Proceedings of the 3rd Joint
Sympostuim on Neural
Computation, La Jolla, CA
Vol. 6, 222-230 (1996)

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 is to systematically compare the performance of a range of dynamic visual features on a speechreading task. We have found that compression by local low-pass filtering works surprisingly better than global principal components analysis (PCA). In addition, pixel-based representations yielded better performance than optical-flow based approaches. We examine these results and explore possible explanations.

1 INTRODUCTION

Visual speech recognition is a challenging task in sensory integration. Psychophysical work by McGurk and MacDonald [7] 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 [11] used feedforward networks to combine gray scale images with acoustic representations of vowels. Wolff, Prasad, Stork, and Hennecke [10] 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 (CV) pairs. A separate TDNN was trained on the acoustic signal. Because humans seem to combine acoustic and visual information in a conditionally independent fashion (Massaro & Cohen [6]), Wolff et al [10] combined the output probabilities multiplicatively. Bregler and Konig [2] 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 [2] combined the acoustic and visual information in the input representation, which gave improved performance in noisy environments, compared with acoustic information alone. Silsbee [9] has explored bimodal integration in automatic speech recognition with the development of a system that can control the relative contribution of the acoustic and visual information.

Surprisingly, the visual signal alone carries a substantial amount of information about spoken words. Garcia, Goldschen, and Petajan [4] used a variety of visual features from the mouth region of a speaker's face to recognize test sentences using hidden Markov models (HMMs). The feature space was reduced through were found through a correlation matrix, the use of principal component analysis, and heuristics. Those found to give the best discrimination tended to be dynamic in nature, rather than static. Mase and Pentland [5] 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 [8] has also explored 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 our current work was succinctly stated by Bregler and Konig [2]: "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 several different dynamic techniques that are based on general purpose processing at the pixel level. We have avoided model-based methods (e.g., contour fitting of lip position) for fear of losing critical information when we reduce dimensionality. We started with gray scale form information that is combined with a delta image. We then investigated a PCA reduction of this form and delta information. Our next three 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. We then combined optical flow information with the acceleration of

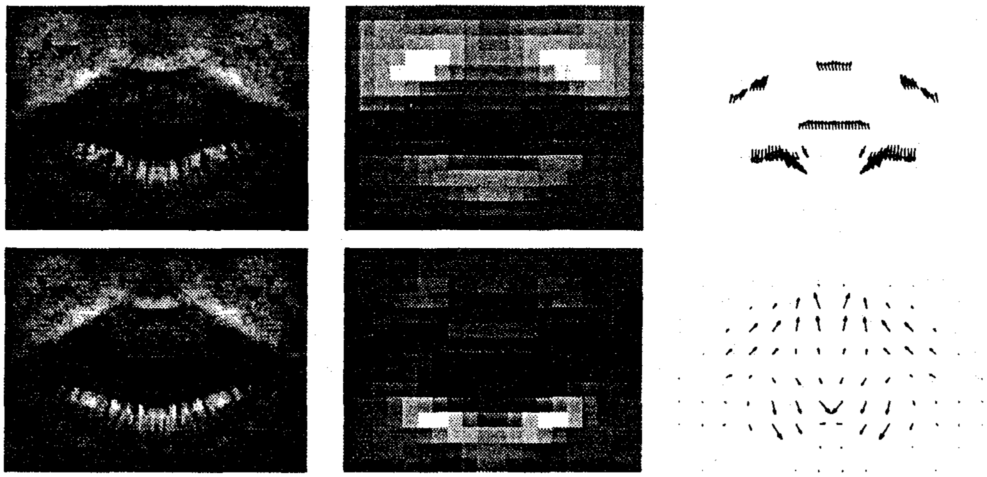

Figure 1: Image processing techniques. Left column: Two successive video frames (frames 1 and 2) from a subject saying the digit "one". These images have been made symmetric by averaging left and right pixels relative to the vertical midline. Middle column: The top panel shows gray scale form information 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.

visual lip features — the difference between subsequent optical flow vectors. Cells in area MST of visual cortex are known to selectively respond to specific patterns of optical flow, including acceleration (Duffy & Wurtz [3]). Finally, we conjoined form information with optical flow output.

