ICSV14
Cairns • Australia
9-12 July, 2007

LOUD: A 1020-NODE MICROPHONE ARRAY AND ACOUSTIC BEAMFORMER

Eugene Weinstein[1*], Kenneth Steele[2†], Anant Agarwal[2 3‡], James Glass[3§]

[1]Courant Institute of Mathematical Sciences
251 Mercer Street, New York, NY 10012, USA

[2]Tilera Corporation
1900 West Park Drive, Suite 290, Westborough, MA 01581, USA

[3]MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139, USA

[*]eugenew@cs.nyu.edu
[†]ken@tilera.com
[‡]agarwal@csail.mit.edu
[§]glass@csail.mit.edu

Abstract

Recording speech and other sound is difficult in environments with a large amount of noise and/or crosstalk. In these environments, array microphones are needed in order to obtain a clean recording of desired speech. In this work, we have designed, implemented, and tested LOUD, a 1020-node microphone array. To the best of our knowledge and as documented by Guinness World Records [6], this is currently the largest microphone array in the world. We have implemented an acoustic beamforming algorithm for sound source amplification in a noisy environment, and have obtained preliminary results demonstrating the efficacy of the array. From one to 1020 microphones, we have shown a 13.7dB increase in peak SNR for a representative utterance, an 87.2% drop in word error rate (WER) with interferer present, and an 91.3% drop in WER without an interferer.

1. INTRODUCTION

Speech recognition, and sound recording in general, in the presence of significant noise or crosstalk is difficult. When sound is recorded in a noisy environment through a single microphone, proximity of the microphone to the speaker's mouth is essential for audio of sufficient quality for speech recognition. This proximity can not be achieved without tethered close-talking microphones. However, human friendly pervasive computing environments such as CMU's Aura [4] or MIT's Oxygen [15] are characterized by mobile users going about their daily business and preclude the use of tethered microphones. Arrays of microphones have a spatial extent that can be exploited along with the propagation qualities of a sound wave to detect, separate, amplify, and track speech sources in a noisy environment.

We have created a modular microphone array called the Large acOUstic Data (LOUD) array, which currently consists of 1020 microphones. Our motivation for building a large array is that the performance of a microphone array improves linearly as the size of the array grows. This is well established in the theoretical literature on microphone arrays (e.g., [2, 23]), and our experimental results given in Section 5 confirm this in practice. LOUD is also novel in that it utilizes a new scalable parallel processing architecture developed in our lab called Raw [22], which is specifically designed to handle large volumes of streaming data. The details of LOUD's use of Raw are beyond the scope of this paper; this information may be found in our technical report [25].

We begin the paper with an overview of related work in Section 2. We then outline the details of our microphone array implementation in Section 3. Section 4 presents the setup and methods used in our experiments to evaluate the array. In Section 5, we present and discuss the results of our experiments. In Section 6, we conclude and outline our plans for future work

2. RELATED WORK

Sensor arrays have been extensively explored in the past half-century, initially as a tool for radar-based tracking of objects [18], and then for a number of other applications including radio astronomy [8], sonar systems [14], and seismology [9]. Over the past two decades, arrays of microphones (i.e., acoustic sensors in air) have been increasingly used for sound source separation and amplification, and since the late 1980s have been explored as a tool for capturing audio in difficult acoustic environments [3, 12].

Microphone arrays have quickly become popular as an aide for speech recognition, and several recent projects report significant improvements in recognition performance when using a microphone array when compared to a single omnidirectional microphone. For instance, Moore *et al.* [13] report a near three-fold decrease in recognition error rates using a circular array of eight microphones in a conference room, and Sullivan [21] reports similar gains with an eight-microphone linear array. However, all of these used substantially smaller arrays than the one presented in this paper.

There is a handful of larger arrays in existence. Intermediate sized arrays of 32, 64, and 64 microphones, have been studied by Wilson *et al.*, Havelock, and Stanford, respectively [26, 7, 19]. Flanagan *et al.* describe a 400-microphone square array two meters on a side that has been used to record speech in a large auditorium [3]. Finally, the array of 512 microphones described by Silverman *et al.* [17], has to our knowledge been the largest microphone array to date. However, the publications stemming from this work mostly use a 16-microphone subset of the large array. For instance, Adcock showed improved recognition performance for this smaller array size [1]. Based on investigation of past work, it appears that there is little or no published work on speech recognition experiments using microphone arrays with numbers of microphones close to that of the array presented in this paper.

