Review of Neural Networks for Speech Recognition

Richard P. Lippmann*
MIT Lincoln Laboratory, Lexington, MA 02173, USA

The performance of current speech recognition systems is far below that of humans. Neural nets offer the potential of providing massive parallelism, adaptation, and new algorithmic approaches to problems in speech recognition. Initial studies have demonstrated that multilayer networks with time delays can provide excellent discrimination between small sets of pre-segmented difficult-to-discriminate words, consonants, and vowels. Performance for these small vocabularies has often exceeded that of more conventional approaches. Physiological front ends have provided improved recognition accuracy in noise and a cochlea filter-bank that could be used in these front ends has been implemented using micro-power analog VLSI techniques. Techniques have been developed to scale networks up in size to handle larger vocabularies, to reduce training time, and to train nets with recurrent connections. Multilayer perceptron classifiers are being integrated into conventional continuous-speech recognizers. Neural net architectures have been developed to perform the computations required by vector quantizers, static pattern classifiers, and the Viterbi decoding algorithm. Further work is necessary for large-vocabulary continuous-speech problems, to develop training algorithms that progressively build internal word models, and to develop compact VLSI neural net hardware.

1 State of the Art for Speech Recognition

Speech is the most natural form of human communication. Compact implementations of accurate, real-time speech recognizers would find widespread use in many applications including automatic transcription, simplified man-machine communication, and aids for the hearing impaired and physically disabled. Unfortunately, current speech recognizers perform poorly on talker-independent continuous-speech recognition tasks that people perform without apparent difficulty. Although children learn to understand speech with little explicit supervision and adults take speech recognition ability for granted, it has proved to be a difficult task

*This work was sponsored by the Department of the Air Force. The views expressed are those of the author and do not reflect the official policy or position of the U.S. Government.

Neural Computation 1, 1–38 (1989)

to duplicate with machines. As noted by Klatt (1986), this is due to variability and overlap of information in the acoustic signal, to the need for high computation rates (a human-like system must match inputs to 50,000 words in real time), to the multiplicity of analyses that must be performed (phonetic, phonemic, syntactic, semantic, and pragmatic), and to the lack of any comprehensive theory of speech recognition.

The best existing speech recognizers perform well only in artificially constrained tasks. Performance is generally better when training data is provided for each talker, when words are spoken in isolation, when the vocabulary size is small, and when restrictive language models are used to constrain allowable word sequences. For example, talker-dependent isolated-word recognizers can be trained to recognize 105 words with 99% accuracy (Paul 1987). Large-vocabulary talker-dependent word recognition accuracy with sentence context can be as high as 95% for 20,000 words from sentences in office memos spoken with pauses between words (Averbuch et al. 1987).

Accuracy for a difficult 997-word talker-independent continuous-speech task using a strong language model (an average of only 20 different words possible after any other word) can be as high as 96% (Lee and Hon 1988). This word accuracy score translates to an unacceptable sentence accuracy of roughly 50%. In addition, the word accuracy of this high-performance recognizer when tested with no grammar model is typically below 70% correct. Results such as these illustrate the poor low-level acoustic-phonetic matching provided by current recognizers. These recognizers depend heavily on constraining grammars to achieve good performance. Humans do not suffer from this problem. We can recognize clearly spoken but contextually inappropriate words in anomalous sentences such as "John drank the guitar" almost perfectly (Marslen-Wilson 1987).

The current best performing speech recognition algorithms use Hidden Markov Model (HMM) techniques. Good introductions to these techniques and to digital signal processing of speech are available in (Lee and Hon 1988; Parsons 1986; Rabiner and Juang 1986; Rabiner and Schafer 1978). The HMM approach provides a framework which includes an efficient decoding algorithm for use in recognition (the Viterbi algorithm) and an automatic supervised training algorithm (the forward-backward algorithm). New neural-net approaches to speech recognition must have the potential to overcome the limitations of current HMM systems. These limitations include poor low-level and poor high-level modeling. Poor low-level acoustic-phonetic modeling leads to confusions between acoustically similar words while poor high-level speech understanding or semantic modeling restricts applications to simple situations where finite state or probabilistic grammars are acceptable. In addition, the first-order Markov assumption makes it difficult to model coarticulation directly and HMM training algorithms can not currently learn the topological structure of word and sub-word models. Finally, HMM theory does not

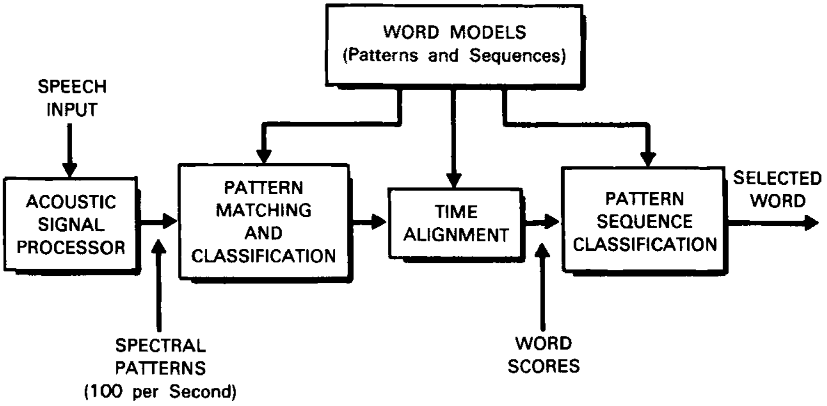
WORD MODELS
(Patterns and Sequences)
SPEECH INPUT
ACOUSTIC SIGNAL PROCESSOR
PATTERN MATCHING AND CLASSIFICATION
TIME ALIGNMENT
PATTERN SEQUENCE CLASSIFICATION
SELECTED WORD
SPECTRAL PATTERNS (100 per Second)
WORD SCORES

Figure 1: Block diagram of an isolated word recognizer.

specify the structure of implementation hardware. It is likely that high computation and memory requirements of current algorithms will require new approaches to parallel hardware design to produce compact, large-vocabulary, continuous-speech recognizers.

2 The Potential of Neural Nets

Neural nets for speech recognition have been explored as part of the recent resurgence of interest in this area. Research has focused on evaluating new neural net pattern classification and training algorithms using real speech data and on determining whether parallel neural net architectures can be designed which perform the computations required by important speech recognition algorithms. Most work has focused on isolated-word recognition.

A block diagram of a simple isolated word recognizer is shown in figure 1. Speech is input to this recognizer and a word classification decision is output on the right. Three major operations are required. First, a preprocessor must extract important information from the speech waveform. In most recognizers, an input pattern containing spectral information from a frame of speech is extracted every 10 msec using Fast Fourier Transform (FFT) or Linear Predictive Coding (LPC) (Parsons 1986; Rabiner and Schafer 1978) techniques. Second, input patterns from the preprocessor must be compared to stored exemplar patterns in word models to compute local frame-to-frame distances. Local distances are used in a third step to time align input pattern sequences to stored exemplar pattern sequences that form word models and arrive at whole-word matching scores. Time alignment compensates for variations in talking rate and pronunciation. Once these operations have been performed, the selected word to output is that word with the highest whole-word matching score.

This paper reviews research on complete neural net recognizers and on neural nets that perform the above three operations. Auditory preprocessors that attempt to mimic cochlea and auditory nerve processing are first reviewed. Neural net structures that can compute local distance scores are then described. Classification results obtained using static speech patterns as inputs are then followed by results obtained with dynamic nets that allow continuous-time inputs. Techniques to integrate neural net and conventional approaches are then described followed by a brief review of psychological and physiological models of temporal pattern sequence recognition. The paper ends with a summary and suggestions for future research. Emphasis throughout is placed on studies that used large public-domain speech data bases or that first presented new approaches.

3 Auditory Preprocessors

A preprocessor extracts important parameters from the speech waveform to compress the amount of data that must be processed at higher levels and provide some invariance to changes in noise, talkers, and the acoustic environment. Most conventional preprocessors are only loosely modeled on the cochlea and perform simple types of filtering and data compression motivated by Fourier analysis and information theory. Recent physiological studies of cochlea and auditory nerve responses to complex stimuli have led to more complex physiological preprocessors designed to closely mimic many aspects of auditory nerve response characteristics. Five of these preprocessors and the VLSI cochlea filter listed in table 1 are reviewed in this section. Good reviews of many of these preprocessors and of response properties of the cochlea and auditory nerve can be found in (Greenberg 1988a; 1988b).

The five preprocessors in table 1 rely on periodicity or synchrony information in filter-bank outputs. Synchrony information is related to the short-term phase of a speech signal and can be obtained from the arrival times of nerve spikes on the auditory nerve. It could increase recognition performance by supplementing the spectral magnitude information used in current recognizers. Synchrony information is typically obtained by filtering the speech input using sharp bandpass filters with characteristics similar to those of the mechanical filters in the cochlea. The resulting filtered waveforms are then processed using various types of time domain analyses that could be performed using analog neural net circuitry.

Study	Processing	Comments
Deng and Geisler (1987)	Cross-Channel Correlation of Neural Outputs	Physiologically Plausible (Untested for Speech Recognition)
Ghitza (1988)	Create Histogram of Time Intervals Between Threshold Crossings of Filter Outputs	Improved Speech Recognition In Noise
Hunt and Lefèbvre (1988)	Periodicity and Onset Detection	Improved Speech Recognition In Noise and with Spectral Tilt
Lyon and Mead (1988)	Tapped Transmission Line Filter with 49 Outputs	Implemented Using Micropower VLSI Techniques
Seneff (1988)	Provides Periodicity and Spectral Magnitude Outputs	Synchrony Spectrograms Provide Enhanced Spectral Resolution (Untested for Speech Recognition)
Shamma (1988)	Lateral Inhibition Across Cochlea Filter Outputs	Physiologically Plausible (Untested for Speech Recognition)

Table 1: Recent Physiological Preprocessors.

Spectrograms created using physiological preprocessors for steady state vowels and other speech sounds illustrate an improvement in ability to visually identify vowel formants (resonant frequencies of the vocal tract) in noise (Deng and Geisler 1987; Ghitza 1988; Seneff 1988; Shamma 1988). Comparisons to more conventional front ends using existing speech recognizers have been performed by Beet (Beet et al. 1988), Ghitza (1988), and by Hunt and Lefèbvre (1988). These comparisons demonstrated significant performance improvements in noise (Ghitza 1988; Hunt and Lefèbvre 1988) and with filtering that tilts the input spectrum up at high frequencies (Hunt and Lefèbvre 1988). Extensive comparisons have not, however, been made between physiological preprocessors and conventional preprocessors when the conventional preprocessors incorporate current noise and stress compensation techniques. Positive results from such comparisons and more detailed theoretical analyses would do much to foster the acceptance of these new and computationally intensive front ends.

