Dynamic features for visual speechreading: A systematic comparison

Michael S. Gray[1,3], Javier R. Movellan[1], 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, P. O. Box 85800
San Diego, CA 92186-5800
Email: mgray, jmovellan, tsejnowski@ucsd.edu

Abstract

Humans use visual as well as auditory speech signals to recognize spoken words. A variety of systems have been investigated for performing this task. The main purpose of this research was to systematically compare the performance of a range of dynamic visual features on a speechreading task. We have found that normalization of images to eliminate variation due to translation, scale, and planar rotation yielded substantial improvements in generalization performance regardless of the visual representation used. In addition, the dynamic information in the difference between successive frames yielded better performance than optical-flow based approaches, and compression by local low-pass filtering worked surprisingly better than global principal components analysis (PCA). These results are examined and possible explanations are explored.

1 INTRODUCTION

Visual speech recognition is a challenging task in sensory integration. Psychophysical work by McGurk and MacDonald [5] first showed the powerful influence of visual information on speech perception that has led to increased interest in this

area. A wide variety of techniques have been used to model speech-reading. Yuhas, Goldstein, Sejnowski, and Jenkins [8] used feedforward networks to combine gray scale images with acoustic representations of vowels. Wolff, Prasad, Stork, and Hennecke [7] explicitly computed information about the position of the lips, the shape of the mouth, and motion. This approach has the advantage of dramatically reducing the dimensionality of the input, but critical information may be lost. The visual information (mouth shape, position, and motion) was the input to a time-delay neural network (TDNN) that was trained to distinguish among consonant-vowel pairs. A separate TDNN was trained on the acoustic signal. The output probabilities for the visual and acoustic signals were then combined multiplicatively. Bregler and Konig [1] also utilized a TDNN architecture. In this work, the visual information was captured by the first 10 principal components of a contour model fit to the lips. This was enough to specify the full range of lip shapes ("eigenlips"). Bregler and Konig [1] combined the acoustic and visual information in the input representation, which gave improved performance in noisy environments, compared with acoustic information alone.

Surprisingly, the visual signal alone carries a substantial amount of information about spoken words. Garcia, Goldschen, and Petajan [2] used a variety of visual features from the mouth region of a speaker's face to recognize test sentences using hidden Markov models (HMMs). Those features that were found to give the best discrimination tended to be dynamic in nature, rather than static. Mase and Pentland [4] also explored the dynamic information present in lip images through the use of optical flow. They found that a template matching approach on the optical flow of 4 windows around the edges of the mouth yielded results similar to humans on a digit recognition task. Movellan [6] investigated the recognition of spoken digits using only visual information. The input representation for the hidden Markov model consisted of low-pass filtered pixel intensity information at each time step, as well as a delta image that showed the pixel by pixel difference between subsequent time steps.

The motivation for the current work was succinctly stated by Bregler and Konig [1]: "The real information in lipreading lies in the temporal change of lip positions, rather than the absolute lip shape." Although different kinds of dynamic visual information have been explored, there has been no careful comparison of different methods. Here we present results for four different dynamic techniques that are based on general purpose processing at the pixel level. The first approach was to combine low-pass filtered gray scale pixel values with a delta image, defined as the difference between two successive gray level images. A PCA reduction of this gray-scale and delta information was investigated next. The final two approaches were motivated by the kinds of visual processing that are believed to occur in higher levels of the visual cortex. We first explored optical flow, which provides us with a representation analogous to that in primate visual area MT. Optical flow output was then combined with low-pass filtered gray-scale pixel values. Each of these four representations was tested on two different datasets: (1) the raw video images, and (2) the normalized video images.

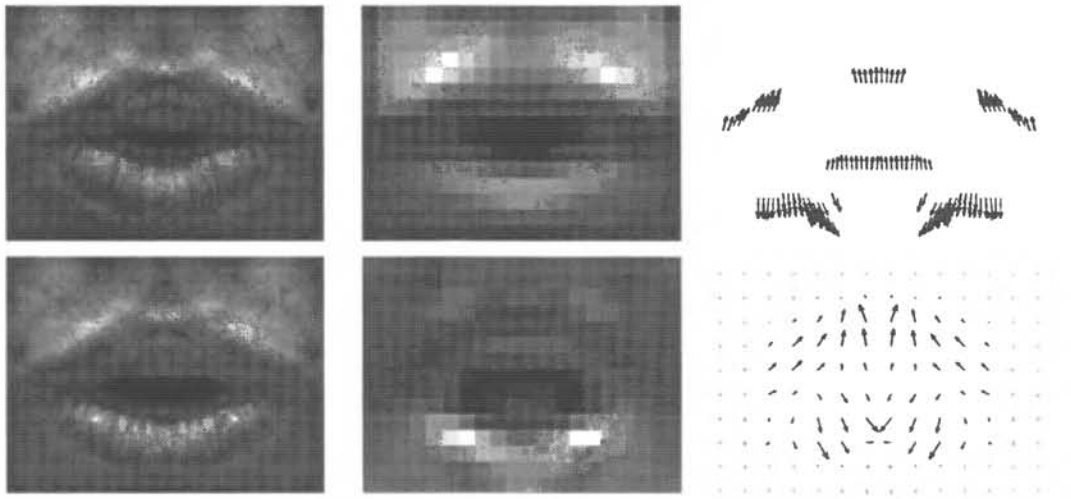

Figure 1: Image processing techniques. Left column: Two successive video frames (frames 1 and 2) from a subject saying the digit "one". These images have been made symmetric by averaging left and right pixels relative to the vertical midline. Middle column: The top panel shows gray scale pixel intensity information of frame 2 after low-pass filtering and down-sampling to a resolution of 15 x 20 pixels. The bottom panel shows the delta image (pixel-wise subtraction of frame 1 from frame 2), after low-pass filtering and downsampling. Right column: The top panel shows the optical flow for the 2 video frames in the left column. The bottom panel shows the reconstructed optical flow representation learned by a 1-state HMM. This can be considered the canonical or prototypical representation for the digit "one" across our database of 12 individuals.

2 METHODS AND MODELS

2.1 TRAINING SAMPLE

The training sample was the Tulips1 database (Movellan [6]): 96 digitized movies of 12 undergraduate students (9 males, 3 females) from the Cognitive Science Department at UC-San Diego. Video capturing was performed in a windowless room at the Center for Research in Language at UC-San Diego. Subjects were asked to talk into a video camera and to say the first four digits in English twice. Subjects could monitor the digitized images in a small display conveniently located in front of them. They were asked to position themselves so that their lips were roughly centered in the feed-back display. Gray scale video images were digitized at 30 frames per second, 100 x 75 pixels, 8 bits per pixel. The video tracks were hand segmented by selecting a few relevant frames before the beginning and after the end of activity in the acoustic track. There were an average of 9.7 frames for each movie. Two sample frames are shown in the left column of Figure 1.

2.2 IMAGE PROCESSING

We compared the performance of four different visual representations for the digit recognition task: low-pass + delta, PCA of (gray-scale + delta), flow, and low-pass