AN IMPROVED TDOA-BASED LOCATION ESTIMATION ALGORITHM FOR LARGE APERTURE MICROPHONE ARRAYS

Ying Yu and Harvey F. Silverman

LEMS, Division of Engineering, Box D, Brown University, Providence, Rhode Island 02912

ABSTRACT

Location estimation is a vital aspect of a real-time large-aperture array system. To date, only time-difference-of-arrival(TDOA)-based algorithms run sufficiently fast for a real-time system, so we have been using this kind of locator, LEMSalg, in our real-time environment. In this paper we present an improved location estimation algorithm. In the new method, microphone-pair selection is made dynamic, rather than fixed *a priori*. Dynamic selection requires derivation of parameters and criteria. Here we use theoretical analysis, real system statistics and experimental results to obtain: 1) a suitable range for microphone-pair separation, 2) an ideal length for the TDOA vector, and 3) an effective TDOA quality assessment criterion. Experimental results based on real speech recorded in a moderately noisy environment at 5 different SNR levels are given to show the performance improvements.

1. INTRODUCTION

A large array of microphones (HMA) is being studied as a possible means of acquiring data in offices, conference rooms, and auditoria without requiring close-talking microphones. An array that surrounds all possible sources has a *large aperture* and such arrays have attractive properties for accurate spatial resolution and significant signal-to-noise enhancement [1]. Our recent paper [2] presented all the details of "LEMSalg", a real-time, source-location estimation algorithm based on TDOAs derived from a phase transform (PHAT) [3, 4] generalized cross-correlation (GCC) [5] (see Figure 1). LEMSalg is based on finding the TDOAs between 16 pairs of microphones using the PHAT weighting of the GCC and then searching in the three-dimensional room space using the simplex method [6] for the point-source location that had the maximum likelihood of having produced the measured TDOA vector [2]. In all cases, a subset of microphones(24 in LEMSalg) was selected for location estimation from four adjacent panels on two orthogonal walls (128 microphones). Orthogonal panels are used to get higher resolution in both directions parallel to each of the walls. We have found that success is strongly dependent upon the details of the implementation.

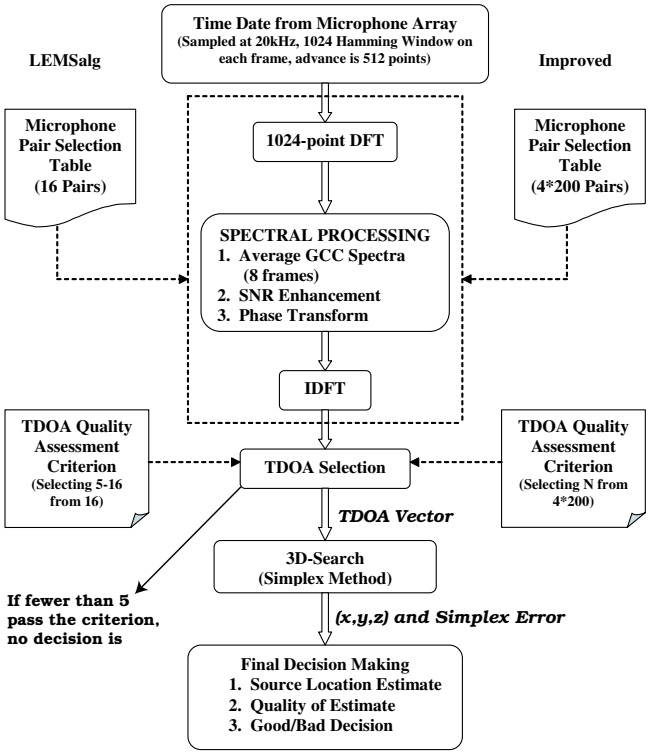

Fig. 1. Diagram of LEMS Location Estimation system

In [2], LEMSalg was compared to the SRP-PHAT algorithm [7], and an augmented LEMSalg [2], in which each microphone signal was replaced by the output of a corresponding four-microphone local beamformer. SRP-PHAT proved to be very accurate under all SNR conditions, but was extremely costly to implement using a restricted grid search, as gradient search methods often did not converge. Augmented LEMSalg was computationally feasible, but required *a priori* knowledge of the source location. In addition to the circularity of the algorithm, this can be a problem for moving talkers, new talkers, or after a long quiet period.