Lyon and Mead (1988) describe a filter bank that could be used in a physiological preprocessor. This filter bank was carefully modeled after the cochlea, provides 49 analog outputs, and has been implemented using micropower analog VLSI CMOS processing. Extra circuitry would be required to provide synchrony or spectral magnitude information for a speech recognizer. This recent work demonstrates how preprocessors can be miniaturized using analog VLSI techniques. The success of this approach is beginning to demonstrate that ease of implementation using VLSI techniques may be more important when comparing alternative neural net approaches than computational requirements on serial Von Neuman machines.

4 Computing Local Distance Scores

Conventional speech recognizers compute local frame-to-frame distances by comparing each new input pattern (vector of parameters) provided by a preprocessor to stored reference patterns. Neural net architectures can compute local frame-to-frame distances using fine-grain parallelism for both continuous-observation and discrete-observation recognizers. New neural net algorithms can also perform vector quantization and reduce the dimensionality of input patterns.

Local distances for continuous-observation recognizers are functions related to log likelihoods of probability distributions. Simple log likelihood functions such as those required for independent Gaussian or binomial distributions can be calculated directly without training using single-layer nets with threshold-logic nonlinearities (Lippmann 1987; Lippmann et al. 1987). More complex likelihood functions can be computed using multilayer perceptrons (Cybenko 1988; Lapedes and Farber 1988; Lippmann et al. 1987), hierarchical nets that compute kernel functions (Albus 1981; Broomhead and Lowe 1988; Hanson and Burr 1987; Huang and Lippmann 1988; Moody 1988; Moody and Darken 1988), or high-order nets (Lee et al. 1986; Rumelhart et al. 1986a). Training to produce these complex functions is typically longest with multilayer perceptrons. These nets, however, often provide architectures with fewer nodes, simpler nodal processing elements, and fewer weights. They also may develop internal hidden abstractions in hidden layers that can be related to meaningful acoustic-phonetic speech characteristics such as formant transitions and that also could be applied to many different speech recognition tasks.

Discrete-observation recognizers first perform vector quantization and label each input with one particular symbol. Symbols are used to calculate local distances via look-up tables that contain symbol probabilities for each reference pattern. The look-up table calculation can be performed by simple single-layer perceptrons. The perceptron for any reference pattern must have as many inputs as there are symbols. Weights must equal symbol probabilities and all inputs must be equal to zero except for that corresponding to the current input symbol. Alternatively, a multilayer perceptron could be used to store probabilities for symbols that have been seen and interpolate between these probabilities for unseen symbols. The vector quantization operation can be performed using an architecture similar to that used by Kohonen's feature-map net (Kohonen 1984). Inputs to the feature-map net feed an array of codebook nodes containing one node for each symbol. Components of the Euclidean distance between the input and the reference pattern represented by weights to each node are computed in each node. The codebook node with the smallest Euclidean distance to the input is selected using lateral inhibition or other maximum-picking techniques (Lippmann et al. 1987). This process guarantees that only one node with the minimum Euclidean distance to the input has a unity output as required. Weights used in this architecture can be calculated using the feature-map algorithm or any other standard vector quantization algorithm based on Euclidean distances such as k-means clustering (Duda and Hart 1973).

Kohonen's feature-map vector quantizer is an alternative sequentially-trained neural net algorithm. It has been tested successfully in an experimental speech recognizer (Kohonen 1988; Kohonen et al. 1984) but not evaluated with a large public speech data base. A version with a small number of nodes but including training logic has been implemented in VLSI (Mann et al. 1988). Experiments with a discrete-observation HMM recognizer (Mann et al. 1988) and with a template-based recognizer (Naylor and Li 1988) demonstrated that this algorithm provides performance similar to that provided by conventional clustering procedures such as k-means clustering (Duda and Hart 1973). The feature-map algorithm incrementally trains weights to a two-dimensional grid of nodes such that after training, nodes that are physically close in the grid correspond to input patterns that are close in Euclidean distance. One advantage of this topological organization is that averaging outputs of nodes that are physically close using nodes at higher levels corresponds to a probability smoothing technique often used in speech recognizers called Parzen smoothing (Duda and Hart 1973). This averaging can be performed by nodes with limited fan-in and short connections.

The auto-associative multilayer perceptron (Elman and Zipser 1987; Hinton 1987) is a neural net algorithm that reduces the dimensionality of continuous-valued inputs. It is a multilayer perceptron with the same number of input and output nodes and one or more layers of hidden nodes. This net is trained to reproduce the input at the output nodes through a small layer of hidden nodes. Outputs of hidden nodes after training can be used as reduced dimensional inputs for speech processing as described in (Elman and Zipser 1987; Fallside et al. 1988). Recent theoretical analyses have demonstrated that auto-associative networks are closedly related to a standard statistical technique called principal components analysis (Baldi and Hornik 1989; Bourlard and Kamp 1988). Auto-associative nets are thus not a new analytical tool but instead a technique to perform the processing required by principal components analysis.

5 Static Classification of Speech Segments

Many neural net classifiers have been applied to the problem of classifying static input patterns formed from a spectral analysis of pre-segmented words, phonemes, and vowels. Table 2 summarizes results of some representative studies. Introductions to many of the classifiers listed in this table and to neural net training algorithms are available in (Cowan and Sharp 1988; Hinton 1987; Lippmann et al. 1987). Unless otherwise noted, error rates in this and other tables refer to talker-dependent training and testing, multilayer perceptrons were trained using back-propagation (Rumelhart et al. 1986a), and systems were trained and tested on different data sets. The number of tokens in this and other tables refers to the total number of speech samples available for both training and testing and the label "multi-talker" refers to results obtained by testing and training using data from the same group of talkers. The label "talker-independent" refers to results obtained by training using one group of talkers and testing using a separate group with no common members.

Input patterns for studies in table 2 were applied at once as one whole static spectrographic (frequency versus time) pattern. Neural nets were static and didn't include internal delays or recurrent connections that could take advantage of the temporal nature of the input for real-time processing. This approach might be difficult to incorporate in real-time speech recognizers because it would require long delays to perform segmentation and form the input patterns in an input storage buffer. It would also require accurate pre-segmentation of both testing and training data for good performance. This pre-segmentation was performed by hand in many studies.

Multilayer perceptrons and hierarchical nets such as the feature-map classifier and Kohonen's learning vector quantizer (LVQ) have been used to classify static patterns. Excellent talker-dependent recognition accuracy near that of experimental HMM and commercial recognizers has been provided by multilayer perceptrons using small sets of words and digits. Hierarchical nets have provided performance similar to that of

Study	Network	Speech Materials	Error Rate
Elman and Zipser (1987)	Multilayer Perceptron (MLP) 16 × 20 Inputs	1 Talker, CV's /b,d,g/ /i,a,u/ 505 Tokens	Cons. - 5% Vowels - 0.5%
Huang and Lippmann, (1988)	MLP, Feature Map Classifier (FMC) 2 Inputs	67 Talkers 10 Vowels 671 Tokens	Gaussian, FMC, MLP ≈ 20% FMC Trains Fastest
Kammerer and Kupper (1988)	MLP 16 × 16 Inputs	11 Talkers 20 Words 5720 Tokens	Talker Dep. - 0.4% Talker Indep. - 2.7%
Kohonen (1988)	Learning Vector Quantizer (LVQ) 15 Inputs	Labeled Finish Speech 3010 Tokens	Gaussian - 12.9% kNN - 12.0% LVQ - 10.9%
Lippmann and Gold (1987)	MLP 11 × 2 Inputs	16 Talkers 7 Digits 2,912 Tokens	Gaussian - 8.7% kNN - 6% MLP - 7.6%
Peeling and Moore (1987)	MLP 19 × 60 Inputs	40 Talkers 10 Digits 16,000 Tokens	Talker Dep. - 0.3% Multi Talker - 1.9%

Table 2: Recognition of Speech Patterns Using Static Neural Nets.

multilayer perceptrons but with greatly reduced training times and typically more connection weights and nodes.

5.1 Multilayer Perceptrons. Multilayer perceptron classifiers have been applied to speech problems more often than any other neural net classifier. A simple example from Huang and Lippmann (1988) presented in figure 2 illustrates how these nets can form complex decision regions with speech data. Input data obtained by Peterson and Barney (1952) consisted of the first two formants from vowels spoken by men, women, and children. Decision regions shown in the right side of figure 2 were formed by the two-layer perceptron with 50 hidden nodes trained using back-propagation shown on the left. Training required more than 50,000 trials. Decision region boundaries are near those that are typically drawn by hand to separate vowel regions and the performance of this net is near that provided by commonly used conventional k-nearest neighbor (kNN) and Gaussian classifiers (Duda and Hart 1973).

A more complex experiment was performed by Elman and Zipser (1987) using spectrographic-like inputs. Input patterns formed from 16 filter-bank outputs sampled 20 times over a time window of 64 msec were fed to nets with one hidden layer and 2 to 6 hidden nodes. The analysis time window was centered by hand on the consonant voicing onset. Networks were trained to recognize consonants or vowels in consonant-vowel (CV) syllables composed of the consonants /b,d,g/ and the vowels /i,a,u/. Error rates were roughly 5% for consonant recognition and 0.5% for vowel recognition. An analysis indicated that hidden nodes often become feature detectors and differentiate between important subsets of sound types such as consonants versus vowels. This study demonstrated the importance of choosing a good data representation for speech and of normalizing speech inputs. It also raised the important question of training time because many experiments on this small data base required more than 100,000 training trials.

