Neural Network Models of Sensory Integration for Improved Vowel Recognition

BEN P. YUHAS STUDENT MEMBER, IEEE, MOISE H. GOLDSTEIN, JR., SENIOR MEMBER, IEEE, TERRENCE J. SEJNOWSKI, MEMBER, IEEE, AND ROBERT E. JENKINS, MEMBER, IEEE

Automatic speech recognizers currently perform poorly in the presence of noise. Humans, on the other hand, often compensate for noise degradation by extracting speech information from alternative sources and then integrating this information with the acoustical signal. Visual signals from the speaker's face are one source of supplemental speech information. We demonstrate that multiple sources of speech information can be integrated at a subsymbolic level to improve vowel recognition. Feedforward and recurrent neural networks are trained to estimate the acoustic characteristics of the vocal tract from images of the speaker's mouth. These estimates are then combined with the noise-degraded acoustic information, effectively increasing the signal-to-noise ratio and improving the recognition of these noise-degraded signals. Alternative symbolic strategies, such as direct categorization of the visual signals into vowels, are also presented. The performances of these neural networks compared favorably with human performance and with other pattern-matching and estimation techniques.

I. Introduction

We usually can communicate by using the acoustic speech signal alone, but often communication also involves visible gestures from the speaker's face and body. In situations where environmental noise is present or the listener is hearing impaired, these visual sources of information become crucial to understanding what has been said. Our ability to comprehend speech with relative ease under a wide range of environmental circumstances is due largely to our ability to fuse multiple sources of information in real time. Loss of information in the acoustic signal can be compensated for by using information about speech articulation from the movements around the mouth, or by using semantic information conveyed by facial expressions and other gestures. At the same time, the listener can use knowledge of linguistic constraints to further compensate for ambiguities remaining in the received speech signals.

Speech perception can be improved greatly by watching the face of the speaker [1], [2]. Normal hearing subjects tested on isolated word recognition in noise for a limited vocabulary were able to improve their performance from an initial 13% correct to a 90% performance level when given visual access to the speakers in addition to the noise-degraded acoustic speech signal [3]. This produced an effective gain of 15 dB in the signal-to-noise ratio (S/N). Even when the acoustic signal is completely absent, as in the profoundly deaf, the visual signal alone is able to provide significant speech information through lipreading [4], [5]. Multimodal sensory integration can occur during speech recognition, but it is not clear how or at what level of processing this integration takes place [6].

In contrast to human performance, the performance of automatic speech recognition systems are not as robust and tend to degrade rapidly in noisy environments [7]. Efforts have been made to reduce the noise in the acoustic signal [8] and much work has been done to formalize linguistic constraints [9], but few have attempted to use additional external information sources. One notable exception is a system built by Eric Petajan [10] for isolated digit recognition that used vector-quantized binary images of the speaker's mouth. In this system, the acoustic and visual speech information were independently encoded into symbol strings, and a set of rules was used to reconcile conflicting interpretations. They symbolic intermediates were needed to perform the necessary processing and integration in real time on the serial digital computers available.

The massively parallel architecture of artificial neural networks make it feasible to explore subsymbolic alternatives to Petajan's system. The use of many-dimensional representations allows information from several sources to be combined "softly," before being reduced to discrete symbols. In addition, learning algorithms provide a means of training networks to fuse these signals without explicit rules or restrictive *a priori* models.

In this paper, visual speech signals are preprocessed with a neural network to improve automatic speech recognition.

Manuscript received October 16, 1989; revised March 15, 1990. This work was supported by the Air Force Office of Scientific Research under grant AFOSR-86-0246 and by Internal Research and Development of the Applied Physics Laboratory of Johns Hopkins University.

B. P. Yuhas is with Bell Communications Research (BELLCORE), 445 South Street, Morristown, NJ 07962, USA.

M. H. Goldstein is with the Speech Processing Laboratory, Dept. of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA.

T. J. Sejnowski is with the Computational Neurobiology Laboratory, the Salk Institute and with the Dept. of Biology, University of California at San Diego, La Jolla, CA 92037 USA.

R. E. Jenkins is with the Applied Physics Laboratory, Johns Hopkins University, Laurel, MD 20707, USA.

IEEE Log Number 9039183.

The approach taken here is to use the visual speech signals to clean up the acoustic signal. In effect, we are building a better microphone. Neural networks are trained to estimate the associated acoustic structure from the concurrent visual speech signal. This acoustic estimate is then fused with the noise-degraded acoustic information. By combining the visual and acoustic sources of speech information, we demonstrate that the visual signal can be used to improve the performance of automatic vowel recognition in the presence of noise. This approach does not require categorical preprocessing or explicit rules. The results described here are based on vowels spoken by a single speaker.

II. Speech

There are many ways to characterize speech. At one level, there are linguistic descriptions using abstract symbols. These representations are highly compact and efficiently represented on digital computers. At another level, there are acoustic descriptions of speech based on continuous, analog signals. They make minimal assumptions about the structure of speech, but require extensive storage and are difficult to work with on serial digital computers. Within the speech research community, the degree to which speech contains symbolic and subsymbolic structures remains controversial [11].

The speech representation that is appropriate depends on many factors, including the tools one has available. Neural network models have aspects that allow for symbolic and subsymbolic representations. Information can be represented locally by associating a concept with a single unit. While the individual unit may represent a discrete category, its level of activation can be continuous. At the same time, information can also be distributed across a whole set of units, with a concept being represented by the joint activation of a group of units. These characteristics allow network models for speech processing to have representations that extend across acoustic and linguistic levels.

A. Speech as Symbols

In a linguistic description, the *phoneme* is the shortest distinguishing unit of a given language. For example, the words *beet* and *neat* are distinguished by the phonemes /b/ and /n/, and *boot* and *beet* are distinguished by the phonemes /u/ and /i/. While the phonemes /u/ and /i/ are linguistic abstractions, the speech sounds themselves are identified as *phones* and represented in brackets, [u] and [i]. Phones are descriptive of a set of speech sounds [12], whereas phonemes are functional characterizations that can distinguish one word from another. When the same word is pronounced differently by two individuals, then the same phoneme in that work may be represented by two different phones.

The visual correlate of the phoneme is the *viseme*: the smallest visibly distinguishing unit of a given language [13]. 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 [14].

