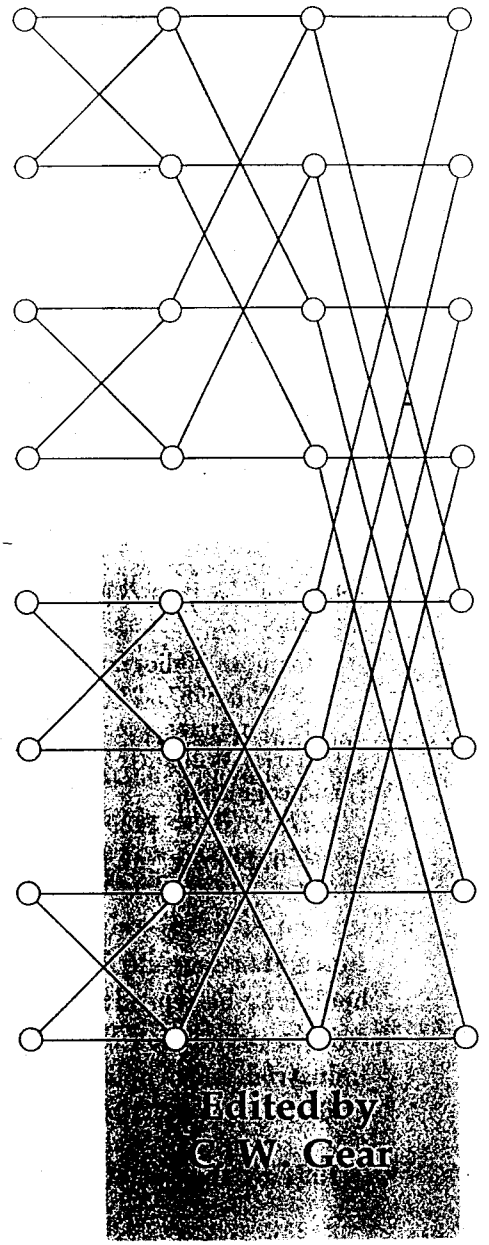
COMPUTATION & COGNITION

Proceedings of the First NEC Research Symposium



Society for Industrial and Applied Mathematics
Philadelphia

Library of Congress Cataloging-in-Publication Data

NEC Research Symposium (1st : 1989 : Princeton, N.J.)
Computation & cognition : proceedings of the First NEC Research Symposium / edited by C.W. Gear.
p. cm.
Includes bibliographical references.
ISBN 0-89871-272-6
1. Computer science–Research–Congresses. I. Gear, C. William (Charles William), 1935– . II. Title. III. Title: Computation and cognition.
QA76.27.N43 1989
004'.072–dc20 91-18523

Chapter 4
Mappings Between High-Dimensional Representations of Acoustic and Visual Speech Signals

Terrence J. Sejnowski* Ben P. Yuhas†

1 Introduction

The continued dramatic improvements in the speed and accessibility of general purpose digital computers has made it possible to model the brain at many different levels of investigation, from biophysical mechanisms to neural systems (Figure 1). However, the disparity between the computational power of even the fastest digital computers and that of the brain limits present simulations to a tiny fraction of the entire brain (Sejnowski [13]), or to creatures with only a few neurons (Selverston [15]). It is possible, for example, to model the information flow inside the dendrites and axons of single neurons, taking into account realistic anatomical and physiological details (Koch & Segev [8]). The study of networks of model neurons is just beginning. Some progress has been made by simplifying the details of single neurons in the network; such "neural network" models are primarily concerned with how the architecture of a network affects its capacity to perform a task and how the size of the network scales with the complexity of the task. In addition, systems-level architectural constraints from the brain at the level of columns of neurons, maps of neurons, and processing hierarchies can also be explored by modeling studies (Sejnowski & Churchland [14]).