Lippmann and Gold (1987) performed another early study to compare multilayer perceptrons and conventional classifiers on a digit classification task. This study was motivated by single-talker results obtained

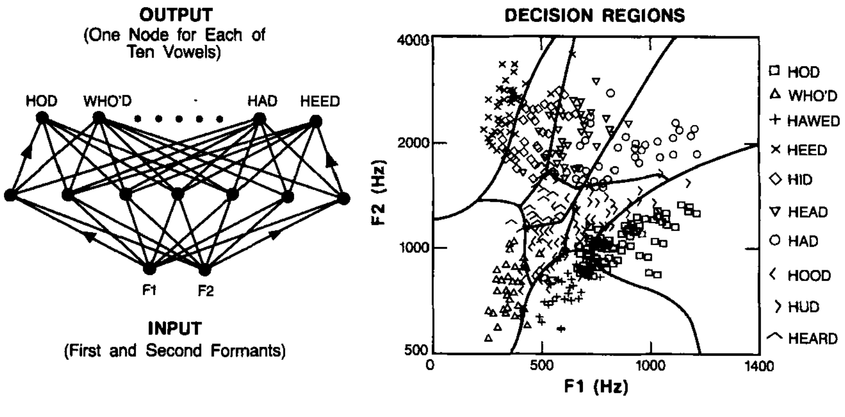

Figure 2: Decision regions formed by a 2-layer perceptron using back-propagation training and vowel formant data.

by Burr (1988a). Inputs were 22 cepstral parameters from two speech frames located automatically by finding the maximum-energy frame for each digit. One- to three-layer nets with from 16 to 256 nodes in each hidden layer were evaluated using digits from the Texas Instruments (TI) 20-Word Speech Data Base (Doddington and Schalk 1981). Multilayer perceptron classifiers outperformed a Gaussian but not a kNN classifier. Hidden layers were required for good performance. A single-layer perceptron provided poor performance, much longer training times, and sometimes never converged during training. Most rapid training (less than 1000 trials) was provided by all three-layer perceptrons. These results demonstrate that the simple hyperplane decision regions provided by single-layer perceptrons are sometimes not sufficient and that rapid training and good performance can be obtained by tailoring the size of a net for a specific problem. The digit data used in these experiments was also used to test a multilayer perceptron chip implemented in VLSI (Raffel et al. 1987). This chip performed as well as computer simulations when down-loaded with weights from those simulations.

Kammerer and Kupper obtained surprisingly good recognition results for words from the TI 20-word data base (Kammerer and Kupper 1988). A single-layer perceptron with spectrogram-like input patterns performed slightly better than a DTW template-based recognizer. Words were first time normalized to provide 16 input frames with 16 2-bit spectral coefficients per frame. Expanding the training corpus by temporally distorting training tokens reduced the error slightly and best performance was provided by single and not multilayer perceptrons. Talker-dependent error rates were 0.4% (14/3520) for the single-layer perceptron and 0.7% (25/3520) for the DTW recognizer. These error rates are better than all but one of the commercial recognizers evaluated in (Doddington and Schalk 1981) and demonstrate good performance for a single-layer perceptron without hidden nodes. Talker-independent performance was evaluated by leaving out the training data for each talker, one at a time, and testing using that talker's test data. Average talker-independent error rates were 2.7% (155/5720) for the single-layer perceptron and 2.5% (145/5720) for the DTW recognizer. Training time was 6 to 25 minutes per talker on an array processor for the talker-dependent studies and 5 to 9 hours for the talker-independent studies.

Peeling and Moore (1987) obtained extremely good recognition results for digit classification. A multilayer perceptron with one hidden layer and 50 hidden nodes provided best performance. Its talker-dependent performance was low and near that provided by an advanced HMM recognizer. Spectrogram-like input patterns were generated using a 19-channel filter-bank analyzer with 20 msec frames. Nets could accommodate 60 input frames (1.2 seconds) which was enough for the longest duration word. Shorter words were padded with zeros and positioned randomly in the 60 frame input buffer. Nets were trained using different numbers of layers and hidden units and speech data from the *RSRE*

40-speaker digit data base. Multi-talker experiments explored performance when recognizers were tested and trained using data from all talkers. Error rates were near zero for talker-dependent experiments 0.25% (5/2000) and low for multi-talker experiments 1.9% (78/4000). Error rates on an advanced HMM recognizer under the same conditions were 0.2% (4/2000) and 0.6% (25/4000) respectively. The computation required for recognition using multilayer perceptrons was typically more than five times less than that required for the HMM recognizer.

The good small-vocabulary word recognition results obtained by both Kammerer and Kupper (1988) and Peeling and Moore (1987) suggest that back-propagation can develop internal feature detectors to extract important invariant acoustic events. These results must be compared to those of other experiments which attempted to classify digits without time alignment. Burton, Shore, and Buck (Burton et al. 1985; Shore and Burton 1983) demonstrated that talker-dependent error rates using the TI 20-Word Data Base can be as low as 0.3% (8/2560) for digits and 0.8% (40/5120) for all words using simple vector-quantization recognizers that do not perform time alignment. These results suggest that digit recognition is a relatively simple task where dynamic time alignment is not necessary and talker-dependent accuracy remains high even when temporal information is discarded. The good performance of multilayer perceptrons is thus not surprising. These studies and the multilayer perceptron studies do, however, suggest designs for implementing computationally-efficient real-time digit and small-vocabulary recognizers using analog neural-net VLSI processing.

5.2 Hierarchical Neural Nets that Compute Kernel Functions. Hierarchical neural net classifiers which use hidden nodes that compute kernel functions have also been used to classify speech patterns. These nets have the advantage of rapid training and the ability to use combined supervised/unsupervised training data.

Huang and Lippmann (1988) described a net called a feature-map classifier and evaluated the performance of this net on the vowel data plotted in figure 2 and on difficult artificial problems. A block diagram of the feature-map classifier is shown in figure. 3. Intermediate codebook nodes in this net compute kernel functions related to the Euclidean distance between the input and cluster centers represented by these nodes. The lower feature map net is first trained without supervision to form a vector quantizer and the upper perceptron-like layer is then trained with supervision using a modified version of the LMS algorithm. This classifier was compared to the multilayer perceptron shown in figure 2 and to a kNN classifier. All classifiers provided an error rate of roughly 20%. The 2-layer perceptron, however, required more than 50,000 supervised training trials for convergence. The feature map classifier reduced the amount of supervised training required by three orders of magni-

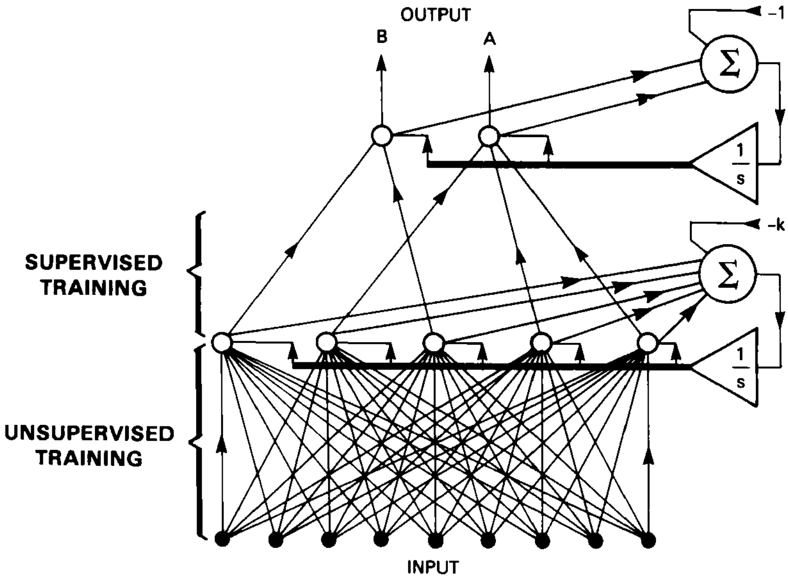

Figure 3: Block diagram of the hierarchical feature-map classifier.

tude to fewer than 50 trials. Similar results were obtained with artificial problems.

Kohonen and co-workers (Kohonen et al. 1988) compared a neural-net classifier called a learning vector quantizer (LVQ) to Bayesian and kNN classifiers. The structure of the learning vector quantizer is similar to that of the feature-map classifier shown in figure 3. Training differs from that used with the feature-map classifier in that a third stage of supervised training is added which adjusts weights to intermediate codebook nodes when a classification error occurs. Adjustments alter decision region boundaries slightly but maintain the same number of codebook nodes.

Bayesian, kNN and LVQ classifiers were used to classify 15-channel speech spectra manually extracted from stationary regions of Finnish speech waveforms. All classifiers were tested and trained with separate sets of 1550 single-frame patterns that were divided into 18 phoneme classes (Kohonen et al. 1988). A version of the LVQ classifier with 117 codebook nodes provided the lowest error rate of 10.9% averaging over results where training and testing data sets are interchanged. The Bayesian classifier and kNN classifiers had slightly higher error rates of 12.9% and 12.0% respectively. Training time for the LVQ classifier was roughly 10 minutes on an IBM PC/AT. These results and those of

Huang and Lippmann (1988) demonstrate that neural nets that use kernel functions can provide excellent performance on speech tasks using practical amounts of training time. Other experiments on artificial problems described in (Kohonen et al. 1988) illustrate trade-offs in training time. Boltzmann machines provided near optimal performance on these problems followed by the LVQ classifier and multilayer perceptrons. Training times were 5 hours on an array processor for the Boltzmann machine, 1 hour on a Masscomp MC 5600 for the multilayer perceptron, and roughly 20 minutes on the Masscomp for the LVQ classifier.

Two recent studies (Niranjan and Fallside 1988; Bridle 1988) have begun to explore a hierarchical net where nodes in a hidden layer compute kernel functions called radial basis functions (Broomhead and Lowe 1988). These nets are similar to previous classifiers that use the method of potential functions (Duda and Hart 1973). They have an advantage over multilayer perceptrons in that once the locations of the kernel functions are established, weights to the output nodes are determined uniquely by solving a least squares problem using matrix-based approaches. Initial results with small amounts of speech data consisting of vowels (Niranjan and Fallside 1988) and words (Bridle 1988) have been encouraging. Further work must explore techniques to assign the locations of kernel functions and adjust scale factors that determine the range of influence of each kernel function.

6 Dynamic Classification of Speech Segments

New dynamic neural net classifiers that incorporate short delays, temporal integration, or recurrent connections have been developed specifically for speech recognition. Spectral inputs for these classifiers are applied to input nodes sequentially, one frame at a time. These classifiers could thus be integrated into real time speech recognizers more easily than static nets because accurate pre-segmentation is typically not required for good performance and only short delays are used.

Both multilayer nets with delays and nets with recurrent connections have been used to classify acoustically similar words, consonants, and vowels. Excellent performance has been obtained using time delay nets in many studies including those by Lang and Hinton (1988) and by Waibel et al. (1987; 1988). Performance for small vocabularies often slightly exceeded that provided by high-performance experimental HMM recognizers. Techniques have also been developed to scale nets up for larger vocabularies and to speed up training times both for feed-forward and recurrent nets. Rapid training has been demonstrated using a hierarchical learning vector quantizer with delays and good performance but extremely long training times has been provided by Boltzmann machines.

6.1 Time-Delay Multilayer Perceptrons. Some of the most promising neural-net recognition results have been obtained using multilayer perceptrons with delays and some form of temporal integration in output nodes (Lang and Hinton 1988; Waibel et al. 1987; Waibel et al. 1988). Table 3 summarizes results of six representative studies.

