A METHOD FOR WAVEFRONT CURVATURE RANGING OF SPEECH SOURCES

Eneyew Adugna
Department of Electrical Engineering
Addis Ababa University

ABSTRACT

A new approach for estimating the location of a speech source in a reverberant environment is presented. The approach also implements focusing using a linear array of microphones to reduce the effects of reverberation. The algorithm estimates the curvature of the incident wavefront of the source with regard to its images by using an estimated signal-to-interference ratio (SIRe) function. The location which maximizes the SIRe function is used to estimate the source location. A delay-and-sum technique is then used to focus on the source in order to reduce the effects of reverberation and noise. Simulation results are presented for a room of size 8 × 6 m and the first 23 planer images of the source. The results show that the new algorithm is effective in locating a speech source and minimizing multipath effects

INTRODUCTION

This paper addresses the problem of automatically locating a speech source in a reverberant room using a linear array of microphones. The problem of locating a speech source is complicated by the presence of correlated signals, due to multipath effects, and other acoustic sources as well as the wideband nature of the speech signal.

Source location using an array of transducers has many applications in navigation, surveillence, aerospace, and geophysics [6]. Signals radiating from a source arrive at the sensors of an array with relative time delays which depend on the source sensor geometry. Differences in the relative times of arrival to each sensor can be used to infer the location of the source.

Most approaches to source location use a plane wave approximation [1,4,5]. That is the source is assumed to lie in the far-field of the array. Methods which apply to the near-field case for speech applications have also been suggested [7,13]. The approach taken by Flanagan *et al* [7] uses a ray tracing approach. This approach requires exact knowledge of the room geometry and is computationally complex. An alternate approach is to use a Green's function which models the room more accurately than a plane wave but not as exact as ray tracing. The use of a free-space Green's function has been suggested by Abe *et al* [13]. Here a cost function is defined based on the free-space Green's function and the maximum SIRe used to find the location. Variance bounds on estimated range and bearing, using wavefront ranging methods, are suggested in [9,10].

In this paper we take a similar approach to Abe *et al* [13] with several fundamental differences. A narrow-band (NB) filter is used to increase the SNR of the measured signal and to reduce the effects of reverberation. The SIRe function is introduced and used to find the range as opposed to the focused function used by Abe. The SIRe measures the relative deviation from a plane wave of the focused function at the location in question. This is shown to yield better performance in the reverberation or multi-source situation than Abe's approximation. Fast search algorithms and multi resolution strategies can be used to find the primary source in real time.

The paper is organized as follows. An optimal single source location estimation algorithm is presented and the limitations of this optimal approach, for use in multi-source environment, are discussed in terms of spatial frequency overlap. Finally a new heuristic algorithm, effective in multi-source environments, is presented.

RANGING USING TIME DELAY ESTINATION

Consider a point source in the near-field of a linear array (M sensors d cm apart). The source location parameters r_i and θ_i are shown in Fig. 1. Let the NB source frequency be ω. Note that for omnidirectional sensors, the received signal is independent of ϕ. Hence, we will consider sources in the x - z plane only. For a sensor spacing d, z nd,

$$\sin(\theta_i) \qquad \frac{z_i}{r_i} \tag{1}$$

the signal received at a sensor can be written as

$$s(\tau,t) = \frac{1}{r}\cos(\omega t - \omega\tau) \qquad (2)$$

where τ is the delay parameter and c is the speed of sound in air. The relative delays, τ_n, are computed from relative phase computations using the analytic signal representation of the signals in each sensor. Analytically, the relative delay at sensor n is given by

$$\tau_n = \frac{r_n - r_0}{c}, \qquad (3)$$

where

$$r_n^2 = r^2 + n^2d^2 - 2ndr_0\sin(\theta_0). \qquad (4)$$

The above two equations are combined to give an equation for each n, i.e.,

$$\begin{aligned} C_n &= B_n r_0 \sin(\theta_0) + D_n r_0 \\ &= B_n z_0 + D_n r_0 \end{aligned} \qquad (5)$$

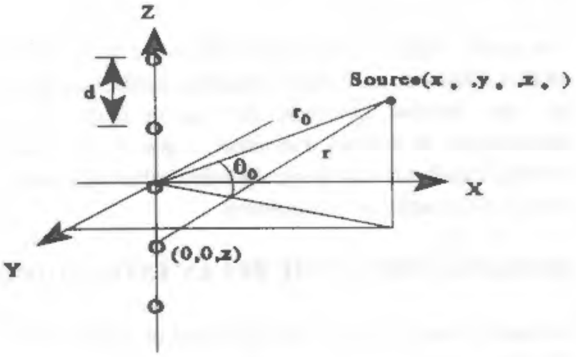


Figure 1 Linear array and source parameters

where $-M/2 < n < M/2$. Hence, a system of M equations in the unknowns r_0 and θ_0 is formed. This system of equations can be written compactly as

$$y = Hx + \eta \qquad (6)$$

where y is a vector of the known values C_n, H is a matrix of known parameters, η is the measurement error and x is a vector of the unknown source parameters. The least squares solution of Eq. 6 is given by

$$x_{ls} = (H^TH)^{-1}H^Ty \qquad (7)$$

This solution yields the maximum likelihood estimate of r_0 and θ_0 under a gaussian noise assumption. The least squares method is a simple and an efficient way of locating a single NB source using a single linear array of sensors. It is computationally inexpensive and accurate. The results using this method for a single source can be used as a bound of performance for the proposed multi-source location algorithms. However, we shall see that this method is not applicable in practice for multiple source problems.

