Integration of Acoustic and Visual Speech Signals Using Neural Networks

Ben P. Yuhas
Moise H. Goldstein, Jr.
Terrence J. Sejnowski

AUTOMATIC SPEECH RECOGNITION SYSTEMS rely almost exclusively on the acoustic speech signal and, consequently, these systems often perform poorly in noisy environments [1]. Attempts to clean up the acoustic input have had limited success [2]. Another approach is to use other sources of speech information, such as visual speech signals. The perception of acoustic speech by humans can be affected by the visible speech signals [3–5]. Specifically, when the acoustic signal is degraded by noise, the visual signal can provide supplemental speech information that improves speech perception [6–8]. When no acoustic signal is available, as for the profoundly deaf, the visual signal alone can provide speech information through lip reading [9–11]. Here we answer two questions: Can the speech information conveyed by visual speech signals be extracted automatically? How can this information be combined with information from the acoustic signal to improve automatic speech recognition?

For a limited vocabulary, Petajan demonstrated that visual speech signals can be used to significantly improve automatic speech recognition compared to acoustic recognition alone.

The only speech recognition system that has extensively used visual speech signals was developed by Eric Petajan [12] [13]. For a limited vocabulary, Petajan demonstrated that visual speech signals can be used to significantly improve automatic speech recognition compared to acoustic recognition alone. The system relied upon a codebook of images that were used to translate incoming images into corresponding symbols. These symbol strings were then compared to stored sequences representing different words in the vocabulary. This process is computationally intensive and requires efficient image encoding to perform a reasonable number of comparisons. The early encoding and categorization of continuous speech signals resulted in the loss of relevant speech information. The overall performance of the system was degraded by this early encoding.

The need for early categorization of speech signals can be traced to the computational limitations of currently available hardware. On a digital computer, the inherently analog visual speech signals must first be converted to digital format. Next, a significant amount of preprocessing and encoding must be performed before these signals can be compared to a stored set of patterns. Finally, the symbolic descriptions of the segmented visual signal stream are combined with the auditory symbol stream, using rules that require a significant amount of programming. The von Neumann architecture requires that all of these steps be performed sequentially. We propose an alternative method for processing visual speech signals, based on analog computation in a distributed network architecture. By using many interconnected processors working in parallel, large amounts of data can be handled concurrently. In addition to speeding up the computation, this approach does not require segmentation in the early stages of processing; rather, analog signals from the visual and auditory pathways would flow through networks in real time and would be combined directly in the final analog Very Large-Scale Integration (VLSI) implementation.

Results are presented from a series of experiments that use neural networks to process the visual speech signals of a male talker. In these preliminary experiments, the results are limited to static images of vowels. We demonstrate that these networks are able to extract speech information from the visual images, and that this information can be used to improve automatic vowel recognition. The first section of this article reviews the structure of speech, and its corresponding acoustic and visual signals. The next section describes the specific data that was used in our experiments along with the network architectures and algorithms. In the final section, we present the results of integrating the visual and auditory signals for vowel recognition in the presence of acoustic noise.

The Visual and Acoustic Signals of Speech

Symbol Strings

Continuous speech signals are traditionally treated as a sequence of discrete components [14] [15]. As such, the phoneme

0163-6804/89/0011-0065 $01.00 © 1989 IEEE

is the shortest acoustically distinguishing unit of a given language. For example, boot and beat are distinguished by the phonemes /u/ and /i/, which are abstractions corresponding to the "oo" and "ea" sounds in those particular words. The sounds themselves are called phones and designated [u] and [i] to distinguish them from the abstract linguistic units /u/ and /i/. While the phonemes are functional and distinguish one word from another, phones are descriptive units and describe speech sounds.

The visual correlative of the phoneme is the viseme, which is the smallest visibly distinguishing unit of a given language [16]. The mapping between the phonemes and visemes is generally many to one; for example, the phonemes /p/, /b/, and /m/ are usually visibly indistinguishable and treated as a single viseme.

Sub-Symbolic Structure

The acoustic speech signal that is emitted from the mouth can be modeled as the response of the vocal tract filter to a sound source [17] [18]. The configuration of the articulators define the shape of the vocal tract and the corresponding resonance characteristics of the filter [19] [20]. The resonances of the vocal tract are called formants [21] and they often appear as peaks in the short-time power spectrum. The formants are used to help identify the individual vowels [22–24], while the complete amplitude of the short-time spectra of the acoustic speech wave contains much of the information necessary for speech perception [25], intelligibility [2], and automatic recognition [26].

While some of the articulators are visible on the face of the speaker (e.g., the lips, the teeth, and sometimes the tip of the tongue), others are not. The contribution of the visible articulators to the acoustic signal result in speech sounds that are much more susceptible to acoustic noise distortion than are the contributions from the hidden articulators [13], and therefore, the visual speech signal tends to complement the acoustic signal. The most visibly distinct speech sounds, such as /b/ and /k/, are among the first pairs to be confused when presented acoustically in the presence of noise. Similarly, those phonetic segments that are visibly indistinguishable, such as /p/, /b/, and /m/, are among the most resistant to confusion when presented acoustically [27] [28]. Because of this complementarity, the perception of speech in noise is greatly improved when both speech signals are present.

