AN EXPERIMENT THAT VALIDATES THEORY WITH MEASURMENTS FOR A LARGE-APERTURE MICROPHONE ARRAY

Harvey F. Silverman, William R. Patterson III and Joshua M. Sachar

Laboratory for Engineering Man/Machine Systems (LEMS)
Division of Engineering, Brown University
Providence, RI 02912 [*email: hfs,wrp,or jms @lems.brown.edu]*

ABSTRACT

Poor sound pick up by remote microphones in multimedia applications, conference rooms and auditoria has traditionally hampered speech recognition and communication among spatially-separated groups. The problems are reverberation, acoustic noise, and the variability of the radiation pattern of unconstrained talkers. One potential solution that is becoming increasingly practical is to use an array of microphones and sophisticated signal processing. In this paper a brief description of a large, real-time, working system is presented and its measured beamforming performance is compared to what is predicted from a mathematical model. A combination of a synchronized test signal/system and a careful mathematical model results in the performances matching surprisingly well. From this match of theory to practice, we are able to draw some important inferences about future system improvements.

1. INTRODUCTION

There is now much commercial activity using very small microphone arrays for controlling teleconferencing cameras and for speech recognition entry. Several special workshops have been held on array technology [1, 2, 3, 4]. For the last year or so, we have been investigating the behavior of a large-aperture microphone array system [5, 6, 7, 8] to better understand its properties.

In this paper, we present some recent measurements from a real-time system of 256 microphones and compare them with results computed from an idealized mathematical model. The array is mounted in a highly reverberent room with a significant amount of background noise so, to compare mathematics with measurement, a special test-signal system had to be developed and the mathematical model had to be adapted to accommodate expected behavior.

The Huge Microphone Array (HMA), fully described in [5, 6], can support up to 512 microphones. Our results are based on the current set of algorithms that 1) locate a talker, 2) decide whether the determined location is "good" in some sense, 3) apply an interpolating time-domain, delay-and-sum beamformer, and 4) frequency shape the output. The latency of the system is currently about 125ms; this value is psychoacoustically intolerable for applications such as the pickup and amplification of actors in a play, but is adequate for remote applications such as video conferencing, recording the proceedings of a meeting or providing input to a speech recognizer.

2. THE ARRAY AND TESTING ENVIRONMENT

An array of 256 microphones is currently being used in a square room of 8.4M on a side. The floor is hard tile and the cement ceiling is three meters above the floor. The ceiling has regular rectangular boxlike cavities that are about 4Mx1Mx0.3M.

The array system has eight 1.34Mx0.67M panels. Each panel is an aluminum-framed piece of foam that is 6cm thick onto which 32 omnidirectional electret microphones have been placed in a random pattern. To insure a minimum separation, the pattern is random subset of the nodes of a 3cmx3cm grid. The panels are hung on walls or suspended from the ceiling in the pattern shown in Figure 1. The 256 microphones span three adjacent walls in the rectangular room. A photograph showing one corner of the

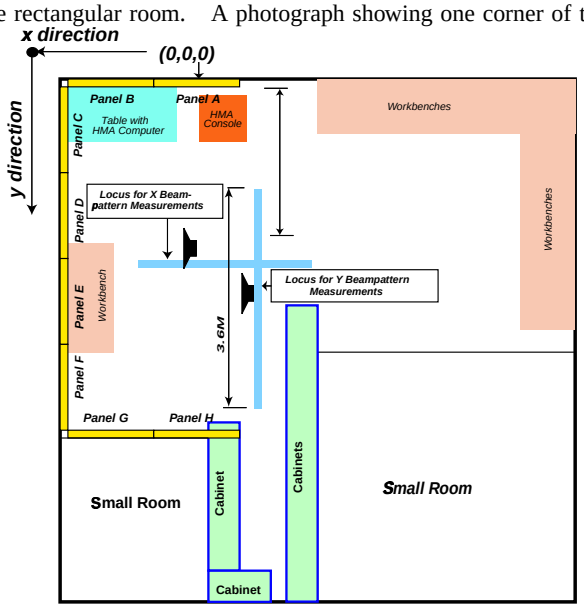

Fig. 1. Top View of the Acoustic Environment showing the Positions of the Panels and of Test Sources

array and the support console is shown in Figure 2. Using the person as a referent, one can judge the size of the room and panels and observe the ceiling pattern and the equipment of the normal laboratory environment.