The physical realities of speech signals are often difficult to reconcile with these linguistic units, and consequently it is often impossible to find invariant features that define these speech segments. To provide a transition between these levels, a hierarchy of descriptive languages has evolved. Phonemes can be represented as sets of binary distinctive features [15]. For example, the difference between the sounds [z] and [s] is the absence or presence of the feature voicing. At an even lower level, the binary distinctive features can be represented by the continuously changing locations, movements and relative timing of the speech articulators [16]. Here the description of speech becomes closely related to the acoustic signal itself.

B. Speech as Signals

Acoustic speech signals are often represented by the magnitude of their short-term power spectrum. This representation assumes that the signals are approximately stationary over a short time and that the phase information is not essential. Early experiments in machine synthesis indicated that the phase component of the spectrum does not play an important role is speech recognition [17]. It has also been found that phase information contributes little to speech intelligibility [8]. The ability to read spectrograms has been used as further evidence that the short-term power spectrum carries the necessary information to convey speech information [18]. Today, some form of the short-term power spectrum serves as a basic unit for most automatic speech recognition systems [9], [19].

The acoustic speech signal emitted from the mouth can be modeled as the response of the vocal-tract filter to a switchable sound source [20], [21]. In a first-order vocal-tract model, the configuration of the articulators (such as the mouth opening, lips, teeth, tongue, velum, and glottis) defines the shape of the vocal-tract filter, which then determines the filter's frequency response. The resonances of the vocal-tract filter appear as peaks in the envelope of the short-term power spectrum of the acoustic signal and are called formants.

A simplified model of the vocal tract for non-nasalized speech consists of a series of tubes of uniform length with different diameters. The acoustic characteristics of this model can be represented as an all-pole filter using linear predictive coding (LPC) [22], which allows for a compact representation of the vocal-tract filter using only a few time-varying coefficients. Speech signals are routinely encoded, stored, and resynthesized by using LPC coefficients along with a characterization of the driving source.

C. The Audio–Visual Interaction

Although some of the articulatory features are often visible (for example, the lips, the teeth, and sometimes the tongue), other components of the articulatory system, such as the glottis and velum, are not. Those articulators that are visible tend to modify the acoustic signal in ways that are more susceptible to acoustic distortion than are those effects due to the hidden articulators [6], [10]. For example, the quasi-periodic sound produced by the glottis is rather resistant to noise degradation. The information in the visual speech signal tends to complement the information in the acoustic signal. Consequently, phonemes, such as /b/ and /k/, that are produced in visibly distinct manners, have acoustic correlates that are among the first pairs to be confused in the presence of noise. Conversely, phonetic segments that are visibly indistinguishable, such as /p/, /b/, and

/m/, are among the most resistant to confusion when presented acoustically [23], [14]. This complementary structure demonstrates how these two speech signals can interact to improve the perception of speech in noise.

The visible and acoustic speech signals combine using a common representation that lies somewhere between the abstract linguistic segment and the continuous analog signal. At one end, the two signals can be symbolically interpreted and the visual signal provides a linguistic-level constraint. At the other, the visual signal can provide an independent estimate of the vocal-tract transfer function and serve as a low-level acoustic constraint. Neural networks provide a computational framework within which one can explore this full range of representations.

II. Neural Networks

The architecture of artificial neural networks is motivated by the computational style found in biological nervous systems. The key features are a large number of relatively simple nonlinear processing units and high degree of connectivity between these units. A unit performs a nonlinear transformation on the sum of its inputs to produce an output signal. When this output signal travels across a connection to another unit, the signal is attenuated or amplified by the weight associated with that connection. Computation is performed by the interaction of these units and signals. Rather than having an explicit program, the computation is defined by the properties of the individual units and their interconnects.

In terms of architectural abstraction, these models differ from actual neural networks found in the nervous systems. For example, the processing units used in this study simply add their weighted inputs and have a static sigmoidal nonlinear output function, while neurons in real nervous systems have more complex spatiotemporal nonlinearities and are capable of much more complex discriminations [24]. Nevertheless, in terms of architecture, these networks provide alternative approaches to difficult computational problems. The architecture and weights needed to solve a particular problem can be either predefined or found using learning algorithms [25], [26]. These algorithms iteratively adjust the weights to reduce some error measure defined on a set of training examples. Neural networks have been constructed to solve a variety of problems, such as optimization problems, mapping text to speech, associative memories, and pattern classification [27]–[30].

A. Architecture

Feedforward network architectures were used in this study. The units in a feedforward network are arranged in layers, with connections only allowed between layers, and only in one direction. The units that receive inputs from outside the network are referred to as input units, and those that are observed from outside the network are output units. The remaining units are referred to as hidden, because they only exchange signals with other parts of the network. The units themselves use a nonlinear sigmoid squashing function to transform the sum of their inputs. The standard multilayered feedforward networks with arbitrary squashing functions are a class of universal approximators [31]. Moreover, any nonlinear mapping can be learned by a network if there are sufficient data to characterize the mapping and if the number of parameters in the network matches the information content of the data [32], [33].

B. Training

A modified backpropagation algorithm was used to train feedforward networks [26]. The gradient was calculated in the standard manner, but instead of using steepest descent, a conjugate-gradient algorithm was used to update the weights. In addition, the fixed-step size and momentum term associated with backpropagation were replaced with a line-search minimization [34].[1]

The number of adjustable weights in a neural network can often exceed the number of training patterns. In these cases, the networks have too many free parameters and are subject to the problem of overfitting or overlearning the training data. The effects of overlearning can be minimized by increasing the size of the training data set, by reducing the number of hidden units, or by stopping the training before the network has completely converged.

IV. The Speech Signals

The speech signals used in this study were obtained from video recordings of a seated speaker facing a camera under well-lit conditions. The visual and acoustic signals were stored on a laser disc [35] where the individual frames and their corresponding speech segments were indexed. The NTSC video standard of 30 frames/s was used and each frame had 33 ms of speech associated with it. Phonemes usually are shortened or dropped altogether during fluent speech, so single video frames often span more than one phoneme. To avoid this problem, we selected speech samples such as stressed vowels in isolated words or consonant-vowel-consonant (CVC) nonsense syllables that change relatively slowly. In these contexts, the vowels often were steady state over periods of 50–100 ms.