2 METHODS AND MODELS

2.1 TRAINING SAMPLE

The training sample consisted of 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

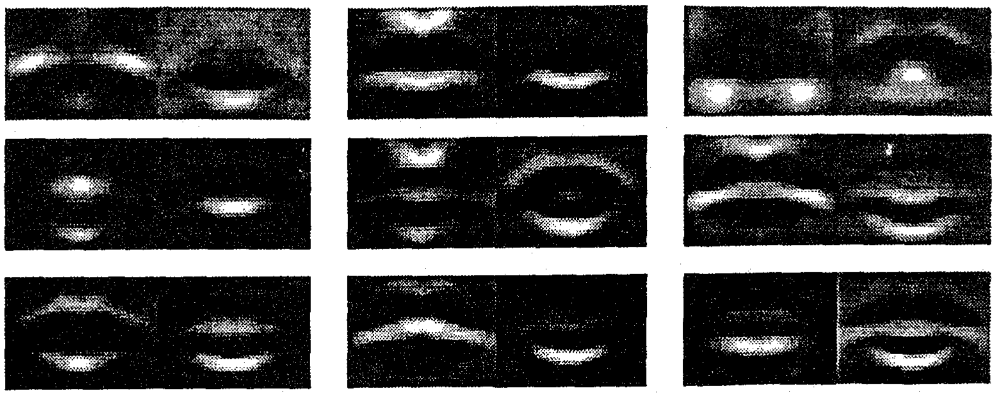

Figure 2: The first 9 principal components of the lip images, starting in upper left and proceeding in normal English reading order. The left half of each image contained form information, and the right half represented the delta image.

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 5 different visual representations for the digit recognition task: form + delta, form + delta PCA, flow, flow + acceleration, and form + flow. For all representations, we first made our images symmetric by averaging pixels from the left and right side of the image (Figure 1, left column). The form + delta representation (Movellan [8]) consisted of 2 parts (Figure 1, middle column). We low-pass filtered the images, and downsampled to a resolution of 15 x 20 pixels. Delta images were formed from the pixel-by-pixel difference between subsequent time frames, and then low-pass filtered and downsampled to 15 x 20 pixels. Because the images were symmetric, we used only half of the form and delta images, resulting in a 300-dimensional input vector.

Our form + delta PCA representation was derived from input images and delta images at a resolution of 45 x 60 pixels (1200 inputs). We took the projections of the first 300 (of 1200) principal components (PCs) to match the dimensionality of the form + delta representation. These first 300 PCs accounted for more than 99% of the variance in the original images. The first 9 PCs are illustrated in Figure 2.

The delta image representation captures information about changes in the lip image over time, but does not signal the direction in which lip features are moving. To get directional information, we computed optical flow. Our computation was based on the standard *brightness constraint* equation, followed by thresholding. We then low-pass filtered and downsampled the resulting flow field to obtain our flow representation: a 140-dimensional input vector (70 for left/right motion, and 70 for up/down motion), which is illustrated in Figure 1 (right column, top panel). Experimentation with more sophisticated 2nd-order optical flow techniques (Barron, Fleet, & Beauchemin [1]) resulted in extremely noisy output, presumably due to violation of the rigidity constraint. In addition to optical flow, we computed acceleration of the lip features as the difference between subsequent optical flow fields. This acceleration information, combined with optical flow, formed our flow + acceleration representation. Finally, we joined the form information with optical flow to make the form + flow inputs.

2.3 RECOGNITION ENGINE

The different visual representations described above formed the input to hidden Markov models which were separately trained for each word category. The images were modeled as mixtures of Gaussian distributions in pixel space. The initial state probabilities, transition probabilities, mixture coefficients, and mixture centroids were optimized using the EM algorithm. Because the probability of images rapidly approached zero when using the EM algorithm with Gaussian mixtures, we constrained the variance parameters for all the states and mixtures to be equal. In addition, the centroids of the mixtures were initialized using a linear segmentation followed by k-means clustering.