3. IMPLEMENTATION

3.1. Hardware

We have opted to create small microphone modules to ensure LEGO-like modularity in the design of our array. The 1020-node microphone array (Figure 1) consists of 510 printed circuit

boards (PCBs) each of which contains two Panasonic WM-54BT Electret Condenser microphones. Each PCB also contains one stereo A-to-D converter (Cirrus Logic CS53L32A), and a small CPLD (Xilinx Coolrunner XCR3032XL). The A-to-D converter samples at 16 KHz, generating 24-bit serial data for each microphone. Our decision to place two microphones on one PCB was mainly due to the fact that the A-to-D converter is able to accommodate two channels of audio. The two-microphone boards are connected in chains of 16 boards (32 microphones), and each chain plugs into a connector board. The data are streamed through the chain and into the connector board using time-division multiplexing. Each connector board takes eight chains, and four connector boards are used to accommodate a maximum of 1024 microphones in total.

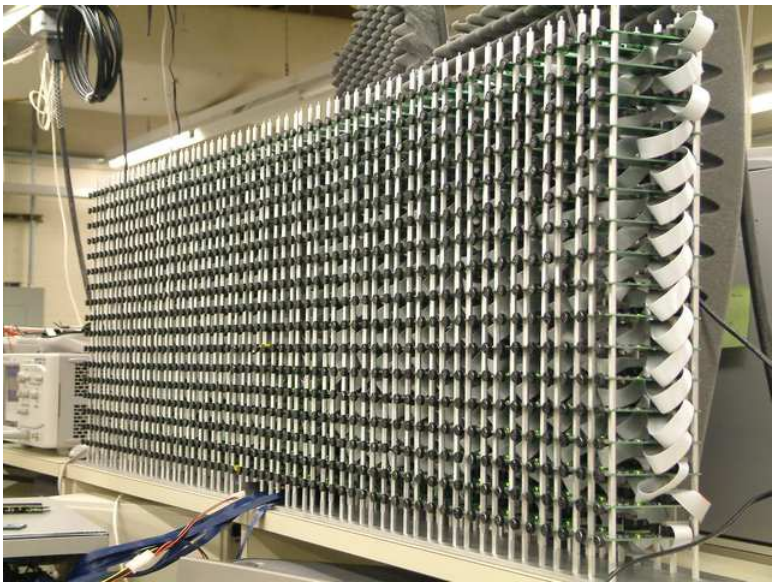

Figure 1. A picture of the LOUD 1020-node microphone array

3.2. Algorithm

In order to selectively amplify sound coming from a particular source or multiple sources, we use a technique called beamforming. Beamforming algorithms filter the sound signal spatially by taking advantage of the properties of sound propagation through space. Currently, we are using a delay-and-sum beamforming algorithm [24, 2], the simplest way of computing the beam. As the name suggests, delay-and-sum beamforming applies a delay to the sound signal recorded at each microphone, and adds the shifted signals together. Mathematically, if $x_i[n]$ is the sample recorded at the ith microphone at time n, τ_i is the delay for the ith microphone, and M is the number of microphones in the array, the output of delay-and-sum beamforming at time n is computed as

$$y[n] = \sum_{i=1}^{M} x_i[n - \tau_i] \tag{1}$$

Each delay τ_i is exactly the time required for sound to travel from the target source to microphone i. These delays can be empirically measured or calculated from the array geometry and are different for each sound source location. By delaying the signal from each microphone

by the propagation delay, we selectively amplify sound coming from a particular direction. Sound coming from other directions is attenuated, with the amplification and attenuation characteristics dictated by the array geometry. More information on this technique is available in many sources [2, 23, 24]. Our delay-and-sum beamforming algorithm runs in real time on the Raw processor (see our technical report [25] for details).

For the work presented in this paper, we assume that the position of the speaker is known in advance (however, Section 6 outlines our work in source localization). Hence, it is necessary to determine the sound propagation delay from the source to each microphone in the array. This is done in advance using the following procedure. A broadband "chirp" (frequency sweep) is played through a small loudspeaker located at the point marked as the sound source location. A reference recording is obtained with a single microphone at the loudspeaker position. The audio captured by each microphone is also captured and stored on disk. The recordings are then upsampled by a factor of 50 to obtain sub-sample precision in the calculation. A cross-correlation function (basically a dot-product at every possible time offset) is then calculated between the reference recording and the signal from each microphone. The time shift that maximizes the cross-correlation is taken as the propagation delay for that microphone. If the sound recorded by microphone of interest and the reference microphone at sample n are given by $x[n]$ and $r[n]$, respectively, then the time shift between them is determined as

$$\tau = \arg\max_{n} \sum_{m=-\infty}^{\infty} x[m+n]r[m] \tag{2}$$

Anecdotally, this method has proven more accurate than calculating microphone delays from array and point geometry. This can likely be attributed to some variability in microphone positions on the boards.

