AUTODIRECTIVE MICROPHONE SYSTEMS FOR NATURAL COMMUNICATION WITH SPEECH RECOGNIZERS

J. L. Flanagan and R. Mammone
CAIP Center, Rutgers University, New Brunswick, New Jersey

G. W. Elko
AT&T Bell Laboratories, Murray Hill, New Jersey

Abstract

Two technological advances support new sophistication in sound capture; namely, high-quality low-cost electret microphones and high-speed economical signal processors. Combined with new understanding in acoustic beamforming, these technologies permit spatially-selective transduction of speech signals several octaves in bandwidth. Spatial selectivity mitigates the effects of noise and reverberation, and digital processing provides the capability for speech-seeking, autodirective performance. This report outlines the principles of autodirective beamforming for acoustic arrays, and it describes two experimental implementations. It also summarizes the direction and emphasis of continuing research.

Introduction

In many applications of automatic speech recognition, it is desirable for the talker to have hands and eyes free for concurrent tasks. Typical examples include parcel sorting, product assembly and inspection, voice dialing for cellular telephones, and data plotting and manipulation in a situation room. The user frequently needs to move around in the workspace, which often is noisy and reverberant, while issuing commands to the speech recognizer. Electrical tethers, close-talking microphones and body-worn sound equipment represent undesirable encumbrances. Ideally, one would like an acoustic system able to capture high-quality sound from natural conversational exchanges in the work space.

Speech-seeking autodirective microphone arrays enable unencumbered freedom of movement, while providing sound pickup quality approaching that of close-talking microphones. Low-cost high-quality electret microphones, in combination with economical signal processing, permit sophisticated beamforming and dynamic beam positioning for tracking a moving talker. Multiple beam formation permits "track while scan" performance, similar to phased-array navigational radars, so that multiple sound sources can be monitored and algorithmic decisions made about the signals [1, 2]. Beamforming has been found to be more useful than adaptive noise filtering for sound pickup in noisy, reverberant enclosures [3].

This report mentions the acoustic principles involved in dynamic beamforming and the design factors governing the ability of steered arrays to combat noise and room reverberation. It discusses the as-yet rudimentary algorithms for sound source location and speech/non-speech detection. It then describes an initial application of an autodirective array and a limited-vocabulary connected-word speech recognizer for voice control of a video/audio teleconferencing system. It concludes by indicating the directions for research needed to refine further the capabilities of hands-free natural sound pickup.

Acoustic Beamforming

The signal output H from an arbitrary array of N discrete omnidirectional acoustic sensors due to a time-harmonic plane wave with wavevector $\mathbf{k}$ is

$$H(\mathbf{k}, \mathbf{r}) = \sum_{n=0}^{N-1} a_n \, e^{-j\mathbf{k} \cdot \mathbf{r}_n}, \qquad (1)$$

where a_n is the amplitude weighting of sensor n, $\mathbf{r}_n$ is the position vector of sensor n with respect to some defined origin, and the bold case indicates a vector quantity. The time-harmonic term is omitted for compactness.

The array can be steered to wave arrivals from different directions by introducing a variable time delay τ_n for each sensor element. The response of the steered array is

$$H(\mathbf{k}, \mathbf{r}) = \sum_{n=0}^{N-1} a_n e^{-j(\mathbf{k}\cdot\mathbf{r}_n + \omega\tau_n)}, \qquad (2)$$

where $\omega = 2\pi f$ is the radian frequency. It is convenient to make a change of variables and define $\mathbf{k}'$ as $\mathbf{k}' = \frac{\omega}{c}\hat{\mathbf{k}}'$, where $\hat{\mathbf{k}}'$ is the unit vector in the wavevector $\mathbf{k}'$ direction, c is the speed of sound, and

$$\mathbf{r}_n \cdot \hat{\mathbf{k}}' = -c\tau_n . \qquad (3)$$

Equation (2) can then be rewritten as

$$H(\mathbf{k}, \mathbf{r}) = \sum_{n=0}^{N-1} a_n e^{-j\mathbf{k}''\cdot\mathbf{r}_n}, \qquad (4)$$

where $\mathbf{k}'' = \mathbf{k} - \mathbf{k}'$. Equation (4) shows that the array response is maximum when $|\mathbf{k}''|$ is 0, or when the delays have been adjusted to co-phase the wave arrival at all sensors. The received *spatial* frequency is 0 (or DC), and the array has a maximum response which is equal to $\sum_{n=0}^{N-1} a_n$. For waves propagating from directions other than $\mathbf{k}'$ the response is diminished.