For a given phoneme, a preliminary list of candidate words was identified from a transcription of the laser disc. Each word was then played acoustically to confirm the suspected pronunciation. A representative frame for the vowel was then isolated by alternately dropping a frame and then listening until the surrounding consonants were removed. The number of frames that remained after this process depended upon the degree to which that particular vowel was stressed. Stressed vowels, for example, can last up to 132 ms or 4 frames, while an unstressed vowel in continuous speech will often not last the full 33 ms of a single frame. The acoustic signals of the remaining frames were digitized and visually examined to ensure that each signal was approximately in steady state. From this set, a single frame was selected only if the periodic waveform appeared relatively stable, neither increasing nor decreasing in amplitude.

This paper describes results obtained using data from a single male speaker. A data set was constructed of 108 images of 9 different vowels in 12 sets. The vowels were

[1]Our neural networks were simulated on a MIPS M/120 computer and an ANALOGIC AP5000 array processor. Because of the conjugate-gradient learning algorithm, the time it took to perform on backpropagation step varied depending upon the number of evaluations required in the line-minimization search. For a network with 2559 weights it took the MIPS M/120 approximately 35 msec to perform one evaluation.

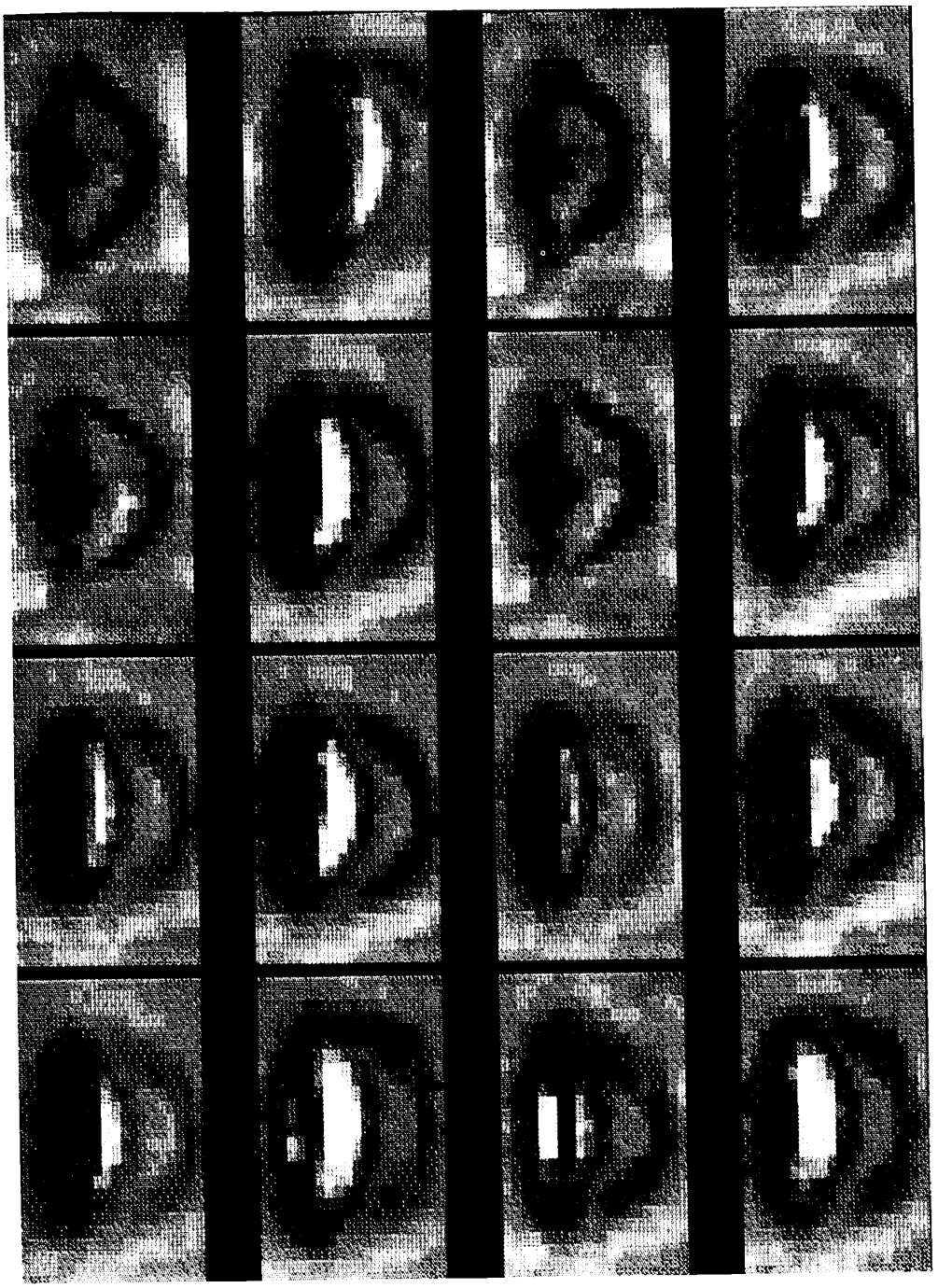


Fig. 1. Typical images used to train the neural networks.

taken from words and CVCs. Because these words and syllables were spoken deliberately and in isolation, these vowels were isolated easily. In other experiments not presented here data from a female speaker were also studied [36].

A. Preprocessing the Images

Instead of searching for an optimal encoding of the input images, we chose a simple representation that seemed to contain the relevant information. A rectangular area-of-interest was automatically defined and centered about the mouth. The image was further reduced to produce an image that could be comfortably handled by out network simulations. Within the rectangle, the average value of each 4 × 4 pixel squares was computed to produce a topographically accurate gray-scale image of 20 × 25 pixels (Fig. 1). Rather than attempting to extract special features, this encoding represented a form that could be obtained easily through an array of analog photoreceptors.

Two methods of processing these images of the speaker's mouth were explored. In the first approach, we treated the images categorically and attempted to make hard phonemic decisions directly from the images. Such linguistic identifications can be used to constrain the linguistic interpretation of a noise-degraded acoustic signal. In the second approach, we obtained acoustic information directly from the images by estimating the transfer function of the vocal tract. These independent estimates were then used to constrain the acoustic interpretation of the noise-degraded acoustic signal directly.

V. Categorization

Neural networks were trained to identify the vowel directly from the image. The images were presented across 500 input units, and the output consisted of 9 output units, each representing one of the nine vowels in the data. An input image was correctly categorized when the activation value of the correct vowel unit was larger than all the other output units. The data set of 108 images was split into a test set and a training set of 54 images, each containing a balanced set of vowels. The number of hidden units varied.