4. EVALUATION

We have conducted preliminary experiments with the LOUD array. The task was to recognize the speech of a person in a very noisy hardware laboratory. The main noise sources were several tens of cooling fans for computers and custom hardware, and a loud air conditioner. The subject read a series of random digit strings, and the speech was simultaneously recorded with the LOUD microphone array and a high-quality noise-canceling close talking microphone (Sennheiser HMD-410). In some of the trials, another person served as the interferer, reading a text passage (the "Rainbow Passage") at the same time as the main speaker was speaking. The interferer scenario models a situation where several people in a room are talking, but we are interested in recording the voice of only one person, such as a conference or a surveillance situation.

The experimental setup was as follows. The array was positioned on a counter 145 cm from the ground. The main speaker stood in line with the left edge of the array, 88.5 cm to the left of the center of the array, 137 cm in front of the array, with mouth position approximately 25 cm above the bottom row of the array. The interferer stood at a mirror image point in line with the right edge of the array, or 88.5 cm to the right of center, and 137 cm in front.

On each recording, we performed a beamforming run for each of 23 different microphone configurations, ranging from one microphone to all 1020 microphones, and stored the resulting audio streams to disk. This process simulated simultaneously recording the same audio stream

with arrays of varying sizes, allowing us to evaluate the effect of the number of microphones on array performance. Due to hardware constraints at the time of data collection, it was not practical to record the output of each of the 1020 microphones; however, this capability is now in place.

For this initial round of experiments, we recorded 150 utterances from two male native English speakers with an interferer, and 110 utterances from the same speakers without interferers. In order to provide a baseline for the speech recognition experiments, we also simultaneously recorded 80 utterances with interferer using a close-talking microphone. In the interferer trials, the person not serving as the subject served as the interferer. Certainly, much more extensive testing is necessary in order to evaluate the microphone array in sufficient detail (see section 6).

The MIT SUMMIT recognizer [5] trained on a combination of clean and noisy speech from the Aurora digits corpus [10] was applied to the recorded speech. We note that due to the channel differences between the close-talking microphone used to record the Aurora data and the LOUD microphone array, it is unrealistic to expect the array test data to match accuracy rates previously reported for Aurora.

5. RESULTS

Figure 2 gives approximate peak SNRs for a representative utterance, displaying the trend of improvement as the number of microphones is increased. The close-talking microphone, with an SNR level of 35.0dB, serves as the baseline. The SNR improves from 17.2dB with one microphone to 30.9dB with all 1020 microphones.

Figure 3 gives the WERs for the experimental data that we have collected, for all the array sizes ranging from one microphone to all 1020. Our baseline WER is for a close-talking microphone, at 1.2%. This is consistent with results from the Aurora corpus when tested on clean speech [10], meaning that our speech recognizer performs on a level consistent with other state-of-the-art recognizers. The WER drops from 97.0% for one microphone to 12.4% for the full array (an 87.2% drop) in the presence of an interferer, and from 92.9% to 8.0% (a 91.3% drop) without an interferer.

5.1. Discussion

The results in Figures 2 and 3 demonstrate the benefit of using arrays of this size. In this work, we focused on the design of the system, and did not implement sophisticated beamforming algorithms or other signal processing software components. However, even with the simplest beamformer possible, we were able to obtain increasing SNRs and gains in recognition accuracy from one to 1020 microphones.

The most drastic jump in the recognition accuracy curve is seen when the number of microphones goes from 32 to 60. This could be because this completes the full line of the array (60 microphones), making the beam width almost twice as narrow as with 32 microphones. After this point, adding more microphones does not make the array wider, just taller. Another reason for the jump is that the landmark component of the recognizer [5] was not optimized for low SNRs. We note that the WER even with 1020 microphones (12.4% and 8.0%) is clearly significantly short of the 1.2% baseline from the close-talking microphone; and this is consistent with the SNRs noted in the recordings. However, with more complicated signal processing and beamforming algorithms and a better match between the recognizer training and test conditions