Simulation Results

The optimum performance, for the case of a linear array with *M = 63*, *d = 4.25 cm*, and a single narrow-band source (*at f = 1 kHz.*) in free space, has been considered to evaluate the performance of the least squares solution. Some representative results are shown in Table 1 and Figs. 2 and 3. The performance of the algorithm detriorates in noise at lower SNR values, as shown in Fig. 3.

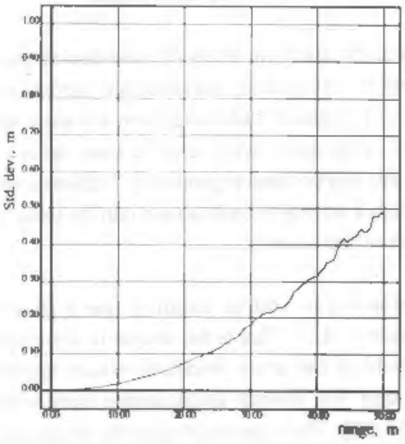


Figure 2 Range standard deviation at *SNR* = 40 dB and $\theta_0 = 0°$.

Table 1 : Computed range and angle values using the least squares method. Distance is in *cm* and angle is in *degrees*.

Expected		Computed		% Error	
θ_o	r_o	θ_o	r_o	θ_o	r_o
0	850	0.00	849.99	.00	.001
0	1700	0.00	1699.98	.00	.001
03	8500	0.00	8503.62	.00	.043
0	212	29.99	212	.03	.001
30	850	30.00	849.99	.00	.001
30	3400	29.99	3400.31	.03	.009
45	425	44.99	425.00	.02	.001
45	850	45.00	850.03	.00	.004
45	1700	44.99	1699.73	.02	.017
60	850	60.00	849.99	.00	.001
60	1700	59.99	1699.75	.02	.015
60	3400	59.99	3403.21	.02	.094

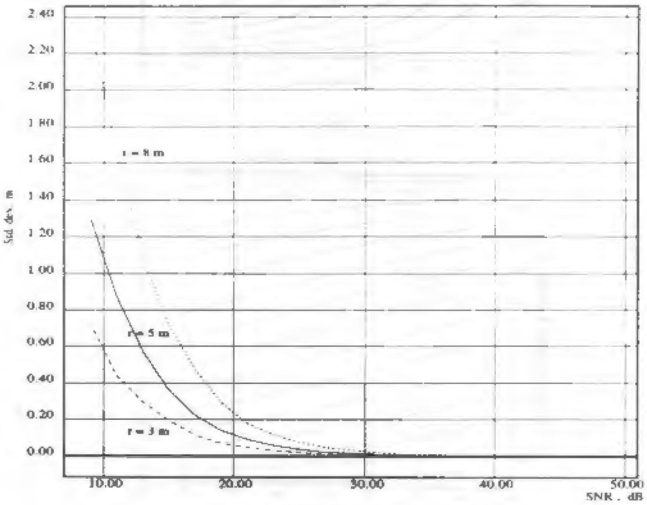


Figure 3 Range variance with SNR at r_0 r_0 $= 3, 5, 8\ m.$

SPECTRAL BROADENING DUE TO NEAR-FIELD SOURCE

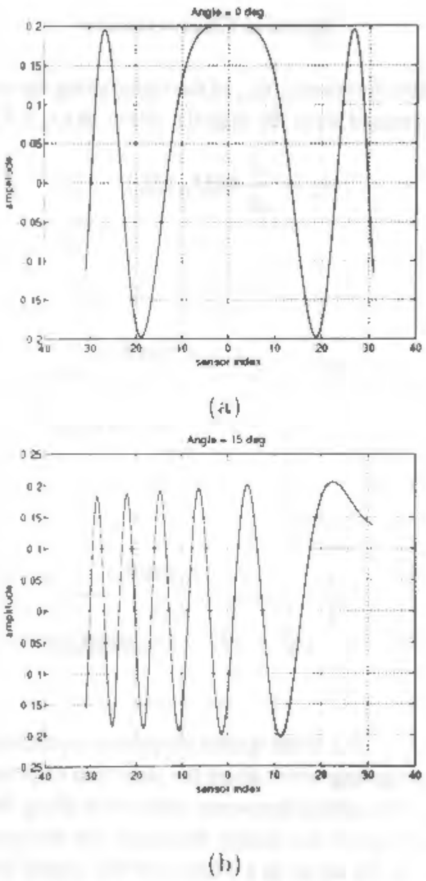
In this section we shall show that the multi-source situation introduces overlap of the spatial spectrum. This overlap causes problems for multiple-source location using linear least squares.