Automatic Interpretation

Speech recognition systems generally have a "front-end" signal processor, followed by categorization and symbolic computation [1] [29] [30]. In these systems, continuous acoustic signals are often segmented into discrete units that will hopefully correspond to phonemes. Even if the segmentation is correct, the identification process is complicated by the variability with which a phoneme can be spoken. The description of the speech wave is significantly reduced when the continuous signal is converted to a symbolic representation. At every successive level of encoding, additional information about the original speech signal is lost. Constraints from low-level speech production models, phonotactic rules, morpheme order, syntax, grammar, and semantics can compensate for these losses, but the overall performance of the system depends upon the correctness of the early encoding of the signals.

The most successful speech recognition systems have avoided low-level phonemic identification [31] [32] and have attempted to define units based more closely on the actual signal structure. Recent work with hidden Markov models indicates that the most improvement in speech recognition is obtained by extracting more information from the input signals [32], rather than depending upon higher-level constraints. These findings suggest that enhancing the information in the earliest stages of processing can improve the overall performance of a speech recognition system.

The visual speech signal is a secondary source of speech information. Can the information in the visual speech be fused with the acoustic signal at an early stage in the recognition process? In previous approaches, the information from the visual signal has been incorporated into the recognition system at levels beyond the categorization stage [13]. In our approach, visual signals will be used to resolve ambiguities in the acoustic signal before either is categorized. By combining these two sources of information at an early stage of processing, it is possible to reduce the number of erroneous decisions made and increase the amount of information passed to later stages of processing [10]. The additional information provided by the visual signal can serve to constrain the possible interpretations of an ambiguous acoustic signal, or it can serve as an alternative source of speech information when the acoustical signal is heavily noise-corrupted. In either case, a massive amount of computation must be performed on the raw data. New massively parallel architectures based on neural networks and new training procedures may make this approach feasible, even when scaled up to the full phonetic set.

The description of the speech wave is significantly reduced when the continuous signal is converted to a symbolic representation.

Neural Network Architecture

Layered feed-forward networks were used in this study. The image was presented in the bottom layer of units, which then passed the signals to a layer of hidden units, which in turn projected to an output layer. As a signal traveled from unit i to unit j, it was multiplied by the weight, w_{ij}, associated with that connection. These weights have continuous values that can be positive (excitatory), negative (inhibitory), or zero. The internal activation of the ith unit was then obtained by summing all of its inputs

$$u_i = \sum_j w_{ij} f(u_j) + I_i \tag{1}$$

which included the weighted sum of the outputs of all units feeding into it, and any additional input from outside the network, I_i. Note that the jth unit's output is a function of its activation, u_j where the function f can be any continuous linear or nonlinear transformation. The nonlinear logistic function was used:

$$f(u) = \frac{1}{1 + e^{-u}} \tag{2}$$

The networks were simulated on an MIPS M/120 RISC computer and on an ANALOGIC AP5000 array processor. These simulations would run much faster on parallel hardware.

A network is programmed to solve a problem by specifying the pattern of connectivity and the connection strengths or weights [33] [34]. Learning algorithms have been developed that iteratively adjust these weights given a set of examples

[35–37]. The ability of these networks to solve problems has been demonstrated in speech recognition [38], sonar target identification [39], and text-to-speech [40] [41]. The back-propagation procedure used here is an efficient optimization procedure that allows error gradients to be computed with $O(N)$ complexity, where N is the number of adjustable weights [36].

We used the back-propagation technique to compute the error gradients for the weights [42]. The steepest descent algorithm for updating the weights with fixed step size was unfortunately unable to learn the high-dimensional mappings needed in this study, as is typical of scale-up problems. However, with a line search minimization and a conjugate gradient constraint, the networks were able to find the necessary weights to solve the problems in this article [43].

The Speech Data

The speech signals used in these experiments were obtained from a male speaker who was videotaped while seated facing the camera, under well-lit conditions. The visual and acoustic signals were then transferred and stored on laser disc [44], which allowed the access of individual video frames and the corresponding sound track.

Selecting the Images

The NTSC video standard is based upon 30 frames/s and each video frame corresponds to 33 ms of the acoustic speech signal. In this format, individual words are preserved as a series of frames on the laser disc. A data set was constructed of 12 examples of nine different vowels, for a total of 108 images per talker. The phonemes represented in this sample are:

$$i,\ I,\ e,\ \varepsilon,\ æ,\ \alpha,\ \Lambda,\ o,\ u$$

While stressed vowels can last up to 132 ms or four frames, an unstressed vowel in continuous speech can often be shorter than the 33 ms of a single frame. The vowels were selected from a list of words read from the modified rhyme test, where the vowels are usually stressed. A preliminary list of candidate words was identified from a transcription of the video-disc corpus for each vowel. Each word was then played acoustically to confirm the suspected pronunciation. The vowel within each word was surrounded by two consonants. The individual consonants were removed by alternately dropping the end frames and then listening to those remaining. After the vowel was isolated, the acoustic signal was digitized and examined to be sure that it was stationary. Frames were rejected if the periodic wave appeared to be increasing or decreasing in amplitude.

