An Algorithm for Determining Talker Location using a Linear Microphone Array and Optimal Hyperbolic Fit

Harvey F. Silverman
Laboratory for Engineering Man/Machine Systems (LEMS)
Division of Engineering
Brown University
Providence, RI 02912

Abstract

One of the problems for all speech input is the necessity for the talker to be encumbered by a head-mounted, hand-held, or fixed position microphone. An intelligent, electronically-aimed unidirectional microphone would overcome this problem. Array techniques hold the best promise to bring such a system to practicality. The development of a robust algorithm to determine the location of a talker is a fundamental issue for a microphone-array system. Here, a two-step talker-location algorithm is introduced. Step 1 is a rather conventional filtered cross-correlation method; the cross-correlation between some pair of microphones is determined to high accuracy using a somewhat novel, fast interpolation on the sampled data. Then, using the fact that the delays for a point source should fit a hyperbola, a best hyperbolic fit is obtained using nonlinear optimization. A method which fits the hyperbola directly to peak-picked delays is shown to be far less robust than an algorithm which fits the hyperbola in the cross-correlation space. An efficient, global nonlinear optimization technique, Stochastic region Contraction (SRC) is shown to yield highly accurate (>90%), and computationally efficient, results for a normal ambient.

Introduction

One of the problems for all speech input is the necessity for the talker to be encumbered by a head-mounted, hand-held, or fixed position microphone, or, perhaps, a technician-controlled mechanical unidirectional microphone. Whether for teleconferencing [1], speech recognition [2], or large-room recording or conferencing [3], an intelligent, electronically-aimed unidirectional microphone would overcome this problem. Array techniques hold the best promise to bring such a system to practicality.

Algorithms for passive tracking -- the determination of range, bearing, speed, and signature as a function of time for a moving object -- have been studied for nearly 100 years particularly for radar and sonar systems. While there is currently much activity involved with the tracking of multiple sources using variants of the eigenvalue-based decomposition MUSIC algorithm, [4], [5], [6], [7], [8], most systems still use correlational techniques [9], [10], [11].

The method presented here is also based on correlation. First, a coarse, normalized cross-correlation function is computed over the delay range of interest. It turns out that, even for the relatively high sampling rate of 20kHz, the 50μs resolution of the time-delay estimates causes derived locations to be unsatisfactory. However, the latter may be refined by nearly two orders of magnitude through accurate interpolation techniques which can be attained for a relatively small computational using multirate filtering[12].

For M microphones, one can estimate $M-1$ independent relative delays. As, theoretically, only two relative delays are needed to triangulate a source, for $M>3$, the system is overspecified. However, since noise is always present in a real system, this extra information can be profitably used to overcome some of the effects of the noise. In fact, the geometry of the array constrains the vector of relative delays. For example, a simple linear array, with all the microphones on the axis, $y=0$, has delays constrained to be on a particular hyperbola with a focus on the target. Therefore, errors in the estimation of the delays may be corrected by fitting the *best* hyperbola. Two methods for doing so are presented here.

In the first method, **Time-Delay Estimation, Hyperbolic Fit (TDEHF)**, peak-picking is used on the results of the interpolated cross-correlations to estimate the individual time delays. Then, constrained nonlinear optimization is used to fit the best hyperbola through the sparse time-delay estimations. As the data turn out to be pretty much unimodal, gradient techniques [13] were used to minimize a least-squares functional. TDEHF suffers when original time-delay estimates exhibit large, and

often "dumb" errors. TDEHF is introduced in Section 4.

The second (and more robust) method **Interpolated Cross-correlation Hyperbolic Fit (ICHF)**, fits the best hyperbola to the actual output of the interpolated cross-correlations. As reasonable cross-correlations are always positive, the sum of the cross-correlations across all the microphones for a given hyperbola is used as a functional to maximize. As the functional surface is multimodal, results for a hierarchical grid search and for application of Stochastic Region Contraction (SRC), [14] , [15], a new method for efficient global nonlinear optimization, are presented.