Spatial Characteristics

The basic features of the received signal across the array due to a NB source placed in the array's near-field are presented in this section. The nonquadratic phase term in Eq. 5 (due to *r*) in the near-field case has not been treated exhaustively. The Fresnel approximation [14,15] for the near-field case is sometimes used to approximate the phase to a quadratic. However, the exact form of the phase expression results in a much more complex spectrum. In either case, as a result of the curvature of the wave at the array, we find a broadening of the spectral bandwidth. We shall show analytically that the signal is nonlinearly frequency modulated and that the extent of this frequency variation depends on such parameters as r_o, θ_o, λ, and length of the array.

Consider a single NB harmonic source at $x(t) = A\ cos\ (\omega t)$ in the near-field of the array at (θ_o, r_o) on the x-z plane. Then the signal received at the array at time *t* is

$$s(z,\theta_o,t) = \frac{A}{r}\cos(\omega t - \frac{\omega}{c}(r_o^2 + z^2 - 2r_o z\sin(\theta_o))^{\frac{1}{2}}) \quad (8)$$



(a)

(b)

Figure 4 Array signal at $r_0 = 5$ m and $f = 3$ kHz, a) $\theta_o = 0°$, b) $\theta_o = 15°$,

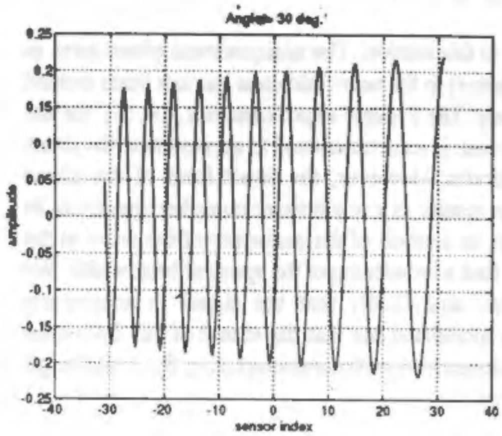


Figure 4 contd. c) $\theta_o = 30°$

where $-(M/2)d \leq z \leq (M/2)d$. Fig. 4 shows some representative snapshots with $M = 63$, $d = 4.25\ cm$, $f = 3\ kHz$, and three values of θ_0 for a given value of r_0.

Spectral Characteristic

The spatial frequency, ω_a, of the signal along the array can be determined from the signal's phase $\varphi(t,r_0,z,\theta_0)$ as

$$\omega_a = \frac{d}{dz}\varphi(t,r_0,z,\theta_0) \tag{9}$$

i.e.,

$$\omega_a = -\frac{\omega}{c}\frac{z - r_0\sin(\theta_0)}{(r_0^2 + z^2 - 2r_0 z\sin(\theta_0))^{\frac{1}{2}}} \tag{10}$$

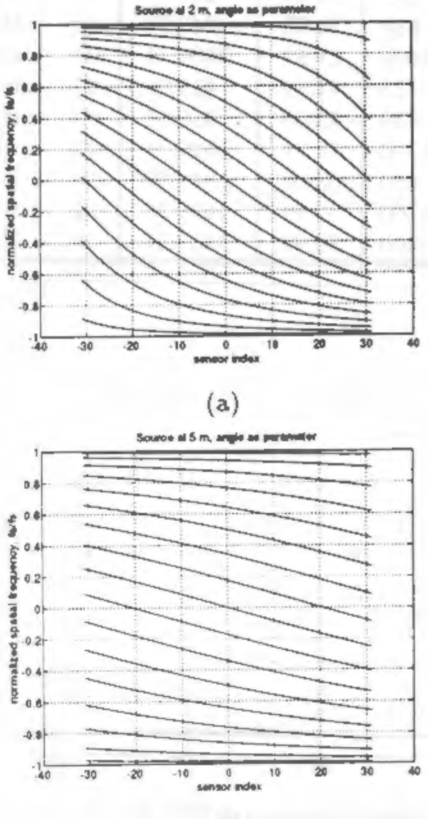
or

$$f_a = -\frac{1}{\lambda}\frac{z - r_0\sin(\theta_0)}{(r_0^2 + z^2 - 2r_0 z\sin(\theta_0))^{\frac{1}{2}}} \tag{11}$$

where $f_s = 1/\lambda$ is the spatial frequency (cycles/meter) of the propagating wave along the direction of propagation and f_a is the spatial frequency of the wave along the array. Eq. (12) gives the spatial frequency for the snapshot at point z on the array as a fraction of the spatial frequency of the wave along its direction of propagation, f_s. Note that **f_a and f_s have a nonlinear relationship. The normalized spatial frequency then becomes,**

$$\omega_{an} = -\frac{z - r_0\sin(\theta_0)}{(r_0^2 + z^2 - 2r_0 z\sin(\theta_0))^{\frac{1}{2}}} \tag{12}$$