3. ALGORITHMS AND SOFTWARE

Current software tracks the location of a talker in real time and aims the beamformer continuously, or the beamformer may be

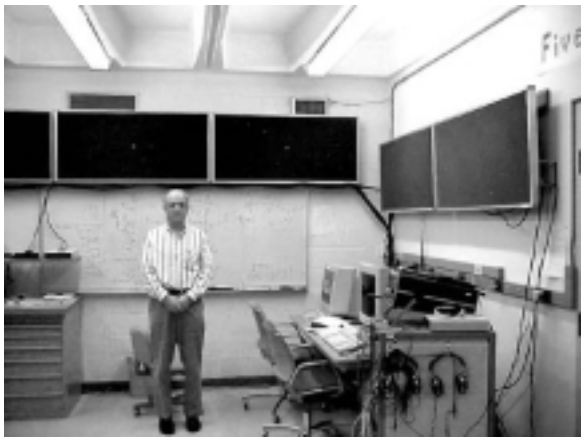

Fig. 2. Photograph of one Corner of the Array and Room

aimed at a fixed point. The output signal of the beamformer, $\hat{S}(t)$, that is supposed to be a close replica of the original speech signal $S(t)$, is computed by delay-and-sum beamforming the 256 microphone signals as given by

$$\hat{S}(t) = \sum_{i=1}^{256} W_i \cdot M_i(t - \tau_i) \qquad (1)$$

where M_i is the microphone signal for microphone i, W_i is a weight for microphone i, and τ_i a time-delay for microphone i. The time-delays and weights may each be functions of the source location and the aiming point. When new microphone-weight and source-location data are available, the software tests to see whether the source position has changed. If a new aiming location has been detected, then new specific delays are computed based upon the differences in the time of arrival expected for each microphone relative to a selected referent microphone. Using the new delays, a lowpass filter that interpolates with a 64:1 improvement in time resolution is used to generate the delayed data sequence at 20kHz for each microphone. These are summed together in phase. Additional spectral processing may be performed on the array output to eliminate noise that the beamformer has not attenuated. The current delay in the beamforming system is 3.5 frames, or 89.6ms.

The locationing algorithm is shown in Figure 3. It was suggested by Prof. Michael Brandstein of Harvard University and is based on the *phase transform* method of [9]. This concept has been adapted to the new array and computational environment.

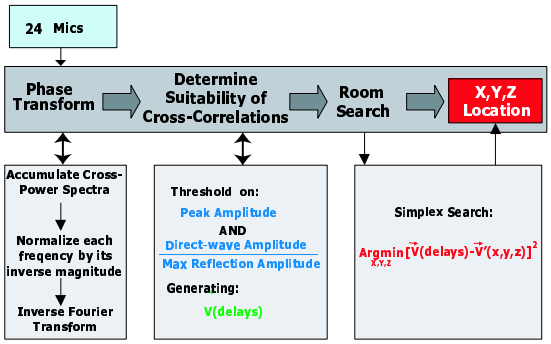

Fig. 3. The Locationing System

The location determination algorithm uses three processors and makes estimates of the source location every 200ms. Only 24 microphone signals are used in the current implementation. The 24 microphones are selected in eight groups of three, each group of three (or triangle) located in a half of one of the 1.34Mx0.67M panels. This implies that four panels are used; of the four panels, two are next to each other on one wall, and two are on an adjacent orthogonal wall.

Weights , W_i, for microphone signals, M_i, may be computed as: 1) **Uniform Weights** - the weights are the same, $W_i = 1/M$, 2). **Inverse-Distance Weights** - Under the assumption all the microphones suffer from about the same amount of uncorrelated noise, a known ideal is $W_i = \frac{1}{d_i}$ where d_i is the distance from microphone i to the source, and 3) **Signal-to-Noise-Squared Weights** - If separate estimates of per-channel signal and noise power are availablei, the optimal weighting is, $W_i = \frac{\bar{signal}}{\bar{noise}^2}$. The SNR is determined by keeping the 128 most recent values of the power for the overlapped 1024-point intervals, or about 3.2768 seconds; the signal level is taken as the sum of the highest three values, and the noise level is taken as the sum of the lowest three values.

4. THE SYNCHRONIZED TEST SIGNAL SYSTEM (STSS)

A calibration mechanism is essential for any large-aperture microphone array system in order to precisely determine the locations of the microphones in a new environment. For this purpose, we constructed a synchronized test signal system(STSS) so that a test signal could be played out of a transducer precisely synchronized (jitter of about 50ns) to the HMA. The STSS system also is a vital tool for measuring the performance of the array. The STSS design insured that time data for many synchronized, pulsed signals could be averaged and uncorrelated noise essentially eliminated from the measurement. Moreover, as the test signal was precisely windowed, the beginning of response was due to just the direct wave of the signal allowing reverberation effects to be eliminated from the measurement.