This principle has been used to design one-dimensional and two-dimensional arrays of sensors spaced by d distance. The element spacing dictates the highest frequency for which spatial aliasing (or, ambiguity in directivity) does not occur. This frequency also depends upon the steering parameters but has a lower bound of $f_{\text{upper}} = c/2d$. Alternatively the spacing is chosen as $d = \lambda_{\text{upper}}/2$. The lowest frequency for which useful spatial discrimination occurs depends upon the overall dimensions of the array.

For speech pickup applications, the desired bandwidth of the array is greater than three octaves. The magnitude of $\mathbf{k}''$ in (4) is proportional to frequency, hence the beamwidth and directivity are inversely proportional to frequency.

A design artifice to combat this frequency dependence is to use "harmonic nesting" [1,2] of the sensors, so that different harmonically-spaced groups of sensors are used to cover contiguous octaves. Some sensors in the nest serve every octave band. Figure 1 shows a nested two-dimensional array of sensors, its directivity index as a function of frequency, and its beam pattern when the a_n's of (4) are Chebyshev weighted for −30 dB sidelobes.

Using these relations one-dimensional and two-dimensional arrays have been designed for conferencing and voice-control applications (see Fig. 2). Digitally-addressable bucket brigade chips on each sensor provide the delay steering under control of a 386 computer.

Algorithms for Speech-Seeking Autodirective Performance

Because of limited computational power in the control computer, algorithms for sound-source location and speech detection are, as yet, rudimentary. Sources are located by a blind search and energy detection, and speech/non-speech decisions are made by waveform heuristics. Beams can be positioned in less than a millisecond, but speech decisions require about twenty milliseconds in a given position.

Full digital designs are in progress having enough signal processing power to make computations of correlations and cepstral coefficients. This will enable more sophistication in both source location and speech detection.

Experimental Applications

The large two-dimensional array, consisting of over 400 electret microphones, has been in use for the past year and a half for interlocation conferencing from an auditorium seating more than 300 persons. Performance greatly surpasses the traditional isolated microphones in the room, and speech quality comparable to Lavalier pickups can be achieved (Fig. 3a).

The small one-dimensional array, consisting of 21 pressure-gradient elements, is being used for an experimental multimedia conferencing system (HuMaNet) designed for ISDN telephone communications [4], (Fig. 3b).

Research Directions

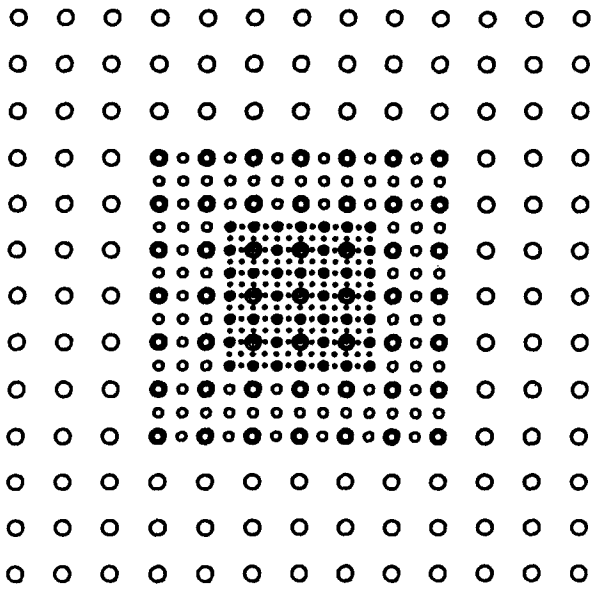
With continued progress in arithmetic capability and economy of single-chip digital signal processors, substantial refinement and expanded performance are possible for autodirective microphone systems. Four areas in particular are receiving research effort. They are:

- accurate spatial location of multiple sound sources

- reliable speech/non-speech discrimination
- spatial volume selectivity in sound capture (and projection)
- characterization of array performance in noisy reverberant enclosures

Properties of three-dimensional microphone arrays appear to provide advantages in some of these areas, and are presently being studied. In particular, 3D arrays can be delay-steered to beamform over 4 pi steradians without spatial ambiguity and with beamwidth independent of steering direction [5].