Early results on consonant and vowel recognition were obtained by Waibel and co-workers (Waibel et al. 1987) using the multilayer percep-

Study	Network	Speech Materials	Error Rate
Lang and Hinton (1988)	Time Delay MLP 16 Inputs	100 Talkers "B,D,E,V" 768 Tokens	Multi Talker - 7.8%
Unnikrishnan, Hopfield, and Tank (1988)	Time Concen- tration Net 32 Inputs	1 Talker Digits 432 Tokens	0.7%
Waibel et al. (1987)	Time Delay MLP 16 Inputs	3 Japanese Talkers, /b,d,g/, Many Contexts $> 4,000$ Tokens	/b,d,g/ - 1.5%
Waibel, Sawai, and Shikano (1988)	Time Delay MLP 16 Inputs	1 Japanese Talker, 18 Cons., 5 Vowels $> 10,000$ Tokens	/b,d,g,p,t,k/ - 1.4% 18 Cons. - 4.1% 5 Vowels - 1.4%
Watrous (1988)	Temporal Flow Structured MLP 16 Inputs	1 Talker Phonemes, Words $> 2,000$ Tokens	/b,d,g/ - 0.8% rapid/rabid - 0.8% /i,a,u/ - 0.0%
McDermott and Katagiri (1988)	Time Delay LVQ 16 Inputs	3 Japanese Talkers, /b,d,g/ $> 4,000$ Tokens	/b,d,g/ - 1.7%

Table 3: Recognition of Speech Using Time-Delay Neural Nets.

tron with time delays shown in figure 4. The boxes labeled τ in this figure represent fixed delays. Spectral coefficients from 10 msec speech frames (16 per frame) are input on the lower left. The three boxes on the bottom thus represent an input buffer containing a context of three frames. Outputs of the nodes in these boxes (16 × 3 spectral coefficients) feed 8 hidden nodes in the first layer. Outputs from these nodes are buffered across the five boxes in the first hidden layer to form a context of five frames. Outputs from these boxes (8 × 5 node outputs) feed three hidden nodes in the second hidden layer. Outputs from these three nodes are integrated over time in a final output node.

In initial experiments (Waibel et al. 1987), the time-delay net from figure 4 was trained using back-propagation to recognize the voiced stops /b,d,g/. Separate testing and training sets of 2000 voiced stops spoken by three talkers were excised manually from a corpus of 5260 Japanese words. Excised portions sampled the consonants in varying phonetic contexts and contained 15 frames (150 msec) centered by hand around the vowel onset. The neural net classifier provided an error rate of 1.5% compared to an error rate 6.5% provided by a simple discrete-observation HMM recognizer. Training the time-delay net took several days on a four-processor Alliant computer. More recent work (Waibel et al. 1988) has led to techniques that merge smaller nets designed to recognize small sets of

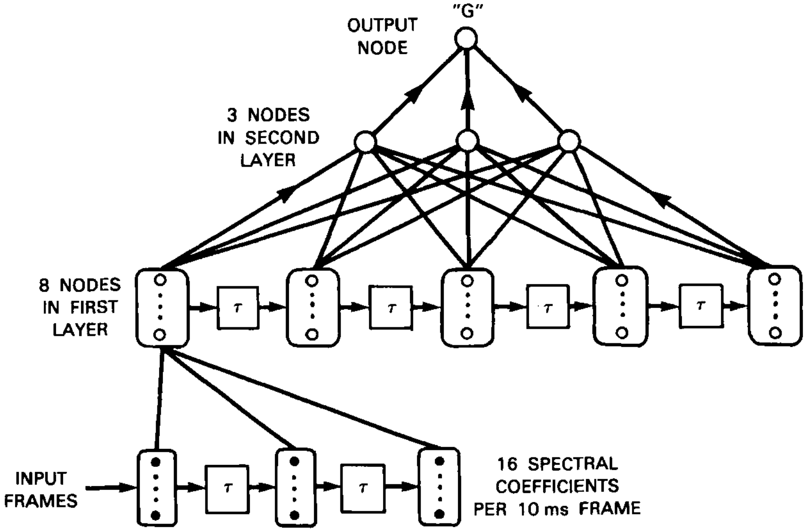

Figure 4: A time-delay multilayer perceptron.

consonants and vowels into large nets which can recognize all consonants at once. These techniques greatly reduce training time, improve performance and are a practical approach to the scaling problem. Experiments resulted in low error rates of 1.4% for the consonants /b,d,g,p,t,k/ and 1.4% for the vowels /i,a,u,e,o/. The largest net designed from smaller subnets provided a talker-dependent error rate for one talker of 4.1% for 18 consonants. An advanced discrete-observation HMM recognizer provided an error rate of 7.3% on this task. These two studies demonstrate that good performance can be provided by time-delay nets when the network structure is tailored to a specific problem. They also demonstrate how small nets can be scaled up to solve large classification problems without scaling up training times substantially.

Lang and Hinton (1988) describe an extensive series of experiments that led to a similar high-performance time-delay net. This net was designed to classify four acoustically similar isolated words "B", "D", "E", and "V" that are the most confusable subset from the spoken alphabet. A multi-talker recognizer for 100 male talkers was first trained and tested using pre-segmented 144 msec speech samples taken from around the vowel onset in these words. A technique called multi-resolution training was developed to shorten training time. This involved training nets with smaller numbers of hidden nodes, splitting weight values to hidden nodes to create larger desired nets, and then re-training the larger nets. A multiresolution trained net provided an error rate of 8.6%. This result, however, required careful pre-segmentation of each word. Presegmentation was not required by another net which allowed continuous speech input and classified the input as that word corresponding to the output node whose output value reached the highest level. Training used simple automatic energy-based segmentation techniques to extract 216 msecs of speech from around the vowel onset in each word. This resulted in an error rate of 9.5%. Outputs were then trained to be high and correct for the 216 msec speech segments as before, but also low for counter-example inputs selected randomly from the left-over background noise and vowel segments. Inclusion of counter-examples reduced the error rate to 7.8%. This performance compares favorably with the 11% error rate estimated for an enhanced HMM recognizer on this data base and based on performance with the complete E-set (Bahl et al. 1988; Lang and Hinton 1988).

Watrous (1988) also explored multilayer perceptron classifiers with time delays that extended earlier exploratory work on nets with recurrent connections (Watrous and Shastri 1987). These multilayer nets differed from those described above in that recurrent connections were provided on output nodes, target outputs were Gaussian-shaped pulses, and delays and the network structure were carefully adjusted by hand to extract important speech features for each classification task. Networks were tested using hand-segmented speech and isolated words from one talker. Good discrimination was obtained for many different recognition tasks.

For example, the error rate was 0.8% for the consonants /b,d,g/, 0.8% for the word pair "rapid/rabid," and 0.0% for the vowels /i,a,u/. Watrous has also explored the use of gradient methods of nonlinear optimization to decrease training time (Watrous 1986).

Rossen et al. (1988) recently described another time delay classifier. It uses more complex input data representations than the time-delay nets described above and a brain-state-in-a-box neural net classifier to integrate information over time from lower-level networks. Good classification performance was obtained for six stop consonants and three vowels. Notable features of this work are training to reject noise inputs as in (Lang and Hinton 1988) and the use of modular techniques to build large nets from smaller trained modules as in (Waibel et al. 1988). Other recent work demonstrating good phoneme and syllable classification using structured multilayer perceptron nets with delays is described in (Harrison and Fallside 1988; Homma et al. 1988; Irino and Kawahara 1988; Kamm et al. 1988; Leung and Zue 1988).

Unnikrishnan, Hopfield, and Tank (1988) obtained low error rates on digit classification using a time-concentration neural net that does not use only simple delays. This net, described in (Tank and Hopfield 1987), uses variable length delay lines designed to disperse impulsive inputs such that longer delays result in more dispersion. Impulsive inputs to these delay lines are formed by enhancing spectral peaks in the outputs of 32 bandpass filters. Outputs of delay lines are multiplied by weights and summed to form separate matched filters for each word. These matched filters concentrate energy in time and produce a large output pulse at the end of the correct word. Limited evaluations reported in (Unnikrishnan et al. 1988) for digit strings from one talker demonstrated good performance using a modified form of back-propagation training. A prototype version of this recognizer using discrete analog electronic devices was also constructed (Tank and Hopfield 1987). Tests performed by Gold with a large speech data base and a hierarchical version of the time concentration net that included both allophone and word models yielded performance that was no better than that of an existing HMM recognizer (Gold 1988).

6.2 Hierarchical Nets that Compute Kernel Functions. McDermott and Katagiri (1988) used Kohonen's LVQ classifier on the same /b,d,g/ speech data base used by Waibel et al. (1987). They were able to obtain an error rate of 1.7% which is not statistically different from the 1.5% error rate obtained by Waibel et al. using the time-delay net shown in figure 4 (Waibel et al. 1987). Inputs for the LVQ classifier consisted of a 7-frame window of 16 filterbank outputs. The nearest of 150 codebook nodes were determined as the 15-frame speech samples were passed through this 7-frame window. The normalized distances between nearest nodes and 112-element input patterns were integrated over time and used to classify speech inputs. The error rate without the final stage of LVQ training was high (7.3%). It dropped to 1.7% after LVQ training was complete. This result demonstrates that nets with kernel functions and delays can perform as well as multilayer perceptrons with delays. These nets train faster but require more computation and memory during use. In this application, for example, the LVQ classifier required 17,000 weights which was more than 30 times as many required for the time-delay net used in (Waibel et al. 1987). If memory is not an important limitation, rapid search techniques such as hashing and k-d trees described in (Omohundro 1987) can be applied to the LVQ classifier to greatly reduce the time required to find nearest-neighbors. This would make the differences in computation time between these alternative approaches small on existing serial Von Neuman computers.

6.3 Nets with Recurrent Connections. Nets with recurrent connections have not been used as extensively for speech recognition problems as feed-forward nets because they are more difficult to train, analyze, and design. Table 4 summarizes results of three representative studies. Initial work explored the use of recurrent Boltzmann machines. These nets typically provided good performance on small problems but required extremely long training times. More recent studies have focused on modified back-propagation training algorithms described in (Almeida 1987; Jordan 1986; Pineda 1987; Rohwer and Forrest 1987; Rumelhart et al. 1986a; Watrous 1988) that can be used with recurrent nets and time varying inputs.

Study	Network	Speech Materials	Error Rate
Anderson, Merrill, and Port (1988)	Recurrent Net 36 Inputs	20 Talkers, CV's /b,d,g,p,t,k/, /a/ 561 Tokens	Talker Indep. - 13.1%
Prager, Harrison, and Fallside (1986)	Boltzmann Machine 2048 Inputs	6 Talkers 11 Vowels 264 Tokens	Multi Talker - 15%
Robinson and Fallside (1988b)	Recurrent Net 20 Inputs	7 Talkers 27 Phonemes 558 Sentences	Multi Talker - 30.8% Talker Dep. - 22.7%

Table 4: Recognition of Speech Using Recurrent Neural Nets.