LEMSalg is being used successfully in an acoustically-

harsh moderate-sized room environment where microphone SNRs are below 0dB. We shall use LEMSalg as a baseline. In this paper, we introduce the idea of selecting microphone pairs dynamically (LEMSalg has 16 microphone pairs fixed *a priori*). Our method of dynamic selection is well-suited for using a fixed-length TDOA vector as input to the simplex search function[6, 2]. (LEMSalg may reduce the size of the TDOA vector from 16 to as few as 5, based on a "quality of TDOA" criterion [2]). Microphone-pair selection and TDOA quality assessment criteria are the two defining aspects of of this work that impact the performance of the location-estimation algorithm.

The improvements to the algorithm all result from changing the method for selecting the microphone pairs. As shown in Figure 1, in LEMSalg, the selection of the 16 pairs of microphones (thus 16 GCCs) is done *a priori* by hand, and ultimately from 5 to 16 TDOAs that have met some TDOA quality assessment criteria (if fewer then 5 per frame have met the criteria, it is considered a "no estimate made" frame) will comprise a TDOA vector to be used by the search algorithm. Thus the TDOA vector length varies and we use only a small portion of all the microphones available. In the improved algorithm, we use all the microphones in the array and dynamically select microphone pairs that have a separation distance within an "optimal" range. We also choose some fixed number, N, of TDOAs given by the most reliable correlations according to a TDOA quality assessment criterion.

2. PARAMETERS AND CRITERIA FOR DYNAMIC MICROPHONE-PAIR SELECTION

Given a single source with noise in a real room, if two microphones are close to each other, there should be a high level of cross-correlation between their signals. Thus there is a higher probability of getting a "correct" TDOA. However, closely-spaced microphones have very small TDOA's even for sources placed nearly end-fire. This implies quantization error in TDOA can cause large errors in a spatial position estimate. On the contrary, if a pair of microphones is widely separated, their signals from the source would be more poorly correlated, which implies a smaller probability of getting a "correct" TDOA, albeit the error due to quantization would be much smaller.

Let us look at the error due to quantization as a function of microphone separation distance. Refer first to Figure 2, a 2D illustration, in which S denotes the source position, A and B denote the positions of the pair of microphones, and θ is the angle of the source off the normal to the microphone pair. Suppose we get a "correct" TDOA from the pair of microphones. The TDOA determines one branch of a hyperbola given by

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \qquad (1)$$

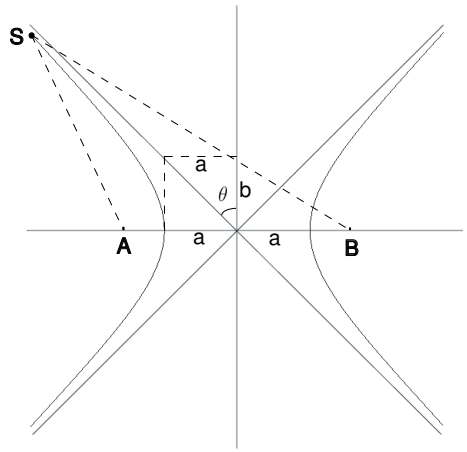

Fig. 2. 2D Illustration of a Microphone Pair TDOA

in the space. The foci of the hyperbola are the positions of A and B, and $a^2 + b^2 = (d/2)^2$, where d is the distance between the two microphones ($|AB|$), and $2a = TDOA \times c$ (speed of sound), which is the distance difference of arrival. Notice that the asymptote is the straight line with angle $\theta = arcsin(\frac{a}{d/2})$. In most conditions, $|SA|$ and $|SB|$ are much bigger than $|AB|$. That is to say that the source is more than 2 or 3 $|AB|$. In these cases, only θ, the angle to the source, is significant. Quantization produces error in estimating θ. Here the sampling rate is $20kHz$. When the GCC is a good correlation, the error of the GCC peak position (which gives the TDOA in samples) will uniformly distribute in the range over a sample. This means that the standard deviation error of the TDOA will be $E_{time} = \frac{50\mu sec}{\sqrt{12}} = 14.4\mu sec.$ The corresponding standard deviation in distance difference of arrival ($2a$), is $E_{dist} = E_{time} \times c$. Then the error in angle θ due to quantization may be considered:

$$E_\theta = |arcsin(\frac{a + \frac{1}{2} \times E_{dist}}{d/2}) - arcsin(\frac{a}{d/2})| \qquad (2)$$

which simplifies to:

$$E_\theta = |arcsin(sin(\theta) + \frac{E_{dist}}{d}) - \theta| \qquad (3)$$

Figure 3 shows E_θ versus the microphone separation distance parameterized by θ, for 20kHz sampling.

The relationship between the microphone separation distance and the correlation quality for some source is more difficult to model and quantify theoretically. If two microphones are close to each other, they have similar background, propagation channel and orientation angle from the source. Thus they are more likely to produce a good correlation, in which the highest peak position is the correct TDOA. On the contrary, two microphones that are far from

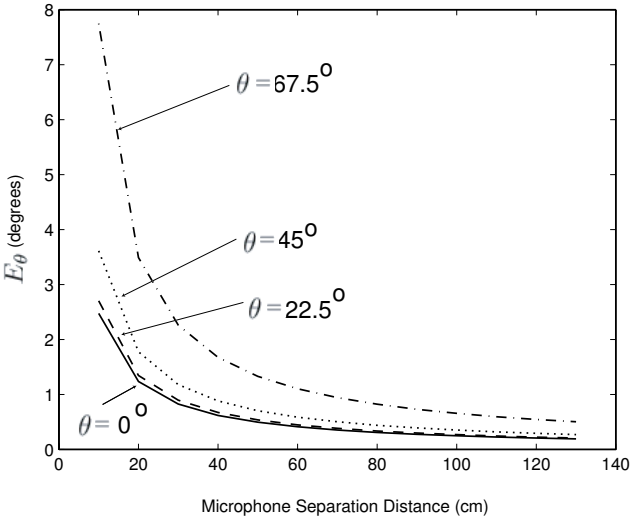

Fig. 3. E_θ vs. Microphone Separation Distance

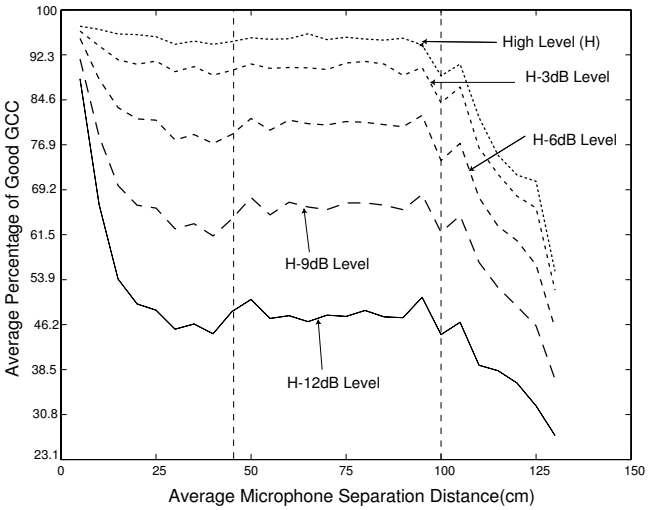

Fig. 4. Average Percentage of Good Correlation vs. Average Microphone Separation Distance.

each other are less likely to produce a good correlation. Using speech data from a transducer at a known source position we can determine the percentage of good TDOA prediction by the highest peak as a function of the microphone separation distance, with 5 different SNR-levels of speech (Figure 4).