Preprocessing the Images

A reduced area-of-interest in the image was automatically defined and centered around the mouth (see Figure 1), and the resulting sub-image was sampled to produce a topographically accurate image of 20 × 25 pixels. This particular encoding was chosen to reduce the amount of data while trying to approximate what one might obtain by observing an image through an array of sensors. It is definitely not the most efficient encoding one could use; however, it is faithful to the parallel approach to computation advocated here. More sophisticated preprocessing would be required to operate over a wide range of lighting conditions and parallel implementation would be needed to achieve real-time performance.

Preprocessing the Acoustic Data

A vowel is identified primarily by the shape of its acoustic spectrum. In particular, the perception of vowels by humans is largely determined by the location of the peaks in the spectral envelope [22–24]. The spectral shape can be calculated from the short-term power spectrum of the acoustic signal. Each video frame on the laser disc has associated with it 33 ms of acoustic speech. Low-pass filtering the acoustic signal to 5 kHz provides sufficient bandwidth for speech intelligibility, and in particular vowel identification [45], while allowing us to sample the signal at 10 kHz. After applying a Hamming window [46], the short-term power spectrum was calculated. The cepstrum of the resulting power spectrum was computed and values above 3.2 ms were zeroed [47] [48]. The inverse Fourier transform of the remaining data produced a smooth envelope of the original power spectrum that could be sampled at 32 frequencies.

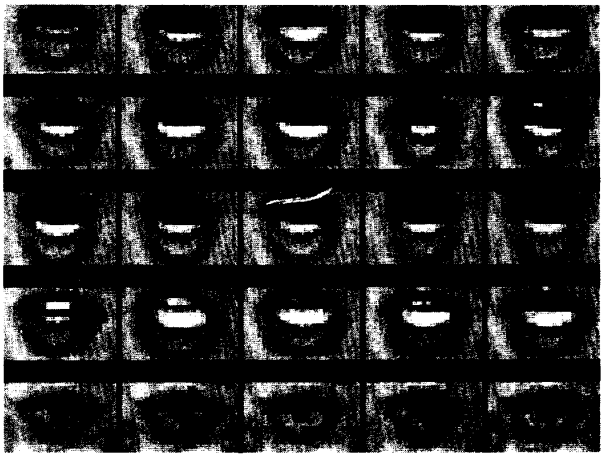

Fig. 1. Typical images presented to the network.

Integrating Visual and Auditory Speech Signals

Estimating acoustic structure from the visual speech signals alone is an ill-posed problem. The visual signals provide only a partial description of the vocal tract transfer function—and that description is usually ambiguous. For a given visual signal, there are many possible configurations of the full vocal tract, and consequently many possible corresponding acoustic signals. The goal is to define a good estimate of that acoustic signal from the visual signal and then use that estimate in conjunction with any residual acoustic information. Combined, these two sources of speech information result in better automatic recognition rates than were obtained from either source alone.

We chose to map the visual signal into an acoustic representation closely related to the vocal tract's transfer function [49]. Given such a mapping, the visual signal could be converted and then integrated with the acoustic signal prior to any symbolic encoding. The first step was to obtain this acoustic representation directly from the visual signal.

Training the Network

A feed-forward neural network was trained to estimate the Short-Time Spectral Amplitude Envelope (STSAE) of the acoustic signal from the corresponding visual signals emitted from around the mouth of the talker. The visual signal was pre-

sented as input to the network (see Figure 2) and the network was trained to produce the amplitude envelope of the 256-point short-time power spectrum of the corresponding acoustic signal. The error gradients were computed using the back-propagation algorithm, as described in the previous section on neural networks.

For a given visual signal, there are many possible configurations of the full vocal tract, and consequently many possible corresponding acoustic signals.

As with any estimation technique, the number of free parameters in the network is a trade-off between desired accuracy, computational efficiency, and the potential problem of over-fitting the data. The number of hidden units chosen for the network restricts the bandwidth between the input and output patterns. If this bandwidth is not wide enough, then the network is unable to pass the relevant information in the images to the output units and the network performs poorly. But if the bandwidth is too wide, the network begins to memorize the idiosyncratic details of the training set, and may fail to develop general rules that apply to images in the test set. Preliminary parametric experiments found that networks with five hidden units provide the necessary bandwidth while minimizing the effects of over-learning.