Consider the chirp in Figure 4. The original chirp, 8.2ms in length with the frequency varying from 2kHz to 6.3kHz in this time, is seen in a) and is degraded only by the sampling. The chirp

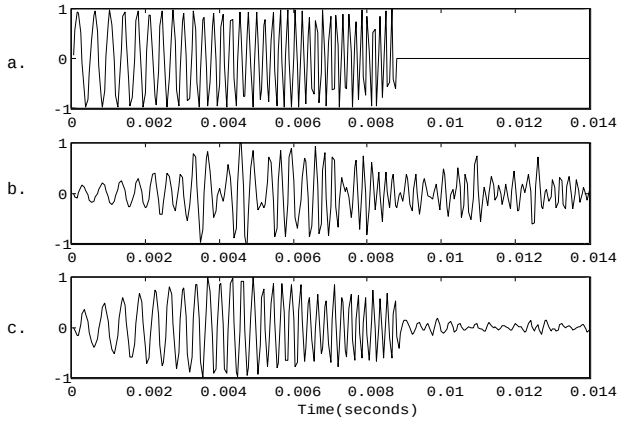

Fig. 4. a) Original Chirp, b) Received Signal at a Single Microphone, after Averaging over 10,000 Repetitions, c) Beamformed Composite of the single Microphone Signals

as heard by one of the 256 microphones is shown in b). Reverberations start after about 3ms and continue for a long time. The direct

wave, affected only by the response of the radiating transducer, is seen for the first 3ms. Ten-thousand chirp repetitions, each separated in time by about 1.2s to guarantee that any residual power from one chirp was essentially zero by the time of the next chirp were used to obtain the beamformed output in c). This technique reduced any effects from background noise by at least 40dB. It is clear that the beamforming significantly reduces the reverberations both during and after the chirp, and that the change in shape of the chirp is virtually all due to the response of the radiating transducer. This test signal very clearly demonstrates, therefore, the positive effects of the weighted-delay-and-sum beamforming, even for the case when uncorrelated noise is inconsequential. The reverberent quality of the desired signal has been reduced significantly.

The reverberation response of the room is shown in Figure 5. The reverberation time of a room, T_{60} is defined as the time

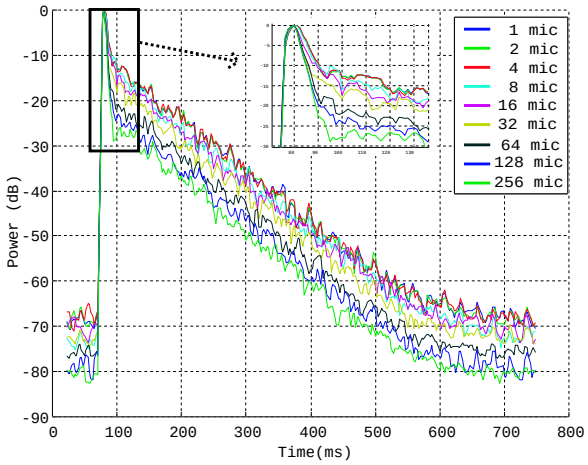

Fig. 5. Power Response for Various Numbers of Microphones for Large-Aperture Array

for the sound pressure level to decrease by 60dB after the source has been cut off. The reverberation time for a single microphone is about 450ms, while that for the 256-microphone beamformed signal is about 320ms. Moreover, as may be seen in the inset, post-chirp reverberant energy is about 20dB below that of a single microphone, an important and evident factor when listening to the real-time system.

5. MATHEMATICAL MODEL FOR THE BEAMPATTERN

The beampattern may be thought of as the magnitude of the power output of an array system that has been aimed to a single location $\vec{r}_A$ as the source is moved through the room to location $\vec{r}_S$. For a single frequency, the beampattern is manifest in the transfer function derived by exciting Equation 1 with a complex exponential and $t_A(i) \equiv \frac{|\vec{r}_A - \vec{x}(i)|}{c}$, $t_S(i) \equiv \frac{|\vec{r}_S - \vec{x}(i)|}{c}$ with c the speed of sound and $\vec{x}(i)$ the location of microphone i,

$$B(\vec{r}_S, \vec{r}_A, \omega) \equiv |\sum_{i=1}^{256} W_i \cdot e^{j\omega(t_A(i) - t_S(i))}|^2, \qquad (2)$$