Let us now combine the two findings. First, consider a nominal situation for a microphone pair, in which the average distance from the source to the microphone pair is about $2m$, and a near end-fire angle θ is 67.5^o. Furthermore, we assume the error we allow in estimating the source location is $5cm$, which is an angle of 1.43^o. According to the curve labelled $\theta = 67.5^o$ in Figure 3, microphone separation distance needs to be larger than $40cm$ to ensure this. In Figure 4, we can see that for all 5 different SNR-levels, the general trend is: the closer the separation distance, the better the correlation quality. $40cm$ to $100cm$ is a steady period, and after $100cm$ performance falls off sharply. Combining the two, we conclude that the suitable microphone separation distance should be between $40cm$ and $100cm$.

A second important issue is the assessment of the quality of a TDOA estimate prior to using it in the simplex search. We investigated three different methods as measures of TDOA quality:

1) Highest-peak criterion: we use the N TDOAs from the GCCs with the highest main peaks;

2) Highest-ratio criterion: we use the N TDOAs from the GCCs with the highest main peak/secondary peak ratios;

3) Combined criterion: we first screen the GCCs by the criterion "main peak/secondary peak ratio ≥ 1.4", then pick the N TDOAs from the remaining GCCs with highest main peaks.

Finally, in the improved system, it is simpler to fix the length N of the TDOA vector. Thus, we need to ascertain an ideal value of N, i.e., the value of N that gives the best, *predictable* performance. Predictable performance means the percentage of correct location estimates that 1) a threshold on residual error of the simplex search predicts and 2) in fact **are** correct location estimates.

We have observed that the simplex search algorithm finds a source position with acceptable error if and only if ***all*** TDOA's in the vector are "correct" for the true source. This implies that, given a particular TDOA quality assessment criterion, the smaller N ($N > 3$) is, the more likely the simplex algorithm will produce a result with acceptable error. However, if N is too small, typical TDOA errors will have significant impact, potentially causing an error in the source position estimate sufficient to make it unacceptable.

We used speech data to plot the percentage of frames having all TDOAs in their TDOA vector correct as a function of five SNR-levels, using the highest-ratio criterion. In Figure 5 four different values of N are shown. The statistics for LEMSalg are also shown. The figure validates that smaller values of N are indeed better.

Table 1 lists the predictability of the results for different TDOA lengths N. Clearly larger values of N are better. We conclude that $N = 8$ is the best compromise, considering both factors, with very high predictability and the second highest percentage of frames with all TDOAs correct.

3. EXPERIMENT AND RESULTS

In all cases the testing environment used was very difficult, having individual microphone signal-to-reverberation energy in the range [-2dB, -12dB]. To simplify comparisons

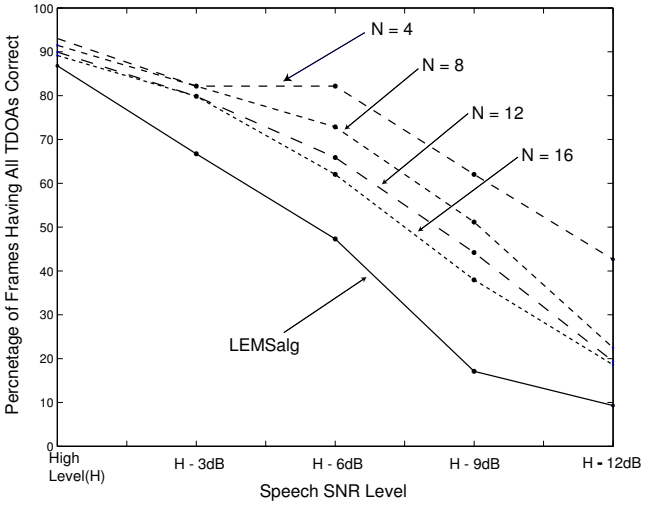

Fig. 5. TDOA Vector Length N vs. Percentage of Frames with All N TDOAs Correct

Level(dB)	LEMSalg	Improved Algorithm with Various N Value N=4	N=8	N=12	N=16
(H)igh	94.6 %	75.8 %	96.8 %	100 %	100 %
H-3	90.7%	69.8%	98.1%	100 %	100%
H-6	90.2%	72.6%	100%	100 %	100%
H-9	100%	71.3%	98.5%	100 %	100%
H-12	100%	50.9%	100%	100 %	100%

Table 1. Percentage of Predicted Correct Location Results that are Actually Correct. (Correct means within $10cm$ of the true source position.)

between algorithms that cannot all execute in real time, we recorded data from the HMA array and did all performance measurements with off-line computation. The data was generated using a recording of a male, native speaker of American English played back at different amplitudes through a domed tweeter in a known location. For details of the testing environment and the procedure, please refer to [2]. Table 2 shows the performance of the improved algorithm ($N = 8$) using the three different TDOA quality assessment criteria for 5 different SNR-levels of speech.

