专 题 汇 编

IJCNN' 语声处理

SPEECH RECOGNITION IN NOISE USING A SELF-STRUCTURING NOISE REDUCTION MODEL AND HIDDEN CONTROL MODELS

Helge B.D. Sorensen

Speech Technology Centre, Institute of Electronic Systems
Aalborg University, DK-9220 Aalborg, DENMARK

1. ABSTRACT

This paper describes how speech recognition in the presence of F-16 jet cockpit noise can be performed using a sequence of three units - an auditory model and two neural models. A method for noise reduction in the cepstral domian based on a self-structuring universal approximator is proposed and tested on a large database of isolated words contaminated with jet noise. This approach is a potential alternative to traditional recognition methods for noisy speech and makes noise reduction possible in all three models as in the system in [1]. The first model performs a spectral analysis of the input speech signal. The second model is a Self-structuring Neural Noise Reduction (SNNR) model, which is an alternative to the noise reduction model [1] presented at IJCNN91. The noise reduced output from the SNNR network is propagated through the speech recognizer consisting of a set of Hidden Control Neural Networks (HCNN).

2. INTRODUCTION

Signals from digital signal processing systems are often corrupted by stationary or non-stationary noise and noise reduction is thus necessary. An interesting new non-parametric approach to noise reduction is the use of connectionist models [1],[2]. Noise reduction can be considered as a non-linear mapping from a noisy signal space to a noise-free signal space. Hornik, Stinchcombe and White [5] have proved that multilayer feedforward neural network models are universal approximators i.e., they are capable of representing arbitrarily accurate approximations to arbitrary mappings, if the models are sufficiently complex. Halbert White [6] has proved that these approximations are learnable using universal approximators. The theoretical conclusion is thus that a complex non-linear mapping from a noisy signal space to a noise-free signal space should be possible and learnable using a universal approximator.

The application described in this paper is speech recognition in noise. As an introduction two connectionist approaches for noise reduction will be described. Noise reduction can be performed directly on the noisy speech samples using Tamura and Waibels Time Domain Noise Reduction (TNR) network [2]. At IJCNN91 we presented an alternative approach [1], which is based on a concatenation of an auditory model and two neural models for speech recognition of isolated words in noise. Noise is reduced using the Cepstral Noise Reduction multilayer neural (CNR) network. The output from the CNR network is input to a neural network classifier. The CNR network is a general noise reduction model usable in both isolated and continuous noisy speech recognition systems.

The above mentioned noise reduction systems are based on training neural models with fixed model architectures, which are selected before training. A neural model with a self-structuring architecture established during training would be a better solution. This paper presents a noisy speech recognition method based on three models - an auditory model, the SNNR network and a Hidden Control Neural Classifier. During training the SNNR network builds a network architecture and this can be controlled by the performance function, which measures the difference between targets and target estimates. Thus a near optimal network architecture can be constructed during training, see the following sections.

0-7803-0559-0/92 $3.00 © 1992 IEEE

3. SPEECH RECOGNITION IN NOISE USING THE SNNR NETWORK

The speech recognition and noise reduction system is illustrated in figure 1 and described in the following. The system consists of a concatenation of three models - an auditory model, a SNNR network and a Hidden Control Neural Classifier.

3.1 Homomorphic preprocessing

LPC-based cepstral analysis is a widely used preprocessing technique in speech recognition systems. This method and cepstral analysis based on an auditory model were compared in [1] as preprocessing for the CNR network, and we found that the auditory model was more noise robust than the traditional LPC-based cepstral analysis module, presumably due to the critical-band filtering in the auditory model. Therefore we only apply the auditory model in this work, see figure 1. Hermansky's auditory model [7] was selected. The noise corrupted speech signal $\tilde{s}(n)$ sampled at 8 kHz is analysed every 10 msec and 256 samples is input to the first stage, which performs a critical-band filtering. The output from stage three is an "auditory spectrum" $\underline{a}$ consisting of 30 filter outputs, which are transformed into "cepstral" coefficients using four steps. Details can be found in [1]. From these coefficients 12 delta cepstral coefficients are calculated. Thus the input to the SNNR network is a homomorphic vector consisting of 12 cepstral and 12 delta cepstral coefficients. The output from the preprocessing module for each word pattern is a sequence of noisy homomorphic vectors. The sequence is time normalized into a sequence of K (= 40) noisy homomorphic vectors, where j is an index for the j'te input pattern:

$$M_j = [\underline{\tilde{m}}(1, j)...\underline{\tilde{m}}(k, j)...\underline{\tilde{m}}(K, j)] \quad (1)$$

The corresponding sequence of K noise-free homomorphic vectors is given by:

$$M_j = [\underline{m}(1, j)...\underline{m}(k, j)...\underline{m}(K, j)] \quad (2)$$

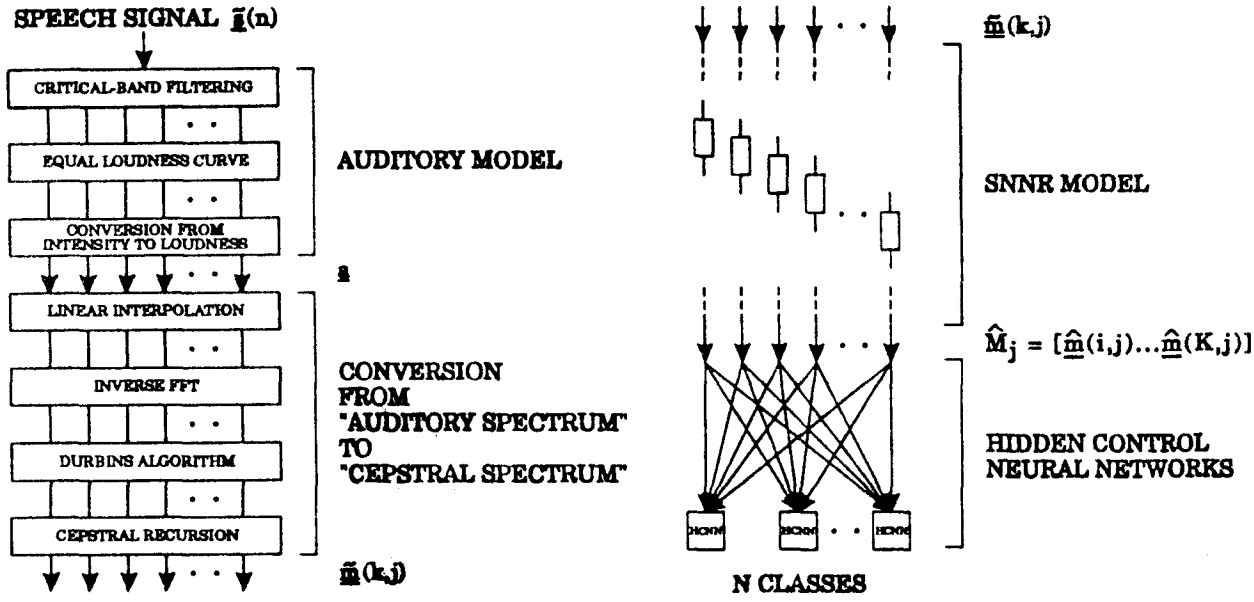


Figure 1: The noise reduction and speech recognition system. $\underline{\tilde{m}}(k,j)$ is a noisy homomorphic vector from (1). $\hat{M}_j$ is an estimate of the sequence M_j of noise-free homomorphic vectors. The SNNR network is described in details in figure 2.