A network was trained until the categorization of all 54 images in the training set was perfect. Overtraining was minimized by immediately terminating the training at this point, before the output units were driven to saturation. In a few cases, the network would learn all but one or two tokens and then take an excessive amount of time to learn the last few cases. Often this additional training would result in poorer performance on the test set. The training was therefore stopped at 500 epochs whether or not all the training data were categorized correctly. After the network was trained, it then was tested on the second set of 54 images from the same speaker.

A. Results

The results reported here are based on networks with five hidden units; fewer than five hidden units produced worse results. Performance levels were averaged across eight networks initialized with different random weights. The networks were trained on 54 patterns. For half of the networks, the training and test sets were reversed. The eight networks trained on the male data obtained an average performance of 76% correct categorizations for the images in the test set.

This performance compared favorably with the traditional categorization technique of nearest neighbor. A nearest-neighbor classifier (NN) was constructed using the training data as the set of stored templates. The individual images from the test set were correlated with the stored templates, and the image was classified according to its closest match. The process was repeated, but with the test and training sets reversed. The NN classifier correctly classified and male data set with an average accuracy of 79%.

The performance of the network also compared favorably with two human subjects tested and trained on the same data. After 5 training sessions, the two subjects obtained an average of 70% on the images in the test set, with performances in some follow-up sessions approaching 80%. In Fig. 2, the types of errors made by the human subjects in these experiments are compared to those made by the network.

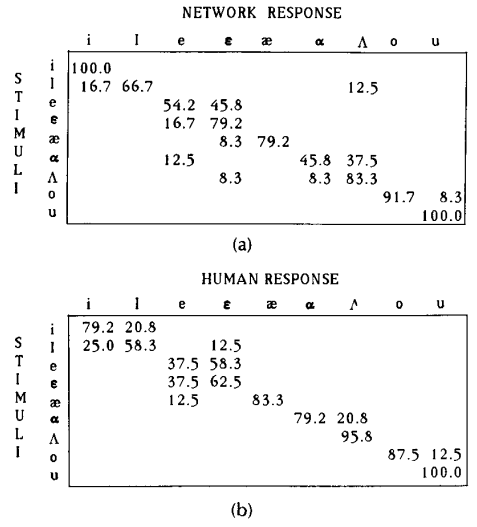


Fig. 2. Confusion matrices for a) networks trained to categorize individual images by vowel and b) well-trained human subjects categorizing the same images. The networks results are accumulated from four different networks. The human responses are accumulated from four trials by two subjects. Percentages less than 4.3% were omitted in order to simplify the matrix.

B. Discussion

Since steady-state vowels are relatively easy to identify acoustically, why is the performance less than perfect on the test images? Was it the lack of information in the images, or was it the lack of clear categories? To address the second part of this question, the short-term spectra of the corresponding acoustic data were examined. Networks were trained to categorize the acoustic spectra using the same procedures as used for the visual speech signals. The performance on the acoustic signals was almost identical to that of the visual signals, with the network obtaining 82% on the testing set. This suggests that some of the discrepancy between the performance on the training and test sets can be attributed to inherent ambiguity of the categories.

It is difficult, if not impossible, to correctly identify isolated speech segments out of context. The particular pro-

nunciation of a given vowel varies depending upon the particular context in which it is produced. This phenomena, called *coarticulation*, produces large variations in the production of a given vowel. For example, the same vowel takes on quite different physical realizations depending upon the surrounding phonemic context. In automatic speech recognition, the best performances have been obtained when commitment to a categorical decision is delayed to a higher level of processing that takes into account more contextual information [7], [9]. This allows additional constraints from both internal and external sources to be introduced and to assist in the decision making.

The problem becomes more acute when the speech segments are taken from continuous speech. When a vowel is in a stressed position within an isolated word, the coarticulation is significantly less dramatic than when the vowel is extracted from the middle of continuous discourse. The effect of these differences is seen in machine recognition, where unstressed words are harder to recognize than are stressed words [9]. To demonstrate this, we constructed a second data set of 108 images taken from continuous speech spoken by a second speaker. The data had the same distribution of vowels and the networks were trained in the same manner as before. On this data, the networks were able to achieve only 40% on the test set.[2]

VI. Subsymbolic Processing

Summerfield has proposed and evaluated a variety of ways in which information in the acoustic and visual signals might merge [6]. He concluded from psychoacoustic experiments that information from the two modalities must be integrated before phonetic or lexical categorization takes place. One striking observation was that an auditorially presented larynx-frequency pulse can be used to improve lipreading even though there is not enough information in the pulse alone to phonetically segment the signal.

The assumption made is that the acoustic and visual signal streams share a common representation at their conflux [6]. In this section, we propose that the vocal tract transfer function can serve as this common representation, and we show that networks can be designed for integrating visual and acoustic speech signals using this representation. An estimate of the vocal tract's acoustic characteristics are obtained directly from images of the speaker's mouth. This estimate then serves as an independent source of acoustic information and is used to constrain the interpretation of the acoustic signal.

A. The Corresponding Acoustic Signal

The acoustic speech signal is produced by a source signal that passes through the vocal tract and is emitted from the mouth [20]. For voiced speech, the driving signal is a quasi-periodic pulse train convolved with the glottal waveform. This driving signal's contribution to the short-term acoustic spectrum is a series of harmonics reducing in amplitude by −12 dB per octave. This reduction is partially compensated by the radiation of the acoustic signal from the lips, which produces an effective gain of +6 dB per octave. The spectral envelope of the short-term spectrum that remains after these two effects are removed is the frequency response of the vocal-tract filter. The transfer function of the vocal tract can be estimated by measuring the short-term spectral amplitude envelope (STSAE) of the acoustic signal.

There is not enough information in the visual speech signal to completely specify the vocal-tract transfer function. Many different acoustic signals can be produced by vocal-tract configurations that correspond to the same visual signal. Thus, the visual signals can provide only a partial description of the vocal-tract filter. Nonetheless, it may be possible to obtain a *good* estimate of the vocal-tract transfer function if additional constraints are considered. Neural networks with the architecture shown in Fig. 3 were trained to estimate the STSAE of the acoustic signal directly from the visual signals around the mouth. The estimate of the STSAE was then combined with estimates from acoustic information to improve the S/N ratio prior to recognition.