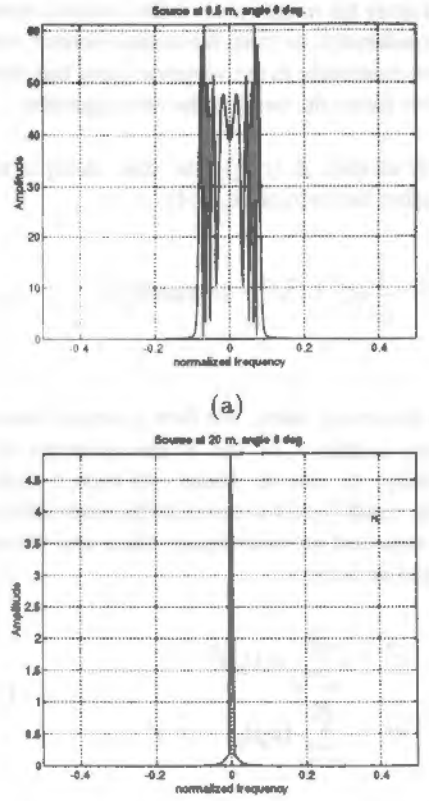
The variation of ω_{an} with angle of arrival, for r_0 constant, is shown in Fig. 5. The figures also show that the spatial frequency along the array, f_a, does not exceed the source spatial frequency, f_s, at any value of z. The variation of spectral bandwidth with source location parameters and temporal frequency of the source is shown in Figs. 6 and 7. Note the decrease in the FM bandwidth as r_0 increases.



(a)

(b)

Figure 5 Spatial frequency modulation across the array at a) $r_0 = 2\ m$, b) $r_0 = 5\ m$, with θ_0 as a parameter varying from $-\pi/2$ to $\pi/2$.



(a)

(b)

Figure 6 Spectral width of signal at $\theta_o = 0°$ and temporal frequency of 3 kHz for a) $r_o = 0.5$ m and b) $r_o = 20$ m. Note the decrease in FM bandwidth as r_o increases.

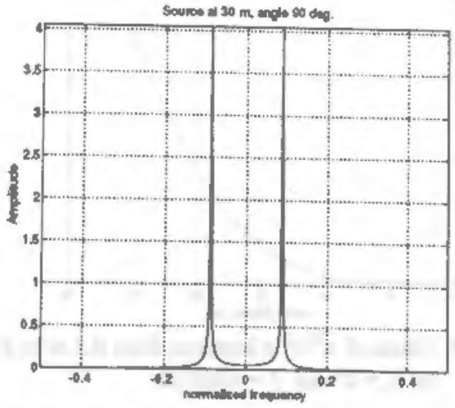


Figure 7 Spectral width of signal for $r_o = 30$ meters at $\theta_o = 90°$ and temporal frequency of 3 kHz.

As can be seen in Fig. 6 the wide band nature of the array signal makes it difficult to spatially filter a signal from a given direction. The problem gets worse with wideband signals, such as speech. In the next section we present a new algorithm for finding the location of a source for speech signals. The new method uses a Green's function to demodulate the FM modulation and thus collapse the spectral bandwidth of the source. The collapsed source spectrum is then filtered out 'purely' from the interfering spectrum.

SPEECH SOURCE LOCATION IN A REVERBERANT ENVIRONMENT

The discussions given in the previous section indicate that a spatial frequency filtering approach cannot be employed to separate near-field sources. We present in this section a different approach to locate a speech source in a multiple source environment.

Issues on array size and sensor spacing and their relationships to source bandwidth [1, 11] will be made use of in the new algorithm.

The New Algorithm

The new algorithm has two major components. The first component deals with estimating the location of the speech source from a series of snapshot vectors. The second component makes use of the location information to focus on the source. Information on source location is updated regularly.

Consider a speech source at location (r_o, θ_0), an array with $2M+1$ uniformly spaced microphones, and N image sources. Let $s(t)$ represent the source signal and $x_i(t)$ represent the received signal at the i^{th} microphone. The signal at the i^{th} microphone may be expressed as

$$x_i(t) = \sum_{n=0}^{N} \frac{A}{r_{in}} s(t - \frac{r_{in}}{c}), \quad i = 1,2,...,M \qquad (13)$$

where

i - is sensor index,
n - is the image index, n = 0 being for the source,
r_{in} - is the distance from the i^{th} sensor to the n^{th} source,
c - is the speed of sound in space, and
A - is the source signal strength.

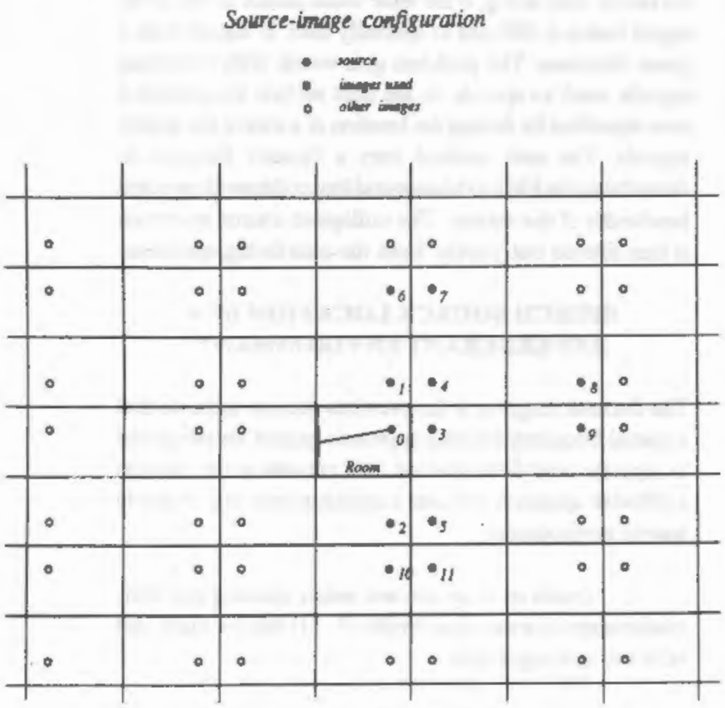


Figure 8 Location of a source and its images, in a rectangular room.