3.2 The Self-structuring Neural Noise Reduction (SNNR) model

The noise reduction model in [1] is non-self-structuring, because the network architecture is selected before training. As a starting point for an alternative approach we have selected the Cascade-Correlation network [3]. We have tried different model modifications described in the following. The proposed architecture in figure 1 is denoted as the Self-structuring Neural Noise Reduction (SNNR) network.

3.2.1 Model training and test

Input to the SNNR network architecture is 24 noise contaminated coefficients and before training the network has the architecture illustrated in figure 2A. The number of output (linear) Processing Elements (PEs) are equal to the number of input terminals, because the goal is to learn the network to perform a non-linear autoassociative mapping between each pair of noisy homomorphic vector in (1) and the corresponding noise-free homomorphic vector in (2).

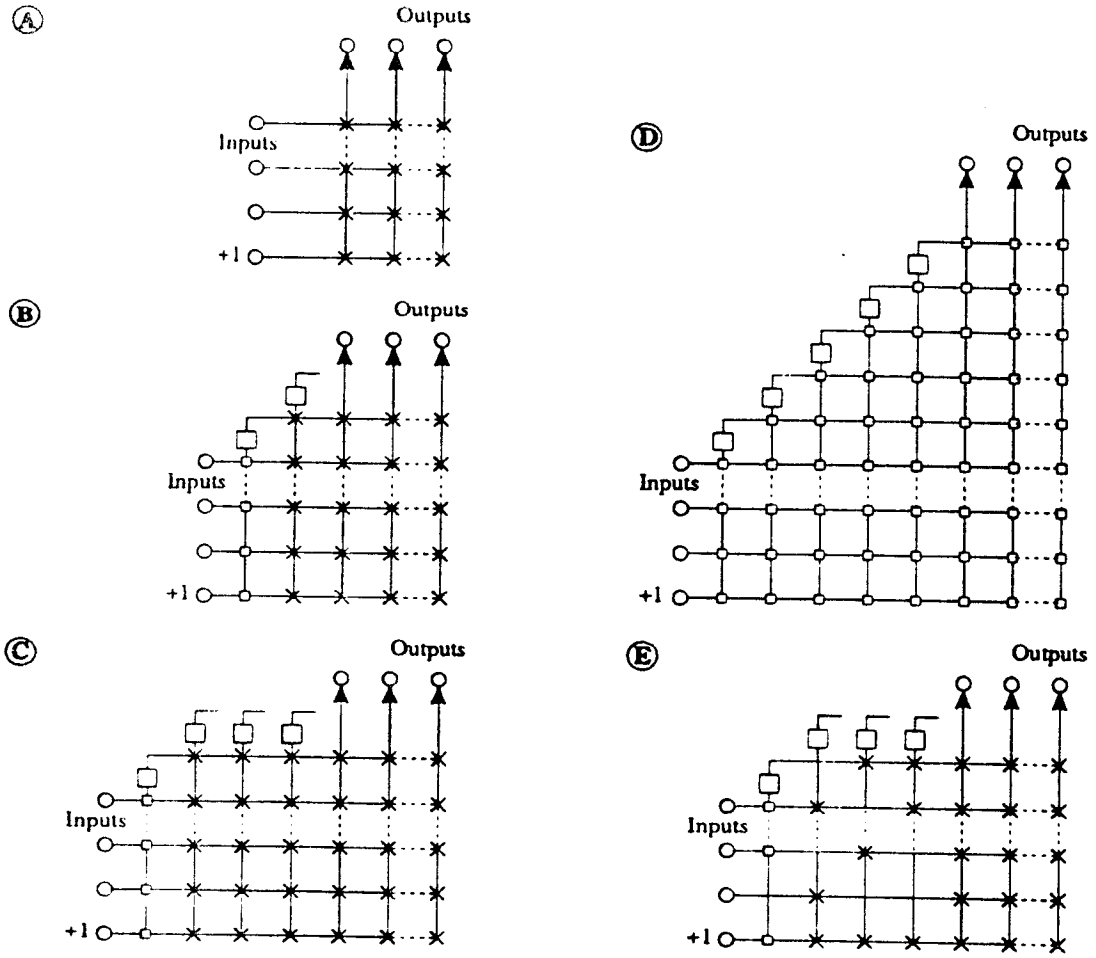


Figure 2: The self-structuring network. Boxed connections are frozen and x connections are trained repeatedly. (A) Network architecture before training. (B) Training has started. One hidden PE has been added and a candidate PE is tested. (C) A pool of PE candidates (D) After training. One of the SNNR networks we have applied. (E) Example of fan-in reduction. No x indicate no weight.

Before training the SNNR network architecture is equal to the minimal network in figure 2A. Then the network automatically trains and adds new hidden PEs one by one constructing a cascade multilayer structure as indicated in figure 2B. When a hidden PE is added to the network, its input weights are frozen. Based on all output PEs the error function E is calculated as in multilayer perceptron networks. For each candidate PE the correlation between V, the candidate output value and, E_o, the residual output error observed at an output PE is calculated. The correlation function S is the summation of the correlation values for all training patterns and for all output PEs. The pseudo code in this section summarizes the training procedure [3].

```
Initialize weights
While E large do
    While E continues to decrease do
        Train output weights
    Create a pool of new PE candidates
    While S continues to increase do
        Train each candidate weights
    Select the PE candidate with the largest S
    Install this PE and freeze weights
    Initialize this PE's output weights
    with minus the correlation
```

Figure 2C illustrates how a pool of new PE candidates looks. This pool is PEs with different input-output relations e.g. sigmoidal or gaussian and thus a larger area of the weight space is investigated during training. The network determines its own size and topology based on the performance measure E and the definition of the stop criteria in the loop 'While E large do'. The training of the network is fast due to the fact that weights are frozen, thus there is no need to back-propagate through hidden PEs. Therefore we can use a single-layer training algorithm for this multilayer network! After training the SNNR network this is applied by feed forward recall i.e. forward pass of the noisy input. Figure 2D gives an example of one of the SNNR networks we have applied using only 5 hidden PEs.