Coarse Cross-Correlation

Consider a linear microphone array having M microphones, each located on the line $y=0$ at a distinct point $(z_m,0)$ in the x,y plane. A simple case is to be considered in this paper in which a single source (talker) is located at some point (x,y) in front of the array, although there will be ambient noise. Without loss of generality, microphone 1 is selected as the reference. It is assumed that the signal at each microphone is appropriately sampled at some reasonable rate, R and that each microphone thus receives a signal of time (indexed by j), $\rho_m^R(j)$. As sources might be separable in the frequency domain, one can, in general, filter each received signal using a zero-phase FIR filter; this is the only reasonable choice as delay estimation is yet to be performed. This implies,

$$\hat{r}_m^R(j) \equiv \sum_{j'=-J}^{J} f_m(j') \cdot \rho_m^R(j-j'), \qquad (2.1)$$

where $f_m(j)$ is a $2J+1$ element symmetric FIR filter. It is advantageous, as will be seen later, to define rectangularly-windowed data, referenced to time index k', for the correlations as,

$$r_m^R(k'+l) \equiv \begin{cases} \hat{r}_m^R(k'+l) & 0 \le l \le L-1 \\ 0 & otherwise \end{cases} \qquad (2.2)$$

Each of the $M-1$ independent cross-correlations for a delay of k samples each of duration $1/R$ may be defined,

$$C_m^R[k,k'] \equiv \frac{A_m(k')}{L-|k|} \cdot \sum_{l=0}^{L-1} r_1^R(k'+l) \cdot r_m^R(k'+l+k), \qquad (2.3)$$

where $A_m(k')$ is a normalizing factor. A reasonable normalization is to make the autocorrelation of the unshifted reference signal have a value of unity for any particular time reference k',

$$C_1^R[0,k'] \equiv 1.0. \qquad (2.4)$$

Combining (2.3) and (2.4) gives,

$$A_1(k') = \frac{L}{\sum_{l=0}^{L-1} (r_1^R(k'+l))^2}, \qquad (2.5)$$

which generalizes to,

$$A_m(k') = \frac{L}{\left[\sum_{l=0}^{L-1} (r_1^R(k'+l))^2\right]^{1/2} \cdot \left[\sum_{l=0}^{L-1} (r_m^R(k'+l))^2\right]^{1/2}}. \qquad (2.6)$$

Computational Considerations for the Cross-Correlations

An important consideration is the selection of L, the number of points in the cross-correlation. When autocorrelations are taken for LPC analysis, the length is limited by the assumption that the vocal tract is essentially stationary over the interval. As one is not doing this pseudo-stationary modeling of the vocal tract, this fact does not limit L here. Rather, the tradeoff between information content -- tending to make one increase L -- and computational load -- tending to make one decrease L -- governs this decision. For the typical human talker, computing a position about five times per second is sufficient. With no redundancy, selecting L to correspond to 100-200ms of data is reasonable, as the experimental data show.

The range of the correlations, $[-K_-, K_+]$, may be determined from the sample rate and the geometry shown in Figure 1 for a one-dimensional array. For a symmetric arrangement in a room, $K_- = K_+$ and

$$K_- = K_+ = \left[Length \cdot \cos(\theta) \cdot \frac{R}{c}\right], \qquad (2.7)$$

where c is the speed of sound with value about 342M/s.