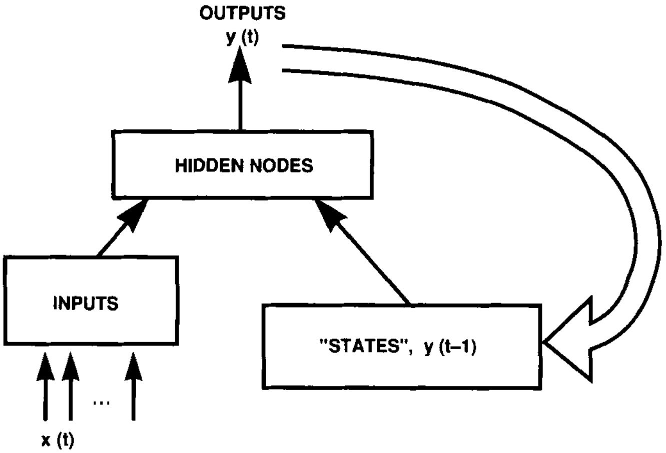

Figure 5: A recurrent neural net classifier.

Prager, Harrison, and Fallside (Prager et al. 1986) performed one of the first experiments to evaluate the use of Boltzmann machines for speech recognition. At the time this study was performed, the Boltzmann machine training algorithm described in (Ackley et al. 1985) was the only well-known technique that could be used to train nets with recurrent connections. This training algorithm is computationally intensive because simulated annealing procedures (Kirkpatrick et al. 1983) are used to perform a probabilistic search of connection weights. Binary input and output data representations were developed to apply Boltzmann machines to an 11-vowel recognition task. One successful net used 2048 input bits to represent 128 spectral values and 8 output bits to specify the vowel. Nets typically contained 40 hidden nodes and 7320 links. Training used 264 tokens from 6 talkers and required 6 to 15 hours of processing on a high-speed array processor. The resulting multi-talker error rate was 15%.

Prager, Harrison, and Fallside (Prager et al. 1986) also explored the use of a Boltzmann machine recognizer inspired by single-order Markov Model approaches to speech recognition. A block diagram of this recurrent net is presented in figure 5. The output of this net is delayed and fed back to the input to "carry" nodes that provide information about the prior state. This net was trained to identify words in two sentences spoken by one talker. Training time required 4 to 5 days of processing on a VAX 11/750 computer and performance was nearly perfect on the training sentences. Other recent work on Boltzmann machines (Bengio and De Mori 1988; Kohonen et al. 1988; Prager and Fallside 1987) demonstrates that good performance can be provided at the expense of excessive training time. Preliminary work on analog VLSI implementations of the training algorithm required by Boltzmann machines has demonstrated practical learning times for small hardware networks (Alspector and Allen 1987).

Many types of recurrent nets have been proposed that can be trained with modified forms of back-propagation. Jordan (1986) appears to have been the first to study nets with recurrent connections from output to input nodes as in figure 5. He used these nets to produce pattern sequences. Bourlard and Wellekens (1988) recently proved that such nets could be used to calculate local probabilities required in HMM recognizers and Robinson and Fallside (1988a) pointed out the relationship between these nets and state space equations used in classical control theory. Nets with recurrent self-looping connections on hidden and output nodes were studied by Watrous and Shastri (1987) for a speech recognition application. Nets with recurrent connections from hidden nodes to input nodes were studied by Elman (1988) and by Servan-Schreiber, Cleeremans, and McClelland (1988) for natural language applications.

Two recent studies have explored recurrent nets similar to the net shown in figure 5 when trained with modified forms of back-propagation. Robinson and Fallside (1988b) used such a net to label speech frames with one of 27 phoneme labels using hand-marked testing and training data. Training used an algorithm suggested by Rumelhart et al. (1986a) that, in effect, replicates the net at every time step during training. Talker-dependent error rates were 22.7% for the recurrent net and 26.0% for a simple feed-forward net with delays between input nodes to provide input context. Multi-talker error rates were 30.8% for the recurrent net and 40.8% for the feed-forward net. A 64 processor array of transputers provided practical training times in these experiments.

Anderson, Merrill, and Port (1988) also explored recurrent nets similar to the net in figure 5. Stimuli were CV syllables formed from six stop consonants and the vowel /a/ that were hand segmented to contain 120 msecs of speech around the vowel onset. Nets were trained on data from 10 talkers, tested on data from 10 other talkers, and contained from one to two hidden layers with different numbers of hidden nodes. Best performance (an error rate of 13.1%) was provided by a net with two hidden layers.

7 Integrating Neural Net and Conventional Approaches

Researchers are beginning to combine conventional HMM and DTW speech recognition algorithms with neural net classification algorithms and also to design neural net architectures that perform computations required by important speech recognition algorithms. This may lead

Study	Approach	Comments
Bourlard and Wellekens (1987)	MLP Provides Allophone Distance Scores for DTW Recognizer	Good Performance on 918-Word, Talker-Dependent, Continuous-Speech Task
Burr (1988a)	MLP Classifier After Energy-Based DTW	Tested on Single-Talker E-Set
Huang and Lippmann (1988)	Second-Stage MLP Discrimination After HMM Recognizer	Improved Performance for "B,D,G" from TI Alpha-Digit Data Base
Lippmann and Gold (1987)	"Viterbi-Net" Neural Net Architecture for HMM Viterbi Decoder	Same Good Performance on Large Data Base as Robust HMM Recognizer
Sakoe and Iso (1987)	MLP Provides Distance Scores for DTW Recognizer	No Hand Labeling Required, Untested

Table 5: Studies Combining Neural Net and Conventional Approaches.

to improved recognition accuracy and also to new designs for compact real-time hardware. Combining the good discrimination of neural net classifiers with the automatic scoring and training algorithms used in HMM recognizers could lead to rapid advances by building on existing high-performance recognizers. Studies that have combined neural net and conventional approaches to speech recognition are listed in table 5. Many (Bourlard and Wellekens 1987; Burr 1988b; Huang et al. 1988; Sakoe and Iso 1987) integrate multilayer perceptron classifiers with conventional DTW and HMM recognizers and one (Lippmann and Gold 1987) provides a neural-net architecture that could be used to implement an HMM Viterbi decoder. One study (Bourlard and Wellekens 1987) demonstrated how a multilayer perceptron could be integrated into a DTW continuous-speech recognizer to improve recognition performance.

7.1 Integrating Multilayer Perceptron Classifiers with DTW and HMM Recognizers. At least three groups have proposed recognizers where multilayer perceptrons compute distance scores used in DTW or HMM recognizers (Bourlard and Wellekens 1987; Burr 1988a; Sakoe and Iso 1987). Bourlard and Wellekens (1987) demonstrated how the multilayer perceptron shown in figure 6 could be used to calculate allophone distance scores required for phoneme and word recognition in a DTW discrete-observation recognizer. One net had inputs from 15 frames of speech centered on the current frame, 50 hidden nodes, and 26 output nodes. Outputs corresponded to allophones in a 10-digit German vocabulary. Inputs were from 60 binary variables per frame. One input bit was on in each frame to specify the codebook entry that represented that frame. The multilayer perceptron was trained using hand-labeled training data to provide a high output only for that output node corresponding to the current input allophone. Recognition then used dynamic time warping with local distances equal to values from output nodes. This provides good discrimination from the neural net and integration over time from the DTW algorithm. Perfect recognition performance was provided for recognition of 100 tokens from one talker.

Bourlard and Wellekens (1987) also used a multilayer perceptron with contextual input and DTW to recognize words from a more difficult 919-word talker-dependent continuous-speech task. The net covered an input context of 9 frames, used one of 132 vectors to quantize each frame, had 50 or 200 hidden nodes, and had 50 output nodes corresponding to 50 German phonemes. This net was trained using 100 hand-segmented sentences and tested on 188 other sentences containing roughly 7300 phonemes. The phoneme error rate was 41.6% with 50 hidden nodes and 37% with 200 hidden nodes. These error rates were both lower than the 47.5% error rate provided by a simple discrete-observation HMM recognizer with duration modeling and one probability histogram per phoneme. Bourlard and Wellekens suggested that performance could be improved and the need for hand-segmented training data could be eliminated by embedding multilayer perceptron back-propagation training in an iterative Viterbi-like training loop. This loop could progressively improve segmentation for DTW or HMM recognizers. Iterative Viterbi training was not performed because the simpler single-pass training required roughly 200 hours on a SUN-3 workstation. As noted above, Bourlard and Wellekens (1988) also recently proved that recurrent neural nets could calculate local probabilities required in HMM recognizers.

Sakoe and Iso (1987) suggested a recognition structure similar to that of Bourlard and Wellekens (1987) where a multilayer perceptron with delays between input nodes computes local distance scores. They, however, do not require output nodes of the multilayer perceptron to represent sub-word units such as phonemes. Instead, a training algorithm is described that is similar to the iterative Viterbi-like training loop suggested by Bourlard and Wellekens (1987) but for continuous input parameters. No results were presented for this approach.

Burr (1988a) gave results for a recognizer where words were first aligned based on energy information to provide a fixed 20 input frames of spectral information. These inputs were fed to nine outputs representing members of the E-set ("B,C,D,E,G,P,T,V,Z"). This recognizer was trained and tested using 180 tokens from one talker. Results were nearly perfect when the initial parts of these words were oversampled.

Huang and Lippmann demonstrated how a second-stage of analysis using a multilayer perceptron could decrease the error rate of an HMM recognizer (Huang and Lippmann 1988). The Viterbi backtraces from an HMM recognizer were used to segment input speech frames and average HMM log probability scores for segments were provided as inputs to single- and multilayer perceptrons. Performance was evaluated using the letters "B,D,G" spoken by the 16 talkers in the TI alpha-digit data base. Ten training tokens per letter were used to train the HMM and neural net recognizer for each talker and the 16 other tokens were used for testing. Best performance was provided by a single-layer perceptron which almost halved the error rate. The error rate dropped from 7.2% errors with the HMM recognizer alone to 3.8% errors with the neural net postprocessor.