3.2.2 Model reduction

We consider the self-structuring neural model as a very powerful model, compared to non-self-structuring models. The only disadvantage is in applications demanding many hidden PEs, the network then becomes deep and the PEs get high fan-in. We have defined and tested a set of possible strategies to minimize fan-in. Some of the possible heuristic modifications of the training procedure are: (**Method 1**: Full fan-in.) **Method 2**: For the selected candidate PE: Install only weights larger than w_l. **Method 3**: Set the fan-in for all candidate PEs to f_l. Fan-in from some of the first hidden PEs are thus eliminated. **Method 4**: Define fan-in for each candidate PE to f_l randomly selected input connections. See figure 2E.

We assume that a test of these methods on the well known and difficult "two-spirals" benchmark [3] is a good evaluation of these. The solution of the two-spirals problem is to separate two interlocking spirals. For each method 50 models were trained to solve the problem and the results in table 1 are the average performance.

Deleting small weight connections is the best method and this result seems reasonable, because these connections assumably transmit the smallest amount of information. These simulations indicate that simplified model reduction without considerable performance loss is possible. The training time for each method was approximately the same as for the network with full fan-in. Future experiments with the SNNR network should include model reduction, if deep networks are developed during training.

Method no.	Definition of fan-in	Average no. of PE's	Average no. of weights
1	Full fan-in	19	298
2	$w_1 = 0.4$	18	218
2	$w_1 = 1.0$	20	187
3	$f_1 = 11$	25	398
4	$f_1 = 11$	20	247

Table 1: For each method 50 models were generated and average number of PEs and weights were calculated. All weights below w_1 were removed before installation of a candidate PE using method 2. f_1 = constant number of fan-in connections to candidate PEs using method 3 or 4.

3.3 Speech recognition

In our previous noisy speech recognition work [1] we selected a multilayer neural network for the recognition task. Speech is output of a non-linear time-varying system. The Hidden Control Neural Network [9] is a more powerful recognizer, because it can model linear/non-linear and time-varying systems in contrast to standard multilayer neural networks capable of modelling linear/non-linear time-invariant systems. The speech recognizer consists of 10 small HCNNs, one for each word pattern, see figure 1.

4. EXPERIMENTS AND RESULTS

The databases applied, experiments performed and results achieved are presented in the following sections. We would like to emphasize that these results are preliminar and can be improved as indicated in the last section.

The training and test database for the experiments were selected from a multi-speaker speech database containing approx. 1600 isolated Danish words, i.e. the digits 0-9 pronounced by males and females. The training database contained 880 words from 22 females and 22 males. The test database contained 800 words from 20 females and 20 males. The noise was extracted from a CD-ROM, produced by IP-TNO [4]. Non-stationary F-16 jet cockpit noise was chosen and selected randomly from the CD-ROM for the experiments. The speech and noise were added at five different signal-to-noise ratios: 21, 15, 9, 3 and 0 dB. Only results with SNR of ∞, 21 and 0 dB are presented in the following. Initially training of the SNNR network was based on only one version of the training database with a specific SNR, see table 2, Method A.

SNR in dB	∞	21	0
- SNNR model	94.8	66.4	15.3
+ SNNR model (Method A)	-	81.9	40.4
+ SNNR model (Method B)	-	91.5	79.8

Table 2: The first two rows show the recognition results for the system in figure 1 on the test database in percentage without and with the SNNR network. The training of the SNNR network was based on only one version of the training database with a specific SNR. The third row shows the improved results achieved when training the SNNR network on five different SNR versions of the training database. Notice that the HCNNs were trained with noise-free data. If the HCNNs are trained with noise-reduced data then the recognition rate can be improved furthermore.

Training of the same network with different SNR (21, 15, 9, 3 and 0 dB) versions of the training database resulted in a better noise reduction and thus a better recognition, see table 2, Method B. The network architecture for the SNNR network in all experiments were (F_A, F_B, F_C, F_D, F_E, F_F, F_G) = (24, 1, 1, 1, 1, 1, 24). F_A = 24 indicate that the input field has 24 inputs and F_B = 1 is the number of hidden PEs in the first hidden field. Two different training strategies were applied for the HCNNs. The first is to train the HCNNs using sequences of noise-free homomorphic vectors and the second is to train the HCNNs using sequences of noise reduced homomorphic vectors from the SNNR network. The latter strategy is the most natural, because it is this kind of input the HCNNs will recieve on-line. Table 2 presents some of our preliminary results.

5. RELATIONS TO OTHER METHODS

A comparison of the system in figure 1 and the speech recognition and noise reduction system in [1] is relevant due to the fact that they both apply a connectionist approach to both the noise reduction and the speech recognition problem. The following aspects indicate some of the improvements: The noise reduction network in figure 1 is self-structuring and can provide a more optimal architecture, few hidden PEs are necessary, faster learning, dynamic inputs i.e. delta cepstral coefficients and HCNN is a better speech recognizer. The results in table 2 were achieved using only 5 hidden PEs in the noise reduction network. In our previous work [1] we applied 64 hidden PE! The noise used in this work was selected randomly compared to the noise applied in [1].

6. CONCLUSION

We believe that the SNNR network is a very powerful method for noise reduction in general and that the preliminary results presented above can be improved. Some of the possibilities for improvement we are investigating at the moment are: Optimal weighting of the cepstral coefficients. A better training stop criteria for the SNNR network. Larger SNNR network. More representative training database for the SNNR network. Comparison of the performance of different speech recognizers in the system in figure 1. Training of the speech recognizer with noise reduced data instead of noise-free data.

7. ACKNOWLEDGEMENTS

Thanks to Professor Paul Dalsgaard, Head of the Speech Technology Centre at the Institute of Electronic Systems, Professor Uwe Hartmann from the Institute of Electronic Systems, A. Jørgensen, M. Drejer and J. Nonboe for supporting this work. Thanks to S.E. Fahlman and R.S. Crowder from Carnegie Mellon University for a copy of the Cascade-Correlation architecture.

REFERENCES

[1] H.B.D. Sorensen, "Noise-Robust Speech Recognition Using a Cepstral Noise Reduction Neural Network Architecture", in Proc. IJCNN91, Seattle, USA, July 8-12, 1991.

[2] S. Tamura and M. Nakamura, "Improvements to the noise reduction neural network", in Proc. ICASSP90, pp. 825-828, Albuquerque, USA, April 1990.

[3] S.E. Fahlman, C. Lebiere, "The Cascade-Correlation Learning Architecture", Advances in Neural Information Processing Systems 2, Edited by D.S. Touretzky, Morgan Kaufmann Publishers, San Mateo, California, 1990.

