Optical Flow for Visual Speech Recognition

Michael S. Gray[1,3] **and Javier R. Movellan**[1] **and Terrence J. Sejnowski**[2,3]
Departments of Cognitive Science[1] and Biology[2]
University of California, San Diego
La Jolla, CA 92093
and
Howard Hughes Medical Institute[3]
Computational Neurobiology Lab, The Salk Institute
La Jolla, CA 92037
{mgray,jmovellan,tsejnowski}@ucsd.edu

Introduction

Visual speech recognition is a challenging task in sensory integration. Psychophysical work by McGurk and MacDonald (1976) first showed the powerful influence of visual information on speech perception that has led to increased interest in this area. For example, they presented the speech sound $/na-na/$ with the visible articulation of $/pa-pa/$. When subjects were asked to identify the sound that they *heard*, many reported $/ma-ma/$. This sound represents the best compromise of these two conflicting sources of information because $/ma-ma/$ is similar to $/pa-pa/$ visually, and $/ma-ma/$ is similar to $/na-na/$ acoustically. More generally, the visual signal provides good information about place of articulation, but voicing and nasality are more difficult to determine. The acoustic signal, on the other hand, has good information about voicing, but is ambiguous with respect to place of articulation.

Massaro (1987) used a fuzzy logic model to show that the best explanation of the data was obtained when the visual and acoustic information were treated as independent factors. In other words, human response probabilities for the different combinations of visual and acoustic signals were best described as the result of a process in which each modality makes an independent, multiplicative contribution. This property of the acoustic and visual signals is often referred to as *conditional independence*. Movellan (in press) further tested this conditional independence assumption using hidden Markov models. He compared models that were constrained to utilize visual and acoustic information independently with models that were unconstrained. Because the optimal constrained model did not perform any worse than the optimal unconstrained model, we can assume that the conditional independence assumption holds.

Current Directions

Movellan (1995) recently explored the ability of hidden Markov models to recognize spoken digits using visual information alone. The input representation for the model consisted of smoothed pixel intensity information at each time step, as well as a delta image that showed the pixel by pixel difference between subsequent time steps. Peak performance of this model (89%) closely matched the results of untrained human subjects.

In the current work, an optical flow representation rather than the delta image was used. Cells in area MST of visual cortex selectively respond to specific patterns of optical flow (Duffy & Wurtz, 1991). This flow information has been interpreted as representing egomotion, but may also be valuable for segmenting independently moving objects (Zemel & Sejnowski, 1995) and recognizing different patterns of lip movements. This higher-level visual representation may be more resistant to varying illumination conditions, and other forms of noise, than the pixel-based delta image.

Our optical flow computation was based on the standard *brightness constraint* equation, followed by thresholding. Experimentation with more sophisticated 2nd-order optical flow techniques resulted in extremely noisy output, presumably due to violation of the rigidity constraint. The optical flow representation formed the input to an HMM which was trained on spoken digits from a database of 12 speakers. Peak performance of 61% was obtained with a 9-state model. Adding information about the acceleration of lip features (differences in optical flow) resulted in an additional 10% improvement.

References

Duffy, C.J. & Wurtz, R.H. (1991). The sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large field stimuli. *Journal of Neurophysiology*, 65, 1329-1345.

Massaro, D.W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 126-130.

Movellan, J.R. (1995). Visual speech recognition with stochastic networks. In G. Tesauro, D.S. Touretzky, & T. Leen (Eds.), *Advances in Neural Information Processing Systems 7* (pp. 851-858). Cambridge, MA: MIT Press.

Movellan, J.R. (in press). Channel separability in the audio visual integration of speech: A Bayesian approach.

Zemel, R.S. & Sejnowski, T.J. (1995). Grouping components of three-dimensional moving objects in area MST of visual cortex. In G. Tesauro, D.S. Touretzky, & T. Leen (Eds.), *Advances in Neural Information Processing Systems 7* (pp. 165-172). Cambridge, MA: MIT Press.