One major difference between digital computers and brains is in the organization of memory. In conventional digital computers, memory is separated from the central processing unit and there is a communications bottleneck between them. Memory is expensive so that there is a premium on exploring algorithms that minimize the need for information storage. In contrast, the brain has a much greater capacity for storing information. The nervous system of man has approximately 10^{12} neurons and 10^{16} synapses, or connections between pairs of neurons. If every synapse could store only 1 bit

*The Salk Institute and University of California, San Diego, La Jolla, CA.
†Bell Communications Research, Morristown, New Jersey.

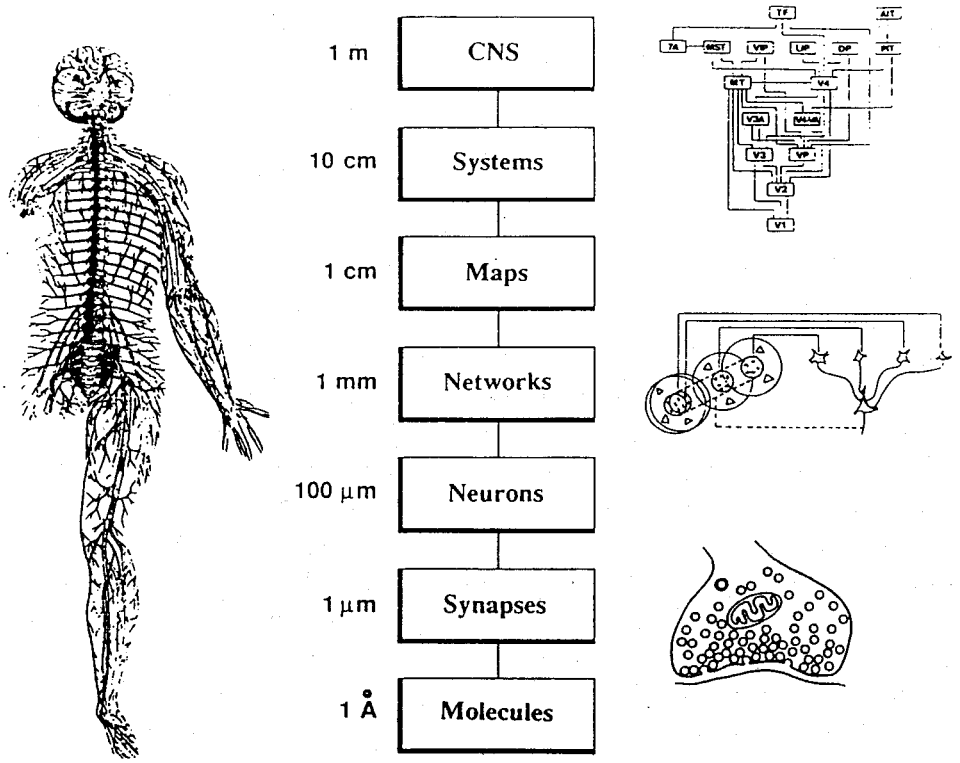


Figure 1:
Levels of investigation in the nervous system. The spatial scale on the left of the figure is related to the anatomical structures represented by each box. The schematic diagrams on the right illustrate the hierarchy of visual processing centers (top), the integration of input from the retina to form oriented receptive fields in the visual cortex (center), and the structure of a chemical synapse (bottom).

of information, the total capacity of the brain would still be over 10 million megabytes. Furthermore, all of this information is available on line, since the nonlinear processing elements in neurons are intertwined with the storage elements. These numbers probably underestimate the complexity of the brain because they do not take into account the continuous variables that are dynamically active during the processing within neurons and synapses. For example, there are biochemical mechanisms within synapses that modulate the strengths of synapses on a short time scale, over milliseconds, and others that produce changes that can last for years. It is these biochemical mechanisms that allow us to remember what happened a few minutes ago and to recall events in our childhood (Squire [16]).