and

$$r_{in}^2 = r_n^2 + i^2d^2 - 2r_i d\sin(\theta_0) \qquad (14)$$

where

r_n - is n^{th} image distance from array center,
θ_n - is n^{th} image angle, measured from the array normal,
d - is sensor spacing, and
i - is the sensor index.

The received snapshot vectors, x_k, form columns of a matrix X. This matrix of snapshot vectors, X, is formed for K snapshots, where

$$X = [x_1\ x_2\ ...\ x_k\ ...\ x_K] \qquad (15)$$

For an array geometry matched to the wavefront from the source, the received signal vector will be a constant vector (within a scalefactor due to spherical spreading of the wave). Signal components from other than the source location will have a non-constant signalvector contribution. Thus, in a focused array the mean value of the received steered vector is considered to be from the desired source while other sources contribute to the variance about this mean. This property forms the basis of the new algorithm.

For a source located at $(r_s\theta_s)$, the time delay to each sensor is, using the form of Eq. (14),

$$\tau_i^2 = \frac{1}{c^2}(r_s^2 + i^2d^2 - 2r_i d\sin(\theta_s)) \qquad (16)$$

where i is the sensor index. We form a steered vector x by applying a delay τ_i to the i^{th} row of matrix **X**. A common delay τ_0 can be added to each$_i$ τ without changing the result. Vector **x** represents the desired focused snapshot vector and its mean square value and variance are computed as follows:

$$\begin{aligned} \overline{x^2} &= (\sum_{n=-M}^{M} x_n(t_0))^2 \\ \sigma^2 &= \sum_{n=-M}^{M} (x_n(t_0) - \bar{x})^2. \end{aligned} \qquad (17)$$

From our previous discussions, $\overline{x^2}$ represents the desired signal where as σ^2 represents the interference from the other sources. Thus, we define the signal-to-interference ratio estimate (SIRe) parameter as

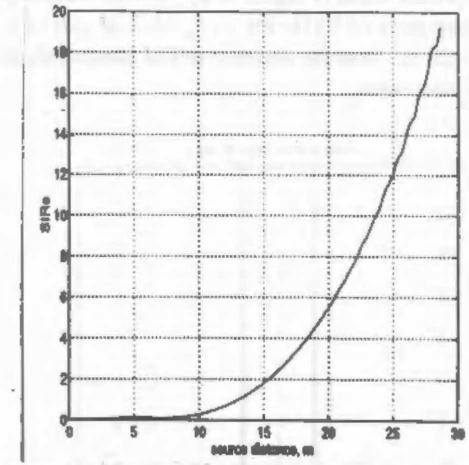


Figure 9 Values of $\overline{x^2}/\sigma^2$ at locations from *0.5 m* to *30 m* for $\theta_s = 0°$ and $f = 3000\ Hz$.

$$SIRe = \frac{\bar{x}^2}{\sigma^2} \qquad (18)$$

This is the figure of merit which is maximized in the new algorithm. A maximum value of the parameter indicates the location of the source. For the single source case the SIRe goes to infinity at the location of the source. In multiple source cases the SIRe value has been found to be maximum at the location of a source.

For $\theta_o = 0°$, Fig. 9 shows that the SIRe gets larger as the source location moves away from the array, even when the array geometry does not match the progressing wave's wavefront. Appropriate delays are applied to each sensor to simulate an array focused to a source point. The SIR estimate should theoretically go to infinity for an array steered to the location of a source. However, in practice σ^2 does not go to zero in a multiple source situation, but the value becomes very small. If, however, the virtual location is not matched to the source, then the value of SIRe will not be a maximum.

Thus, the strategy used in this algorithm is to search for the location that maximizes the SIR estimate. The search region is divided into small cells of dimension *dx* by *dy*. Take a candidate cell, compute the delay vector for that cell, steer the received signal matrix X, compute the SIRe. Maximize SIRe over several snapshots for the cell. The cell that gives the maximum SIRe value is assumed to contain the source. Simulation results of the algorithm are presented in the next section.

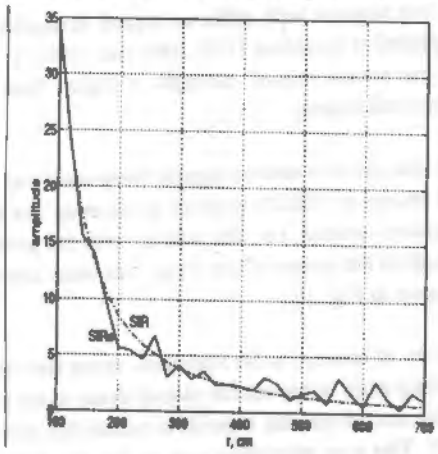


Figure 10 SIRE and SIR as a function of source distance in *cm*. The dashed line represents SIR and the solid line represents SIRe.