3 RESULTS

Each input representation was tested 3 times with 20 different architectures generated by combining different numbers of states (1, 3, 5, 7, 9) with different numbers of Gaussians (1, 3, 5, 7) to represent each state. Each set of simulations took approximately 22 hours on a 300 MHz DEC Alpha processor. Mean performance for each input representation (averaged across all 20 architectures) is shown in Table 1. Results from the form + delta representation matched, as expected, Movellan [8]. The form + delta PCA and flow representations did not perform as well as the form + delta inputs. Performance improved somewhat for the flow + acceleration and form + flow representations. The states learned by the HMM for these flow inputs give us information about the dynamic movement of the lips through time. These learned states (for 5-state HMMs) are shown in Figure 3, and provide an intuitive notion to the kinds of muscular activity in the face that correspond to each digit. Figure 1 (right column, bottom panel) shows the flow learned by a 1-state HMM.

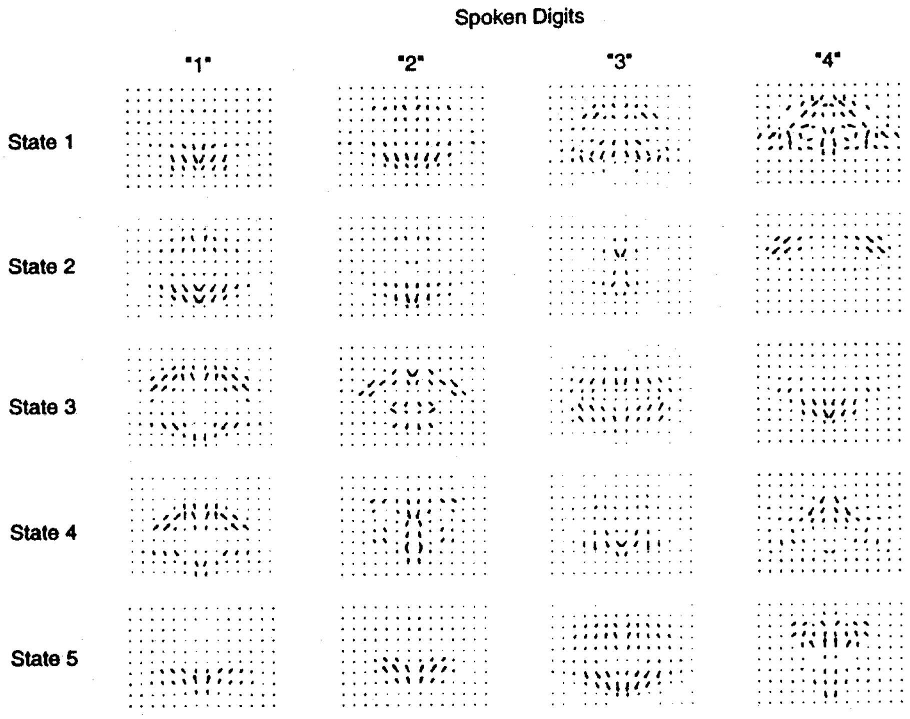

Figure 3: The optical flow representations learned by the 5-state HMMs.

Image Processing	Mean Performance
Form + Delta	78.2
Flow + Acceleration	63.4
Flow + Form	63.0
Flow	51.6
Form + Delta PCA	40.6

Table 1: Performance (% correct) for the different visual input representations, averaged across all 20 architectures.