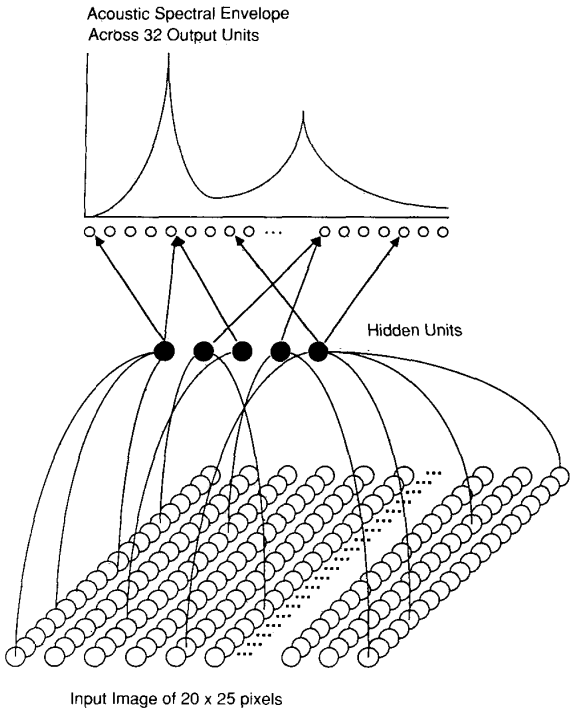

Fig. 2. Network architecture for estimating acoustic spectral shape from lip images.

Various network architectures were trained to estimate the spectral envelope of the acoustic signal from the image. These networks were tested on a second set of data to estimate the ability of the networks to generalize to new images. During the training, the error on the training set decreased asymptotically with the number of iterations. At the same time, the error on the test data would decrease at first but would then begin to increase. When the error begins to increase on the test set, the network is said to be over-learning the training data. The specific location of the upturn depended upon many factors, including the number of training patterns and hidden units, and the transfer function *f(u)*. Over-learning can be minimized by increasing the amount of training data or by reducing the number of hidden units.

The point at which over-learning began was identified by the following procedure. First, the test set was divided into two subsets, one of which was used to track the error during training. Training was terminated when the error of the tracking set started to increase; this was defined as the best performance of the network. The quality of the estimates obtained from the networks compared favorably to those obtained using other estimation techniques.

Influence of Noise on Speech Recognition

Judging the quality of a spectral estimate is significantly more difficult than judging the accuracy of a categorization, largely because intelligibility is not a simple function of the spectrum. To assay the spectral estimates, a vowel recognizer was constructed using a feed-forward network. The network was trained to correctly categorize the STSAE from six examples each of nine different vowels. With no noise present, the trained network could correctly categorize 100% of the training set. The network vowel recognizer was then presented with STSAE through two channels, as shown in Figure 3. The path

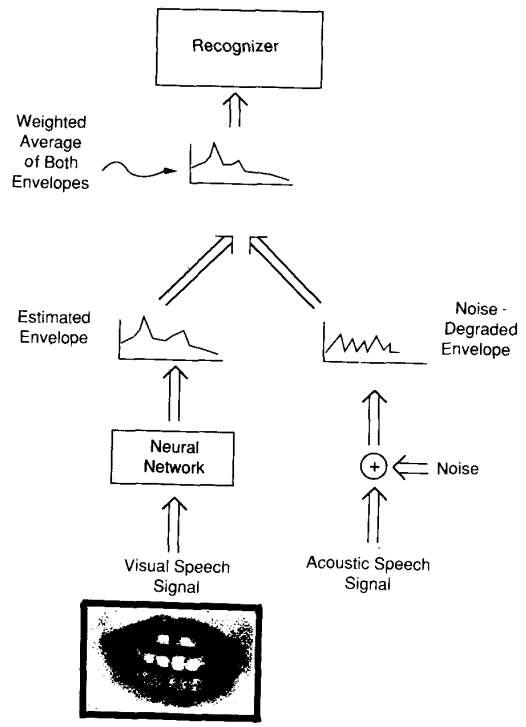

Fig. 3. Simple vowel recognizer operation.

on the right in Figure 3 represents the information obtained from the acoustic signal, while the path on the left provides information obtained from the corresponding visual speech signal.

To assess the performance of the recognizer in noise, clean spectral envelopes were systematically degraded by noise and then presented to the recognizer. In this particular condition, no visual input was given to the network. The noise was introduced by averaging the STSAE with a normalized random vector. Noise-corrupted vectors were produced at 3-dB intervals from -12 dB to 24 dB. At each step, six different vectors were produced, and the performance reported was the average. The lower curve of figure 4 shows the recognition rates as a function of the Speech-to-Noise ratio (S/N). At an S/N of -12 dB, the recognizer was operating at an 11.1% error rate, expected of totally random performance.

Next, a network trained to estimate the spectral envelopes from images was used to provide an independent STSAE input into the recognizer (left side of Figure 3). This network was not trained on any of the data that was used in training the vowel recognizer. For fusing, the estimates obtained from visual signals were averaged together with the noised degraded envelopes of the corresponding acoustic input and then passed on to the recognizer. At an S/N of -12 dB, the recognizer was now performing at 35%.