[4] H.J.M. Steeneken, F.W.M. Geurtsen, "Description of the RSG-10 noise database, report No. IZF 1988-3, TNO Institute for perception, Soesterberg, Netherlands, 1988.

[5] K. Hornik, M. Stinchcombe and H. White, "Multilayer feedforward networks are universal approximators", Neural Networks, Vol. 2, pp. 359-366, 1989.

[6] H. White, "Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings", Neural Networks, Vol. 3, pp. 535-549, 1990.

[7] H. Hermansky, "Perceptually based linear predictive analysis of speech", in Proc. ICASSP85, Tampa, USA, 1985.

[8] S.E. Fahlman, "Faster-learning variations on back- propagation: An empirical study", in Proc. of the 1988 Connectionist models summer school, Morgan Kaufmann.

[9] E. Levin, "Word recognition using Hidden Control neural architecture", in Proc. ICASSP90, pp. 433-436, Albuquerque, USA, April 1990.

submitted to IJCNN-92, Baltimore MD

Neural network lipreading system for improved speech recognition

David G. Stork,* Greg Wolff* and Earl Levine†

* Ricoh California Research Center
2882 Sand Hill Road, Suite 115
Menlo Park CA 94025-7022
stork@crc.ricoh.com

† Department of Electrical Engineering
Stanford University
Stanford CA 94305

Abstract

We designed and trained a modified time-delay neural network (TDNN) to perform both automatic lipreading ("speech reading") in conjunction with acoustic speech recognition in order to improve recognition both in silent environments as well as in the presence of acoustic noise. The speech reader subsystem has a speaker-independent recognition accuracy of 51% (in the absence of acoustic information); the combined acoustic-visual system has a recognition accuracy of 91%, all on a ten-utterance speaker-independent task. Most importantly, with no free parameters, our system is far more robust to acoustic noise and verbal distractors than is a system *not* incorporating visual information. Specifically, in the presence of high amplitude pink noise the low recognition rate in our acoustic only system (43%) is raised dramatically to 75% by the incorporation of visual information. Additionally, our system responds to (artificial) conflicting cross-modal patterns in a way closely analogous to the McGurk effect in humans.

We thus demonstrate the power of neural techniques in several crucial and difficult domains: 1) pattern recognition, 2) sensory integration, and 3) distributed approaches toward "rule-based" (linguistic-phonological) processing. Our results suggest that speech reading systems may find use in a vast array of real-world situations, for instance high noise environments such as factory and shop floors, cockpits, large office environments, outdoor public spaces, and so on.

Corresponding author and presenter

Dr. David G. Stork
Ricoh California Research Center
2882 Sand Hill Road Suite 115
Menlo Park CA 94025-7022
stork@crc.ricoh.com
415-496-5720 (w)
415-854-8740 (fax)

Technical area

1st choice: Speech recognition. 2nd choice: Applications.

Presentation preferred

Oral

Audio Visual requirements

Overhead projector
Video player

0-7803-0559-0 /92 $3.00

submitted to IJCNN-92, Baltimore MD

Neural network lipreading system for improved speech recognition

David G. Stork,* Greg Wolff* and Earl Levine†

* Ricoh California Research Center
2882 Sand Hill Road, Suite 115
Menlo Park CA 94025-7022
stork@crc.ricoh.com

† Department of Electrical Engineering
Stanford University
Stanford CA 94305

Abstract

We designed and trained a modified time-delay neural network (TDNN) to perform both automatic lipreading ("speech reading") and acoustic speech recognition in order to improve recognition both in silent environments as well as in the presence of acoustic noise. The speech reader subsystem has a speaker-independent recognition accuracy of 51% (in the absence of acoustic information); the combined acoustic-visual system has a recogntion accuracy of 91%, all on a ten-utterance speaker-independent task. Most importantly, with no free parameters, our system is far more robust to acoustic noise and verbal distractors than is a system *not* incorporating visual information. Specifically, in the presence of high amplitude pink noise the low recognition rate in an acoustic only system (43%) is raised dramatically to 75% by the incorporation of visual information. Our system responds to (artificial) conflicting cross-modal patterns in a way closely analogous to the McGurk effect in humans.

We thus demonstrate the power of neural techniques in several crucial and difficult domains: 1) pattern recognition, 2) sensory integration, and 3) distributed approaches toward "rule-based" (linguistic-phonological) processing. Our results suggest that speech reading systems may find use in a vast array of real-world situations, for instance high noise environments such as factory and shop floors, cockpits, large office environments, outdoor public spaces, and so on.

Introduction

Automatic artificial speech recognition is notoriously hard, and no computer system approaches the human ability to recognize spoken language amidst variations in speaker (accents), gender, rate, degree of coarticulation — all in the presence of acoustic distractors and noise [20]. Current automated systems are of lower accuracy and higher brittleness than that necessary to fulfill the vast need in computer speech-to-text conversion, automatic translation, speech control, etc. Representative approaches include hidden Markov models (HMM), in which transition probabilities are encoded in links between nodes representing phonemic segments, and blackboard methods, in which multiple special purpose subsystems (phonological, lexical, grammtical, etc.) work synergistically to maximize a recognition score . More recently, neural network techniques have been applied with some success in limited domains [23]. Hybrid systems have tried to incorporate attractive features from several component methods.

Any predictive source of information and constraints that could be incorporated into an artificial system would be desirable, and traditionally most research has focussed on the inclusion of grammatical, syntactic and other higher linguistic information. It is clear that humans can employ information other than just the acoustic signal. In particular, humans, especially the hearing impaired, can utilize visual information — speech reading — for improved recognition accuracy [3,6-9,11-13,19,21,22]. Speech reading can provide direct

0-7803-0559-0 /92 $3.00

information about speech segments and phonemes, as well as rate, speaker gender and identity, and subtle information for segmenting speech from background noise.

The "Cocktail party effect," in which speech corrupted by crowd noise is drastically more intelligible when the talker can be seen, provides strong evidence that humans use visual information in speech recognition [4]. Likewise, the "McGurk effect," in which artificially conflicting bi-modal stimuli are presented, reveals the influence of visual information on the perception of speech [12]. Thus, for example, if a /bi/ (+front) is presented visually and a /gi/ (+back) is presented acoustically, the listener will perceive a /di/ — "averaging" these features to get +middle.

One theoretical analysis gives further impetus for the incorporation of visual information in speech recognition. According to followers of the motor theory of speech perception [14], sound is merely the medium; it is the speech articulations that are the true speech signal. Hence a more direct access to these articulations through vision would be expected to improve perception.