Another major difference between brains and digital computers is in the way that information is manipulated. Algorithms for digital computers exploit the ability of a central processing unit to perform sequences of operations with great accuracy. In contrast, the logical depth that can be implemented by neural systems is not nearly as great—our ability to compute recursive functions without a paper and pencil is relatively weak. Finally, the brain should really be compared to a large number of computers, perhaps several hundred, rather than a single one. Each of these subsystems is dedicated to a different function, but they are able to communicate and cooperate to accomplish difficult tasks in real time. Some of these special purpose computers have neural circuits that can be reorganized by learning to solve new problems. This type of memory is programmed by experience.

What types of algorithms would run efficiently on architectures that resembled those of brains? Because memory is abundant, it would no longer be necessary to form the most compact representation for a problem. Thus, objects in the world and relationships between them could be represented in high-dimensional spaces, and the entire representation could be permanently stored. Even the brain, however, does not have enough memory capacity to handle the complexity of the world if the only representations available were in the space of the stimuli and all possible relationships had to be explicitly stored. For example, the visual representations of objects at the level of pixel intensities is not a good one for expressing categorical relationships. Consequently, visual processing, which comprises nearly half of cerebral cortex, extensively modifies the representation of an object to make it partially invariant to photometric variables and spatial transformations. Similarly, relationships must also be represented in an invariant way. These high-level representations are still very rich ones that contain much information about the physical properties of the stimulus, including properties from other modalities and even information about the use of the object (Damasio [3]). It is these high-level representations that are used by the brain to perform content-addressed retrieval, to make perceptual decisions, and to perform sensory-motor coordination, such as reaching out and grasping an

object.

Representations in artificial intelligence are primarily symbolic. Even when thinking in terms of massively-parallel networks, there is a tendency to use discrete, low-dimensional representations, which have computational as well as conceptual advantages (Feldman [5], Valiant [18]). However, at least as a working hypothesis we would like to explore the possibility that cognitive tasks could be performed in the brain by mappings between high-dimensional spaces that constitute high-level representations of the sensory world and our possible interactions with it. Thus, the goal of our research is to understand a number of interlocking problems: What properties should high-dimensional distributed representations have to make them robust, efficient and flexible? Can mappings between these representations be performed that honor the computational structure of the tasks that must be accomplished? What can we learn from human performance that can help constrain possible network solutions?

In this paper we will explore these questions in the context of a specific problem in speech processing. The traditional approaches to speech recognition start with acoustic signals and end up with symbolic representations of distinctive features, phonemes, syllables, words, phrases and sentences. This approach ignores the speech information contained in other sensory modalities, such as the visual speech signals from the face of the speaker. Other sources of information relevant to speech include gestures, facial expressions, and even face color through stimulation of the autonomic nervous system. If the ultimate goal of a speech system is to extract semantic information from the speech stream, then these alternative sources of information could be important and perhaps make the interpretation of the acoustic signals much easier.

Petajan [11] has explored the visual speech signals for isolated digit recognition. In his system, the acoustic and visual speech information were independently reduced to symbol strings, and a set of rules was used to reconcile conflicting interpretations. The symbolic intermediates were needed to allow the necessary processing and integration to be performed 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 high-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. We will summarize recent results in visual speech recognition based on this new approach (Yuhas et al. [21]).

2 Neural Networks

The primary computational technique we used was mappings between high-dimensional vector spaces implemented with multilayer neural network models. The key features of these models are a large number of relatively simple nonlinear processing units and a high degree of connectivity between these units. A unit performs a nonlinear transformation on the sum of its inputs to produce a continuous 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. These models differ significantly from actual neural circuits 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. Nevertheless, these networks provide a starting point for finding alternative approaches to difficult computational problems.

Feedforward network architectures were used in most of this study (Figure 2). The units in a feedforward network were 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 (Figure 2). The standard multilayered feedforward networks with arbitrary squashing functions are a class of universal approximators (White [19]). 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 (White [20]).