Averaging the two independent sources of information was less than optimal. Using the STSAE estimated from the visual signal alone, the recognizer was capable of 55.6%. However, when this estimate was combined with the noise-degraded acoustic signal, the recognizer was only capable of 35% at an S/N of -12 dB. Similarly, at very high S/Ns, the combined input produced poorer results than the acoustic signal alone provided. To correct for this, the two inputs needed to be weighted according to the relative amount of information available from each source. A weighting factor was introduced, which was a function of speech-to-noise:

$$\alpha S_{Visual} + (1-\alpha) S_{Acoustic} \tag{3}$$

The optimal value for the parameter α was found empirically to vary linearly with the S/N in dB for the range from -12-dB S/N to 24 dB:

$$\alpha = .535 - .022 \frac{S}{N} \tag{4}$$

The results based on using these values for α are shown in the top curve of Figure 4.

Discussion

The results shown in Figure 4 demonstrate that visual and acoustic speech information can be effectively fused at a subcategorical level. The low-level integration of the two speech signals was particularly useful in the range of S/Ns from 3 dB to 15 dB, where the combined signals were recognized with a greater accuracy than either of the two component signals alone. For the set of vowels studied, the results show that the two speech signals can complement each other to improve automatic recognition. An independent categorical decision on each channel would have required additional information in order to produce the same level of performance.

The successful combination of the two speech signals required the introduction of a weighting factor, α. This weight can be interpreted as an attentional parameter. When we are listening to a speaker in a noisy environment, we are more likely to attend to the speaker's mouth. However, if there is no perceptible noise interfering with the acoustic signal, we tend to rely on the acoustic signal.

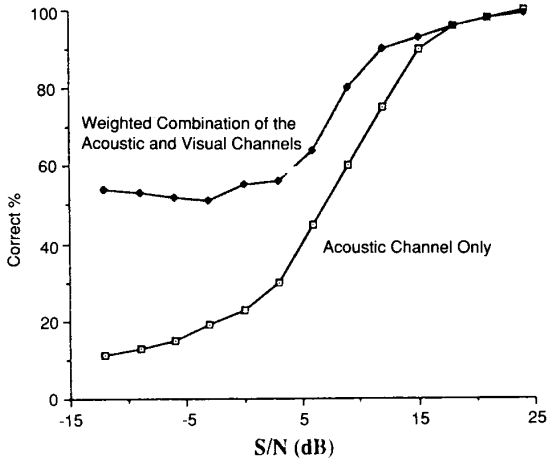

Fig. 4. Improved recognition due to visual augmentation of noise-degraded speech.

The particular values used for α were obtained by systematically trying different values of α between 0 and 1, at the various S/N levels. A line was then fit to the best values for α at each of the S/N samples. We intend to explore the use of neural networks to directly fuse the visually estimated and noise-degraded spectral envelopes without having to make *a priori* assumptions about how best to combine them. Additional improvements may be possible through nonlinear interactions between the two streams of speech information.

Conclusion

Human beings are capable of combining information received through distinct sensory channels with great speed and ease. The use of visual speech signals together with acoustic speech signals is just one example of integrating information across modalities. Sumby and Pollack [6] have shown that the relative improvement provided by the visual signal varies with the signal-to-noise ratio of the acoustic signal. By combining the speech information available from the two speech signals before categorizing, we obtained performance comparable to that demonstrated by humans.

Lip reading research has traditionally focused on the identification and evaluation of visual features [50–52]. Reducing the original speech signals to a finite set of predefined parameters or to discrete symbols can waste a tremendous amount of information. For an automatic recognition system, that information may prove to be useful at a later stage of processing. In our approach, we have used the visual signal to obtain an estimate of the corresponding acoustic spectrum. This allowed us to access speech information in the visual signal without requiring discrete feature analysis or making categorical decisions.

There are a number of improvements that would have to be made to our system in order to make it a practical one. First, all of our studies were performed on speech data off-line. The visual processing needed to prepare the image for the network is computation-intensive on a sequential machine. One way to alleviate these problems would be to design special-purpose parallel hardware for performing the operations in real time. One particularly promising approach is to use analog VLSI. Mead [53] has already fabricated synthetic retinas and

cochleas that perform much of the preprocessing that we would need for a speech recognition system combining auditory and visual information at an early stage of processing. Further advances would be needed to construct parallel hardware for the highly interconnected networks that perform the mapping between sensory modalities.

This line of research has consequences for other problems, such as target identification based on multiple sensors. The same problems arise in designing systems that combine radar and infrared data, for example, that do in combining visual and auditory speech information. Mapping into a common representation using neural network models could also be applied to these problem domains. The key insight is to combine this information at a stage that is prior to categorization. Neural network learning procedures allow systems to be constructed for performing the mappings, as long as sufficient data are available to train the network, with the appropriate architecture and training algorithm.

The same problems arise in designing systems that combine radar and infrared data, for example, that do in combining visual and auditory speech information.

Acknowledgments

This research was supported by grant AFOSR-86-0256 from the Air Force Office of Scientific Research. We would like to thank Robert Jenkins of the Applied Physics Laboratory at Johns Hopkins University for his assistance throughout this project.