Previous speech reading systems

Several speech reading systems have been developed recently. Petajan et al. [16,17] used thresholded images of a talker's face during the production of a word. They used a dictionary of pre-stored labelled utterances and a standard minimum distance classifier for visual recognition. Pentland and Mase [15] used an optical flow technique to estimate four velocity values (upper and lower lips, and the corners of the mouth) from the raw pixel video image of a mouth. They then performed a principal components analysis and standard mimimum distance classifier on three- and four-digit phrases. Yuhas et al. [24,25] trained a neural network using *static* images of the mouth shape for vowel recognition. Moreover, their system employed an omniscient controller (which adjusted the relative weights of visual and auditory contributions) for best recognition in different amounts of acoustic noise.

Our approach differs from these in several ways. Whereas Petajan et al. saved entire video sequences of pixel maps for categorization — an extremely slow and memory intensive procedure, even during recall — we recorded merely the positions of ten markers, a drastic reduction in information. Our system is thereby much faster and speaker *in*dependent as well. Whereas Pentland and Mase preprocessed the raw video image to estimate visual velocities, we used special markers in order to focus on the higher level speech issues. (Presumably a final practical system would employ preprocessing of the raw video image.) We also had more input dimensions then they did, allowing subtler distinctions to be made. Whereas Yuhas et al.'s treated static images of vowels, we treat instead *temporally changing* data in full consonant-vowel or vowel-consonant phonemes. Moreover, whereas Yuhas et al. employed an omniscient controller who adjusts the relative weight of different evidence based on acoustic signal-to-noise ratio, our system is fully autonomous, requiring no such controller. We are unique, too, in our use of the structured TDNN backpropagation architecture.

Data acquisition and preprocessing

Our raw visual data consisted of the positions of ten reflective markers placed on the talker's face and sampled at 60 Hz by means of the two-dimensional motion tracker of Motion Analysis Corporation. This data was preprocessed to yield five numbers at each 1/60 s interval, insensitive to tipping of the talker's face in the frontal plane (Figure 1). (We Fourier transformed several utterances and found that there was insignificant energy above 15 Hz in the visual data, and thus our sampling rate of 60 Hz is more than adequate.) Each of these five values were then normalized to have value 0.0 at the resting position, and 1.0 at the maximum positive excursion, and resampled at 100 Hz (to be compatible with the acoustic data).

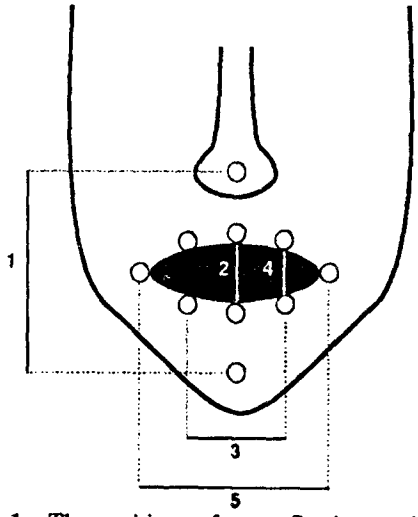


Figure 1: The positions of ten reflective markers on each talker's face were recorded at 60 Hz, then transformed into feature vectors having five components, as shown. Feature 1 is the nose-chin separation, which therefore represents the lowering of the jaw. Feature 2 is the vertical separation of two markers on the top and bottom lip, which therefore represents mouth opening. Feature 3 is the average of the horizontal separations of the two pairs of points (northeast and northwest, and the southeast and southwest quadrants, as shown). Feature 4 is, analogously for the vertical separations of these points.

Acoustic data was taken simultaneously by a cardioid microphone and later performed in software off-line. Acoustic information was preprocessed through a bank of fourteen mel-scale filters and sampled every 10 ms [1]. All patterns were one second in duration, synchronized and segmented by hand in software, off line.

We present here results on just ten spoken letters by five talkers. The particular ten letters were chosen because they illustrate crucial issues in speech reading: b, d, f, m, n, p, s, t, v, z. For instance, /bi/ and /di/ are acoustically very similar, as are /em/ and /en/. Each of these pairs illustrates, however, a significant *visual* difference: /bi/ has closure whereas /di/ does not; /em/ has closure wherease /en/ does not. Converesely, /bi/ and /pi/ are visually identical, though are easily disambiguated by differences in voicing.

Network architecture and training

We explored several network architectures, all based on a modified time delay neural network (TDNN) [23]. Figure 2 shows such a network (***VO***) for performing classification based on five visual features only. In all networks, the input was 1000 ms in duration, sampled at 10 ms intervals; this corresponds to array of input units 100 wide. The ***VO*** network (for video only) has five feature values at each 10-ms interval. Each (sigmoidal) hidden unit received signals from all five features at twenty intervals (200 ms), as shown in black in Figure 2. We found that 200 ms was a typical time scale for visual features. The array of hidden units was 81 wide (in order to "cover" all input units), and four high (chosen for moderate data compression). The next layer consisted of x-units, each having an exponential transfer function. Because the fan-in from the hidden units was chosen to be ten (for broad input coverage), the x-unit array is 72 units wide. It is 10 units high, corresponding to the ten letter categories, as listed. In any interval, the relative activities of the x-units encode the probabilities of the spoken letters at that interval. The final layer consists of just ten p-units (probability units), which encode the relative probabilities of the presence of the letters [10]. Each p-unit receives signals from the entire width of the x-unit array, but from only the row corresponding to its letter category. In the TDNN architecture, weights are "shared," i.e., the pattern of input-to-hidden weights is forced to be the same at each interval. Thus the total number of independent weights in this ***VO*** network is 800. The parameters here and below were chosen somewhat heuristically, and are not necessarily optimal.

The network utilizing only acoustic data, ***AO***, had the same structure as the ***VO*** network shown in Figure 2, but differed in the parameters. The number of input features for each 10-ms interval was fourteen, corresponding to fourteen mel-scale coefficients (activations) from 0 Hz to 5 kHz. As with the ***VO*** network, the ***AO*** network had four

hidden units at each interval. The fan-in from input to hidden unit layer was five units (50 ms), a typical time scale for most important acoustic features (e.g., bursts, transitions, etc.). The fan-in from hidden to x-unit was 25 units wide. This latter value was chosen in order to insure that x-units in both the ***AO*** and ***VO*** networks ultimately received information from the same number of input intervals — corresponding to 300 ms. There were 1280 independent weights in the ***AO*** network.