If we wish to look at the performance of the array over a range of frequencies (e.g., a discrete set of K frequency values), we can weight each power value by a real, positive weight, $F(\omega)$, giving

$$\beta(\vec{r}_S, \vec{r}_A, F) = 10 \cdot \log[\sum_{l=1}^{K} F^2(\omega_k) \cdot B(\vec{r}_S, \vec{r}_A, \omega_k)]. \qquad (3)$$

However, $\beta(\vec{r}_S, \vec{r}_A, F)$ gives the power for the transfer function, or, equivalently by Parseval's theorem, the power in the impulse response. If we want to drive the system from the STSS and look only at the direct wave (the first 3ms), then, for valid comparison, we need to recreate the *response* in the time domain and truncate, taking the power in the first 3ms only. Figure 6 not only shows the need to do this to make a fair comparison, but also shows that the performance measure using only the first 3ms is not a realistic one. In b), the amplitude scale is significantly reduced from that of a), but the major part of the energy is from data outside of the first 3ms! This spreading of the unwanted energy is the normal spread due to time-shifting and has nothing to do with any additional unwanted energy due to reverberations.

Mathematically, this implies that if $q_3(n)$ is defined as the discrete time function for the 3ms chirp, and $Q_3(k)$ its DFT with k the index on frequency, we form

$$Y(\vec{r}_S, \vec{r}_A, k) \equiv B(\vec{r}_S, \vec{r}_A, k) \cdot Q_3(k), \qquad (4)$$

and take its inverse DFT yielding $y(\vec{r}_S, \vec{r}_A, n)$. We then take the energy over the first 3ms (60 points) as

$$\Psi(\vec{r}_S, \vec{r}_A) \equiv 10 \log[\frac{1}{60} \sum_{n=0}^{59} y^2(\vec{r}_S, \vec{r}_A, n)]. \qquad (5)$$

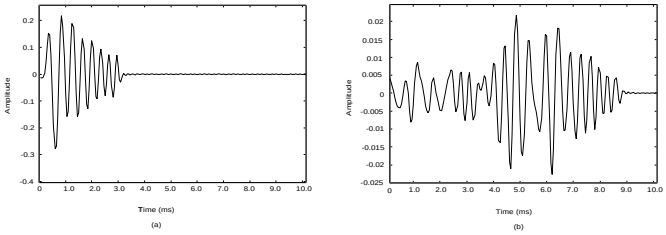

Fig. 6. Ideal Response Driven by 3ms Chirp at (a) the Aiming Point and (b) 2M from the Aiming Point

6. MEASURED VERSUS MATHEMATICAL BEAMPATTERN

Measurements along a line parallel to the Y axis and through the source (see Figure 1) are presented as three of the curves in each of Figures 7 and 8. The array was aimed at a fixed location, and the source transducer – a mid-range, 8cm dome speaker – was moved to each measurement location prior to capturing the beamformer's output. In each figure, the fourth curve is the mathematical computation. The test signal was the same chirp of Figure 4. The three measured curves are: 1) the direct wave only – i.e., the power in the first 3ms, 2) the power in the full 8.9ms chirp – a few early reverberations are included, but not the "ringing" energy, and 3) the power in the first 150ms including most of the "ringing" energy. The difference between the two figures is that in Figure 7 the microphones are uniformly weighted, both for the measurements and for the ideal data, while in Figure 8, each of the microphones is weighted by a suitably derived S/N^2 factor. The most notable

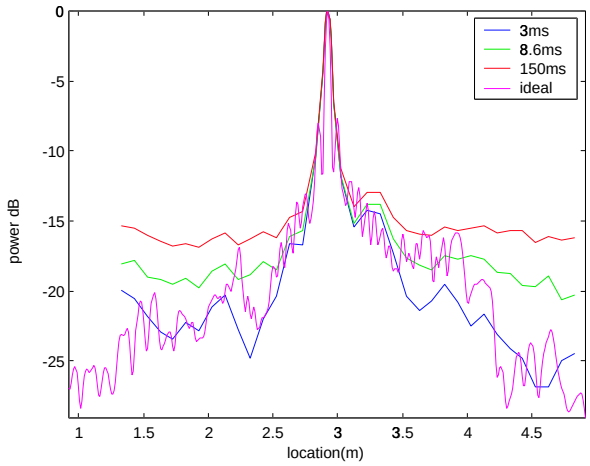

Fig. 7. Measured and Ideal Beampatterns for the Uniformly-Weighted Array over a Line Parallel to the Y Axis and through