References

[1] J. Allen, "A Perspective on Man-Machine Communication by Speech," special issue on man-machine speech communication, *Proc. IEEE*, vol. 73, pp. 1,539–1,550, 1985.

[2] National Research Council, Committee on Hearing, Bioacoustics, and Biomechanics, and Panel on Removal of Noise From a Speech/Noise Signal, *Removal of Noise From Noise-Degraded Speech Signals*, National Academy Press, 1989.

[3] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, vol. 264, pp. 746–748, 1976.

[4] H. McGurk and J. MacDonald, "Visual Influences on Speech Processes," *Perception & Psychophysics*, vol. 24, pp. 253–257, 1978.

[5] P. K. Kuhl and A. N. Meltzoff, "The Bimodal Perceptions of Speech in Infancy," *Science*, vol. 218, pp. 1,138–1,141, 1982.

[6] W. H. Sumby and I. Pollack, "Visual Contribution to Speech Intelligibility in Noise," *J. of the Acoustic Soc. of Amer. (JASA)*, vol. 26, pp. 212–215, 1954.

[7] H. W. Ewersten and H. B. Nielsen, "A Comparative Analysis of the Audiovisual, Auditive and Visual Perception of Speech," *Acta Orolayng*, vol. 72, pp. 201–205, 1971.

[8] N. P. Erber, "Auditory-Visual Perception of Speech," *J. of Speech and Hearing Disorders*, vol. 40, pp. 481–492, 1975.

[9] A. Montgomery and P. L. Jackson, "Physical Characteristics of the Lips Underlying Vowel Lipreading," *JASA*, vol. 73, pp. 2,134–2,144, 1983.

[10] Q. Summerfield, "Use of Visual Information for Phonetic Perception," *Phonetica*, vol. 36, pp. 314–331, 1979.

[11] L. E. Bernstein, S. P. Eberhardt, and M. E. Demorest, "Single-Channel Vibrotactile Supplements to Visual Perception of Intonation and Stress," *JASA*, vol. 85, pp. 397–405, 1989.

[12] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition," *IEEE ComSoc Global Telecom. Conf.*, pp. 265–272, Nov. 26–29, 1984.

[13] E. D. Petajan, "An improved Automatic Lipreading System To Enhance Speech Recognition," Bell Labs. Tech. Report No. 11251-871012-111TM, 1987.

[14] N. Chomsky and M. Halle, *The Sound Pattern of English*, NY: Harper & Row, 1968.

[15] J. Lyons, *Introduction to Theoretical Linguistics*, Cambridge, England: Cambridge University Press, 1971.

[16] C. G. Fisher, "Confusions among Visually Perceived Consonants," *J. of Speech and Hearing Res.*, vol. 11, pp. 796–803, 1968.

[17] G. Fant, *Acoustic Theory of Speech Production*, The Hague, Netherlands: Mouton & Co., 1960.

[18] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, Berlin: Springer-Verlag, 1972.

[19] K. N. Stevens and A. S. House, "Development of a Quantitative Description of Vowel Articulation," *JASA*, vol. 27, pp. 484–493, 1955.

[20] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating Vocal Tract Shapes from Formant Frequencies," *JASA*, vol. 64, pp. 1,027–1,035, 1978.

[21] J. M. Pickett, *The Sounds of Speech Communication: A Primer of Acoustic Phonetics and Speech Perception*, Baltimore, MD: University Park Press, 1980.

[22] H. K. Dunn, "The Calculation of Vowel Resonances, and an Electric Vocal Tract," *JASA*, vol. 22, pp. 740–753, 1950.

[23] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," *JASA*, vol. 24, pp. 175–184, 1952.

[24] R. L. Miller, "Auditory Tests with Synthetic Vowels," *JASA*, vol. 25, pp. 114–121, 1953.

[25] R. A. Cole, A. I. Rudnicky, V. W. Zue, and D. R. Reddy, "Speech as Patterns," *Perception and Production of Fluent Speech*, R. A. Cole, ed., Hillsdale, NJ: Lawrence Erlbaum Assoc., 1980.

[26] D. H. Klatt, "A Digital Filter Bank for Spectral Matching," *Proc. of ICASSP '76*, Philadelphia, PA, 1976.

[27] G. A. Miller and P. E. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants," *JASA*, vol. 27, pp. 338–352, 1955.

[28] B. E. Walden *et al.*, "Effects of Training on the Visual Recognition of Consonants," J. of Speech and Hearing Res., vol. 20, pp. 130–145, 1977.

[29] D. R. Reddy, "Segmentation of Speech Sounds," *JASA*, vol. 40, pp. 307–312, 1966.

[30] D. R. Reddy, "Computer Recognition of Connected Speech," *JASA*, vol. 42, pp. 329–347, 1967.

[31] F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," special issue on man-machine speech communication, *Proc. IEEE*, vol. 73, pp. 1,616–1,624, 1985.

[32] K. F. Lee, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The Sphinx System," Ph.D. dissertation, Computer Science Department, Carnegie Mellon University, 1988.