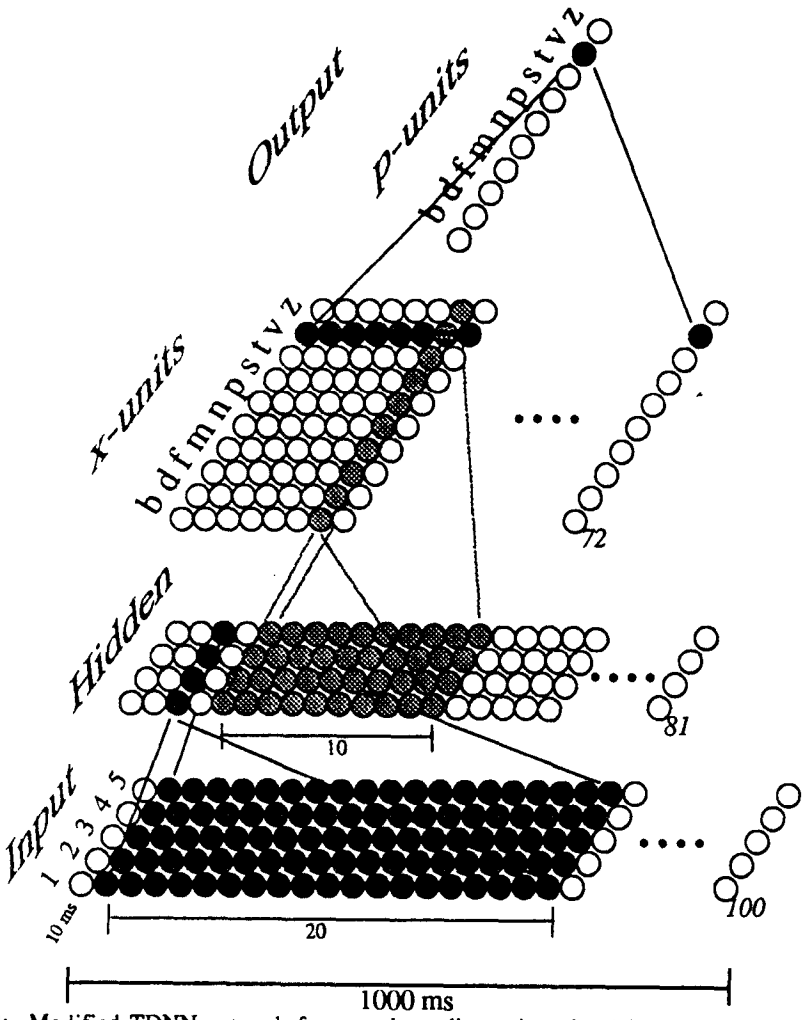


Figure 2: Modified TDNN network for speech reading using visual input only (***VO*** network). The inputs are the five feature values (cf. Figure 1), sampled to give interpolated values at every 10 ms. The hidden units are standard sigmoidal transfer functions and have fan-ins from 5 x 20 input units (200 ms of data, black). The x-units have exponential transfer functions and have fan-ins from 10x4 hidden units (gray). Each of the final ouput units (p-units, or "probability units") have fan-ins from the entire duration of x-units, and hence represent the probability that a letter was spoken at any time within the 1000 ms input window. In experiments incorporating acoustic information, the network had somewhat different parameters (see text).

The ***AxV*** (audio times video) network received both acoustic *and* video input. This network was built from the ***VO*** and the ***AO*** networks (just described) by taking their p-unit outputs to form a new, final set of ten output units. The activity of a final output unit for a category was:

$$p(C|A\&V) = k\ p(C|A)\ p(C|V). \qquad [1]$$

That is, the probability of occurence of a letter category C, given both acoustic and visual evidence, is proportional to the product of the conditional probabilities based on the acoustic and video evidence taken alone. (To insure that the final output probabilities sum to 1.0, we included the normalization constant k.) This "independent opinion pooling" in a Bayesian framework is strictly valid when the acoustic and video information is independent at the level of individual utterances [2], as is achieved in our data. The product given in Eq. 1 can be achieved in a network by sigma-pi neural units. The total number of independent weights in the ***AxV*** network is 2080.

The fourth network we considered, the ***full AV*** network, had the same input-to-hidden connections as the ***AO*** and the ***VO*** networks, hence leading to an array of eight hidden units at each interval (i.e., eight units "high"). Particularly important in our ***full AV*** network was the fact that each x-unit could receive information from *both* the acoustic and the visual hidden units, by means of added connections. The total number of independent weights in this network was 3280. (The most general network would permit integration of video and acoustic information at the *hidden* layers too, but for technical and training time reasons, we did not investigate such networks.)

Error was defined as:

$$E = -ln[\text{output}_c] \qquad [2]$$

where output_c is the activity in the output unit corresponding to the target category. This error is zero when the target unit's output is 1.0. Because of normalization of the output activities, output_i and thus E depend upon the weights to *all* output units. Thus, even though error is defined at a single output unit, *all* weights are updated in this scheme [10].

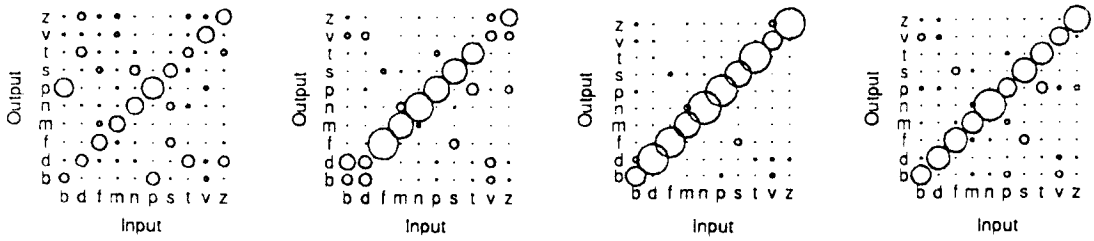
The training of the ***VO*** and ***AO*** networks was by means of backpropagation (learning rate $\eta = 0.001$, momentum $\alpha = 0.9$) [18]. Training on four patterns/talker/letter for five talkers typically required 250 epochs for minimum validation error (determined using one test pattern/talker/letter). The ***AxV*** network required no further training, since the outputs of the trained ***VO*** and ***AO*** networks were merely combined (multiplied) by means of sigma-pi units with equal weights. The ***full AV*** network was constructed by merging the trained ***AO*** and ***VO*** networks (learning parameters: $\eta = 0.0002$, momentum $\alpha = 0.9$). To this large net were added small weight cross-modal connections between the hidden layer to the x-units. Then all weights in the full network were trained to minimum test error.

Results

Figure 3 shows the average output probabilities given by our trained networks. Note espeically the confusion of the /bi/ and /pi/ phonemes in the ***AO*** network. These phonemes differ solely in the invisible distinction of voicing. Note too the acoustic similarity of the /em/ and /en/ phonemes, as represented in the large cross terms; these phonemes differ primarily in the acoustically subtle feature of nasality. These two letters are clearly distinguishable *visually*, as shown in the ***VO*** network. Thus in the ***AxV*** network /em/ and /en/ are disambiguated far better than in the ***AO*** network. (A similar analysis holds for other phonemes.)

