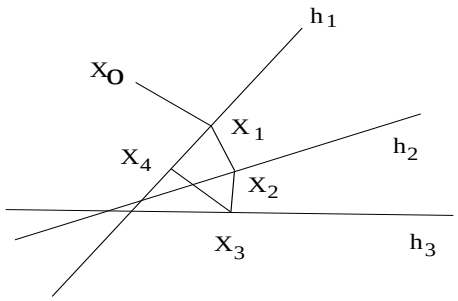

Figure 3: First 4 iterations of the projection method for a 3 equation system.

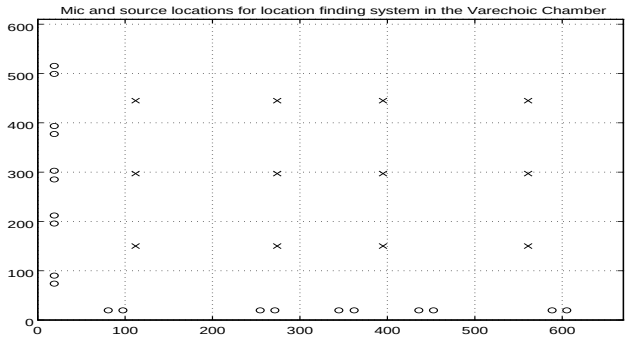

Figure 4: Configuration of source location finding in the Varechoic Chamber. The "o" symbols denote the microphone positions and the "x" symbols denote the loudspeaker positions.

equations is large. Fig. (3) illustrates the first 4 iterations for 3 system equations.

When the system of equations (Equation 4) has been normalized by row vectors, $|\mathbf{a}_i| = 1$, the error measure of Equation 6 is the distance from the projection point $\mathbf{x^i}$ to equation $\mathbf{a}_p = y_p$. After numerous iterations, it is found that the distance from the projection point to each equation converges if none of the equations are too far off. Otherwise, the answer is "stable but not converged". The equation with the maximum distance to the projection point is the possible outlier and will be removed if the error is over the threshold. This algorithm eliminates one equation at a time and is terminated when all the errors are under the threshold. The standard LMS algorithm or equivalent technique is then applied to calculate the final result which is more reliable.

V. Data for real room

Real room data were collected in a digitally controllable variable acoustics facility, the Varechoic Chamber at AT&T Bell Laboratories in Murray Hill, NJ [9]. The walls of the room are constructed of two layered sheets of perforated stainless steel. The chamber is of dimensions 6.7m x 6.1 m x 2.9 m, and consists of 368 independent

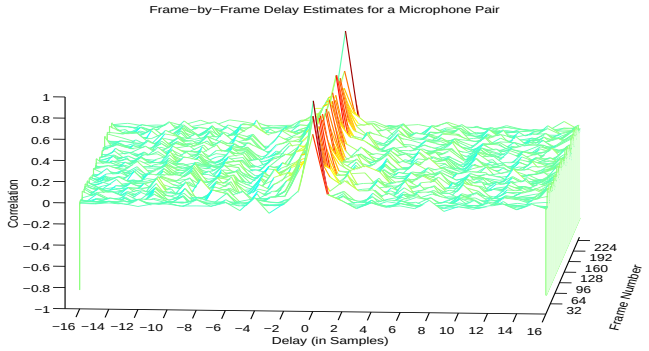

Figure 5: Example of phase correlation between two microphones under slightly reverberant condition. The peak of this function indicates the inter-channel delay.

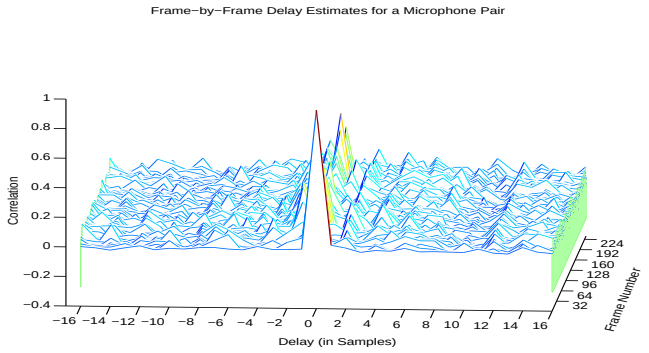

Figure 6: Example of phase correlation between two microphones under highly reverberant condition. The peak of this function indicates the inter-channel delay.

shutters with an area of about $0.38m^2$ each. The holes on each shutter can be controlled in aperture to achieve different acoustic conditions. The reverberation time can be changed between 0.1 and 1.7 seconds by controlling all the shutters from fully open to fully closed.

Speech data were recorded in three acoustic conditions with approximate reverberation times of 0.1, 0.6 and 1.7 seconds. (The shutters were 0%, 50% and 100% closed.) The receivers were two linear arrays on perpendicular walls, each consisting of 10 microphones arranged in 5 pairs. The intra-pair distance was set to 17 cm. A stationary loudspeaker source was positioned at 12 locations in the room. Excitation signals included male speech and pseudo random sequences. The loudspeaker faces toward the microphones on the X-axis. Fig. (4) shows the configuration of the microphone locations and the loudspeaker locations. The data were recorded at a 16 KHz sampling rate.

