Vision-Aided Acoustic Array Processing for Perceptive Environments

Kevin Wilson & Trevor Darrell

Artificial Intelligence Laboratory
Massachusetts Institue 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 [2]. Some work has also been on simple systems for using vision-based tracking systems [1, 3, 4].

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.

Approach: The audio component will consist of dozens of microphones distributed along the edges of the perceptive space. Part of the design process will be to determine a microphone arrangement that will provide a compromise among signal quality, spatial localization capability, and computational efficiency.

By cross-correlating signals from different microphones, a location estimate can be made for a sound source. Information from the vision-based tracking system will be used to improve the audio component's location estimate, and this estimate will be used to extract the desired source from other sound sources and from background noise.

A parallel direction of investigation is to enable the audio system to adapt to slowly changing environmental noises. It should 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 the specific noise source.

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

Placing a large microphone array within the perceptual space will also violate two simplifying assumptions that are often made when microphone arrays are used. 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 system will 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.

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 speakers.

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

Future Work: (This project began September 2000.)

Research Support: Project Oxygen, NTT.

References:

[1] G. Pingali, G. Tunali, and I. Carlbom Audio-Visual Tracking for Natural Interactivity. Proceedings of ACM Multimedia 1999, Orlando, FL.

[2] 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, Vol. 6, No. 3, May 1998.

[3] U. Bub, M. Hunke, and A. Waibel Knowing Who to Listen to in Speech Recognition: Visually Guided Beamforming. In *Proceedings of 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 9-12, 1995.

[4] M. Casey, W. Gardner, and S. Basu Vision Steered Beam-forming and Transaural Rendering for the Artificial Life Interactive Video Environment, (ALIVE) In *Proceedings of the 99th Convention of the Audio Engineering Society*, 1995.