As with linear and planar arrays, harmonic nesting of the receiving elements in 3D arrays can be used to make beamwidth weakly dependent upon bandwidth coverage. For example, a uniform cubic array, shown in Fig. 4, provides unique, constant-width beam patterns over 4 pi steradians. The 3D geometry can also provide range selectivity that goes beyond the point-focusing capabilities of 1D and 2D arrays. These properties are currently under study.



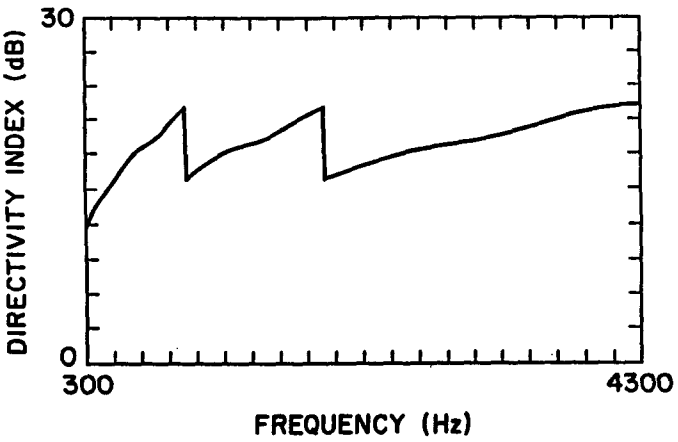
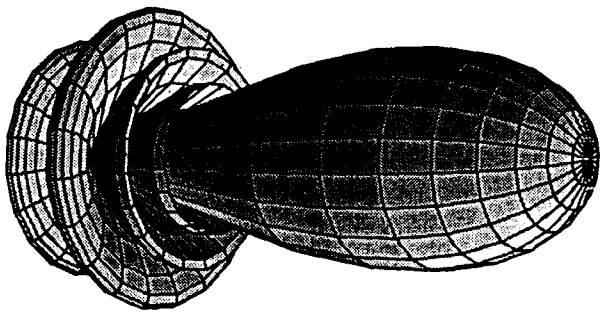
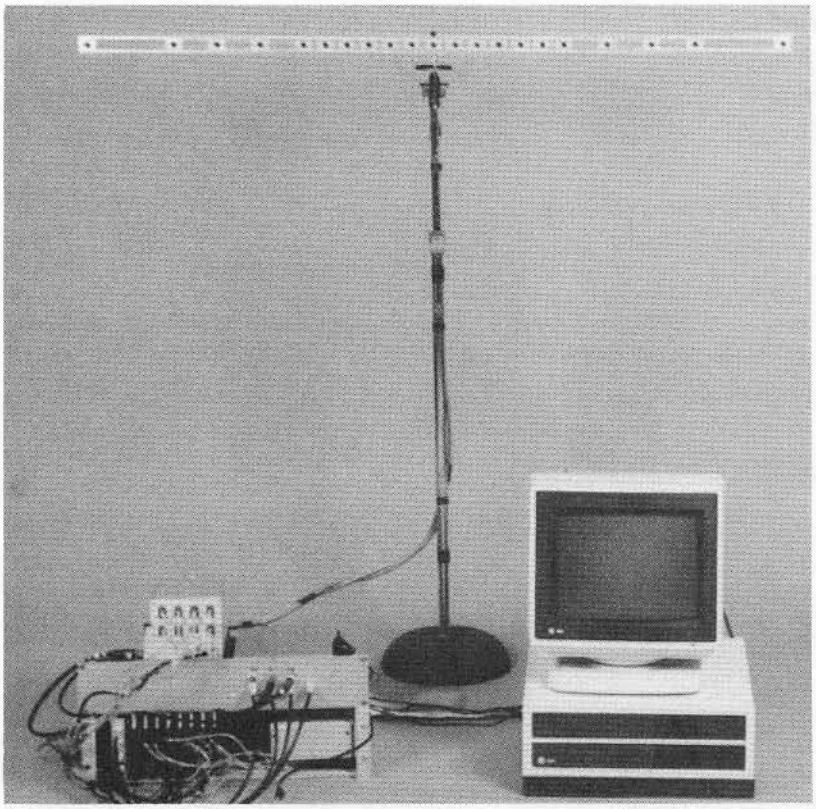


Fig. 1. (a) Harmonic nesting of acoustic sensors for three octaves. Low-frequency elements are shown by the largest circles. Mid and high frequency elements are indicated by smaller and smallest circles, respectively. (b) Directivity index as a function of frequency for nested sensors. (c) Chebyshev weighted beam at broadside (sidelobes are – 30 dB down).



