Signal Level Fusion for Untethered Audio-Visual Interfaces

John Fisher III & Trevor Darrell

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: In a crowded and noisy room, how is that you are able to "hear" a conversation without getting confused by interfering sounds? This phenomenon, commonly called the "cocktail party effect", depends somewhat on your ability to relate observed lip motion to the sounds that you hear. Sounds which are inconsistent with the observed changes are ignored, while sounds which are consistent are "enhanced". The challenge is to quantify and exploit that measure of consistency.

Motivation: The motivation for this work is twofold. In our focused efforts we would like to provide untethered audio-visual input for human-computer interface applications (e.g. speech recognition without attached microphones or wires). The broader goal is to better understand the learning principles for multi-modal fusion.

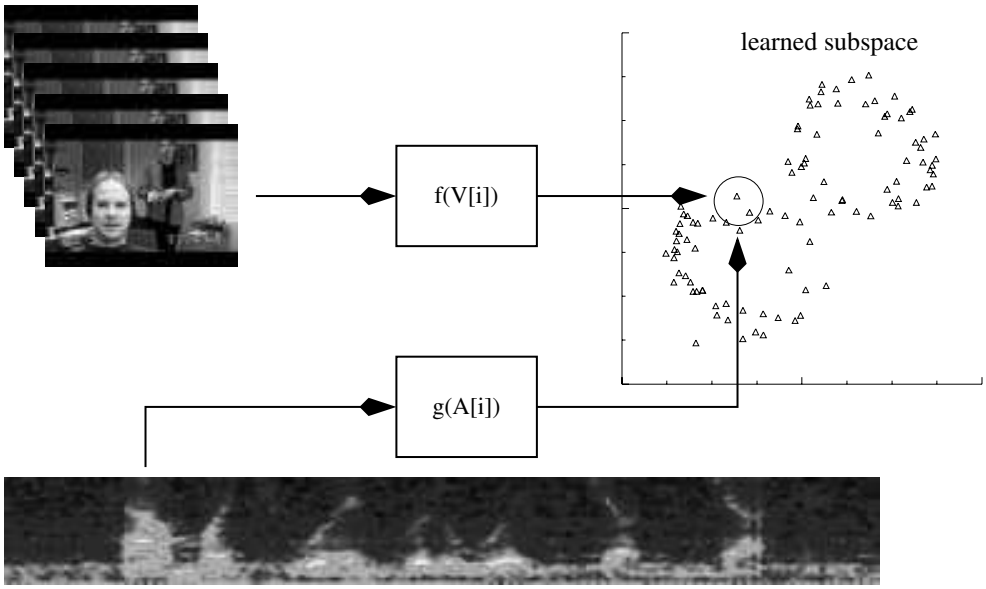

Figure 1: Basic fusion approach. High-dimensional frames from video and audio sources are projected to informative subspaces.

Previous Work: Existing solutions to this problem generally require special microphone configurations, and often assume prior knowledge of the spurious sources. Classical sensor fusion methods typically assume simple joint statistical models (e.g. jointly Gaussian) over the measurements. Such assumptions are not appropriate for audio/video data where the joint statistics are better modeled nonparametrically. Consequently, information theoretic principles naturally arise.

Approach: We are investigating an information theoretic approach to low-level audio-visual data fusion. The method, which is not specific to audio-visual data, is appropriate in that it has the capacity to model complex stochastic relationships. The idea is to learn functional mappings which map high-dimensional multi-modal data to a low dimensional space. There is one mapping for each modality. The mappings are chosen so as to maximize

the mutual information of the derived features.

To date we have demonstrated video localization of a sound source (as in figure 2) as well as the utility of the learned subspace for quantifying audio-visual consistency (i.e. "Did the sound I heard come from the face that I saw?").

Impact: Successful fusion of audio-video data will allow for a variety of alternatives in human-computer interfaces. Because we are concentrating on signal-level fusion, we believe that the results of this research will provide methods which are useful in difficult environments where higher level approaches tend to be brittle.

Future Work: Future work is focused on developing real-time versions of the fusion algorithm.

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

Figure 2: Left: One frame from image sequence. Speaker is in foreground with flickering monitor and moving person in background. Right: Magnitude of learned video projection using our method. Note that high intensities are concentrated around the speaker's mouth and chin.

References:

[1] John W. Fisher III, Trevor Darrell, William T. Freeman, and Paul. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems 13*, 2000.

[2] John W. Fisher III and Jose C. Principe. Information preserving transformations. *(in submission)*, 2001.