A modified backpropagation algorithm was used to train feedforward networks (Rumelhart et al. [12]). The gradient was calculated in the standard manner, but instead of using steepest descent, a conjugate-gradient algorithm was used to update the weights. 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, by adding terms to the cost function that penalize unnecessary weights, or by stopping the training before the network has completely converged.

There is a natural statistical interpretation for the signals carried on the

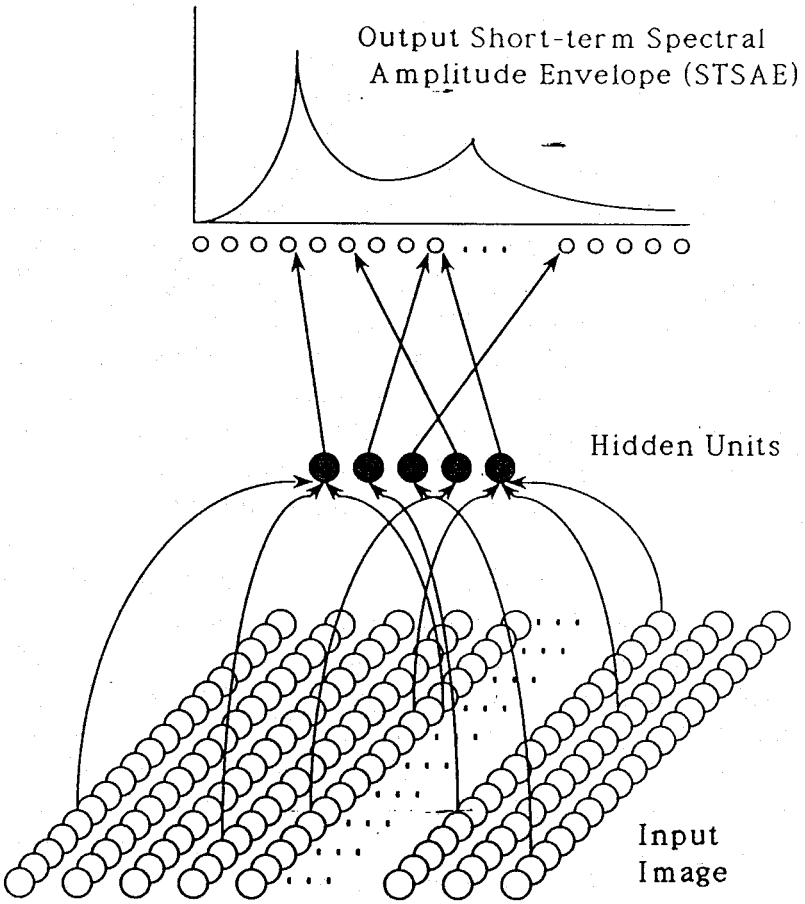


Figure 2:
Network architecture used for estimating the acoustic structure from visual speech signals. The feedforward network has 500 input units all connected to 5 hidden units, each in turn are fully connected to 32 output units. Each output unit represents the amplitude of the vocal tract transfer function at a particular frequency. The processing units in the network have a nonlinear input-output function given by: $f(x) = 1/(1 + exp(-x))$.

output units when the inputs are noisy. If the probability distribution of output units is Gaussian around a desired mean (for a mapping task), then the mean squared error used to train the network is the maximum likelihood cost function if the output units are linear (Bridle [2]; Rumelhart and Durbin, personal communication). A similar result holds for the output units if the probability distribution is binomial (for a binary categorization task) and the output units are sigmoidal. Then the correct maximum likelihood cost function is mutual information or information gain (Hinton & Sejnowski [6]).

3 The Speech Signals

The speech signals used 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 (Bernstein & Eberhardt [1]) where the individual frames and their corresponding speech segments were indexed. The NTSC video standard was used (30 frames/sec) and each frame had 33 milliseconds (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 word or consonant-vowel-consonant (CVC) type nonsense syllables that change relatively slowly. In these contexts, the vowels often were steady state over periods of 50 to 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 acoustic signal was approximately in steady state. From this set, a single frame was selected only if the periodic wave form 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 taken from words and CVCs. Because these words and syllables were spoken deliberately and in isolation, these vowels were isolated easily. Data from a female speaker were also studied.

