Vision-Aided Acoustic Array Processing for Perceptive Environments

Kevin Wilson & Trevor Darrell

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: We seek to create a system that integrates microphone arrays and cameras for use in perceptive environments. The goal is to extract low-noise signals from one or more sound sources in the presence of competing sound sources and environmental noise.

Motivation: Speech recognition software has advanced to the point where it can be employed as a mode of input in perceptive environments; however, current speech recognition techniques require low-noise signals to achieve resonable recognition rates. Such signals can be obtained from close-talking microphones, but we believe that it is possible to obtain similar signal quality from a large microphone array.

Previous Work: Much work has been done to evaluate signal processing approaches for processing data from microphone arrays [3]. Some work has also been on simple systems for using vision-based tracking systems [4, 1, 2].

Most previous work has concentrated on linear microphone arrays tracking distant sound sources. In cases where vision is incorporated, most systems use only vision data for tracking, ignoring localization information that may be derived from the audio signals.

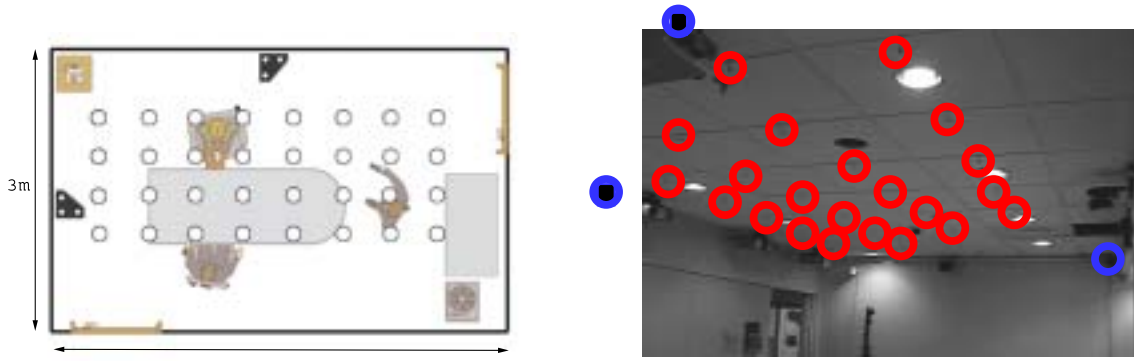

Figure 1: The test environment. On the left is a schematic view of the environment with stereo cameras represented by black triangles and microphones represented by empty circles. On the right is a photograph of the environment with microphones and camera locations highlighted.

Approach: To develop and test our audio and video processing algorithms, we have implemented a test system in a conference room in the AI Lab. The audio component consists of thirty-two microphones distributed across the ceiling of the perceptive space, and the video component consists of a three stereo cameras on the walls of the room (Figure). Together, these sensors allow us to localize both audio and video in space. We are now beginning to explore new algorithms for audio source separation using audio and video sensors.

Challenges: The linear microphone arrays used by most systems have only limited ability to localize sounds. We use a larger, two-dimensional array to improve localization performance. In order to process data from such a large array, new algorithms must be developed to reduce the computational requirements of the processing steps.

Placing a large microphone array within the perceptual space also violates two simplifying assumptions that are often made in array processing. First, sound sources can not be assumed to be located at infinite distance. Second, the space's acoustic characteristics will no longer be static; its characteristics will change whenever a user (or any other object) moves within the space.

Impact: The goal of the system is to be robust enough to replace closetalking microphones as the input device for speech recognition. This will allow users to enter and exit the space without pausing to put on or remove a closetalking microphone. In addition, by combining localization information derived from the audio data with localization information from the vision-based tracker, the system will be able to more robustly track users in the space.

The microphone array will continuously receive signals from the entire space, complementing the role of the cameras with their narrow fields of view but high spatial resolution. When the microphone array detects an unexpected sound source, steerable cameras could be directed toward the noise source to provide additional information.

Future Work: We plan to enable the audio system to adapt to slowly changing environmental noises. It will then be possible for the system to learn the characteristics of noise sources such as air conditioners or computer fans and to compensate for these noise sources through the use of filters that are matched to each noise source.

Research Support: This research is supported by MIT Project Oxygen.

References:

[1] U. Bub, M. Hunke, and A. Waibel. Knowing who to listen to in speech recognition: Visually guided beamforming. In *1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1995.

[2] M. Casey, W. Gardner, and S. Basu. Vision steered beam-forming and transaural rendering for the artificial life interactive video environment, (alive). In *99th Convention of the Audio Engineering Society*, 1995.

[3] C. Marro, Y. Mahieux, and K. U. Simmer. Analysis of noise reduction techniques based on microphone arrays with postfiltering. *IEEE Transactions on Speech and Audio Processing*, 6(3), May 1998.

[4] G. Pingali, G. Tunali, and I. Carlbom. Audio-video tracking for natural interactivity. In *Proceedings of ACM Multimedia 1999*, 1999.