A measure of performance of the new algorithm .is given in (Fig.10), where the SIRe for a source and its five images is plotted in comparison with the true value (SIR) of the source. The value of SIR is very large (ideally infinite), thus a scaling factor has been used for SIR.

Speech Source Location Results

The simulation results in this section are based on the utterance ``*The best way to learn is to solve several problems* " [12]. Different segments of this utterance are employed. Many of the results deal with locating a single source inside the room and its images. The case of two sources in the room has also been considered. The results on location are given in terms of SIRe plots over the search region.

The NB filter frequency is taken at the fourth formant frequency of the frame, for voiced frames, and about the same value for unvoiced segments. The array size is about *2.56 meters*, i.e., $M = 31$ and $d = 8.25$ *cm*. A grid size of *20* x *20 cm* has been used in a room of size *8* x *6 m*.

Single Speech Source

For the single source inside the room case the SIRe variation with position is shown in the various 3D plots, Figs. 11-15.

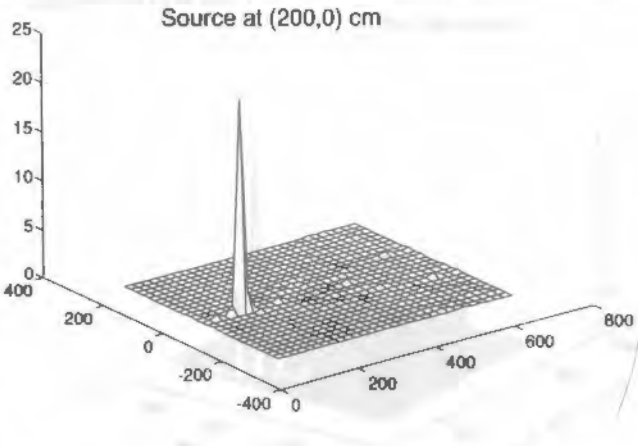


Figure 11 3D plot of SIRe for source at (200,0)cm.

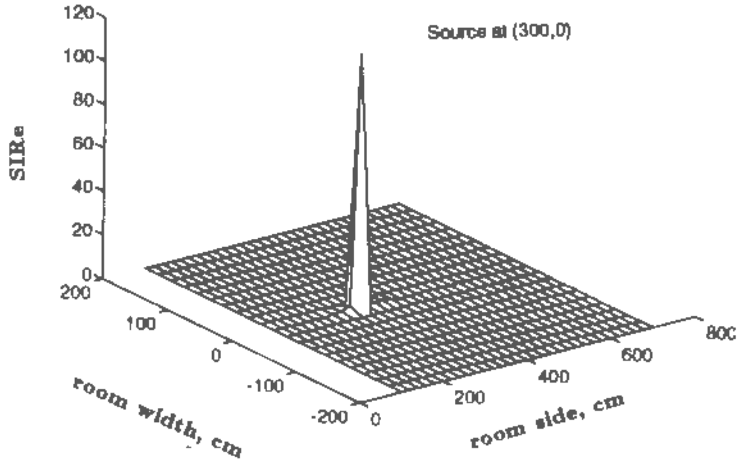


Figure 12 3D plot of SIRe for source at (300,0),cm.

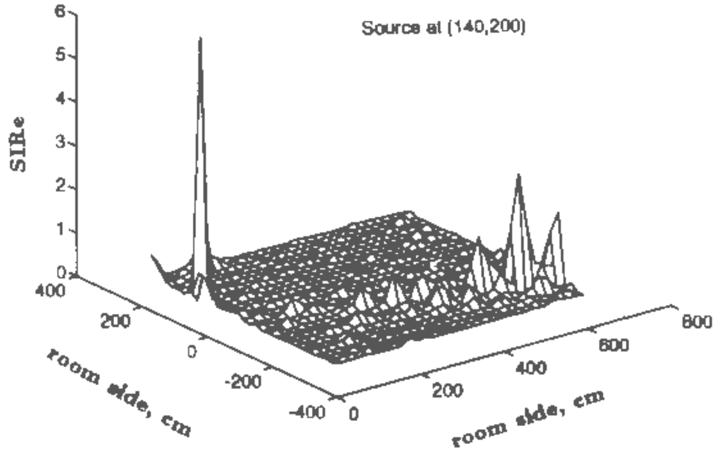


Figure 13 3D plot of SIRe for source at (140,200),cm.