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 our network simulations. Within the rectangle, the average value of each 4 x 4 pixel square was computed to produce a topographically accurate grey-scale image of 20 x 25 pixels. Rather than attempt 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.

The acoustic speech signal emitted from the mouth can be modeled as the response of the vocal-tract filter to a switchable sound source. In a first-order vocal-tract model, the configuration of the articulators (e.g, the mouth opening, the 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. While some of the articulatory features are often visible (e.g., the lips, 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 those effects due to the hidden articulators. This complementary structure can be exploited to improve the perception of speech in noise.

4 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 were 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. After the network was trained, it then was tested on the second set of 54 images from the same speaker.

Performance levels were averaged across eight networks having five hidden units, each 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. A nearest neighbor classifier was constructed using the training data as the set of stored templates and the results compared with the performance of the neural network model. 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 again, but with the test and training sets reversed. The nearest neighbor classifier correctly classified the 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%. The types of errors made by the human subjects in these experiments were similar to those made by the network as judged by comparing the confusion matrices.

5 Precategorical Fusion

Summerfield [17] concluded from psychoacoustic experiments that information from the visual and acoustic modalities must be integrated before phonetic or lexical categorization takes place. The implication was that the acoustic and visual signal streams shared a common representation at their conflux. We have used the vocal tract transfer function as a model for this common representation, and we have shown that networks can be designed for integrating visual and acoustic speech signals using this representation (Yuhas et al. [21]). An estimate of the vocal tract's acoustic characteristics was obtained directly from images of the speaker's mouth. This estimate then served as an independent source of acoustic information and was used to constrain the interpretation of the acoustic signal.

The acoustic speech signal is produced by a source signal that passes through the vocal tract and is emitted from the mouth. For voiced speech, the driving signal is a quasi-periodic pulse train convolved with the glottal wave form. 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 acoustical 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. A feedforward neural network was 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 signal-to-noise ratio prior to recognition. 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. 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.

Vowels are largely identified by their spectral shape, and in particular by the location of their spectral peaks, or formants. 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 is not a well-understood process. To assay our spectral estimates, a simple vowel recognition system 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 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 Figure 3 was for the 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 signal-to-noise 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