4. CONCLUSION

The improved algorithm gives significantly better performance than LEMSalg for all three TDOA quality assessment criteria, especially at mid-level SNRs. Typically, our measured talkers' source level is between $H - 3dB$ and $H - 9dB$. The highest-ratio criterion has the best performance for these levels. (See Table 2.) Relative to LEMSalg, the improvement is 23.2% for $H - 3dB$, 54.1% for $H - 6dB$, and 199.4% for $H - 9dB$.

The level of performance improvement shown in Table 2 is very important for a real system that has talkers at mid-

Level(dB)	Percentage of Frames with Good Results: LEMSalg	LEMS Improved: Highest Peak	Highest Ratio	Combined
(H)igh	86.8%	93.0%	91.5%	92.3%
H-3	66.7%	84.5%	82.2%	86.8 %
H-6	47.3%	58.1%	72.9%	61.2 %
H-9	17.1%	41.1%	51.2%	41.9 %
H-12	9.3%	18.6%	22.5%	20.9%
Level(dB)	**RMS Deviation of Estimates from Known Source Location (cm): LEMSalg**	**LEMS Improved: Highest Peak**	**Highest Ratio**	**Combined**
(H)igh	2.79	3.55	3.90	3.39
H-3	3.11	3.52	3.26	3.51
H-6	3.59	4.09	3.65	4.14
H-9	4.15	4.46	4.18	4.62
H-12	2.57	2.67	3.10	3.09

Table 2. Comparison of LEMS Original Algorithm and LEMS Improved Algorithm as a Function of SNR.

SNR levels. Even with LEMSalg, the real-time system has adequate performance for multiple talkers in a noisy room [1]. However, the computational cost of improvements is high. We estimate the algorithms introduced here require about 10 times the computation of LEMSalg. The more accurate SRP-PHAT algorithm requires an order of magnitude more than this to converge. The augmented LEMSalg, while better performing than the new algorithm and requiring about one-sixth of the computation, still requires knowledge of the source location *a priori*. We believe there are some good ways to reduce the computational cost of the new algorithm and future research will focus on how to do this, as well as developing a more reliable criterion for assessing TDOA quality.

5. REFERENCES

[1] H. F. Silverman, W. R. Patterson III, and J. M. Sachar, "Factors affecting the performance of large-aperture microphone arrays," *Journal of the Acoustical Society of America*, vol. 111, Pt. 1, no. 5, pp. 2140–2157, May 2002.

[2] H. F. Silverman, Ying Yu, J. M. Sachar, and W. R. Patterson III, "Performance of real-time source-location estimators for a large-aperture microphone array," *(In Submission) IEEE Tran. Acoustic, Speech, and Signal Processing and www.lems.brown.edu*, August 2003.

[3] M.Omologo and P. Svaizer, "Acoustic event localization using a crosspower- spectrum phase based technique," in *Proceedings of ICASSP-1994*, Adelaide, Australia, April 1994, pp. II–273 – II–276.

[4] T.Guestafsson, B.D.Rao, and M.Trivedi, "Analysis of time-delay estimation in reverberant environments," in *ICASSP2002*, Dept.of Electr.and Comput.Eng., California Univ. San Diego, May 2002, IEEE, pp. II2097–II2100.

[5] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, August 1976.

[6] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical recipes in C*, Cambridge University Press, Cambridge, UK, 1988.

[7] M. Brandstein and D. Ward (Eds), *Microphone Arrays*, Springer-Verlag, Berlin, 2001.