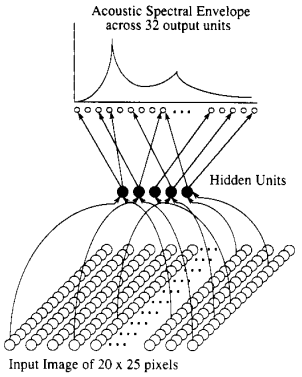


Fig. 3. Network architecture for estimating acoustic structure from visual speech signals. This feedforward network has all units in layer i connected to all units in layer $i + 1$. The output layer consisted of 32 units, each of which represented the amplitude of the vocal tract transfer function at a particular frequency and bandwidth.

The same images of the male speaker used in the categorization experiments were used in these experiments. Each video frame had 33 ms of acoustic speech associated with it. The short-term power spectra of the corresponding acoustic data were calculated and the spectral envelopes were obtained using cepstral analysis [22.]. Each smoothed envelope was sampled at 32 frequencies to produce a vector of scalar values. These vectors were used to represent the vocal-tract transfer functions corresponding to the images.

B. Training

The network shown in Fig. 3 was trained to produce the STSAE across its 32 output units when a visual signal was

[2]Differences in experimental design make it difficult to compare our results with the performance of human lip-readers measured in other studies. In most experiments, human subjects are usually exposed to dynamic information as vowels are presented within a larger context. Berger *et al.* tested the identification of twelve vowels in CV and VC syllables within a context, using live presentation [37]. His data show that lipreaders without training performed at a 53% accuracy level. Jackson *et al.* used 15 vowels and diphthongs in an /h/-V-/g/ context, and found that the average performance level across 10 viewers to be 54% correct [38]. Montgomery and Jackson repeated these experiments and found a mean performance level of 54.2% [4].

presented to the network across the input units. The network had five hidden units as in the categorization experiment. However, there were now 32 output units, each representing the linearly spaced samples of the short-term spectrum's envelope.

One of the consequences of training a network on a continuous mapping, rather than on a discrete categorization, was the problem of deciding when to stop training. We attempted to identify the point at which overlearning began by dividing the test set into two subsets. One subset was used to *track* the error during training by testing the subset after each training epoch as a measure of generalization. When the error of the tracking set started to increase, the training was stopped and the weights in the network were saved. This procedure was unnecessary when the training was simply stopped at 500 iterations.

C. Evaluation

Vowels are largely identified by their spectral shape, and in particular by the location of their spectral peaks, or formants [39], [40]. Nevertheless, evaluating the quality of these spectral estimates is significantly more difficult than judging the accuracy of a categorization because the perceptual processes involved in processing the spectral peaks are not well understood. To assay our spectral estimates, a simple vowel-recognition system was constructed (Fig. 4). The

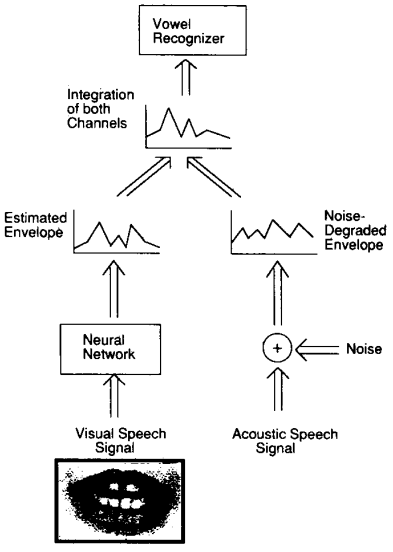


Fig. 4. System used to combine visual and acoustic speech information. A simple vowel recognizer was constructed to receive speech signals from the two modalities. Independent estimates of the vocal tract transfer function were produced and then combined with a weighted average before being passed to the recognizer. A neural network was trained to perform the mapping of the image into the estimated envelope of the acoustic spectra. Noise was introduced into the acoustic speech signal and the improvement due to the visual information was assessed.

vowel recognizer at the top of Fig. 4 was constructed using a simple feedforward network trained to recognize nine vowels from their STSAEs. The network was trained on 6 examples each of 9 different vowels until its performance was 100% on the training data. This network served as a *perfect* recognizer of the noise-free training data and was used to assess the benefit of the visually estimated spectra when combined with the noise-degraded acoustic spectra.

The vowel recognizer was presented with a STSAE through two channels. The path shown on the right in Fig. 4 was for information obtained from the acoustic signal, while the path on the left provided spectral estimates obtained independently from the corresponding visual speech signal.

The first step was to test the performance of the recognizer when the acoustic spectral envelopes were degraded by noise. Zero-mean random vectors were normalized and added to the training STSAEs to produce signals with S/N ratios ranging from −12 dB to 24 dB. Noise-corrupted vectors were produced at 3 dB intervals from −12 dB to 24 dB. At each noise level, 12 different vectors were produced for each of the STSAE in the set. At each level, the performances of the recognizer on the degraded signals were averaged. The overall performance on the training data fell with decreased S/N ratios. At − 12 dB, the recognizer operated at the chance level, which was 11% with nine vowels in the data set (Fig. 5).

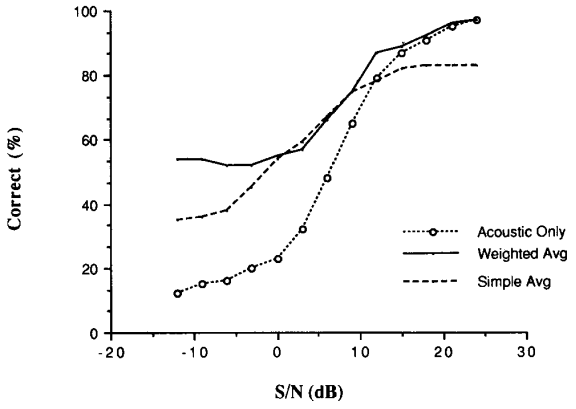


Fig. 5. Intelligibility of noise-degraded speech as a function of speech-to-noise ratio in dB. The lower curve shows the performance of the recognizer under varying signal-to-noise conditions using only the acoustic channel. The intermediate dashed curve shows the performance when the two independent estimates are equally weighted. The top curve shows the improved performance by using a weighting function based on the signal-to-noise. When the visual signal is used alone, the percent correct is 55% across all S/N levels.