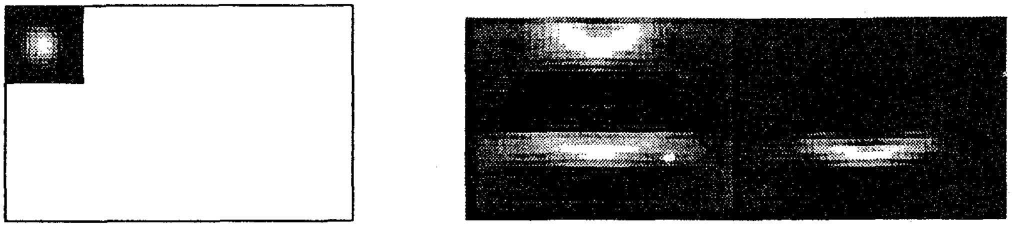

Figure 4: Left: *Local* low-pass filtering. Right: *Global* principal components.

4 DISCUSSION

The purpose of this research was to compare a range of image processing techniques on a visual digit recognition task. We found that local low-pass filtering of gray scale pixel values and the delta image (the form + delta representation) yielded the best performance. Compression of the form and delta images by PCA led to surprisingly poor performance. We believe that this weak performance is due to the fact that PCA is a *global* compression algorithm. What is needed for this task is local image processing. All of the other 4 approaches utilize local blurring with a Gaussian kernel (low-pass filtering). This retains topographic information that may be lost in a PCA representation. The difference between these local and global techniques is illustrated in Figure 4. An additional advantage of this low-pass filtering is that it provides some translation invariance. We are currently exploring methods to perform local PCA compression. In addition, we found that pixel-based representations worked better than methods based on optical flow.

A second area of inquiry is the difference between the information provided by optical flow and the delta image. Both carry dynamic information that signals the difference between the lip images at subsequent time steps. Why should the delta image yield 15% better performance as compared to optical flow, when combined with form information? We believe that part of the reason lies in the assumptions of the optical flow algorithm: rigidity of the objects, and small pixel movement of features between frames (2-4 pixels). Lips certainly violate the rigidity assumption. We also sometimes violate the small movement assumption because we are sampling images at only 30 frames per second. For these reasons, our optical flow tends to be noisy, and must be thresholded. This thresholding leads to an optical flow output that is very sparse, as illustrated in Figure 1. The delta image, on the other hand, contains information at all points in the image. Although the significance of the delta image is not well understood, it does contain local dynamic information at all regions in the image. We will further explore this hypothesis as we extend this work.

The work described here represents an exploration of the kinds of dynamic information that may be valuable for speechreading. In contrast to model-based approaches, we have sought to retain as much information as possible in the lip images by allowing the recognition engine to find relevant features of the input. This effort to combine sophisticated image processing techniques with machine learning algorithms is a valuable approach that will likely lead to new insights in a variety of applications.

References

[1] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.

[2] C. Bregler and Y. Konig. Eigenlips for robust speech recognition. In *Proceedings of IEEE ICASSP*, pages 669–672. Adelaide, Australia, 1991.

[3] C.J. Duffy and R.H. Wurtz. 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, 1991.

[4] O.N. Garcia, A.J. Goldschen, and E.D. Petajan. Feature extraction for optical automatic speech recognition or automatic lipreading. *Technical Report GWU-IIST-9232, Dept. of Electrical Engineering and Computer Science, George Washington University*, 1992.

[5] K. Mase and A. Pentland. Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan*, 22(6):67–76, 1991.

[6] D.W. Massaro and M.M. Cohen. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9:753–771, 1983.

[7] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:126–130, 1976.

[8] J.R. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro, D.S. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 851–858. MIT Press, Cambridge, MA, 1995.

[9] P.L. Silsbee. Sensory integration in audiovisual automatic speech recognition. In *Asilomar Conference on Signals, Systems, and Computers*, volume 28, pages 561–565, 1994.

[10] G.J. Wolff, K.V. Prasad, D.G. Stork, and M. Hennecke. Lipreading by neural networks: Visual preprocessing, learning and sensory integration. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 1027–1034. Morgan Kaufmann, San Francisco, CA, 1994.

[11] B.P. Yuhas, Jr. Goldstein, M.H., T.J. Sejnowski, and R.E. Jenkins. Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, 78(10):1658–1668, 1990.