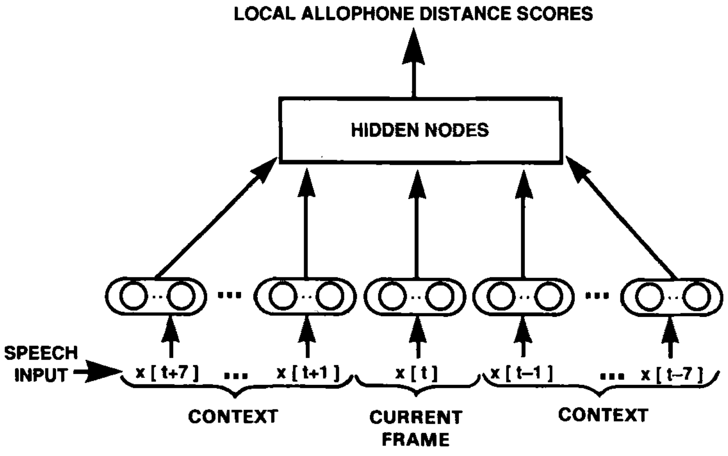

Figure 6: A feed-forward multilayer perceptron that was used to compute allophone distance scores for a DTW recognizer.

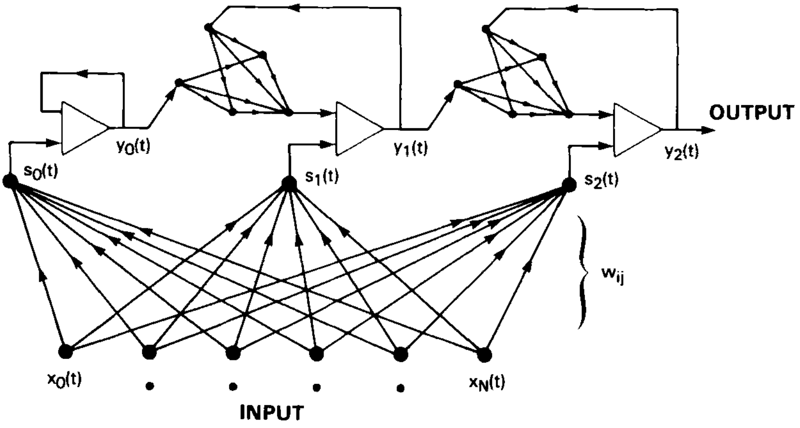

Figure 7: A recurrent neural net called a Viterbi net that performs the calculations required in an HMM Viterbi decoder.

7.2 A Neural Net Architecture to Implement a Viterbi Decoder. Lippmann and Gold (1987) described a neural-net architecture called a Viterbi net that could be used to implement the Viterbi decoder used in many continuous observation HMM recognizers using analog VLSI techniques. This net is shown in figure 7. Nodes represented by open triangles correspond to nodes in a left-to-right HMM word model. Each of these triangles represents a threshold-logic node followed by a fixed delay. Small subnets in the upper part of the figure select the maximum of two inputs as described in (Lippmann et al. 1987) and subnets in the lower part sum all inputs. A temporal sequence of input vectors is presented at the input and the output is proportional to the log probability calculated by a Viterbi decoder. The structure of the Viterbi net illustrates how neural net components can be integrated to design a complex net which performs the calculations required by an important conventional algorithm.

The Viterbi net differs from the Viterbi decoding algorithm normally implemented in software and was thus evaluated using 4000 word tokens from the 9-talker 35-word Lincoln Stress-Style speech data base. Connection strengths in Viterbi nets with 15 internal nodes (one node per HMM model state) were adjusted based on parameter estimates obtained from the forward-backward algorithm. Inputs consisted of 12 mel cepstra and 13 differential mel cepstra that were updated every 10 msec. Performance was good and almost identical to that of current Robust HMM isolated-word recognizers (Lippmann and Gold 1987). The error rate was 0.56% or only 23 out of 4095 tokens wrong. One advantage an analog implementation of this net would have over digital approaches is that the frame rate could be increased to provide improved temporal resolution without requiring higher clock rates.

8 Other Nets for Pattern Sequence Recognition

In addition to the neural net models described above, other nets motivated primarily by psychological and physiological findings and by past work on associative memories have been proposed for speech recognition and pattern sequence recognition. Although some of these nets represent new approaches to the problem of pattern sequence recognition, few have been integrated into speech recognizers and none have been evaluated using large speech data bases.

8.1 Psychological Neural Net Models of Speech Perception. Three neural net models have been proposed which are primarily psychological models of speech perception (Elman and McClelland 1986; MacKay 1987; Marslen-Wilson 1987; Rumelhart et al. 1986b). The COHORT model developed by Marslen-Wilson (1987) assumes a left-to-right real-time acoustic phonetic analysis of speech as in current recognizers. It accounts for many psychophysical results in speech recognition such as the existence of a time when a word becomes unambiguously recognized (recognition point), the word frequency effect, and recognition of contextually inappropriate words. This model, however, is descriptive and is not expressed as a computational model.

Hand crafted versions of the TRACE and Interactive Activation models developed by Elman, McClelland, Rumelhart, and co-workers were tested with small speech data bases (Elman and McClelland 1986; Rumelhart et al. 1986b). These models are based on neuron-like nodes, include both feed-forward and feed-back connections, use nodes with multiplicative operations, and emphasizes the benefits that can be obtained by using co-articulation information to aid in word recognition. These models are impractical because the problems of time alignment and training are not addressed and the entire network must be copied on every new time step. The Node Structure Theory developed by MacKay (1987) is a qualitative neural theory of speech recognition and production. It is similar in many ways to the above models, but considers problems related to talking rate, stuttering, internal speech, and rhythm.

8.2 Physiological Models For Temporal Pattern Recognition. Neural net approaches motivated primarily by physiological and behavioral results have also been proposed to perform some component of the time alignment task (Cohen et al. 1987; Dehaene et al. 1987; Wong and Chen 1986). Wong and Chen (1986) and Dehaene et al. (1987) describe similar models that have been tested with a small amount of speech data. These models include neurons with shunting or multiplicative nodes similar to those that have been proposed in the retina to compute direction of motion (Poggio and Koch 1987). Three neurons can be grouped to form a "synaptic triad" that can be used to recognize two component pattern sequences. This triad will have a strong output only if the modulator input goes "high" and then, a short time later, the primary input goes "high."

Synaptic triads can be arranged in sequences and in hierarchies to recognize features, allophones and words (Wong and Chen 1986). In limited tests, hand crafted networks could recognize a small set of words spoken by one talker (Wong and Chen 1986). More interesting is a proposed technique for training such networks without supervision (Dehaene et al. 1987). If effective, this could make use of the large amount of unlabeled speech data that is available and lead to automatic creation of sub-word models. Further elaboration is necessary to describe how networks with synaptic triads could be trained and used in a recognizer.

Cohen and Grossberg proposed a network called a masking field that has not yet been tested with speech input (Cohen and Grossberg 1987).

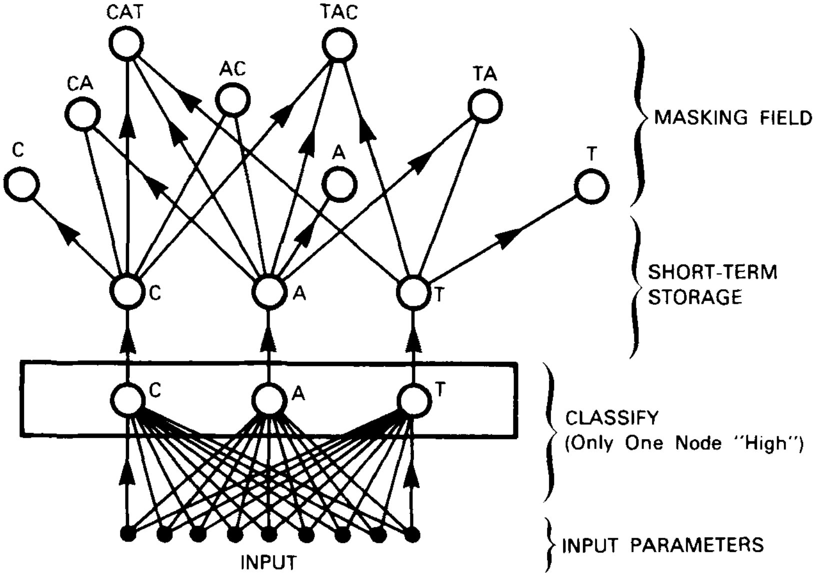

Figure 8: A model called a masking field that can be used to detect pattern sequences.

This network is shown in figure 8. Inputs are applied to the bottom subnet which is similar to a feature map net (Kohonen et al. 1984). Typically, only one node in this subnet has a "high" output at any time. Subnet node outputs feed short-term storage nodes whose outputs decay slowly over time. Different input pattern sequences thus lead to different amplitude patterns in short term storage. For example the input C-A-T sampled at the end of the word will yield an intensity pattern in short-term storage with node C low, node A intermediate, and node T high. The input T-A-C will yield a pattern with node C high, node A intermediate, and node T low. These intensity patterns are weighted and fed to nodes in a masking field with weights adjusted to detect different patterns. The masking field is designed such that all nodes compete to be active and nodes representing longer patterns inhibit nodes representing shorter patterns. This approach can recognize short isolated pattern sequences but has difficulty recognizing patterns with repeated sub-sequences because nodes in short-term storage corresponding to those sub-sequences can become saturated. Further elaboration is necessary to describe how masking fields should be integrated into a full recognizer. Other recent studies (Jordan 1986; Stornetta et al. 1988; Tattersall et al. 1988) have also proposed using slowly-decaying nodes as short-term storage to provide history useful for pattern recognition and pattern sequence generation.

8.3 Sequential Associative Memories. A final approach to pattern sequence recognition is to build a sequential associative memory for pattern sequences as described in (Amit 1988; Buhmann and Schulten 1988; Hecht-Nielsen 1987; Kleinfield 1986; Sompolinsky and Kanter 1986). These nets extend past work on associative memories by Hopfield and Little (Hopfield 1982; Little 1974) to the case where pattern sequences instead of static patterns can be restored. Recognition in this approach corresponds to the net settling into a desired sequence of stable states, one after the other, when driven by an input temporal pattern sequence.

Dynamic associative memory models developed by Amit, Kleinfield, Sompolinsky, and Kanter (Amit 1988; Kleinfield 1986; Sompolinsky and Kanter 1986) use long and short delays on links to generate and recognize pattern sequences. Links with short delays mutually excite a small set of nodes to produce stable states. Links with long delays excite nodes in the next expected stable state. Transitions between states thus occur at predetermined times that depend on the delays in the links. A net developed by Buhmann and Schulten (1988) uses probabilistic nodes to produce sequencing behavior similar to that produced by a Markov chain. Transitions in this net occur stochastically but at some average rate. A final net described by Hecht-Nielsen (1987) is a modified version of Grossberg's avalanche net (Grossberg 1988). The input to this net is similar in structure to Kohonen's feature map. It differs in that nodes have different rise and fall time constants and overall network activity is controlled such that only the outputs of a few nodes are "high" at any time.