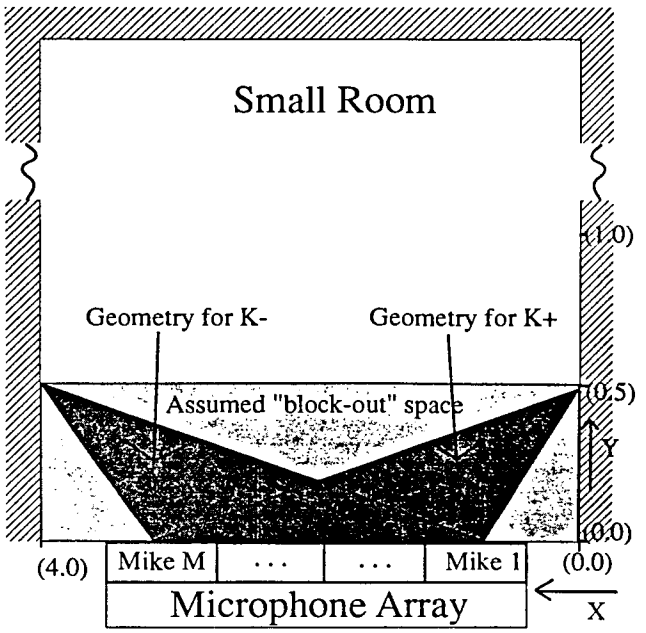

Figure 1: Geometry for Computing K_+, K_-, the Worst-Case delays

As an example, consider a one-dimensional array of length one meter, a room four meters wide, one-half meter of "block-out space" and a sampling rate of 20,000 samples-per-second. For this case, correlations will require 2000 multiplication-addition operations for 100msec of data. As the maximum relative delay may be seen to be $\frac{1 \cdot \cos 14^\circ}{c} = 2.84ms$, Equation (2.7) yields $K_- = K_+ = 57$. Thus, the correlation phase requires

230,000 multiplication-additions per microphone pair if done directly or just under 20ms of computation time using the Analog Devices ADSP-2100A digital signal processor at 12.5MHz clock rate [16]. For eight microphones, about 160 ms would be required, and the location could be computed in real-time for the required five updates per second.

The relative delay between each microphone and its reference could be estimated by selecting the highest positive point in the correlation outputs. i.e.,

$$k_m^* \equiv \underset{-K_- \le l \le K_+}{argmin} \; C_m^R[k,k'], \tag{2.8}$$

$$d_m^R[k'] \equiv \frac{k_m^*}{R}, \tag{2.9}$$

where $d_m^R[k']$ is defined to be the delay, relative to microphone 1, for microphone m. Note that the accuracy is only to that of the sample rate, and that this simple peak-picking algorithm is subject to serious errors when real data are used!

Interpolation for Higher Accuracy

Even for the relatively high (for speech) sampling rate of 20kHz, estimation accuracy of the tracking position is inadequate; a variation of more than one meter in the y dimension is the norm for talkers two meters directly in front of the microphone. Experience has shown that an acceptable region of uncertainty may be achieved for a sampling inteerval of about $1\mu s$.

The most straightforward way to achieve the needed high resolution would be to sample at a much higher rate, R' -- around 1MHz -- and perform the correlations on the data, i.e.,

$$C_m^{R'}[k,k'] = \frac{B_m(k')}{L^{R'}-|k|} \cdot \sum_{l=0}^{L^{R'}-1} r_1^{R'}(k'+l) \cdot r_m^{R'}(k'+k+l) \tag{3.1}$$

where $B_m(k')$ is a normalizing factor and $L^{R'}$ is the number of high-resolution samples in L. Relative to 20kHz sampling, this would force the computation to increase by a factor of $50^2 = 2500$, making the procedure absurd. For an appropriately anti-aliased speech signal, one would be dealing with greatly oversampled signals. Thus, with no loss in accuracy, one could generate the signal at sampling rate R' from the signal sampled at rate R by the simplest standard multirate method if

$$R' \equiv \lambda \cdot R, \tag{3.2}$$

where λ is an integer greater than 1.

The proof for computationally efficient interpolation is given in [17]. The results for computation are:

$$C_m^{R'}[\lambda\sigma_k+\nu_k, k'] = \frac{B_m(k')}{L^{R'} - |\lambda\sigma_k+\nu_k|} \cdot \sum_{\sigma_1=-Q^R}^{Q^R} \sum_{\sigma_2=-Q^R}^{Q^R} \tag{3.3}$$

$$\Phi[\sigma_1,\sigma_2,\nu_k] \cdot \bar{C}_m^R[k+\sigma_1-\sigma_2,k']$$

$$B_m(k') = \frac{L^{R'}}{\sum_{\sigma_1=-Q^R}^{Q^R} \sum_{\sigma_2=-Q^R}^{Q^R} \Phi[\sigma_1,\sigma_2,0] \cdot \bar{C}_m^R[\sigma_1-\sigma_2,k']} \tag{3.4}$$