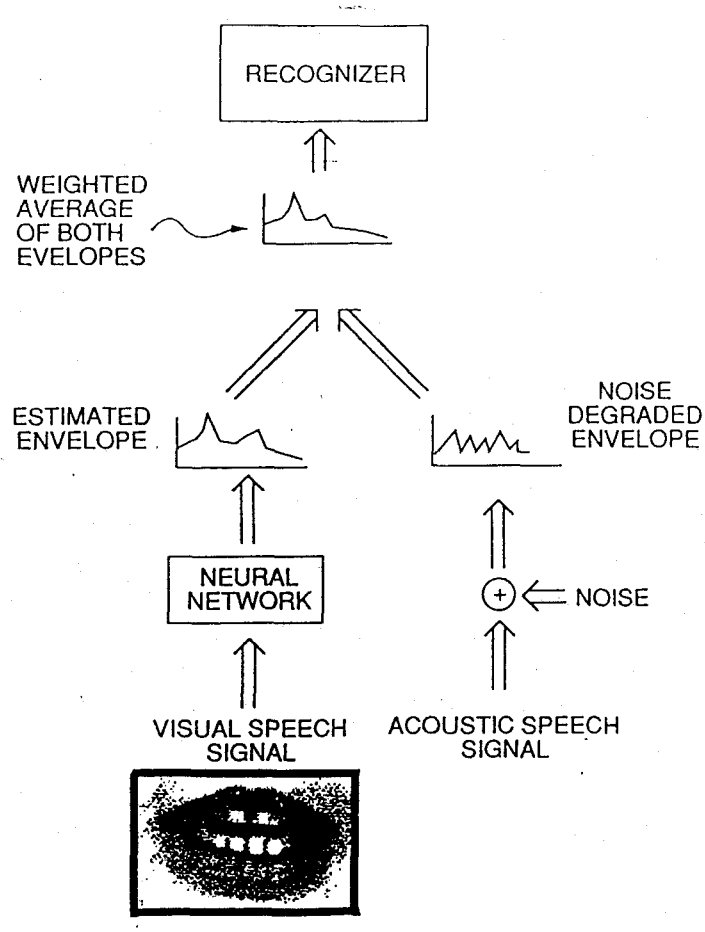


Figure 3:

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.

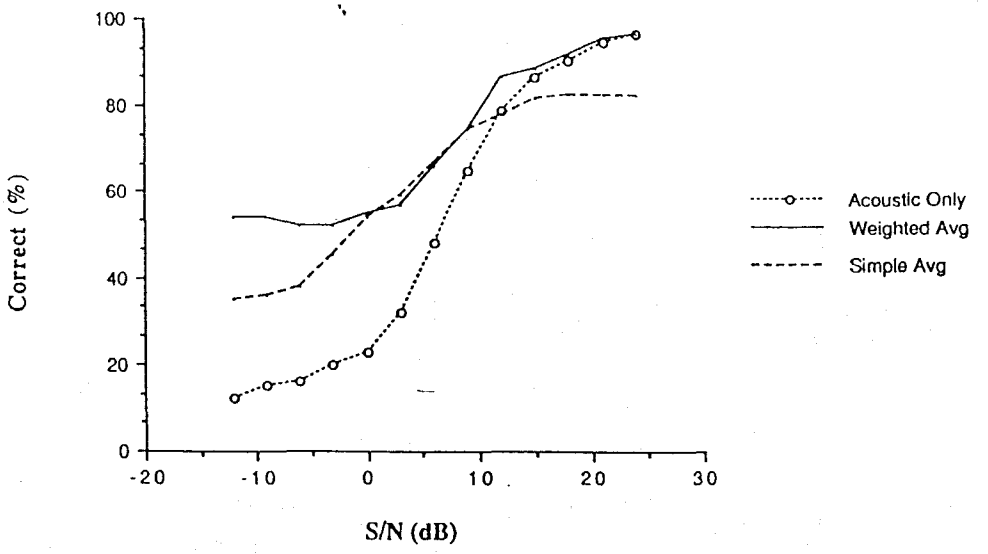


Figure 4:
Intelligibility of noise-degraded speech as a function of speech-to-noise ratio in dB. The lower curve shows the performance of the recognizer under varying signal-to-noise conditions using only the acoustic channel. The intermediate dashed curve shows the performance when the two independent 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.

averaged. The overall performance on the training data fell with decreased signal-to-noise ratios. At -12 dB, the recognizer operated at the chance level, which was 11% with nine vowels in the data set.

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 Figure 3. 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. The acoustic and visual signals were weighted according to their relative information content to compensate for the degraded performance at the signal-to-noise ratio extremes. The optimal value of the weighting was found empirically to vary approximately linearly with the signal-to-noise ratio in dB, from 1 at -12 dB signal-to-noise ratio to 0 at 24 dB. The performance is shown in Figure 4. Another method of fusing the two spectra was accomplished using a sigma-pi neural network (Rumelhart et al. [12]). 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. The results from this method were mixed: while the squared-error between the estimated and actual spectra was significantly lower, their 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. It also raises questions as to the appropriateness of the mean squared-error measure used for training.

The quality of the estimates made by the networks 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. 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 were 46% better on the training set and 12% better on the test set. This comparison shows 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, while the principal components optimized the LSE reconstruction of the images.

6 Dynamics and Speech

In the models described thus far, attention was restricted to static visual images, which were inherently ambiguous because they contain incomplete information about the speech articulators. Speech is a dynamic process and the articulators are physical structures that move. 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.