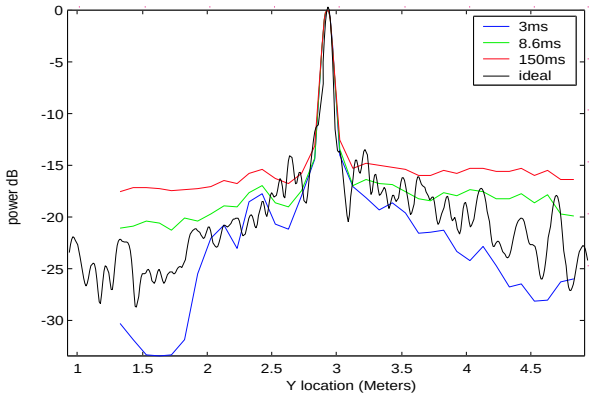

Fig. 8. Measured and Ideal Beampatterns for S/N^2-Weighted Array over a Line Parallel to the Y axis and through the Source

difference between the two weightings is the (expected) slightly larger beamwidth of the S/N^2 weighting.

Neither the measured performance nor the mathematically derived performance in Figure 7 are true measures of how the system behaves in the normal circumstance with background noise and real talkers. Perhaps the realistic performance for our room is characterized best by the curve for 150ms of data in which most of the reverberent energy is considered. However, the measurement and mathematical techniques used to derive the 3ms measured data and the ideal curve do afford us the ability to compare measurement to mathematics under remarkably similar conditions.

From Figures 7 and 8 we can observe:

- The beamwidths of the main lobe for the ideal and all the measurements match very closely.
- The overall shape of the 3ms measured data is quite similar to that of the ideal.
- The ideal curve shows ”oscillations” in the data throughout. We hypothesize that these are due to the assumption of a point source for the ideal computation, rather than a transducer having both a front and back component and an 8cm diameter.
- Both the modest amount of reverberent energy in the 8.6ms curve and the large amount of reverberent energy in the 150ms curve indicate a loss in off-axis attenuation of about 3dB in each case.
- S/N^2 weighting reduces the aperture and widens the beam width near the peak value.

7. CONCLUSIONS

We have shown that a predicted beampattern from an idealized, mathematical model matches measurements of a large-aperture system in a real room for a specialized excitation scenario. We thus can attribute degradations in large-aperture arrays with more confidence. We saw, for the first time, that off-axis performance for the 256-microphone system approached the expected 24dB, when reverberation and background-noise effects were eliminated, but that real performance significantly deteriorated when reverberations alone were considered. The performance also improved with signal-to-noise-squared weighting, particularly when the aiming point was close to some subset of the array, although methods for obtaining good estimates of the signal and the noise need to be developed. Finally, it is apparent that the source radiation pattern needs to be modeled and its effects incorporated into the beamforming system. We have seen large differences from the spherically-radiating point source that will severely impact the performance of a large-aperture array system. However, current listening performance of the HMA system with 256 microphones as described here is starting to approach that of a close-talking microphone.

8. REFERENCES

[1] J. L. Flanagan and H. F. Silverman. Material for international workshop on microphone-array systems: Theory and practice. LEMS Technical Report 113, LEMS, Division of Engineering, Brown University, Providence, RI 02912, October 1992.

[2] J. L. Flanagan and H. F. Silverman. Material for international workshop on microphone-array systems: Theory and practice. Technical report, CAIP, Rutgers University, Piscataway, NJ 08855-1390, October 1994.

[3] J. L. Flanagan and H. F. Silverman. Material for third biennial roundtable of microphone-array technology. LEMS Technical Report 151, LEMS, Division of Engineering, Brown University, Providence, RI 02912, October 1996.

[4] M. S. Brandstein, J. L. Flanagan, and H. F. Silverman. Material for fourth roundtable of microphone-array technology. Technical Report 151, Harvard University, Cambridge, MA, October 2000.

[5] H. F. Silverman, W. R. Patterson III, and J. L. Flanagan. The huge microphone array (HMA)- Part I. *IEEE Transactions on Concurrency*, 6(4):36–46, October-December 1998.

[6] H. F. Silverman, W. R. Patterson III, and J. L. Flanagan. The huge microphone array (HMA) - Part II. *IEEE Transactions on Concurrency*, 7(1):32–47, January-March 1999.

[7] H. F. Silverman, W. R. Patterson III, and J. M. Sachar. Early results for a large-aperture microphone array system. In *Proceedings of SAM2000)*, pages 207–211, Boston, MA, March 1999.

[8] H. F. Silverman, W. R. Patterson III, and J. M. Sachar. First measurements of a large-aperture microphone array system for remote audio acquisition. In *Proceedings of IEEE International Conference on Multimedia and Expo*, New York, NY, July/August 2000. Session TP0, Paper 5.

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