References

1. J. L. Flanagan, J. D. Johnston, R. Zahn, G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms, *J. Acoust. Soc. Amer.* **78**, 1508-1518 (1985).
2. J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E. West, M. M. Sondhi, "Autodirective microphone systems," *Acustica*, February 1991 (in press).
3. M. M. Goulding and J. S. Bird, "Speech enhancement for mobile telephony," *IEEE Trans. Vehic. Tech.* **39**, no. 4, 316-326 (November 1990).
4. J. L. Flanagan, D. A. Berkley, K. L. Shipley, "Integrated information modalities for Human/Machine communications: 'HuMaNet', an experimental system for conferencing," *Jour. Visual Communication and Image Representation* **1**, 113-126 (November 1990).
5. J. L. Flanagan, "Three-dimensional microphone arrays," *J. Acoust. Soc. Amer.* **82**(1), S39 (1987).



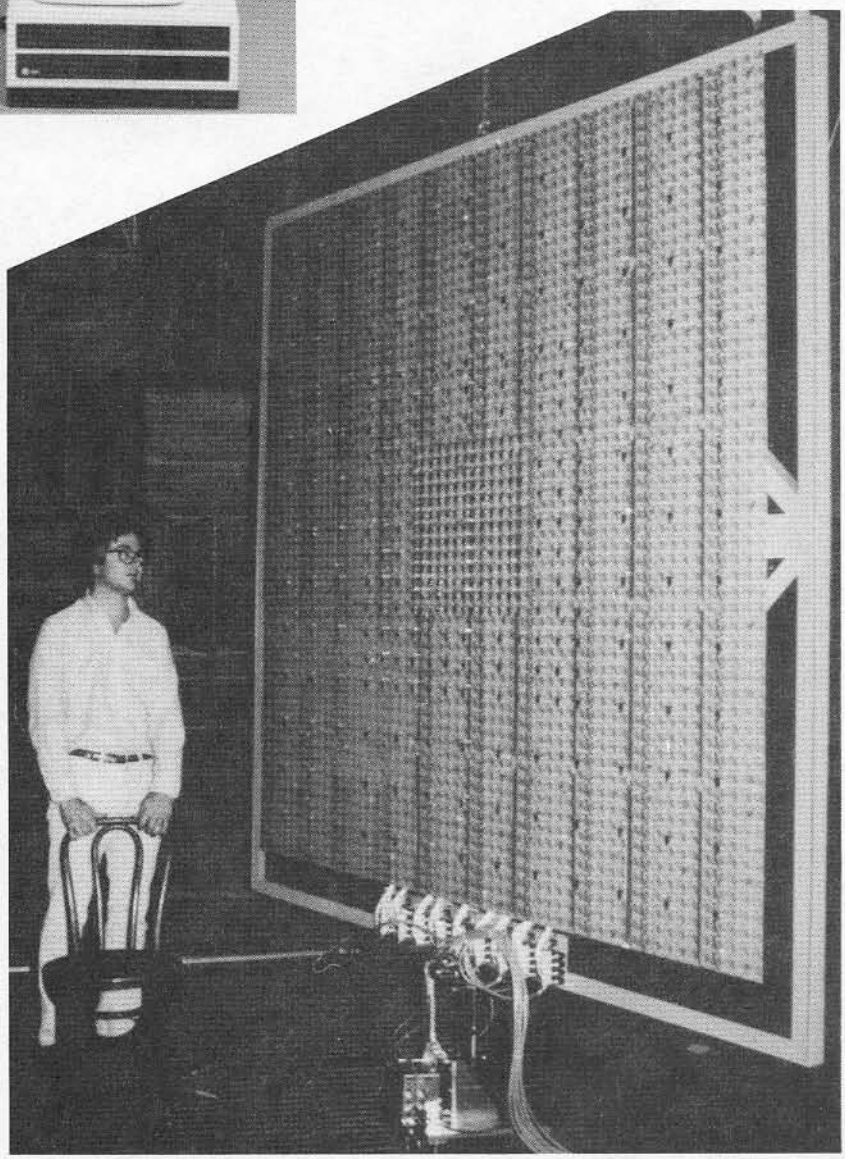


Fig. 2. (a) One-dimensional and (b) two-dimensional nested arrays of electret microphones.