$$\Phi[\sigma_1,\sigma_2,\nu_k] \equiv \sum_{\nu_1=0}^{\lambda-1} \left[f(\lambda\sigma_1+\nu_1) \cdot f(\lambda\sigma_2+\nu_k+\nu_1) \right] \tag{3.5}$$

$$\bar{C}_m^R[k+\sigma_1-\sigma_2,k'] \equiv \frac{L-|k+\sigma_1-\sigma_2|}{A_m(k')} \cdot \tag{3.6}$$

$$C_m^R[k+\sigma_1-\sigma_2,k']$$

Computational Considerations for the Interpolation

One important aspect of the computation of Equation (3.3) is the storage requirement for Φ. Appropriate resolution is achieved for λ=64, R=20kHz and a filter length of 641, implying $Q^R = 5$. Then the range of σ_1 and σ_2 is only 11. Thus (11)(11)(64) = 7744 storage locations are required.

The number of multiplication-additions is $(11)^2 = 121$ to compute the cross-correlation for each interpolated point. One should note that this number is a far cry from the "direct" method in which , for $L = 2000$, (621)(64)(2000) ≈ 80,000,000 operations had to be done to get each interpolated signal and (64)(2000) = 128,000 operations had to be done for each interpolated cross-correlation!

Best Hyperbolic Fit Algorithms

Triangulation

In binaural hearing, both amplitude and phase information is fed to the brain and is used -- expertly -- to determine the location of a sound source. If the phase information -- the delay estimates -- alone were to be used to determine location of a source, a minimum of three microphones is required for this "triangulation" procedure. If microphone 1 is considered to be the reference, and d_2 and d_3 the time delays for microphones 2 and 3 respectively, relative to the arrival at microphone 1, then the estimation of the source location x_0, y_0 may be determined from,

$$x_0 = \frac{c^2 d_2^2 (d_2 - d_3) - d_2(z_3^2 - z_1^2) + d_3(z_2^2 - z_1^2)}{2[d_2(z_1 - z_3) - d_3(z_1 - z_2)]} \tag{4.1}$$

$$y_0 = \left[\left[\frac{(x_0-z_2)^2 - (x_0-z_1)^2 - d_2^2 c^2}{2 d_2 c} \right]^2 - (x_0-z_1)^2 \right]^{1/2}. \tag{4.2}$$

(One should note that these triangulation formulae are normally listed for polar coordinates.) These relatively ugly, nonlinear expressions tend to be very sensitive to variations due to noise in the estimates of d_2 and d_3.

Time-Delay Estimation, Hyperbolic Fit (TDEHF)

For the case of the linear array, where the microphones are all considered to be on $y=0$, the locus of the relative delays for points along this line forms a hyperbola. This is clear from Figure 2 in which the relative delay loci are plotted for various point-source locations (x,y). At $(z_m,0)$, the absolute delay d_m may be computed from the Pythagorean Theorem as

$$d_m \equiv \frac{\sqrt{(x-z_m)^2+y^2}}{c}, \tag{4.3}$$

and, relative to microphone 1,

$$\hat{d}_m = \frac{\sqrt{(x-z_m)^2+y^2}}{c} - d_1. \tag{4.4}$$

Some algebra yields,

$$(\hat{d}_m+d_1)^2 - \frac{(z_m-x)^2}{c^2} = \frac{y^2}{c^2}. \tag{4.5}$$

The points $(z_m,\hat{d}_m)$ lie on a hyperbola parameterized by the speed of sound, c, and the location of the source, (x,y). Thus, there is a one-to-one relationship between a specific hyperbola and a source-point (x,y) located in front of the array -- there is a mirror in back of the array. The task, then, is to fit the best member of this class, the best hyperbola, to the set of relative delay estimates $z_m, d_m^R[k']$, where $m \in [2,M]$.