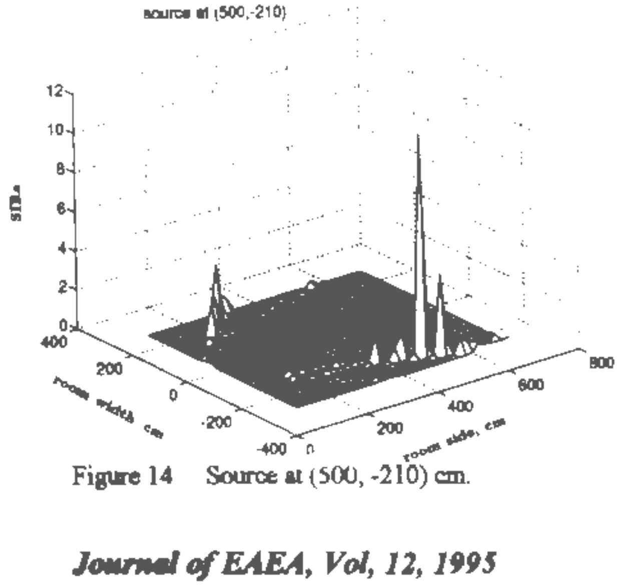


Figure 14 Source at (500, -210) cm.

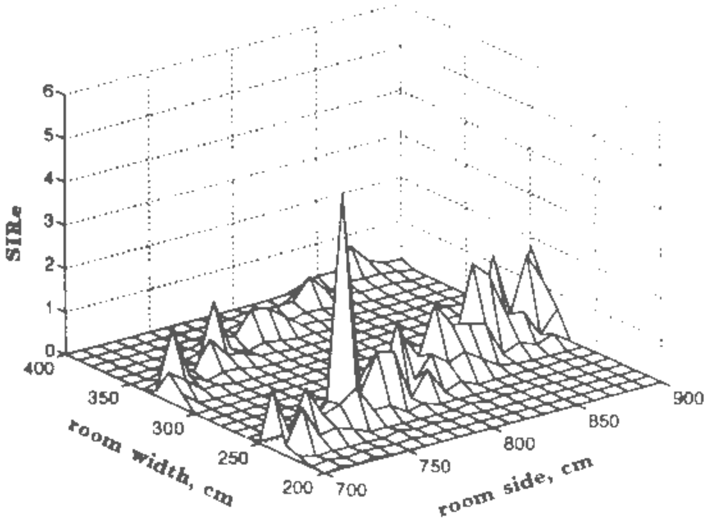


Figure 15 Source at (750,250), 50 cm from each wall near corner at (800,300) cm

Two Speech Sources

The results for two speech sources placed in the room are shown in Figs. 16 and 17. Two cases have been studied:

- the sources are identical in signal strength but are placed at locations (140,200) and (200,0). The second source is closer to the array. The source distance is measured to the center of the array.

- two sources with different signal strengths are placed at locations (140,200) and (200,0). The first source signal strength is higher than the second source.

In the first case the two source signals were identical but the second source, at (200,0), is closer to the array. For this case the closest source, i.e., the source with the greater signal strength at the center of the arrav, has been located as can be seen in Fig. 17.

This result is, of course, to be expected since the other source is acting as an image source placed closer to the true source. In the second case the algorithm locates the source at (140,200). This is an interesting case in that the farthest source has been located. This result may be explained if we consider the effect of the NB filter in each channel. Source II, at (200,0), is a voiced segment and source I, at (140,200), is an unvoiced segment with a lower over all

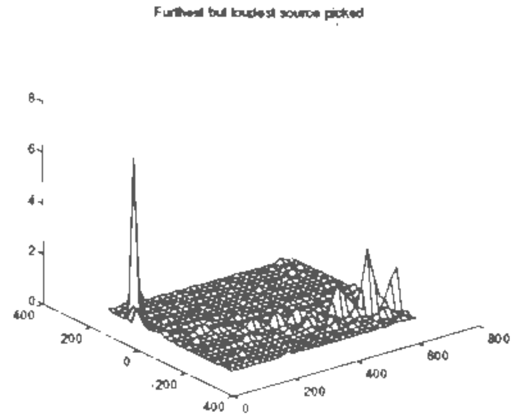


Figure 16 Loudest source at (140,200), cm, is picked

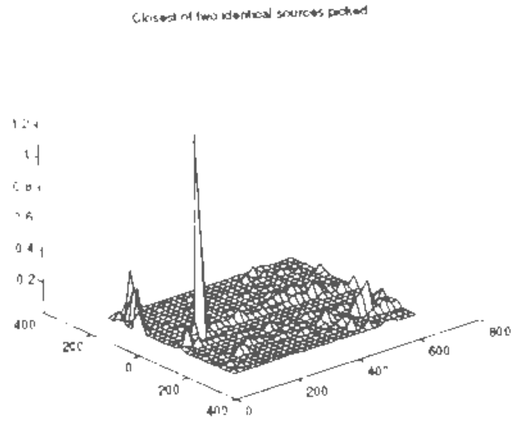


Figure 17 Closest source at coordinate (200,0), cm, picked by algorithm.

signal magnitude. The NB filter frequency is at the fourth formant frequency of the voiced segment. Although source II has a larger overall signal magnitude its energy is concentrated in the low frequency region of its spectrum. At the NB filter frequency the signal magnitude of source II is about *6 dB* lower than that of source I. Hence, at the filter output source I seems to have greater magnitude, and it is this signal that the source location algorithm uses to locate the source. If, however, the first formant was chosen as the NB filter frequency, then the expected result could have been obtained.