VI. Results

Fig. (5) and (6) show the TDE results in slightly and highly reverberant conditions with reverberation time of

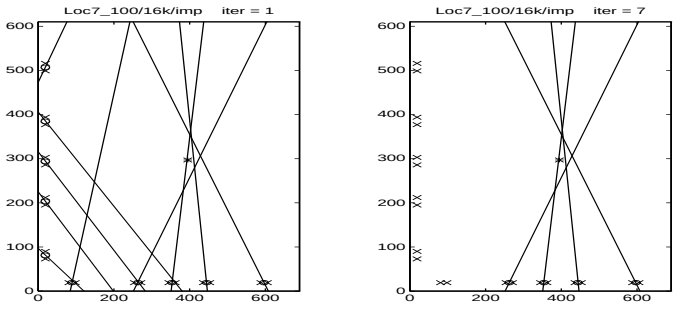

Figure 7: Six delays are removed by the outlier detector. The left and right graphs show before and after outlier detection, respectively.

0.1 sec and 1.7 sec, respectively. Obviously, the TDE is much better under less reverberation. In addition, Fig. (7) demonstrates an example of the line equations before and after outlier elimination in the highly reverberant condition. The loudspeaker is located at center of the room and faces toward the microphones in X-axis. This loudspeaker orientation yields TDE's from the microphone pairs in the Y-axis as somewhat more unstable. The incorrect TDE's are likely due to coincident arrivals of reflected signals. In this case, the TDE estimates a first order image instead of the direct path source.

The source localization algorithm with the outlier detector was evaluated using data under 3 different acoustic conditions. A total of 24 data sets (12 loudspeaker locations, each with a male utterance and a pseudo random sequence as the excitation signals) are included for each condition. The means of the normalized LMS error with and without the outlier detector are 4.0 and 18.2 cm, respectively, without *a priori* information of source location. The corresponding standard deviations are 1.3 and 13.6 cm, respectively. The normalized LMS error is defined by the summation of distance measure from the estimated solution to all equations divided by the number of equations. The outlier detector is able to remove those inconsistent TDE's and improve the system accuracy and confidence.

Using *a priori* information about the source location, the true error measure can be defined as the distance between the source (loudspeaker) and the estimated source location. Table 1 compares errors of 2-D source localization with and without an outlier detector under slightly and moderately reverberant conditions. Also, a total of 24 data sets are evaluated for each acoustic condition. The outlier eliminator improves the source location performance under moderate reverberation conditions.

In highly reverberant environments, some of the data are critically degraded, e.g. more than 6 out of 10 TDE's are incorrect for a given data set implying the data are completely unreliable. More robust TDE and outlier eliminator algorithms are required to resolve this disastrous situation.

Reverberation time (s)	W/O outlier detector (m)	W/ outlier detector (m)
0.1	0.39	0.30
0.6	0.50	0.35

Table 1: Comparison of mean source location error (m) for the system with and without the outlier detector.

VII. Conclusion

The microphone array technique which combines the Time Delay Estimator and outlier detector is found to locate the acoustic source in adverse acoustic environments. Research continues on refining the techniques to provide more consistent estimation in severely reverberant environments.

References

[1] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array", *Computer Speech and Language*, 1995, Vol. 9, pp. 153-169.

[2] J. L. Flanagan, D. A. Berkley, G. W. Elko, and M. M. Sondhi, "Autodirective microphone systems", *Acustica*, 73:58–71, 1991.

[3] J. L. Flanagan, A. C. Surendran and E. E. Jan "Spatially selective sound capture for speech and audio processing", *Speech Communication 13*, 1993, pp. 207-222.

[4] E. E. Jan, "Parallel processing of large scale microphone arrays for sound capture", Ph.D dissertation, Dept. of Electrical Engr. Rutgers University, NJ, May 1995.

[5] E. E. Jan, P. Svaizer, and J. L. Flanagan, "A database for microphone array experimentation", *Eurospeech '95*, Sep. 1995, Madrid, Spain.

[6] R. J. Mammone. *Computational Methods of Signal Recovery and Recognition*, chapter 2, pages 39–42. John Wiley & Sons, Inc., 1992.

[7] M. Omologo, P. Svaizer "Acoustic Event Localization using a Crosspower-Spectrum Phase based Technique", *Proc. ICASSP*, Adelaide 1994, pp. II273-II276.

[8] K. Tanabe. Projection method for solving a singular system of linear equations and its applications. *Numer. Math.*, 17:203–214, 1971.

[9] W. C. Ward, G. W. Elko, R. A. Kubi, and W. C. McDougald "The new Varechoic Chamber at AT&T Bell Labs", *Proceedings of the Wallace Clement Sabine Centennial Symposium*, Acoustic Society of America, pp. 343-346, 1994.