The next step was to compensate for the noise degradation by providing an independent estimate of the STSAE from the visual signal, as shown on the left side of Fig. 4. The network on this pathway was trained to estimate the spectral envelopes corresponding to the input images. The data used to train this network were different from the data used to train the recognizer. The noise-degraded acoustic signal was then combined with the output from the network processing the images to provide a single estimate which is then passed on to the recognizer. When the two estimates were simply averaged together, the recognition rates were improved, as shown by the dashed curve in Fig.

5. At a S/N ratio of − 12 dB the recognizer performed at 35% compared to 11% without the visual signal.

However, averaging the two independent sources of information was far less than optimal at the extremes. Using the STSAE estimated from the visual signal alone, the recognizer was capable of a 55% recognition level. When these estimates were combined with the noise-degraded acoustic signal, the performance fell to as low as 35% at −12 dB S/N. Similarly, at very high S/N ratios, the fused inputs produced poorer results than the acoustic signal presented alone. Clearly, the acoustic and visual signals needed to be weighted according to their relative information content to compensate for the degraded performance at the S/N ratio extremes.

The two estimates of the spectral estimates of the vocal tract transfer function, $S_{\text{visually estimated}}$ and S_{acoustic} were combined with a weighting factor α that depends on the S/N ratio:

$$S_{\text{combined}} = \alpha S_{\text{visually estimated}} + (1 - \alpha) S_{\text{acoustic}} \quad (1)$$

At each S/N ratio, α was varied to optimize performance. The optimal α was found empirically to vary approximately linearly with the S/N ratio to 0 to 24 dB. The improved performance is evident in Fig. 5.

A third method of fusing the two spectra was accomplished using a σ-π neural network. These second-order networks took the estimated STSAE, the noise-degraded acoustic STSAE and a measure of the signal-to-noise ratio as input, and tried to produce a noise-free STSAE as output. In contrast to the simple weighted sum used by first-order units, the units in these second-order networks determine the activation level by summing the weighted product or other units' output [26]. The results from this method were mixed: although the squared-error between the estimated and actual spectra was significantly lower, its categorization was poorer. These results suggest that the vowel recognizer is doing something more complicated than simply making a comparison based upon a squared-error measure. They also raise questions as to the appropriateness of the squared-error measure used for training.

D. Comparing Performance

The quality of the networks' estimates were compared to a combination of two optimal linear-estimation techniques. The first step was to encode the images using a Hotelling or Karhunen-Loeve transform [41]. The images were encoded as five-dimensional vectors defined by the largest principal components of the covariance matrix of the images in the training set. This is an optimal encoding of the images with respect to a least-squared-error (LSE) measure. The next step was to find a mapping from these encoded image vectors to their corresponding short-term spectral amplitude envelopes (STSAEs). The fit was found using a linear least-squares fit.

The estimates obtained by this two-stage process were significantly poorer in overall mean-squared error. The mean-squared error of the estimates made by the networks was 46% better on the training set and 12% better on the test set. The main objective of this comparison was to show that arbitrary encoding of the images may result in a loss of relevant information. In contrast, the network learning algorithm allows the network to produce its own encoding at the hidden layer based upon relevant features. The activation levels of the five hidden units served to encode the image as did the five-dimensional vectors obtained using principal components. The primary difference is that the encoding found by the network optimized the desired output, whereas the principal components optimized the LSE reconstruction of the images.

E. Discussion

The recognition rates of noise-degraded acoustic signals were improved by introducing speech information extracted from the visual speech correlates. The relative improvement provided by fusing these signals varied with the S/N ratio, agreeing with the experimental data presented by Sumby and Pollack in their seminal 1954 paper [3]. This was accomplished without making hard decisions on the separate acoustic and visual sensory channels, and no explicit rules were needed to combine the information. In general, the psychoacoustic evidence suggests that the visual and acoustic speech signals interact in the human perceptual systems even before categorical cognitive processes are activated [42]-[45], [6].

In acoustic speech recognition, significant improvements can be made with existing systems by improving the quality of the signal at the earliest levels [9]. The approach described above provides a means of improving the input to existing speech recognition systems. The strength of this approach is more evident when networks are used on nonsegmental speech structures where categorical identification becomes even more difficult.

VII. Dynamics and Speech

In the work described above, attention was restricted to static visual images, which are inherently ambiguous because they contain incomplete information about the speech articulators. Speech is a dynamic process and the articulators are physical structures that move. As a given moment, their current positions are part of larger dynamic trajectories. These trajectories are constrained by the mechanics of the physical system and by the linguistic rules of the language. Dynamic dependencies could provide additional constraints that can serve to restrict the acoustic interpretation of the visual speech signal. In this section, we outline an approach to introducing dynamic constraints in neural network models.

One way to include temporal constraints is to map time into space, as in NETtalk [28]. In NETalk, consecutive letters were translated into a string of concurrent stimuli and presented to the network in groups of seven. The network was given a sequence of letters and asked to provide a phonetic transcription of the centrally located letter. A similar approach, called a time-delay neural network model, has been effective in acoustic speech-recognition systems [46], [47].

A different way to introduce dynamic constraints is to use feedback connections that provide temporal memory. One approach is to have projections from the output units to the input layer [48]. A second approach is to have projections from hidden units to the input layer [49]. These architectures are based on feedforward networks with a subset of the input units, called *state units*, receiving information from previous time steps. Thus, at time t, the network

receives inputs $I(t)$ and a subset of activations from the upper layers at previous time $t - 1$.

When working with static images, it was possible to use a simple vowel recognizer to test the quality and utility of the acoustic spectra estimated from static images. The success of the vowel recognizer depended on the careful selection of vowels from isolated words or syllables. For continuous speech, however, it is difficult and often impossible to make these definitive identifications of short speech segments taken out of context, so alternative assessments are necessary.

Networks with feedback were used to estimate the STSAE from images within a larger context. The performance of the network on continuous speech was evaluated on its ability to preserve the salient features of the spectral sequences, such as the resonances, or formants, of the estimated vocal-tract filter. The perception of vowels by humans depend upon the location and amplitude of these formants [39], with some of the highest quality machine speech being produced using formant-based synthesizers [50]. To see how well these formats were identified by the network, the sequences of spectra were arranged in a visual display similar to a spectrogram. The spectrogram shown in Fig. 6 was

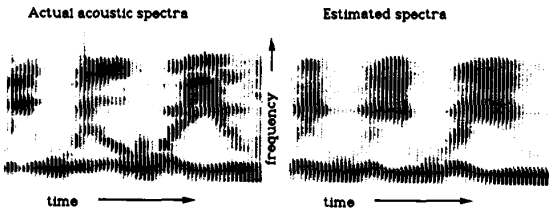


Fig. 6. Spectrograms created from the actual acoustic spectra are compared to visually estimated spectra for the sentence: "We will weigh you." Individual spectral estimates were converted to a grey scale and then aligned by frequency as a function of time. Actual acoustic data from the test set are shown on the left and estimates produced by the feedback neural network model are shown on the right.

created from spectra estimated from a sequence of images not in the training set. In this form, we can observe the changes of energy in the different frequency bands as a function of time. Clearly, much of the acoustic structure was being estimated in these sequences. The ultimate test will be to either resynthesize the acoustic speech signal from these estimated acoustic parameters, or to feed the fused spectra into a full-scale speech recognizer.