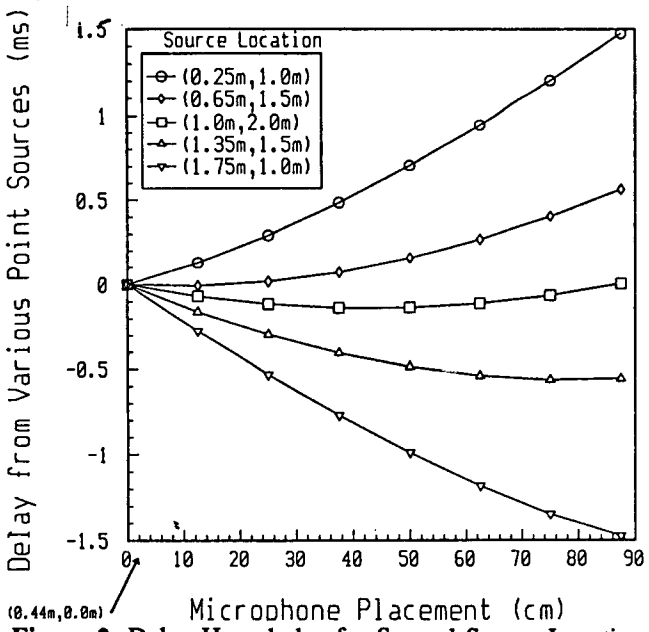

Figure 2: Delay Hyperbolae for Several Source Locations

In TDEHF an estimate of the relative delay for each microphone is obtained by peak-picking as indicated by Equations (2.10) and (2.11). Interpolation is done locally to get a higher resolution estimate, $d_m^R(k')$. While many criteria are possible, a typical squared-error measure is defined as

$$E(k') = \sum_{m=2}^{M} (d_m^R(k') - \hat{d}_m))^2 \tag{4.6}$$

Substituting (4.4) into (4.6), one gets,

$$E(k') = \sum_{m=2}^{M} \left[d_m^R(k') - \frac{\sqrt{(x-z_m)^2+y^2}}{c} - d_1 \right]^2, \tag{4.7}$$

and the esrtimate (x_0,y_0) minimizes $E(k')$. As this surface is normally unimodal, a gradient method [18] has been used.

Interpolated Cross-correlation Hyperbolic Fit (ICHF)

When real data are used, it is often the case that the cross-correlation peak which must be determined in TDEHF is inappropriate. This is due to 1) periodicity in the signal, 2) room reverberations, and 3) noise. A more robust algorithm would clearly result if the specific determination of the delays did not have to be explicitly done. In ICHF, one tries to determine the "optimal-fit" hyperbola in the cross-correlation space itself; thus, no pattern recognition errors are made prior to the optimization.

Plots for real data are presented in Figures 3 and 4. In each case, the data are produced by a loud talker situated at (1M,2M) with low ambient noise. In Figure 3, TDEHF worked well, as the peaks are relatively easy to pick correctly. In Figure 4, however, TDEHF yielded poor results, although it is evident that a hyperbolic fit in the cross-correlation space itself could give the right location.