A few relatively small simulations have been performed to explore the behavior of the sequential associative memories. Simulations have demonstrated that these nets can complete pattern sequences given the first element of a sequence (Buhmann and Schulten 1988) and also perform such functions as counting the number of input patterns presented to a net (Amit 1988). Although this approach is theoretically very interesting and may be a good model of some neural processing, no tests have been performed with speech data. In addition, further work is necessary to develop training procedures and useful decoding strategies that could be applied in a complete speech recognizer.

9 Summary of Past Research

The performance of current speech recognizers is far below that of humans. Neural nets offer the potential of providing massive parallelism, adaptation, and new algorithmic approaches to speech recognition problems. Researchers are investigating:

1. New physiological-based front ends,
2. Neural net classifiers for static speech input patterns,
3. Neural nets designed specifically to classify temporal pattern sequences,
4. Combined recognizers that integrate neural net and conventional recognition approaches,
5. Neural net architectures that implement conventional algorithms, and
6. VLSI hardware neural nets that implement both neural net and conventional algorithms.

Physiological front ends have provided improved recognition accuracy in noise (Ghitza 1988; Hunt and Lefèbvre 1988) and a cochlea filterbank that could be used in these front ends has been implemented using micro-power VLSI techniques (Lyon and Mead 1988). Many nets can compute the complex likelihood functions required by continuous-distribution recognizers and perform the vector quantization required by discrete-observation recognizers. Kohonen's feature map algorithm (Kohonen et al. 1984) has been used successfully to vector quantize speech and preliminary VLSI hardware versions of this net have been built (Mann et al. 1988).

Multilayer perceptron networks with delays have provided excellent discrimination between small sets of difficult-to-discriminate speech inputs (Kammerer and Kupper 1988; Lang and Hinton 1988; Peeling and

Moore 1987; Waibel et al. 1987; Waibel et al. 1988; Watrous 1988). Good discrimination was provided for a set of 18 consonants in varying phonetic contexts (Waibel et al. 1988), similar E-set words such as "B,D,E,V" (Lang and Hinton 1988), and digits and words from small-vocabularies (Kammerer and Kupper 1988; Peeling and Moore 1987; Watrous 1988). In some cases performance was similar to or slightly better than that provided by a more conventional HMM or DTW recognizer (Kammerer and Kupper 1988; Lang and Hinton 1988; Peeling and Moore 1987; Waibel et al. 1987; 1988). In almost all cases, a neural net approach performed as well as or slightly better than conventional approaches but provided a parallel architecture that could be used for implementation and a computationally simple and incremental training algorithm.

Approaches to the problem of scaling a network up in size to discriminate between members of a large set have been proposed and demonstrated (Waibel et al. 1988). For example, a net that classifies 18 consonants accurately was constructed from subnets trained to discriminate between smaller subsets of these consonants. Algorithms that use combined unsupervised/supervised training and provide high performance and extremely rapid training have also been demonstrated (Huang and Lippmann 1988; Kohonen et al. 1988). New training algorithms are under development (Almeida 1987; Jordan 1986; Pineda 1987; Rohwer and Forrest 1987; Watrous 1988) that can be used with recurrent networks.

Preliminary studies have explored recognizers that combine conventional and neural net approaches. Promising continuous-speech recognition results have been obtained by integrating multilayer perceptrons into a DTW recognizer (Bourlard and Wellekens 1987) and a multilayer perceptron post processor has improved the performance of an isolated-word HMM recognizer (Huang et al. 1988). Neural net architectures have also been designed for important conventional algorithms. For example, recurrent neural net architectures have been developed to implement the Viterbi decoding algorithm used in many HMM speech recognizers (Lippmann and Gold 1987) and also to compute local probabilities required in discrete-observation HMM recognizers (Bourlard and Wellekens 1988).

Many new neural net models have been proposed for recognizing temporal pattern sequences. Some are based on physiological data and attempt to model the behavior of biological nets (Dehaene et al. 1987; Cohen et al. 1987; Wong and Chen 1986) while others attempt to extend existing auto-associative networks to temporal problems (Amit 1988; Buhmann and Schulten 1988; Kleinfield 1986; Sompolinsky and Kanter 1986). New learning algorithms and net architectures will, however, be required to provide the real-time response and automatic learning of internal word and phrase models required for high-performance continuous speech recognition. This is still a major unsolved important problem in the field of neural nets.

10 Suggestions for Future Work

Further work should emphasize networks that provide rapid response and could be used with real-time speech input. They must include internal mechanisms to distinguish speech from background noise and to determine when a word has been presented. They also must operate with continuous acoustic input and not require hand marking of test speech data, long internal delays, or duplication of the network for new inputs.

Short-term research should focus on a task that current recognizers perform poorly on such as accurate recognition of difficult sets of isolated words. Such a task wouldn't require excessive computation resources or extremely large data bases. A potential initial problem is talker-independent recognition of difficult E-set words or phonemes as in (Lang and Hinton 1988; Waibel et al. 1988). Techniques developed using small difficult vocabularies should be extended to larger vocabularies and continuous speech as soon as feasible. Efforts should focus on: developing training algorithms to construct sub-word and word models automatically without excessive supervision, developing better front-end acoustic-phonetic feature extraction, improving low-level acoustic/phonetic discrimination, integrating temporal sequence information over time, and developing more rapid training techniques. Researchers should continue integrating neural net approaches to classification with conventional approaches to training and scoring. Longer-term research on continuous-speech recognition must address the problems of developing high-level speech-understanding systems that can learn and use internal models of the world. These systems must be able to learn and use syntactic, semantic, and pragmatic constraints.

Efforts on building neural net VLSI hardware for speech recognition should also continue. The development of compact real-time speech recognizers is a major goal of neural net research. Parallel neural-net architectures should be designed to perform the computations required by successful algorithms and then these architectures should be implemented and tested. Recent developments in analog VLSI neural nets suggest that this approach has the potential to provide the high computation rates required for both front-end acoustic analysis and high-level pattern matching.

All future work should take advantage of the many speech data bases that currently exist and use results obtained with experimental HMM and DTW recognizers with these data bases as benchmarks. Descriptions of some common data bases and comments on their availability are in (Pallett 1986; Price et al. 1988). Detailed evaluations using large speech data bases are necessary to guide research and permit comparisons between alternative approaches. Results obtained on a few locally-recorded speech samples are often misleading and are not informative to other researchers.

Research should also build on the current state of knowledge in neural networks, pattern classification theory, statistics, and conventional HMM and DTW approaches to speech recognition. Researchers should become familiar with these areas and not duplicate existing work. Introductions to current HMM and DTW approaches are available in (Dixon and Martin 1979; Lee and Hon 1988; Parsons 1986; Rabiner and Juang 1986; Rabiner et al. 1978) and introductions to statistics and pattern classification are available in many books including (Duda and Hart 1973; Fukunaga 1972; Nilsson 1965).

Acknowledgments

I would like to thank members of Royal Signals and Radar Establishment including John Bridle and Roger Moore for discussions regarding the material in this paper. I would also like to thank Bill Huang and Ben Gold for interesting discussions and Carolyn for her patience.

References

Ackley, D.H., G.E. Hinton, and T.J. Sejnowski. 1985. A Learning Algorithm for Boltzmann Machines. *Cognitive Science* **9**, 147–160.

Albus, J.S. 1981. *Brain, Behavior, and Robotics*. BYTE Books.

Almeida, L.B. 1987. A Learning Rule for Asynchronous Perceptrons with Feedback in a Combinatorial Environment. *In:* 1st International Conference on Neural Networks. IEEE, II–609.

Alspector, J. and R.B. Allen. 1987. A Neuromorphic VLSI Learning System. *In:* Advanced Research in VLSI: Proceedings of the 1987 Stanford Conference, ed. P Losleben, 313–349. Cambridge: MIT Press.

Amit, D.J. 1988. Neural Networks for Counting Chimes. *Proceedings National Academy of Science, USA* **85**, 2141–2145.

Anderson, S., J. Merrill, and R. Port. 1988. *Dynamic Speech Categorization With Recurrent Networks.* Technical Report 258, Department of Linguistics and Department of Computer Science, Indiana University.

Averbuch, A., L. Bahl, and R. Bakis. 1987. Experiments with the Tangora 20,000 Word Speech Recognizer. *In:* Proceedings IEEE International Conference on Acoustics Speech and Signal Processing, Dallax, TX, 701–704.

Bahl, L.R., P.F. Brown, P.V. De Souza, and R.L. Mercer. 1988. Modeling Acoustic Sequences of Continuous Parameters. *In:* Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, New York, NY, 40–43.

Baldi, P. and K. Hornik. 1988. Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima. *Neural Networks* **2**, 53–58.

Beet, S.W., H.E.G. Powrie, R.K. Moore, and M.J. Tomlinson. 1988. Improved Speech Recognition Using a Reduced Auditory Representation. *In:* Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, New York, NY, 75–78.

Bengio, Y. and R. De Mori. 1988. Use of Neural Networks for the Recognition of Place of Articulation. *In:* Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, New York, NY, 103–106.

Bourlard, H. and Y. Kamp. 1988. Auto-Association by Multilayer Perceptrons and Singular Value Decomposition. *Biological Cybernetics* **59**, 291–294.

Bourlard, H. and C.J. Wellekens. 1988. *Links Between Markov Models and Multilayer Perceptrons.* Technical Report Manuscript M-263, Phillips Research Laboratory, Brussels, Belgium.

———. 1987. *Speech Pattern Discrimination and Multilayer Perceptrons.* Technical Report Manuscript M-211, Phillips Research Laboratory, Brussels, Belgium. Scheduled to appear in the December issue of *Computer, Speech and Language.*

Bridle, J. 1988. Neural Network Experience at the RSRE Speech Research Unit. ATR Workshop on Neural Networks and Parallel Distributed Processing, Osaka, Japan.

Broomhead, D.S. and D. Lowe. 1988. *Radial Basis Functions, multi-variable functional interpolation and adaptive networks.* Technical Report RSRE Memorandum No. 4148, Royal Speech and Radar Establishment, Malvern, Worcester, Great Britain.

Buhmann, J. and K. Schulten. 1988. Noise-Driven Temporal Association in Neural Networks. *Europhysics Letters* **4**, 1205–1209.

Burr, D.J. 1988a. Experiments on Neural Net Recognition of Spoken and Written Text. *In:* IEEE Transactions on Acoustics, Speech and Signal Processing, **36**, 1162–1168.

