

Speech Enhancement by Audio-Visual Fusion in a Graphical Model

John Hershey, Hagai Attias

Machine Perception Laboratory, UCSD,
and
Microsoft Research, Redmond Washington

May 3, 2003

An interfering speaker or noise can significantly decrease speech recognition performance, whereas we know from human psychophysics that seeing the speaker's lips dramatically reverses this effect. Whereas audio-visual speech recognition experiments have confirmed the advantage of using video, the more difficult task of audio-visual speech enhancement has just begun to be explored. Speech enhancement promises application beyond speech recognition to auditory scene analysis and robust perception of paralinguistic speech information. To this end we have developed novel probabilistic graphical models capable of exploiting vision to aid in the enhancement of speech. We employ a variational EM algorithm to learn models of audio-visual speech that build an efficient low-dimensional representation of visual speech, track it spatially as it moves in the video, and correlate it with the acoustic speech signal. Models of noise are also adapted online in this framework. We demonstrate that the combination of the speech and noise models produces a significant improvement in speech quality when using video over the audio models alone.