[33] J. J. Hopfield and D. W. Tank, "Neural Computation of Decisions in Optimizations Problems," *Bio. Cyber.*, vol. 52, pp. 141–152, 1985.

[34] D. Marr and T. Poggio, "Cooperative Computation of Stereo Disparity," *Science*, vol. 194, pp. 283–287, 1976.

[35] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "Learning Algorithm for Boltzmann Machine," *Cog. Sci.*, vol. 9, pp. 147–169, 1985.

[36] F. J. Pineda, "Generalization of Back-Propagation to Recurrent Neural Networks," *Phys. Rev. Lett.*, vol. 59, pp. 229–232, 1987.

[37] D. E. Rumelhart and J. L. McClelland, and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. I*, Cambridge, MA: MIT Press, 1986.

[38] R. P. Lippman, "Review of Neural Networks for Speech Recognition," *Neural Computation*, vol. 1, pp. 1–38, 1989.

[39] R. P. Gorman and T. J. Sejnowski, "Learned Classification of Sonar Targets Using a Massively Parallel Network," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1,135–1,140, 1988.

[40] T. J. Sejnowski and C. R. Rosenberg, "NETtalk: A Parallel Network that Learns to Read Aloud," Johns Hopkins University Department of Electrical Engineering and Computer Science, Tech. Report 8601, 1986.

[41] T. J. Sejnowski and C. R. Rosenberg, "Parallel Networks that Learn to Pronounce English Text," *Complex Syst.*, vol. 1, pp. 145–168, 1987.

[42] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Chapter 8: Learning Internal Representations by Error Propagation," *Parallel Distributed Processing in the Microstructure of Cognition: Vol. 1*, J. McClelland and D. Rumelhart, eds., Cambridge, MA: MIT Press, 1986.

[43] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*, Cambridge University Press, 1988.

[44] L. E. Bernstein and S. P. Eberhardt, *Johns Hopkins Lipreading Corpus I–II*, Johns Hopkins University, Baltimore, MD, 1986.

[45] N. R. French and J. C. Steinberg, *Factors Governing the Intelligibility of Speech Sounds*, vol. 19, pp. 90–119, 1947.

[46] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice-Hall, 1978.

[47] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1975.

[48] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Germany: Springer-Verlag, 1976.

[49] Q. Summerfield, "Some Preliminaries to a Comprehensive Account of Audio-Visual Speech Perception," *Hearing by Eye: The Psychology of Lip Reading*, B. Dodd and R. Campbell, eds., Hillsdale, NJ: Lawrence Erlbaum Assoc., 1982.

[50] V. Fromkin, "Lip Positions in American English Vowels," *Language & Speech*, vol. 7, pp. 215–225, 1964.

[51] N. L. Hesselmann, "Structural Analysis of Lip-Contours for Isolated Spoken Vowels using Fourier Descriptors," *Speech Communication*," vol. 2, pp. 327–340, North-Holland, 1983.
[52] P. L. Jackson, A. A. Montgomery, and C. A. Binnie, "Perceptual Dimensions Underlying Vowel Lipreading Performance," *J. of Speech and Hearing Res.*, vol. 19, pp. 796–812, 1976.
[53] C. Mead, *Analog VLSI and Neural Systems*, NY: Addison-Wesley, 1989.

Biography

Ben P. Yuhas (SM '88) was born in Syracuse, New York on October 3, 1959. He received the B.A. degree in mathematics from the University of Chicago in 1981 and the M.S. degree in electrical engineering and computer science from Johns Hopkins University, Baltimore, Maryland, in 1986.

He is currently completing his doctoral thesis at the Speech Processing Laboratory in the Department of Electrical and Computer Engineering at Johns Hopkins University. At present, his research explores the use of massively parallel architectures to process visual and acoustic speech signals. He is interested in applying distributed processing architectures to signal processing, recognition, and multi-modal integration.

Moise H. Goldstein, Jr. received the Doctor of Science from MIT and was on the Electrical Engineering faculty from 1955 until 1963, when he moved to Johns Hopkins University. He is the Edward J. Schaefer Professor of Electrical Engineering in the Department of Electrical and Computer Engineering, and has a joint appointment in the BME Department in the School of Medicine.

He has done experimental research on stimulus coding in the auditory nervous systems. In 1978, he established the Speech Processing Laboratory, where recent engineering methods of system modelling and electronic device design and realization are used with the aims of better understanding speech processing and developing novel devices for the real-time electronic processing of speech. His personal research interest is in the development of sign and oral language in prelingually deaf infants with the aim of realizing improved and novel diagnostic and communication aids.

Dr. Goldstein is on the Editorial Board of *Brain Research*. He has held a Science Faculty Fellowship from NSF, a Guggenheim Fellowship, and Senior Fellowships from NIH.