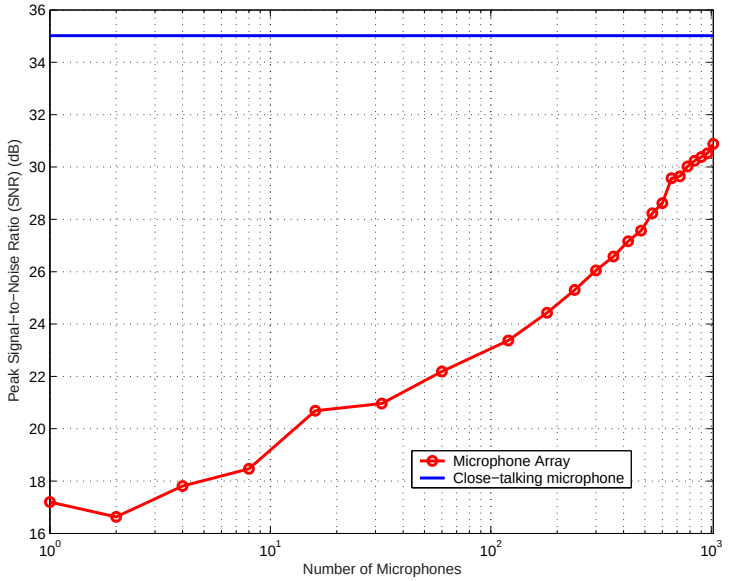

Figure 2. Peak SNRs for one representative recording.

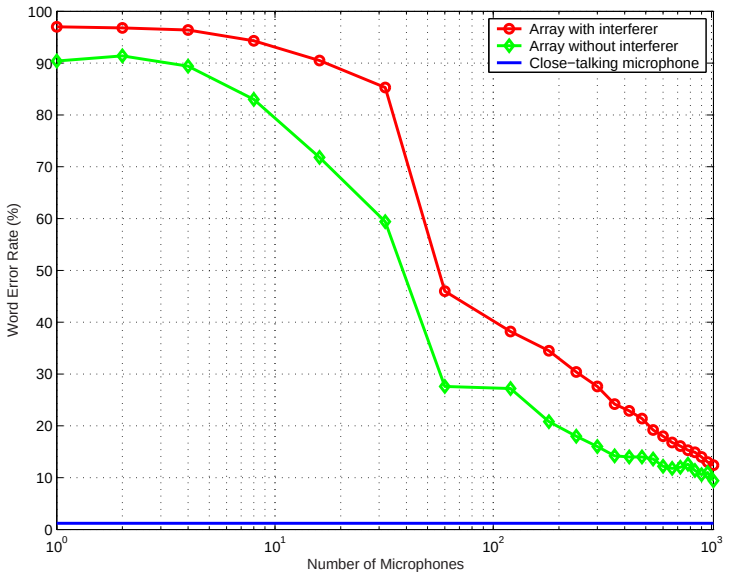

Figure 3. Experimental word error rates.

(see Section 4), we are confident that the recognition accuracy of audio recorded with the array can approach that of a close-talking microphone.

Comparison with past work is difficult for several reasons. One reason is differences in experimental conditions. Our data were collected in a very noisy environment; likely noisier than most of the currently-published results. For instance, Moore *et al.*[13] cite an accuracy rate of 42.4% with a single omnidirectional microphone and one interferer, compared to our 3.0%; Adcock [1] is at 58%; and Sullivan [21] is at 33%. While SNR is one intuitive way of comparing noise levels, it is actually difficult to compare based on SNR, since the various methods for determining SNR can produce very different results. In fact, there is much opinion that SNRs may not be a good measure of speech quality at all [16, 1].

In general, it can be noted that many in the scientific community feel that researching large

arrays is impractical and unbeneficial. While this work by no means constitutes an exhaustive set of experiments, we hope that by demonstrating a consistent improvement with increasing array sizes, we have shown that, at least at the current time, this belief is not entirely accurate. While our accuracy rates even with a 1020-microphone array fall short of that of recordings made with a close-talking high-quality microphone, we believe that this merely serves to motivate future research in adaptive beamforming (e.g. [20]) and other advanced sound source selection techniques (see Section 6).

6. CONCLUSION AND FUTURE WORK

In this work we have introduced LOUD, a 1020-node microphone array and beamformer. We have presented SNRs and recognition accuracy scores for 23 different array sizes, ranging from one to 1020 microphones, showing a steady improvement all the way to 1020 microphones. We believe that with these results we have made the case that large microphone arrays deserve a thorough investigation.

In order to evaluate the quantitative performance of the array further, more speech data must be collected from more speakers, in more points, in different noise environments, with different array configurations and spacings, etc. We plan to produce an array microphone recording corpus, which we will make available for public distribution on the web. Other array corpora have been made available in the past, but only with a smaller number of microphones (e.g. Jan *et al.* [11] have data from 37 and 23 microphones).