Signal Recovery

Most of the results on signal recovery are in the form of audio signals. However, some measure of the similarity, or interference suppression, of the recovered signal with that of the original signal may be obtained by considering a signal-to-noise (or signal-to-interference) ratio between the original and the recovered signals as follows:

$$SIR1 = \frac{\sum_{i=0}^{N} s_i^2}{\sum_{i=0}^{N} (s_i - s_{0i})^2} \tag{19}$$

$$SIR2 = \frac{\sum_{i=0}^{N} s_i^2}{\sum_{i=0}^{N} (s_i - s_n)^2} \tag{20}$$

where

SIR1 - is the signal to interference ratio at the input,
SIR2 - is the signal to interference ratio at the output,
s_i - is the original source signal samples,
s_{oi} - is the received signal samples at array center,
s_n - is the recovered signal at the output of the system,
N - is the frame size.

A measure of the SIR improvement achieved may be computed, in *dB*, as the difference between the SIR2 and SIR1, i.e.,

$$\begin{aligned} SIR_{gain} &= 10\log_{10}\left(\frac{SIR2}{SIR1}\right) \\ &= 10\log_{10}\frac{\sum_{i=0}^{N}(s_i - s_n)^2}{\sum_{i=0}^{N}(s_i - s_{oi})^2} \end{aligned} \tag{21}$$

Values of SIR_{gain} for some source locations are shown in the following table.

These results give a measure of how close the recovered speech signal is to the original signal. They are all taken for locations along the center of the room.

Table 2: SIR gain due to focusing on a source in a reverberant environment.

Position		SNR_{gain} dB
x, cm	*y, cm*	
100	0	14.50
200	0	14.34
300	0	14.42
400	0	12.25
500	0	8.30
600	0	7.44
750	0	3.07

CONCLUSION

A new method has been presented to minimize the effects of reverberation of a speech signal in a room. The method appears to provide listeners an acceptable reproduction of the original speech signal with reduced effects of reverberation and noise. Special emphasis was made on using the source wavefront to determine the location of the source. Utilization of wavefronts to provide source location for autofocusing has been demonstrated by computer simulations.

REFERENCES

[1] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko., *Computer-steered Microphone Arrays for Sound Transduction in Large Rooms*. J. Acoust. Soc. Am. Vol. 78, No. 5, pp 1508-1518, November 1985.

[2] K. Farrell, R.J. Mammone, and J.L. Flanagan, *Beamforming Microphone Arrays for Speech Enhancement*. ICASSP-92, SanFrancisco, CA, March 23-26 1992.

[3] H.F. Silverman. *Some Analysis ofMicrophone Arrays for Speech Data Acquisition*. IEEE Trans. ASSP, Vol. ASSP-35, N0. 12, pp 1699-1711, December 1987.

[4] Don H. Johnson and Dan E. Dudgeon. Array Signal Processing: Concepts and Techniques. Prentice-Hall, Inc., 1993.

[5] Mati Wax and Ilan Ziskind. *On Unique Localization ofMultiple Sources by Passive Sensor Arrays*. IEEE Trans. ASSP Vol. 37, N0. 7, pp. 996-1000, July 1989.

[6] J.O. Smith and J.S. Abel. *Closed-Form Least-Squares Source Location Estimation from Range-DifferenceMeasurements* IEEE Trans. ASSP, Vol. ASSP-35, No 12, pp 1661-1669, December 1987.

[7] J.L. Flanagan, A.C. Surendran, and E.E. Jan. *Spatially Selective Sound Capture for Speech and Audio Processing*. Speech Communication, Vol. 13, pp 207-222, 1993

[8] S.S. Reddi. *Multiple Source Location --- A Digital Approach*. IEEE Trans AES, Vol. AES-15, No. 1, pp 95-105, January 1979.

[9] G. Clifford Carter. *Variance Bounds for Passively Locating an Acoustic Source with a Symmetric Line Array*. J. Aqoust. Soc. Am., Vol. 62, No. 4, October 1977.

[10] K. B. Theriault and R. M. Zeskind. *Inherent Bias in Wavefront Curvature Ranging*. IEEE Trans ASSP., Vol. ASSP-29, No. 3, June 1981

[11] J.L. Flanagan. *Beamwidth and Useable Bandwidth of Delay-Steered Microphone Arrays*. AT&T Technical Journal, Vol. 64, No. 4, pp. 983-995, April 1985.

[12] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshal. *Darpa speech recognition research database: specifications and status*. Proceedings of the DARPA Speech Recognition Workshop, pp 93-99, 1986.

[13] Masato Abe, Kiyohito Fujii, Toshio Sone, and Keniti Kido. *Estimation of Position and Waveform of a Specified Sound Source Decreasing the Effect of Other Sound Sources and Reflection*. Proceedings of ICASSP 91, Vol. M2.7, pp. 2337-2340, 1991.

[14] J.W. Goodman. Introduction To Fourier Optics. McGraw-Hill Book Co., 1968

[15] Albert Macovski. Medical Imaging Systems. Prentice-Hall, Inc., 1983.