Conversely, /di/ and /ti/ are confused in the ***VO*** network, but not the ***AO*** network. This pair, too, is thus disambiguated better in the ***AxV*** than the ***VO*** network. (A similar analysis holds for several pairs of phonemes, as the reader can verify.)

We would expect the performance on the ***full AV*** net to be better than the ***AxV*** net, since the former can learn associations at an earlier level. The fact that we do not find this result suggests that we have insufficient training data; spurious low-level correlations in the training patterns might have been learned by the ***full AV*** net, but these correlations were not in the *test* patterns.



VO 51% *AO* 64% *AxV* 91% *full AV* 87%

Figure 3: Confusion matrixes for the video only (*VO*), acoustic only (*AO*), *AxV* and the *full AV* networks. Each vertical column is labelled by the spoken letter presented as input; each horizontal row represents the output by the network. The radius of each disk in the array is proportional to the output probability given an input letter. The accuracy of each network is shown.

Cluster dendrograms

Our results can be presented in cluster dendrograms, which reveal structural similarities among spoken letters better than does Figure 3. Because the confusion matrixes of Figure 3 do not obey the postulates of a metric space (necessary for clustering algorithms), we converted them. The "distance" between any two categories c_i and c_j, was determined from the confusion matrix C(i,j) as follows:

$$D(c_i,c_j) = C(i,i) + C(j,j) - C(i,j) - C(j,i). \qquad [3]$$

Equation 3 guaranteed, for instance, that $D(c_i,c_i) = 0$ for all categories. Although it does not logically guarantee that $D(c_i,c_j) \geq 0$ for all i and j, in fact we found that only one pair of letters (/bi/ and /di/ in the *AO* network) did not obey this inequality. The distance from cluster to a cluster was the minimum distance among all possible pairs of inter-cluster points.

Summerfield [22] presented acoustic and visual cluster dendrograms for humans based on psychophysical data. The important finding was that the two perceptual modes conveyed complementary information: phonemes that were similar acoustically were *dis*similar visually, and vice versa. There are several reasons for this, including the fact that the voiced-unvoiced distinction is acoustically salient, but invisible; some clearly visible distinctions (e.g. closure in /em/ - /en/ distinction) are acoustically subtle.

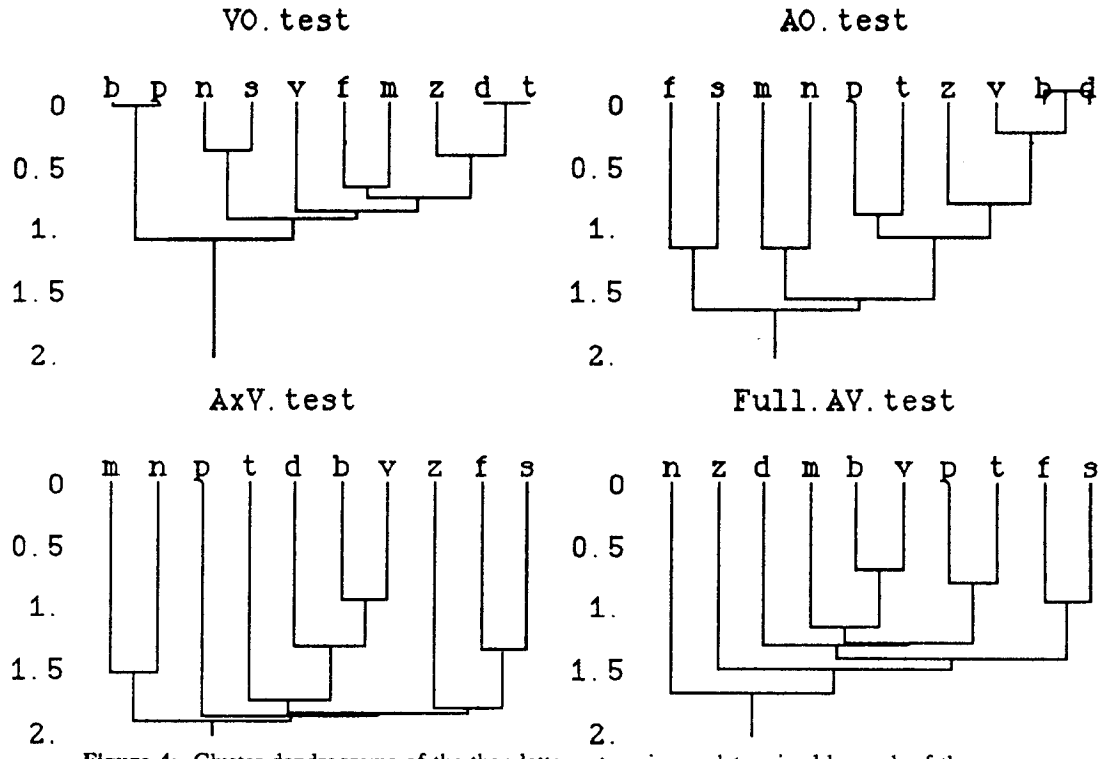


Figure 4: Cluster dendrograms of the then letter categories as determined by each of the networks. Two letter categories are joined at a (vertical) value denoting their distance. The overal scale ($0 \rightarrow 2$) is given by Eq. 3, and kept the same for all four dendrograms, in order to facilitate comparisons.

The first general trend to note in our dendrograms is that overall, letters are a bit more confused visually (***VO***) than acoustically (***AO***), as is to be expected. There is yet greater separability in the combined systems (***AxV*** and ***full AV***). The dendrograms also reveal specific confusions: /bi/ and /pi/ are visually nearly identical, as are /di/ and /ti/ (***VO***). Acoustically, /bi/ and /di/ are quite similar, and so forth. As with Summerfield's dendrograms, we show the complementary nature of acoustica and visual information. We show furthermore the improvement with bi-modal systems.

Robustness in noise

Our system is relatively insensitive to the addition of acoustic noise. Acoustic patterns were normalized to maximum value 1.0. We simulated the addition of pink acoustic noise by adding to each coefficient a random variable (standard deviation = 0.4). Under these conditions, the ***AO*** network had an accuracy of 43%, whereas the ***AxV*** network had an accuracy of 74% — a dramatic improvement. Clearly the visual information led to a significant improvement in recognition in noisy conditions. Note that this is consistent with no extra free parameters, such as a modified ratio of acoustic to visual evidence.