VIII. Conclusions

Under noisy conditions, speech recognition can be aided by extracting information from the visual speech signals and combining it with residual acoustic information. In this article we have examined two representations for the speech information in the visual signal, both of which can be combined with information from the acoustic signal. In the first case the visual signal was treated symbolically, while in the second it was used to provide subsymbolic information about the corresponding acoustic signal. These two cases are two points on a continuum of speech descriptions. Other descriptions, such as description of the articulators themselves [51] could also have been used.

A better understanding of the visual and acoustic sensory systems in humans and other animals will lead to better artificial sensors and their effective integration. Acoustic speech recognition systems, by using models of the human cochlea as a preprocessor, are already benefitting from what is known about the human auditory system [19], [52]. Synthetic cochleas that can process massive amounts of sensory data is real time already have been fabricated in analog VLSI [53]. The output of these chips is a highly distilled, parallel and distributed representation of the acoustic signal. Our results are an encouraging first step toward solving the problem of fusing multiple sources of distributed sensory data. Massively parallel network models could provide the means by which distributed representation can be integrated in real time for producing rapid recognition and decisive actions for automated systems.

IX. Acknowledgment

We would like to thank Nancy Giacobbe for her work on the human subject experiments.

References

[1] H. W. Ewersten, and H. Birk-Nielsen, "A comparative analysis of the audiovisual, auditive and visual perception of speech," *Acta Oto-Larynsol.*, vol. 72, pp. 201–205, 1971.
[2] N. P. Erber, "Auditory-visual perception of speech," *J. Speech Hearing Disorders*, vol. 40, pp. 481–492, 1975.
[3] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 26, pp. 212–215, 1954.
[4] A. Montgomery and P. L. Jackson, "Physical characteristics of the lips underlying vowel lipreading," *J. Acoust. Soc. Am.*, vol. 73, pp. 2134–2144, 1983.
[5] L. E., Bernstein, S. P. Eberhardt, and M. E. Demorest, "Single-channel vibrotactile supplements to visual perception of intonation and stress," *J. Acoust. Soc. Am.*, vol. 85, pp. 397–405, 1989.
[6] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading*. B. Dodd and R. Campbell, Eds. Hillsdale, NJ: Lawrence Erlbaum Assoc. Publishers, 1987.
[7] J. Allen, "A perspective on man-machine communication by speech," *Proc. IEEE*, vol. 73, pp. 1537–1696, 1985.
[8] CHABA (Committee on Hearing, Bioacoustics, and Biomechanics), *Removal of Noise From Noise-Degraded Speech Signals*. Washington, D.C.: National Research Council, National Academy Press, 1989.
[9] K. F. Lee, *Automatic Speech Recognition: The Development of the SPHINX System*. Boston: Kluwer Academic Press, 1989.
[10] E. D. Petajan, *An Improved Automatic Lipreading System to Enhance Speech Recognition*, AT&T Bell Laboratories Technical Report No. 11251-871012-111TM, Murray Hill, NJ, 1987.
[11] B. Lindblom, "Phonetic invariance and the adaptive nature of speech," in *Working Models of Human Perception*, B. Elsendoorn and H. Bouma, Eds. London, England: Academic Press, 1989, pp. 139–173.
[12] J. Lyons, *Introduction to Theoretical Linguistics*. Cambridge, England: Cambridge University Press, 1971.
[13] C. G. Fisher, "Confusions among visually perceived consonants," *J. Speech Hearing Res.*, vol. 11, pp. 796–803, 1968.
[14] B. E. Walden, R. A. Prosek, A. Montgomery, C. K. Scherr, and J. J. Jones, "Effects of training on the visual recognition of consonants," *J. Speech Hearing Res.*, vol. 20, pp. 130–145, 1977.
[15] R. Jacobson and M. Halle, *Fundamentals of Language*. The Hague, Netherlands: Mouton & Co., Publishers, 1956.
[16] O. Fujimura, "Relative invariance of articulatory movements: An iceberg model," in *Invariance and Variability in Speech Processes*, J. Perkell and D. H. Klatt, Eds. Hillsdale, NJ: Lawrence Erlbaum Assoc. Publishers, 1986.
[17] H. K. Dunn, "The calculation of vowel resonances, and an