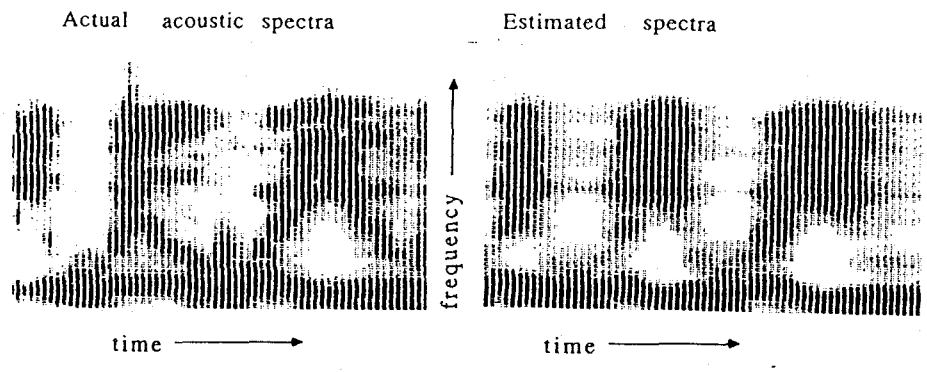


Figure 5:
Spectrograms created from the actual acoustic spectra are compared to visually-estimated spectra for the sentence: "We will weigh you". Individual spectral estimates were converted to a grey scale and then aligned by frequency as a function of time. Actual acoustic data from the test set are shown on the left and estimates produced by the feedback neural network model are shown on the right.

In this section, we outline an approach to introducing dynamic constraints in neural network models. One approach is to have projections from the output units to the input layer (Jordan [7]) or from hidden units to the input layer (Elman [4]).

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. To see how well these formants were identified by the network, the sequences of spectra were arranged in a visual display similar to a spectrogram. The spectrogram shown in Figure 5 was 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.

7. Discussion

Under noisy conditions, speech recognition using acoustic information alone degrades and performance can be aided by extracting information from the visual speech signals and combining it with residual acoustic information. Two representations for the speech information in the visual signal were studied. 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 are two points on a continuum of speech descriptions. Other representations of the speech signals, such as descriptions of the articulators themselves, could also have been used. It would valuable to know what representations are used in the brain. 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.

By combining the visual and acoustic sources of speech information, we have demonstrated that the visual signal can be used to improve the performance of automatic vowel recognition in the presence of noise. This approach did not require categorical preprocessing or explicit rules. The performances of these neural networks compared favorably with human performance and with other pattern-matching and estimation techniques. Our results were based on vowels spoken by single speakers, but this same approach can be extended to multiple speakers and to consonants. Improvements can also be made in the input representations. Synthetic cochleas that can process massive amounts of sensory data in real time already have been fabricated in analog VLSI (Mead [10]). The output of these chips is a highly distilled, parallel and distributed representation of the acoustic signal. These front-ends could improve the overall level of performance of acoustic speech recognition systems, but they would not change our conclusions concerning the need to compensate for noise – they only put off the inevitable.

The results from the specific examples studied in this paper can be generalized to many other problems that depend on the fusion of information from several cues or from several modalities (Lehky et al. [9]). The key idea is to represent the information in a distributed way and to rely on high-dimensional mappings from these cues into a common representation. Learning algorithms can be used to seamlessly combine the two information streams, and to continuously adapt in nonstationary environments. At present, we are only able to guess which representations are likely to be good ones, based in part on what we know about the representations in the brain. We need a deeper understanding of distributed representations that can guide us in these choices. It is also likely that more sophisticated neural architectures will be needed to deal with the fusion of information, especially when there are conflicting sources of information.

Nature has been an inspiration for many mathematical discoveries. Much

of functional analysis grew out of attempts to understand the physical world. The biological world is also a source of inspiration but the complexity of biological systems often exceeds our abilities to develop simple, analyzable, mathematical models. This is especially true in the study of the brain, a biological system with a degree of complexity greater than that of any other known system. As we learn more about the brain, and as we explore the function of the brain with a wide variety of mathematical and computational models, we may begin to develop an understanding of the computational principles of the brain comparable to our mathematical understanding of the physical world.