Hidden unit representations

The hidden unit representations — the patterns of connections from input to hidden layer — are somwhat complex and hard to interpret. Figure 5, however, shows that some "obvious" visual features are found in the *VO* network, such as mouth opening.

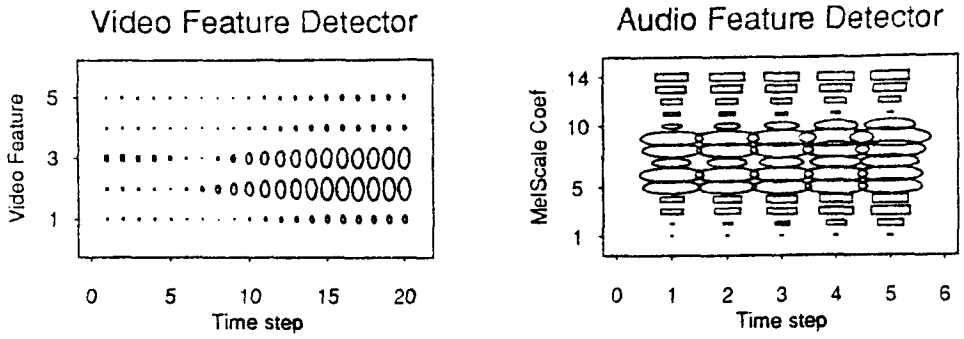


Figure 5. A single hidden unit representation learned by the *VO* network (left) and by the *AO* network (right). The size of the ovals represents the strength of the excitatory connections from the visual features throughout the 200 ms input window; rectangles represent inhibitory connections. The hidden unit in *VO* detects the opening of the mouth (cf. Features 2 and 3 in Figure 1). The acoustic hidden unit detects significant acoustic energy in the middle range of frequencies, roughly in the location of a first formant.

Thus our structural backpropagation networks learn relevant features directly from the data, even using very primitive feature detection.

Cross-modal confusion (McGurk effect)

As with human data, our system exhibits a McGurk effect [12]. In the McGurk effect, the listener-watcher is presented with conflicting cross-modal evidence for phonemes. For instance, if a /bi/ is presented visually (by means of a videotaped talker) while a /gi/ is presented acoustically (by means of a tape recording synchronized with the videotape), the listener-watcher will perceive a /di/, roughly "averaging" the two evidences for +front and +back. To explore the McGurk effect in our *AxV* network, we presented artificial stimuli consisting of all pairs of video letters and acoustic letters to the network. Figure 6 shows typical results.

Visual input

Acoustic input	b	d	f	m	n	p	s	t	v	z
b	d	d	b	p	t	p	t	d	d	d
d	v	d	v	v	v	v	v	v	d	d
f	f	f	f	f	s	f	f	f	f	f
m	m	m	m	m	n	m	m	m	m	m
n	n	m	n	n	n	n	n	n	n	n
p	t	t	t	p	t	p	t	t	p	t
s	v	v	s	s	s	s	s	s	s	v
t	t	t	t	p	t	p	t	t	p	t
v	d	d	b	b	t	p	b	d	d	d
z	z	z	z	z	z	z	z	z	z	z

Figure 6: Cross-modal perception matrix. The vertical columns represent visual patterns; each horizontal row represents a different acoustic pattern for a single speaker. The table entry is the final classification given by the *AxV* network.

Note first that along the diagonal — commensurate presentations — we get typical good recognition results. For conflicting simuli (off the diagonal), the results are more interesting. For instance the categorizations: /pi/ vis. & /di/ acous. → /vi/ as well as /di/ vis. & /pi/ acous. → /ti/ can be described as (rough) feature averaging. (The reader can discern many other examples in the Figure.) Some patterns are particularly salient and dominate categorization. For instance the voiced alveopalatal fricative in the acoustic /zi/ dominates all visual input, and leads to the category /zi/ in all cases. Likewise, the acoustic /em/ dominates all visual patterns except /en/.

Conclusions

Clearly the acoustic component of our networks is very primitive compared to commercial and research speech recognition systems (especially in its choice of input features). The fact that our visual network learns information that complements acoustic information (cf. Figure 3) and improves recognition rate, gives us confidence that a further, refined speech reading system can improve upon state-of-the-art acoustic recognizers, especially in noisy environments.

Finally, the fact that our system mimicks some human psychophysical results — acoustic and visual cluster dendrograms and the McGurk effect — suggests that our system posses structural similarities to human neurobiology, and hence can be used as a research tool for elucidating fundamental cross-modal neurobiological processes in vision and language.

Acknowledgements

We express our thanks to Dan Koff of Motion Analysis Corporation (Santa Rosa CA) for assistance in taking the video data, as well as to David Rumelhart (Stanford University Psychology) for suggestions relating to backpropagation training in our network. Finally, thanks to Mina Nishimura (Ricoh Japan) for background research on both human and machine speech reading.

References

[1] Berg, R. and D. G. Stork, ***The Physics of Sound*** Englewood Cliffs NJ: Prentice-Hall (1982).

[2] Berger, J. O., ***Statistical decision theory and Bayesian analysis*** (2nd ed.) 272-275, New York: Springer-Verlag (1985).

[3] Binnie, C., A. Montgomery and P. Jackson, "Auditory and visual contributions to the perception of consonants," *Journal of Speech and Hearing Research* **17**, 619-630 (1974).

[4] Cherry, E. C., "Some experiments on the recognition of speech, with one and two ears," *Journal of the Acoustical Society of America* **25**, 975-979 (1953).

[5] Cohen, M., S. Grossberg and D. G. Stork, "Speech perception and production by a self-organizing neural network," 615-633 (Chapter 19) in ***Pattern Recognition by Self-Organizing Neural Networks***, S. Grossberg and G. Carpenter (eds.) Cambridge MA: MIT Press (1991).

[6] De Filippo, C. L. and D. G. Sims (eds.), ***New Reflections on Speechreading*** special issue of *The Volta Review* **90**(5), (1988).

[7] Dodd, B. and R. Campbell (eds.), ***Hearing by Eye: The Psychology of Lip-reading*** Hillsdale, NJ: Lawrence Erlbaum Press (1987).

[8] Green, K.P. and J. L. Miller, "On the role of visual rate information in phonetic perception," *Perception and Psychophysics* **38**, 269-276 (1985).

[9] Fisher, C. G. "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research* **11**, 796-804 (1968).

[10] Keeler, J. and D. E. Rumelhart, "Self-organizing segmentation and recognition neural network," ***Advances in Neural Information Processing 4*** (NIPS-4), 1992, in press.

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

[12] McGurk, H. and J. MacDonald, "Hearing lips and seeing voices," *Nature* **264**,746-748 (1976).

[13] Miller, G. A. and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *Journal of the Acoustical Society of America* **27**, 338-352 (1955).