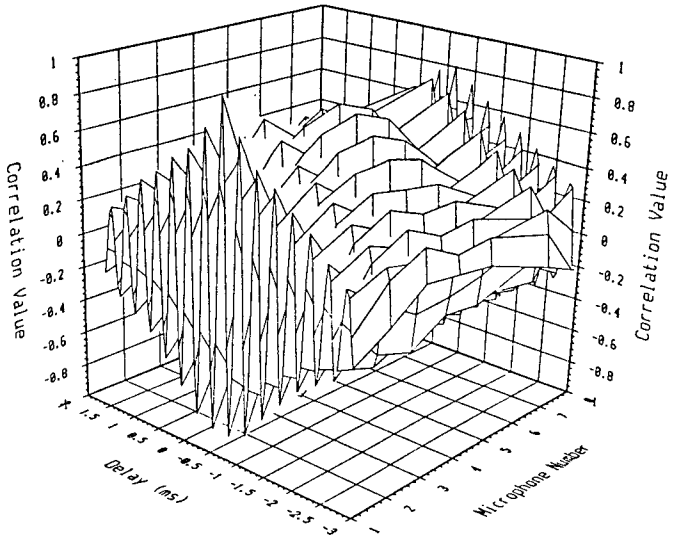

Figure 3: Example of Correlation Space where TDEHF Succeeds

In nonlinear optimization, one must develop a functional that measures "goodness (badness)" as a function of the set of variables over which one wants to optimize. In this case, one wants to develop a measure of the average "goodness" of a particular hyperbola parameterized by (x,y) over the space shown in Figures 3, 4 having independent variables of x, the x spatial variable, and d, the relative delay. Points for the microphones $(z_m,\hat{d}_m)$ may be computed from Equations 4.3 and 4.4; this guarantees they all lie on a unique hyperbola. If a continuous cross-correlation function, $C(x,d)$ were available, then a reasonable functional for maximization would be,

$$\hat{E}(k') \equiv \frac{1}{M} \cdot \sum_{m=1}^{M} C(z_m, \hat{d}_m).$$

$\hat{E}(k')$ represents a measure of the average height of the cross-correlation function measured over the points on the hyperbola taken by the set of microphones. One should note that it would be expected that the value should be positive for reasonable situations, and approaching unity for ideal ones, and thus $\hat{E}(k')$ could also be used to threshold decisions.

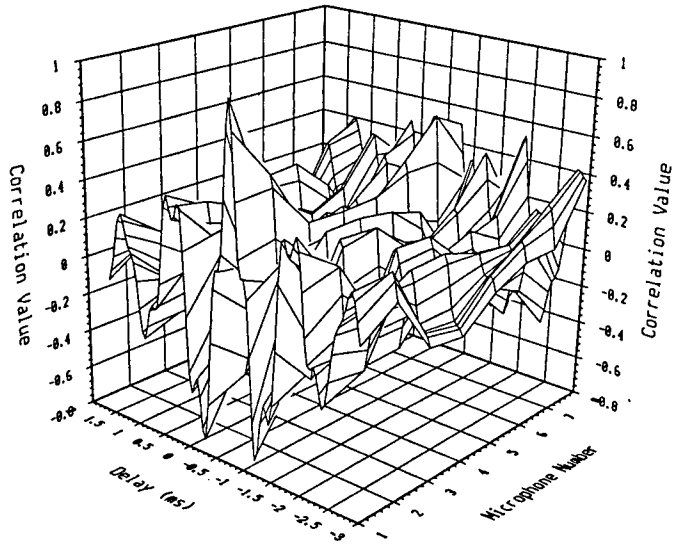

Figure 4: Example of Correlation Space where TDEHF Fails

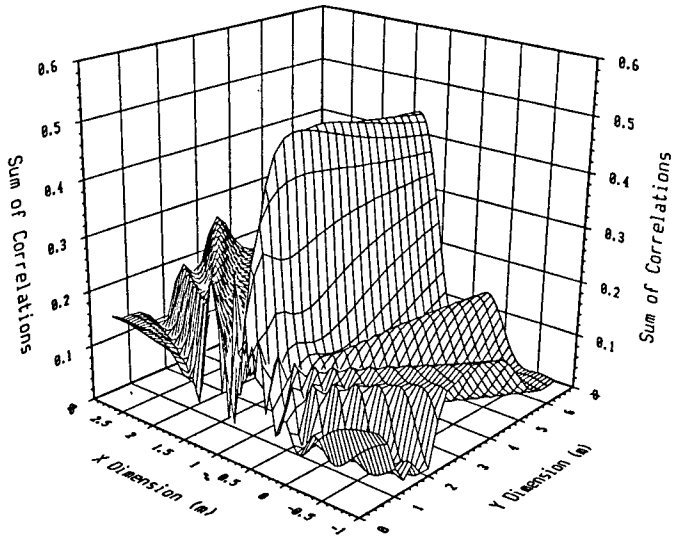

Figure 5: $\hat{E}(k')$ vs (x, y) for Data of Figure 4

As no continuous cross-correlation function is available, one must approximate it. It is assumed that interpolation may be used to achieve an accurate estimate, i. e., one determines σ_m and ν_m from $\hat{d}_m$ using,

$$\sigma_m \equiv \begin{cases} \lfloor R * \hat{d}_m \rfloor & \hat{d}_m \geq 0 \\ \lfloor R * \hat{d}_m - 1.0 \rfloor & \hat{d}_m < 0 \end{cases} \qquad (4.9)$$

$$\nu_m \equiv \lfloor \hat{d}_m * R' - \sigma_m * \lambda + 0.5 \rfloor. \qquad (4.10)$$