We have also implemented an algorithm that allows us to track multiple speakers as they move in the space around the array, and selectively listen to any one of them. A video demonstrating this system is available at
http://cag.csail.mit.edu/mic-array/videos/. The efficacy of this approach has not yet been evaluated in controlled experiments; however anecdotally speaking the performance is promising. Future work will also consider the effect of room acoustics and environmental conditions (other than noise). Array performance can be affected by reverberations and distortions due to the room in which the array is located. We plan to measure array performance with different room configurations in order to understand these effects.

ACKNOWLEDGEMENTS

Many thanks are due to Karen Livescu, who set up the SUMMIT recognizer for use with the Aurora corpus, and provided assistance during this project. The authors also wish to acknowledge Arthur Baggeroer, Patrick Miller, and Kevin Wilson for their valuable advice on microphone arrays. This work was funded in part by DARPA and an industrial consortium supporting the MIT Oxygen alliance.

REFERENCES

[1] J. Adcock. *Optimal Filtering and Speech Recognition With Microphone Arrays*. PhD thesis, Brown University, May 2001.

[2] M. Brandstein and D. Ward, editors. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[3] J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, and M.M. Sondhi. Autodirective microphone systems. *Acustica*, 73(1):58–91, 1991.

[4] D. Garlan, *et al*. Project Aura: Toward distraction-free pervasive computing. *IEEE Pervasive Computing*, April/June 2002.

[5] J. Glass. A probabilistic framework for segment-based speech recognition. *Computer, Speech, and Language*, 17:137–152, 2003.

[6] *Guinness World Records*. 2007.

[7] D.I. Havelock. A large microphone array for outdoor sound propagation studies. In *Proceedings of the Acoustical Society of America*, Austin, Texas, December 1994.

[8] S. Haykin, editor. *Array Signal Processing*, chapter 5. Prentice Hall, Englewood Cliffs, NJ, 1985.

[9] S. Haykin, editor. *Array Signal Processing*, chapter 2. Prentice Hall, Englewood Cliffs, NJ, 1985.

[10] H-G. Hirsch and D. Pearce. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings ISCA ITRW ASR2000*, Paris, France, September 2000.

[11] E.E. Jan, P. Svaizer, and J.L Flanagan. A database for microphone array experimentation. In *Proceedings Eurospeech*, Madrid, Spain, September 1995.

[12] Q. Lin, E. Jan, C. Che, and J.L. Flanagan. Speaker identification in teleconferencing environments using microphone arrays and neural networks. In *Proceedings ESCA Workshop on Speaker Recognition, Identification and Verification*, pages 235–238, Switzerland, October 1994.

[13] D. Moore and I. McCowan. Microphone array speech recognition: Experiments on overlapping speech in meetings. In *Proceedings ICASSP*, pages 497–500, Hong Kong, April 2003.

[14] A. V. Oppenheim, editor. *Applications of Digital Signal Processing*, chapter 6. Prentice Hall, Englewood Cliffs, NJ, 1978.

[15] MIT Project Oxygen. http://oxygen.csail.mit.edu/, 2004.

[16] S.R. Quackenbush, T.P. Barnwell, and M.A. Clements. *Objective Measures of Speech Quality*. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[17] H.F. Silverman, W.R. Patterson, and J.L. Flanagan. The huge microphone array. Technical report, LEMS, Brown University, May 1996.

[18] M.I. Skolnik. *Introduction to Radar Systems*. McGraw-Hill, New York, 1980.

[19] V. Stanford. The NIST Mark-III microphone array - infrastructure, reference data, and metrics. In *Proceedings International Workshop on Microphone Array Systems - Theory and Practice*, Pommersfelden, Germany, May 2003.

[20] D.E. Sturim, *et al*. Tracking multiple talkers using microphone-array measurements. In *Proceedings ICASSP*, Munich, Germany, April 1997.

[21] T.M. Sullivan. *Multi-Microphone Correlation-Based Processing for Robust Automatic Speech Recognition*. PhD thesis, ECE Department, Carnegie Mellon Unvirsity, August 1996.

[22] M.B. Taylor, *et al*. The Raw microprocessor: A computational fabric for software circuits and general purpose programs. *IEEE Micro*, March/April 2002.

[23] H.L. Van Trees. *Optimum Array Processing*. Wiley-Interscience, 2002.

[24] B.D. Van Veen and K.M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5:4–24, April 2002.

[25] E. Weinstein, K. Steele, A. Agarwal, and J. Glass. LOUD: A 1020-node modular microphone array and beamformer for intelligent computing spaces. Technical Report MIT-LCS-TM-642, MIT/LCS, April 2004.

[26] K. Wilson, V. Rangarajan, N. Checka, and T. Darrell. Audiovisual arrays for untethered spoken interfaces. In *Proceedings ICMI*, October 2002.