Audio-Visual Source Separation for Untethered Interface

John Fisher & Trevor Darrell

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Audio-based interfaces usually suffer when noise or other acoustic sources are present in the environment. For robust audio recognition, a single source must first be isolated. In this project we show how multi-modal segmentation can be used to solve a version of the cocktail party problem, separating the speech of multiple speakers recorded with a single microphone and video camera.

Motivation: We would like to provide untethered audio-visual input for human-computer interface applications. Our goal is to support speech recognition with no attached microphone or wires.

Previous Work: Existing solutions to this problem generally require special microphone configurations, and often assume prior knowledge of the spurious sources.

Approach: We have developed new algorithms for segmenting streams of audio-visual information into their constituent sources by exploiting the mutual information present between audio and visual tracks. Automatic face recognition and image motion analysis methods are used to generate visual features for a particular user; empirically these features have high mutual information with audio recorded from that user. We show how utterances from several speakers recorded with a single microphone and video camera can be separated into constituent streams; we also show how the method can help reduce the effect of noise in automatic speech recognition.

Difficulty: This problem is challenging since the number of sources exceeds the number of microphones.

Impact: Unthethered interface with simple, commonly available sensors would be extremely valuable for human-computer interface applications.

Future Work: Adding prior knowledge about object structure directly into the learning framework.

Research Support: Project Oxygen, NTT