(4.8)

Then, $C_m(z_m, \hat{d}_m)$ may be accurately approximated by

$$C_m(z_m, \hat{d}_m) \approx C_m^{R'}[\lambda\sigma_m + \nu_m, k'], \qquad (4.11)$$

which is exactly as derived previously. A three dimensional plot of the surface for $\hat{E}(k')$ is given in Figure 5. Notice the strong peaking due to the hyperbolic-fit transformation.

Results

Some preliminary results for one loud talker standing at (1M,2M) with a low ambient are shown in Figures 6 and 7. A linear array of eight microphones was used for all cases. For these Figures, an algorithm was assumed to have correctly located the talker if it indicated a location within the rectangular region from 1.9M to 2.1M in x and 1.5M to 2.5M in y. As algorithms have improved, the measure of "correctness" is also to be refined in further work. In both TDEHF and ICHF, the tendency is for better performance when larger-size cross-correlations are used, although there seems to be no reason to go beyond 3500 samples (175ms). It is also clear that ICHF is far more robust than is TDEHF. Furthermore, as might be expected, one gets improved performance using bandpass-filtered data. (The filter used is a 61-tap, symmetric FIR filter having transition bands (400Hz -900Hz) and (3300HZ-3800)Hz; stopbands are 50dB down.)

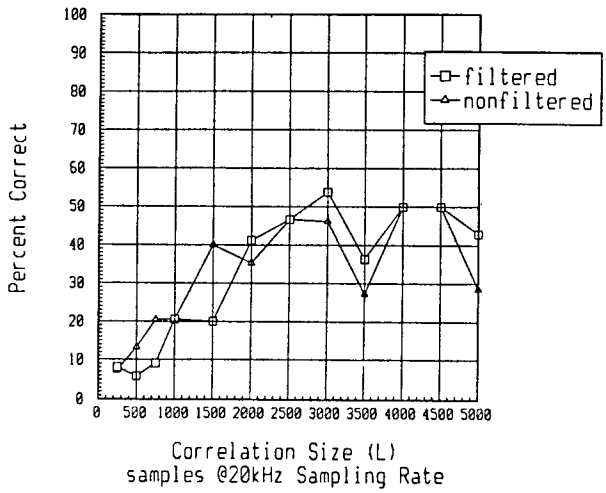

Figure 6: Performance of TDEHF

There is high correlation between "correctness" and the resultant value of $\hat{E}[k']$ for ICHF. Therefore, it is expected that, in regions where the algorithm fails -- perhaps in silence or a high-ambient interval -- the value of $\hat{E}[k']$ would be low and the incorrect location would not be accepted. Given this thresholding, one would expect to almost always get an accurate prediction of a talker's location, providing no other talkers are competing acoustically, a case not yet studied.

Computationally, ICHF is implementable in real-time due to the use of Stochastic Region Contraction [14] for the nonlinear optimization. Relative to a coarse-fine full search, SRC has provided an order-of-magnitude im-

provement with virtually no loss in accuracy.