References

[1] Bernstein, L. E., Eberhardt, S. P., *Johns Hopkins Lipreading Corpus* **I-II**, Johns Hopkins University, Baltimore, MD (1986).

[2] Bridle, J. S., Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman-Soulie (Ed.), *Neuro-Computing: Algorithms, Architectures and Applications*, Springer-Verlag, Berlin (1989).

[3] Damasio, A. R., Category-related recognition as a clue to the neural substrates of knowledge, *Trends in Neuroscience*, **13**, 95-98 (1990).

[4] Elman, J. L., Finding structure in time, *Cognitive Science*, **14**, 179-211 (1990).

[5] Feldman, J., Technical Report TR-189: Neural representation of conceptual knowledge, University of Rochester Department of Computer Science (1986).

[6] Hinton, G., Sejnowski, T., Learning and relearning in Boltzmann machines. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Psychological and Biological Models,* **Vol. 1**, *Foundations* MIT Press, Cambridge, MA, 282-317 (1986).

[7] Jordan, M. I., Supervised learning and systems with excess degrees of freedom, COINS Technical Report 88-27, Computer and Information Science, University of Massachusetts at Amherst (1988).

[8] Koch, C., Segev, I., *Methods in Neuronal Modeling: From Synapse to Networks*, MIT Press, Cambridge, MA (1989).

[9] Lehky, S. R., Pouget, A., & Sejnowski, T. J., Neural models of binocular depth perception. In E. R. Kandel, T. J. Sejnowski, C. F. Stevens,

J. D. Watson (Eds.) *Cold Spring Harbor Symposia on Quantitative Biology: The Brain*, **55**, Cold Spring Harbor, New York, Cold Spring Harbor Press (1990).

[10] Mead, C., *Analog VSLI and neural systems*, Addison-Wesley, Reading, MA (1989).

[11] Petajan, E. D., An improved Automatic Lipreading System To Enhance Speech Recognition, AT&T Bell Laboratories Technical Report No. 11251-871012-111TM, Murray Hill, NJ (1987).

[12] Rumelhart, D. E., Hinton, G. E., & Williams, R. J., Learning internal representations by error propagation. In J. L. McClelland & D. E. Rumelhart (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* **Vol. 1**, *Foundations* MIT Press, Cambridge, MA (1986).

[13] Sejnowski, T. J., Computing with connections: Review of "The Connection Machine" by W. Daniel Hillis. *Journal of Mathematical Psychology*, **31**, 203-210 (1987).

[14] Sejnowski, T. J., Churchland, P. S., Brain and cognition. In M. I. Posner (Ed.), *Foundations of Cognitive Science*, MIT Press, Cambridge, MA (1989)

[15] Selverston, A. I., A consideration of invertebrate central pattern generators as computational data bases, *Neural Networks*, **1**, 109-117 (1988).

[16] Squire, L., *Memory and Brain*, Oxford University Press, Oxford (1987).

[17] Summerfield, Q., Some preliminaries to a comprehensive account of audio-visual speech perception. In B. D. a. R. Campbell (Ed.), *Hearing by Eye: The Psychology of Lip Reading*, Lawrence Erlbaum Assoc., Hillsdale, NJ (1987).

[18] Valiant, L., this book.

[19] White, H., Some asymptotic results for learning in single hidden layer feedforward networks, *Journal of the American Statistical Association*, **85** (1989).

[20] White, H., Learning in artificial neural networks: A statistical perspective, *Neural Computation*, **1**, 425-464 (1990).

[21] Yuhas, B. P., Goldstein, M. H., Jr., Sejnowski, T. J. & Jenkins, R. E., Neural network models of sensory integration for improved vowel recognition, *Proc. IEEE* (October, 1990).