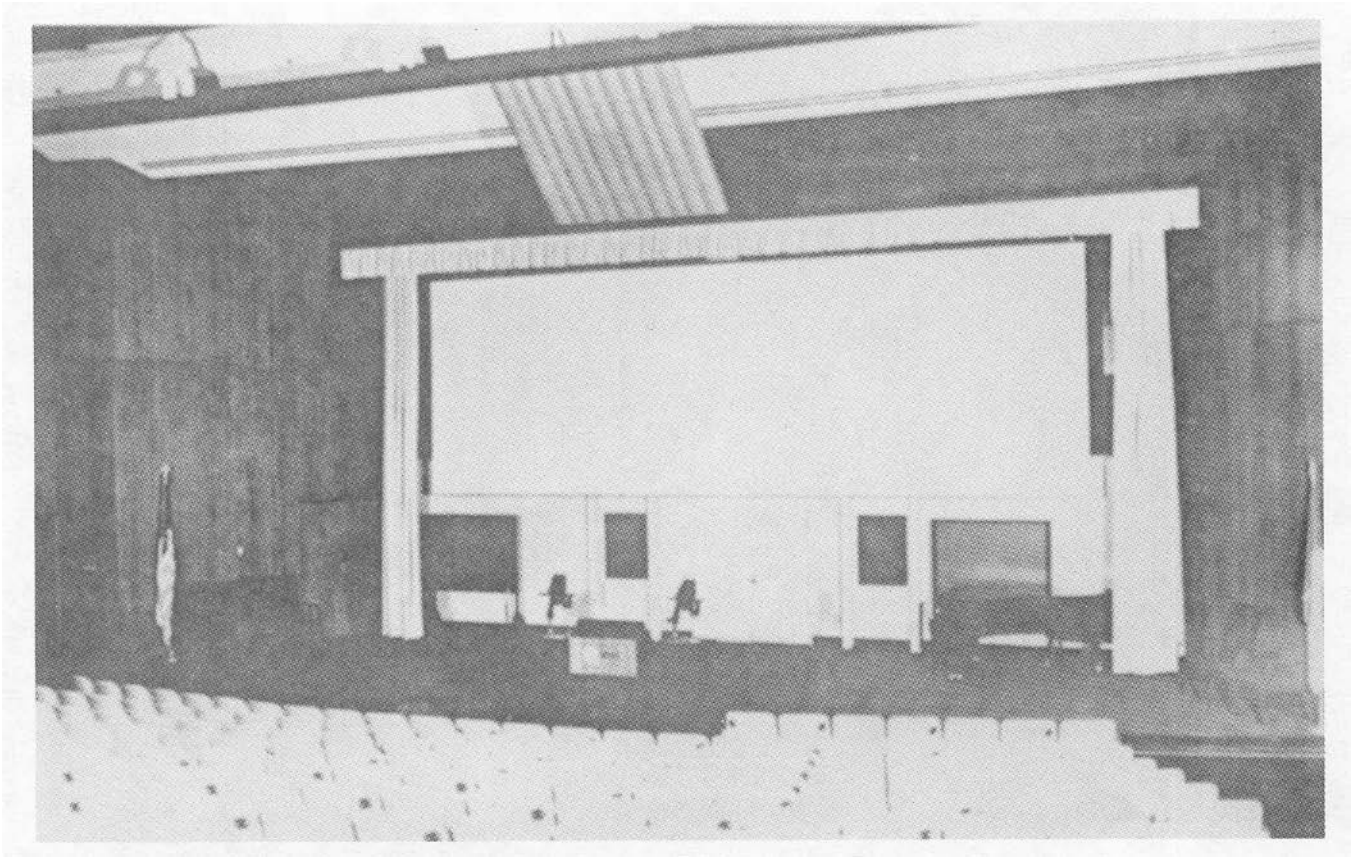
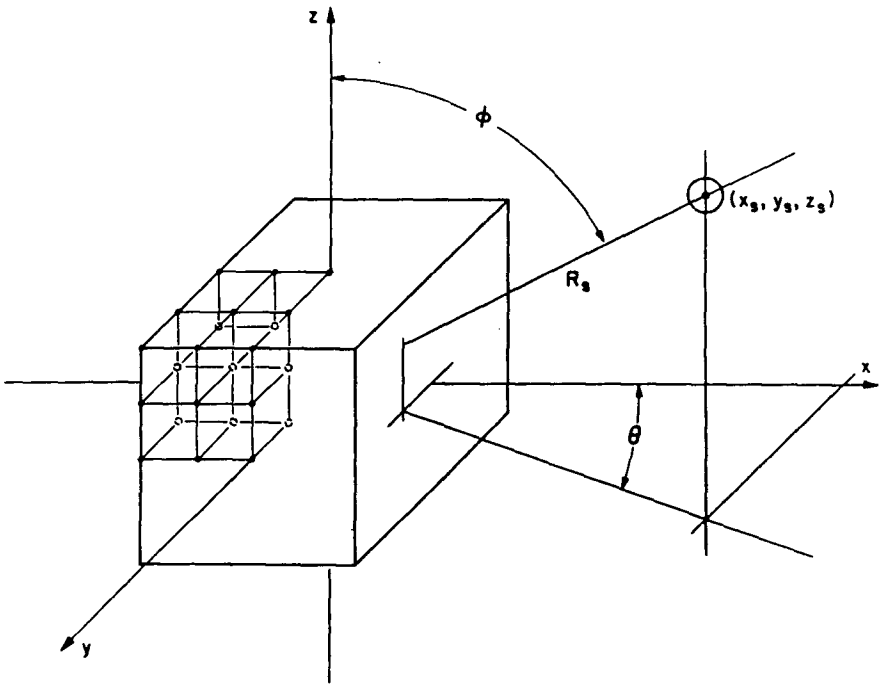
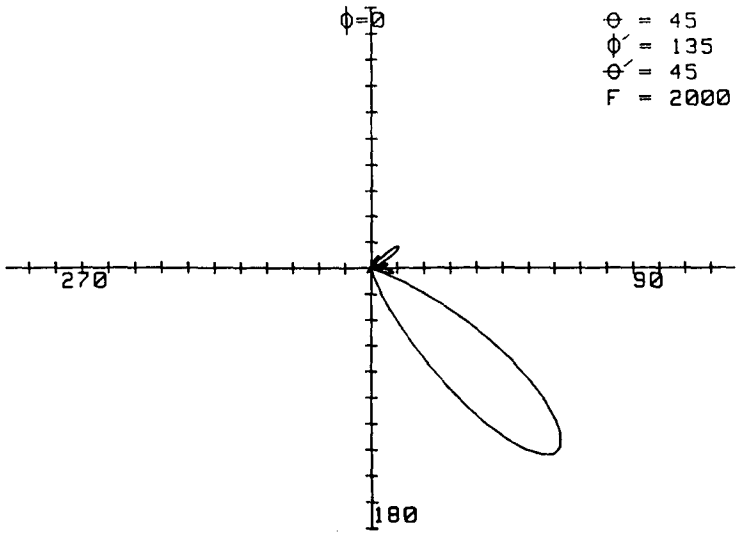