Terrence J. Sejnowski received his Ph.D. in physics from Princeton University in 1978. He was a postdoctoral fellow in the Department of Neurobiology at Harvard Medical School for three years before joining the Department of Biophysics at Johns Hopkins University in 1982. He received a Presidential Young Investigator Award in 1984 and was a Wiersma Visiting Professor of Neurobiology at the California Institute of Technology in 1987. In 1988, he moved to San Diego to found the Computational Neurobiology Laboratory at the Salk Institute and to become a Professor of Biology, Physics, Psychology, Neuroscience, and Cognitive Science at the University of California at San Diego.

Dr. Sejnowski is the Editor-in-Chief of *Neural Computation*, published by the MIT Press. His main research is in understanding how the brain works and the development of massively parallel computers based on the principles of neural computation.

(continued from p.40)

[2] D. Gabor, *Nature*, vol. 161, p. 777, 1948.
[3] D. Psaltis, H. Lee, and X. Gu, "Volume Holographic Interconnections with Maximal Capacity and Minimal Crosstalk," *J. of Appl. Physics*, vol. 65, p. 2,191, Apr. 1989.
[4] F. Rosenblatt, *Principles of Neurodynamics: Perceptron and the Theory of Brain Mechanisms*, Spartan Books, Washington, 1961.
[5] D. Psaltis, D. Brady, and K. Wagner, "Adaptive Optical Networks using Photorefractive Crystals," *Appl. Opt.*, vol. 27, no. 9, pp. 1,752–1,759, May 1, 1988.
[6] J. H. Kim, S. H. Lin, J. Katz, and D. Psaltis, "Monolithically Integrated 2-D Arrays of Optoelectronic Devices for Neural Network Applications," *SPIE*, vol. 1,043, Los Angeles, Jan. 1989.
[7] D. Psaltis, A. Yamamura, M. Neifeld, and S. Kobayashi, "Parallel Readout of Optical Disks," *Proc. of the 3rd OSA Topical Meeting on Optical Computing*, p. 58, Salt Lake City, Ut., Feb. 1989.
[8] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," *Parallel Distributed Processing*, vol. 1, ch. 8, MIT Press, Cambridge, MA, 1986.
[9] Y. Abu-Mostafa and D. Psaltis, "Optical Neural Computers," *Sci. Amer.*, vol. 256, no. 3, pp. 88–95, 1987.
[10] D. Psaltis, D. Brady, X. Gu, and K. Hsu, "Optical Implementation of Neural Computers," *Optical Processing and Computing*, H. Arsenault, ed., Academic Press, 1989.

Biography

Demetri Psaltis received the B.Sc. in electrical engineering and economics in 1974 and the M.Sc. and Ph.D. degrees in electrical engineering in 1975 and 1977, respectively, all from Carnegie-Mellon University, Pittsburg, PA.

After the completion of the Ph.D., he remained at Carnegie-Mellon as a Research Associate, and later as a Visiting Assistant Professor, for a period of three years. In 1980, he joined the faculty of the Electrical Engineering Department at the California Institute of Technology in Pasadena, CA, where he is now Associate Professor and consultant to industry. His research interests are in the areas of optical information processing, holography, radar imaging, pattern recognition, neural network models of computation, optical memories, and optical devices. He has authored, or co-authored, over 170 publications in these areas. Dr. Psaltis is a fellow of the Optical Society of America and served as the first Vice President of the International Neural Networks Society.

Alan A. Yamamura was born in Hampton, VA on March 6, 1965. He received the S.B. degree in electrical engineering, the S.M. degree in electrical engineering and computer science, and the S.B. degree in physics from the Massachusetts Institute of Technology, Cambridge, MA in 1986. He is currently a Ph.D. candidate in the Department of Electrical Engineering at the California Institute of Technology, Pasadena, CA. Mr. Yamamura is a member of Phi Beta Kappa.

Ken Hsu received his Ph.D. in electrical engineering from CalTech in 1989. He is presently an associate professor in electro-optical engineering at National Chiao Tung University in Taiwan.

Steven Lin was born in Tainan, Taiwan on Feb. 11, 1962 and immigrated to the United States in 1977. He received both the B.S. and the M.S. degrees in electrical engineering from the Massachusetts Institute of Technology in Jan. 1985.

From June 1984 to Sept. 1986, he was with the Hewlett-Packard Laboratories in Palo Alto, CA, where he worked on the coupling of single-mode optical fibers to GaAs waveguides and GaAs microstrip and coplanar traveling-wave electro-optic waveguide modulators. Since Sept. 1986, he has been a research assistant at the California Institute of Technology, where he is now working toward his Ph.D. in electrical engineering. His research project involves building arrays of optoelectronic devices on a single GaAs substrate for optical neural network applications. Mr. Lin is a member of Tau Beta Pi and Eta Kappa Nu.

Xiang-guang Gu was born in Shanghai, P. R. China, on Sept. 29, 1963. She received the B.S. degree in physics in 1985 from Fudan University, Shanghai, P.R. China, and the Ph.D. degree in physics from the California Institute of Technology in 1989. She is currently continuing her research at Caltech.

David Brady is a Ph.D. candidate in applied physics at CalTech. He received his B.A. in physics and math from Macalester College in 1984 and his M.S. in applied physics from CalTech in 1986.