

Audio-Visual Sound Separation Via Hidden Markov Models

John Hershey
Department of Cognitive Science
University of California San Diego
jhershey@cogsci.ucsd.edu

May 3, 2002

Abstract

It is well known that under noisy conditions we can hear speech much more clearly when we read the speaker's lips. This suggests the utility of audio-visual information for the task of speech enhancement. We propose a method to exploit audio-visual cues to enable speech separation under non-stationary noise and with a single microphone. We revise and extend HMM-based speech enhancement techniques, in which signal and noise models are factorially combined, to incorporate visual lip information and employ novel signal HMMs in which the dynamics of narrow-band and wide band components are factorial. We avoid the combinatorial explosion in the factorial model by using a simple approximate inference technique to quickly estimate the clean signals in a mixture. We present a preliminary evaluation of this approach using a small-vocabulary audio-visual database, showing promising improvements in machine intelligibility for speech enhanced using audio and visual information.