electric vocal tract," *J. Acoust. Soc. Am.*, vol. 22, pp. 740–753, 1950.
[18] R. A. Cole, A. I. Rudnicky, V. W. Zue, and D. R. Reddy, "Speech as patterns on paper," in *Perception and Production of Fluent Speech*, R. A. Cole, Ed. Hillsdale, NJ: Lawrence Erlbaum Assoc. Publishers, 1980.
[19] F. Jelinek, "The development of an experimental discrete dictation recognizer," *Proc. IEEE*, vol. 73, pp. 1616–1624, 1985.
[20] G. Fant, *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton & Co., Publishers, 1960.
[21] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. Berlin: Springer-Verlag, 1972.
[22] J. D. Markel, and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin: Springer-Verlag, 1976.
[23] G. A. Miller, and P. E. Nicely, "An analysis of perceptual confusions among some English consonants, *J. Acoust. Soc. Am.*, vol. 27, pp. 338–352, 1955.
[24] T. J. Sejnowski, "Open questions about computation in cerebral cortex," *Parallel Distributed Processing in the Microstructure of Cognition*. Vol. 1, *Foundations*. J. McClelland and D. Rumelhart, Eds. Cambridge, MA: M.I.T. Press, 1986.
[25] F. J. Pineda, "Generalization of back-propagation to recurrent neural networks," *Phys. Rev. Lett.*, vol. 59, pp. 2229–2232, 1987.
[26] D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*. Cambridge, MA: M.I.T. Press, 1986, vol. 1.
[27] J. J. Hopfield, and D. W. Tank, "Neural computation of decisions in optimizations problems," *Biol. Cybern.*, vol. 52, pp. 141–152, 1985.
[28] T. J. Sejnowski, and C. R. Rosenberg, "Parallel networks that learn to pronounce English text," *Complex Systems*, vol. 1, pp. 145–168, 1987.
[29] T. Kohonen, *Self-Organization and Associative Memory*. Berlin: Springer-Verlag, 1984.
[30] G. A. Carpenter, "Neural network models for pattern recognition and associative memory," *Neural Networks*, vol. 2, pp. 243–258, 1989.
[31] K. Hornik, M. Stinchcombe, H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.
[32] Y. S. Abu-Mostafa, "The Vapnik-Chervonenkis dimension: Information versus complexity in learning," *Neural Computation*, vol. 1, pp. 312–317, 1989.
[33] H. White, "Learning in artificial neural networks: A statistical perspective," *Neural Computation*, vol. 1, pp. 425–464, 1989.
[34] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*. Cambridge, England: Cambridge University Press, 1988.
[35] L. E. Bernstein, and S. P. Eberhardt, *Johns Hopkins Lipreading Corpus I-II*, Johns Hopkins University, Baltimore, MD, 1986.
[36] B. P. Yuhas, *The Processing of Visual Speech Signals Using Parallel Distributed Processing*. Ph.D. dissertation, Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, 1990.
[37] K. W. Berger, "Vowel confusions in speechreading," *Ohio J. Speech and Hearing*, vol. 5, pp. 123–128, 1970.
[38] P. L. Jackson, A. A. Montgomery, and C. A. Binnie, "Perceptual dimensions underlying vowel lipreading performance," *J. Speech Hearing Res.*, vol. 19, pp. 796–812, 1976.
[39] G. E. Peterson, and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, vol. 24, pp. 175–184, 1952.
[40] R. L. Miller, "Auditory tests with synthetic vowels," *J. Acoust. Soc. Am.*, vol. 25, pp. 114–121, 1953.
[41] R. C. Gonzalez, and P. Wintz, *Digital Image Processing*. Reading, MA: Addison-Wesley, 1977, pp. 104–108.
[42] H. McGurk, and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
[43] H. McGurk, and J. MacDonald, "Visual influences on speech processes," *Perception & Psychophysics*, vol. 24, pp. 253–257, 1978.
[44] P. K. Kuhl, and A. N. Meltzoff, "The bimodal perceptions of speech in infancy," *Science*, vol. 218, pp. 1138–1141, 1982.
[45] E. Spelke, "Infants' intermodal perception of speech events," *Cog. Psychol.*, vol. 8, pp.. 553–560, 1976.
[46] D. W. Tank, J. J. Hopfield, "Neural computation by concentrating information in time," *Proc. Nat. Acad. Sci. USA*, vol. 94, pp. 1896–1900, 1987.
[47] A. Waibel, "Modular construction of time-delay neural networks for speech recognition," *Neural Computation*, vol. 1, pp. 39–46, 1989.
[48] M. I. Jordan, "Supervised learning and systems with excess degrees of freedom," COINS Technical Report 88-27, Computer and Information Science, Univ. of Massachusetts at Amherst, 1988.
[49] J. L. Elman, *Finding structure in time, Cognitive Science*, vol. 14, pp. 179–211, 1990.
[50] D. H. Klatt, "Speech perception: A model of acoustic-phonetic analysis and lexical access," in *Perception and Production of Fluent Speech*, R. A. Cole, Ed. Hillsdale, NJ: Lawrence Erlbaum Assoc. Publishers, 1980.
[51] O. Rioul, and B. S. Atal, personal communication, 1990.
[52] S. Seneff, *Pitch and Spectral Analysis of Speech Based on an Auditory Synchrony Model*, M.I.T. Research Laboratory of Electronics Technical Report 504, Massachusetts Institute of Technology, Cambridge, MA, 1985.
[53] C. Mead, *Analog VLSI and Neural Systems*. New York, NY: Addison-Wesley, 1989.

Ben P. Yuhas (Student Member, IEEE) received the B.A. degree in mathematics from the University of Chicago in 1981. He received the M.S. degree in electrical engineering and computer science in 1986, and the Ph.D. degree in electrical and computer engineering in 1990, both from Johns Hopkins University, Baltimore, MD.

He was an Associate Engineer at the Applied Physics Laboratory of Johns Hopkins University for four years before joining Bell Communications Research (Bellcore) as a Member of Technical Staff (MTS) in 1990. His research interests are in the use of distributed processing architectures for signal processing, recognition, and multimodal integration.

Dr. Yuhas is a member the International Neural Network Society, and the Acoustical Society of America.

Moise H. Goldstein, Jr. (Senior Member, IEEE) received the B.S. degree from Tulane University, New Orleans, LA, in 1949, and the M.S. and Dr.Sci. degrees from M.I.T. in 1951 and 1957, all in electrical engineering.

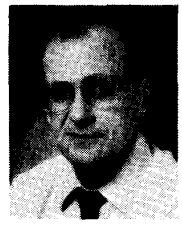
He was a faculty member at M.I.T. from 1955 until 1963, when he moved to Johns Hopkins University. He is Edw. J. Schaefer Professor of Electrical Engineering, and has a joint appointment in the Biomedical Engineering Department at the School of Medicine. The focus of his activities is basic research into speech processing and development of devices to aid profoundly deaf children.

Terrence J. Sejnowski (Member, IEEE) received the 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 the Johns Hopkins University in 1982. He 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, Cognitive Science, and Electrical and Computer Engineering at the University of California at San Diego. He is Director of the Institute for Neural Computation and the Center for Cognitive Neuroscience.

Dr. Sejnowski received a Presidential Young Investigator Award in 1984. He is the Editor-in-Chief of *Neural Computation*, published by the MIT Press. His main research interest is to understand how the brain works and to develop massively parallel computers based on the principles of neural computation.



Robert E. Jenkins (Member, IEEE) received the B.S. degree in engineering and the M.S. degree in physics from the University of Maryland, College Park.

He is a Lecturer in the School of Engineering and a Staff Engineer at the Applied Physics Laboratory, Johns Hopkins University, MD, leading the Space Department IR&D program. Mr. Jenkins is a member of the program committee for the part-time M.A. program in electrical engineering.