Fig. 3. (a) Auditorium installation of a 2D autodirective array. (b) Teleconferencing application of a 1D autodirective array. The array provides input to a connected-word speech recognizer for controlling system features [4].





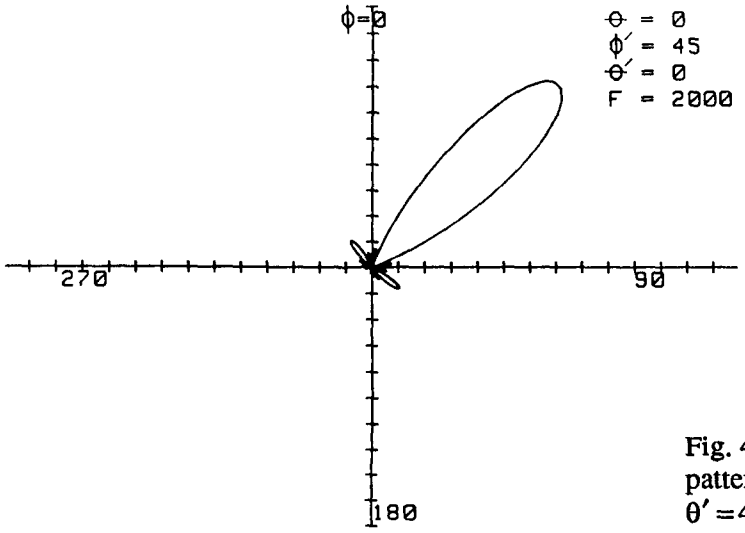


Fig. 4. (a) geometry of a cubic array, unique beam patterns for steering the cubic array to (b) $\phi' = 135°$, $\theta' = 45°$, and (c) $\phi' = 45°$ and $\theta' = 0$, respectively.