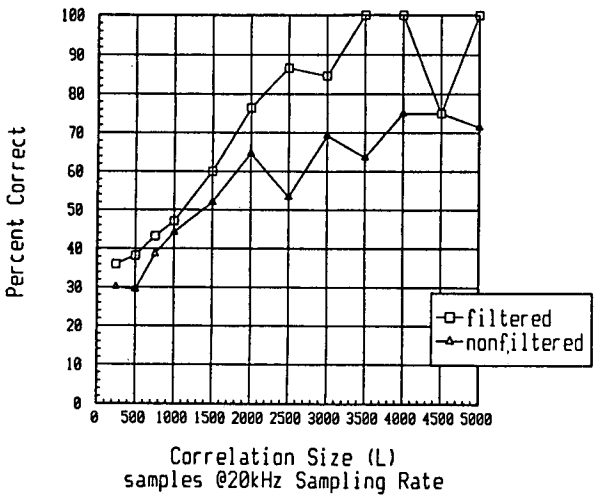

Figure 7: Performance of ICHF

Conclusion

A very promising algorithm for determining the location of a talker in a real acoustic environment has been introduced. In an uncontested acoustic environment, preliminary results from real data indicate that highly accurate performance is achievable. In addition, the SRC method for nonlinear optimization has provided a mechanism for making the algorithm practical in real time. In follow-on work, more data have to be tested, multiple talker and various noise environments need to be explored, and extensions to tracking need to be developed. However, the current level of performance tends to predict that these aspects will go smoothly.

References

[1] Flanagan, J. L., *Bandwidth Design for Speech-seeking Microphone Arrays*, Proc. 1985 ICASSP, Tampa, FL, 3/85, pp. 732-735.

[2] Martin, T. B., *Practical Applications of Voice Input to Machines*, Proceedings IEEE, Vol. 64, 4/76 pp. 487-501.

[3] Flanagan, J. L., Johnston, J. D., Zahn, R., and Elko, G. W., *Computer-steered Microphone Arrays for Source Transduction in Large Rooms*, Journal of the Acoustical Society of America, Vol. 78, No. 5, 11/85, pp. 1508-1518.

[4] Schmidt, R. O., *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*, PhD. Dissertation, Stanford University, Nov. 1981.

[5] Schmidt, R. O., *Multiple Emitter Location and Signal Parameter Estimation*, IEEE Trans. on Antennas and Propagation, Vol. AP-34, No. 3, 3/86, pp. 276-280.

[6] Schmidt, R. O., and Franks, R. E., *Multiple Source DF Signal Processing: An Experimental System*, IEEE Trans. on Antennas and Propagation, Vol. AP-34, No. 3, 3/86, pp. 281-290.

[7] Wax, M. and Kailath, T., *Optimum Localization of Multiple Sources by Passive Arrays*, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-31, No. 5,, 10/83, pp. 1210-1218.

[8] Kesler, S. B., and Shahmirian, V., *Bias Resolution of the MUSIC and Modified FBLP Algorithms in the Presence of Coherent Plane Waves*, IEEE Trans. on Acoustics, Speech and signal Processing, Vol ASSP-36, No. 8, 8/88, pp. 1351-1352.

[9] Knapp, C. H., and Carter, G. C., *The Generalized Correlation Method for Estimation of Time delay*, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-24, No. 4, 8/76, pp. 320-327.

[10] Carter, G. C., *Coherence and Time-Delay Estimation*, Proc. IEEE, Vol. 75, No. 2, 2/87, pp. 236-255.

[11] Bendat, J. S., and Piersol, A. G., *Engineering Applications of Correlation and Spectral Analysis*, John Wiley and Sons, Inc. 1980.

[12] Crochiere, R. E., and Rabiner, L. R., *Multirate Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ 07632, 1983.

[13] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vettering, W. T., *Numerical Recipes in C*, Cambridge University Press, New York, 1988.

[14] Berger, M., and Silverman, H. F., *Microphone Array Optimization by Stochastic Region Contraction*, Technical Report LEMS-62, Division of Engineering, Brown University, August 1989.

[15] Alvarado, V. M., *Talker Localization and Optimal Placement of Microphones for a Linear Microphone Array using Stochastic Region Contraction*, PhD Thesis, LEMS, Division of Engineering, Brown University, May 1990.

[16] Analog Devices, Inc. *ADSP-2100 User's Manual*, Analog Devices, Inc., Norwood, MA, 1989.

[17] Silverman, H. F., and Doerr, K. J., *Talker Location using a Linear Microphone Array and Hyperbolic Fitting* Brown University, Division of Engineering, LEMS Technical Report #73, July 1990.

[18] Fletcher, R. and Powell, M. J. D., *A Rapidly Convergent Descent Method for Minimization*, Computer Journal, Vol. 6, 1963, pp 163-168.