———. 1988b. Speech Recognition Experiments with Perceptrons. *In:* Neural Information Processing Systems, ed. D. Anderson, 144–153. New York: American Institute of Physics.

Burton, D.K., J.E. Shore, and J.T. Buck. 1985. Isolated-Word Speech Recognition Using Multisection Vector Quantization Codebooks. *In:* IEEE Transactions on Acoustics, Speech and Signal Processing, **ASSP-33**, 837–849.

Cohen, M. and S. Grossberg. 1987. Masking fields: A Massively Parallel Neural Architecture for learning, recognizing, and predicting multiple groupings of patterned data. *Applied Optics* **26**, 1866–1891.

Cohen, M.A., S. Grossberg, and D. Stork. 1987. Recent Developments in a Neural Model of Real-Time Speech Analysis and Synthesis. *In:* 1st International Conference on Neural Networks, IEEE.

Cowan, J.D. and D.H. Sharp. 1988. Neural Nets and Artificial Intelligence. *Daedalus* **117**, 85–121.

Cybenko, G. 1988. *Continuous Valued Neural Networks with Two Hidden Layers are Sufficient.* Technical Report, Department of Computer Science, Tufts University.

Dehaene, S., J. Changeux, and J. Nadal. 1987. Neural Networks that Learn Temporal Sequences by Selection. *Proceedings National Academy Science, USA, Biophysics* **84**, 2727–2713.

Deng, Li and C. Daniel Geisler. 1987. A Composite Auditory Model for Processing Speech Sounds. *Journal of the Acoustical Society of America* **82:6**, 2001–2012.

Dixon, N.R. and T.B. Martin. 1979. *Automatic Speech and Speaker Recognition.* New York: IEEE Press.

Doddington, G.R. and T.B. Schalk. 1981. Speech Recognition: Turning Theory into Practice. *IEEE Spectrum*, 26–32.

Duda, R.O. and P.E. Hart. 1973. Pattern Classification and Scene Analysis. New York: John-Wiley & Sons.

Elman, J.L. 1988. *Finding Structure in Time.* CRL Technical Report 8801, University of California, San Diego, CA.

Elman, J.L. and J.L. McClelland. 1986. Exploiting Lawful Variability in the Speech Wave. *In:* Invariance and Variability in Speech Processes, eds. J.S. Perkell and D.H. Klatt. New Jersey: Lawrence Erlbaum.

Elman, J.L. and D. Zipser. 1987. *Learning the Hidden Structure of Speech.* ICS Report 8701, Institute for Cognitive Science, University of California, San Diego, La Jolla, CA.

Fallside, F., T.D. Harrison, R.W. Prager, and A.J.R. Robinson. 1988. A Comparison of Three Connectionist Models for Phoneme Recognition in Continuous Speech. ATR Workshop on Neural Networks and Parallel Distributed Processing, Osaka, Japan.

Fukunaga, K. 1972. *Introduction to Statistical Pattern Recognition.* New York: Academic Press.

Ghitza, O. 1988. Auditory Neural Feedback as a Basis for Speech Processing. *In:* Proceedings IEEE International Conference on Acoustics Speech and Signal Processing, New York, NY, 91–94.

Gold, B. 1988. A Neural Network for Isolated Word Recognition. *In:* Proceedings IEEE International Conference on Acoustics Speech and Signal Processing, New York, NY, 44–47.

Greenberg, S. 1988a. The Ear as a Speech Analyzer. *Journal of Phonetics* **16**, 139–149.

———. 1988b. Special Issue on "Representation of Speech in the Auditory Periphery." *Journal of Phonetics* **16**.

Grossberg, S. 1988. Nonlinear Neural Networks: Principles, Mechanisms, and Architectures. *Neural Networks* **1**, 17–61.

Hanson, S.J. and D.J. Burr. 1987. *Knowledge Representation in Connectionist Networks.* Technical Report, Bell Communications Research, Morristown, New Jersey.

Harrison, T.D. and F. Fallside. 1988. *A Connectionist Structure for Phoneme Recognition.* Technical Report CUED/F-INFENG/TR.15, Cambridge University Engineering Department.

Hecht-Nielsen, R. 1987. Nearest Matched Filter Classification of Spatiotemporal Patterns. *Applied Optics* **26**, 1892–1899.

Hinton, G.E. 1987. *Connectionist Learning Procedures.* Technical Report CMU-CS-87-115, Carnegie Mellon University, Computer Science Department.

Homma, T., L.E. Atlas, and R.J. Marks. 1988. An Artificial Neural Network for Spatio-Temporal Bipolar Patterns: Application to Phoneme Classification. *In:* Neural Information Processing Systems, ed. D. Anderson, 31–40. New York: American Institute of Physics.

Hopfield, J.J. 1982. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences, USA* **79**, 2554–2558.

Huang, W.M. and R.P. Lippmann. 1988. Neural Net and Traditional Classifiers. *In:* Neural Information Processing Systems, ed. D. Anderson, 387–396. New York: American Institute of Physics.

Huang, W.M., R.P. Lippmann, T. Nguyen. 1988. Neural Nets for Speech Recognition. *In:* Conference of the Acoustical Society of America, Seattle WA.

Hunt, M.J. and C. Lefèbvre. 1988. Speaker Dependent and Independent Speech Recognition Experiments With an Auditory Model. *In:* Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing **1**, New York, 215–218.

Irino, T. and H. Kawahara. 1988. A Study on the Speaker Independent Feature Extraction of Japanese Vowels by Neural Networks. ATR Workshop on Neural Networks and Parallel Distributed Processing, Osaka, Japan.

Jordan, M.I. 1986. *Serial Order: A Parallel Distributed Processing Approach.* Institute for Cognitive Science Report 8604, University of California, San Diego.

Kamm, C., T. Landauer, and S. Singhal. 1988. Training an Adaptive Network to Spot Demisyllables in Continuous Speech. ATR Workshop on Neural Networks and Parallel Distributed Processing, Osaka, Japan.

Kammerer, B. and W. Kupper. 1988. Experiments for Isolated-Word Recognition with Single and Multi-Layer Perceptrons, Abstracts of 1st Annual INNS Meeting, Boston. *Neural Networks* **1**, 302.

Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi. 1983. Optimization by Simulated Annealing. *Science* **229**, 671–679.

Klatt, K.H. 1986. The Problem of Variability In Speech Recognition and Models of Speech Perception. *In:* Invariance and Variability in Speech Processes, eds. J.S. Perkell and D.H. Klatt, 300-324. New Jersey: Lawrence Erlbaum.

Kleinfield, D. 1986. Sequential State Generation by Model Neural Networks. *Proceedings National Academy Science, USA, Biophysics* **83**, 9469–9473.

Kohonen, T. 1988. An Introduction to Neural Computing. *Neural Networks* **1**, 3–16.

Kohonen, T. 1984. *Self-Organization and Associative Memory.* Berlin: Springer-Verlag.

Kohonen, T., G. Barna, and R. Chrisley. 1988. Statistical Pattern Recognition with Neural Networks: Benchmarking Studies. *In:* IEEE Annual International Conference on Neural Networks, San Diego, July.

Kohonen, T., K. Makisara, and T. Saramaki. 1984. Phonotopic Maps — Insightful Representation of Phonological Features for Speech Recognition. *In:* IEEE Proceedings of the 7th International Conference on Pattern Recognition.

Lang, K.J. and G.E. Hinton. 1988. *The Development of the Time-Delay Neural Network Architecture for Speech Recognition.* Technical Report CMU-CS-88-152, Carnegie-Mellon University.

Lapedes, A. and R. Farber. 1988. How Neural Nets Work. *In:* Neural Information Processing Systems, ed. D. Anderson, 442–456. New York: American Institute of Physics.

Lee, Kai-Fu and Hsiao-Wuen Hon. 1988. Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM. *In:* Proceedings IEEE

International Conference on Acoustics, Speech and Signal Processing **1**, 123–126.

Lee, Y.C., G. Doolen, H.H. Chen, G.Z. Sun, T. Maxwell, H.Y. Lee, C.L. Giles. 1986. Machine Learning Using a Higher Order Correlation Network. *Physica D*, 276–306.

Leung, H.C. and V.W. Zue. 1988. Some Phonetic Recognition Experiments Using Artificial Neural Nets. *In:* Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing **1**.

Lippmann, R.P., B. Gold, and M.L. Malpass. 1987. *A Comparison of Hamming and Hopfield Neural Nets for Pattern Classification.* Technical Report TR-769, MIT Lincoln Lab.

Lippmann, R.P. 1987. An Introduction to Computing with Neural Nets. *IEEE ASSP Magazine* **4:2**, 4–22.

Lippmann, R.P. and Ben Gold. 1987. Neural Classifiers Useful for Speech Recognition. *In:* 1st International Conference on Neural Networks, IEEE, IV–417.

Little, W.A. 1974. The Existence of Persistent States in the Brain. *Mathematical Biosciences* **19**, 101–120.

Lyon, R.F. and C. Mead. 1988. An Analog Electronic Cochlea. *IEEE Transactions on Acoustics, Speech and Signal Processing* **36**, 1119–1134.

MacKay, D.G. 1987. *The Organization of Perception and Action*, New York: Springer Verlag.

Mann, J., J. Raffel, R. Lippmann, and B. Berger. 1988. A Self-Organizing Neural Net Chip. Neural Networks for Computing Conference, Snowbird, Utah.

Marslen-Wilson, W.D. 1987. Functional Parallelism in Spoken Word-Recognition. *In:* Spoken Word Recognition, eds. U.H. Frauenfelder and L.K. Tyler. Cambridge, MA: MIT Press.

McDermott, E. and S. Katagiri. 1988. Phoneme Recognition Using Kohonen's Learning Vector Quantization. ATR Workshop on Neural Networks and Parallel Distributed Processing, Osaka, Japan.

Moody, J. 1988. Speedy Alternatives to Back Propagation. Neural Networks for Computing Conference, Snowbird, Utah.

Moody, J. and C. Darken. 1988. *Learning with Localized Receptive Fields.* Technical Report YALEU/DCS/RR-649, Yale Computer Science Department, New Haven, CT.

Naylor, J. and K.P. Li. 1988. Analysis of a Neural Network Algorithm for Vector Quantization of Speech Parameters, Abstracts of 1st Annual INNS Meeting, Boston. *Neural Networks* **1**, 310.

Nilsson, Nils J. 1965. *Learning Machines.* New York: McGraw Hill.

Niranjan, M. and F. Fallside. 1988. *Neural Networks and Radial Basis Functions in Classifying Static Speech Patterns.* Technical Report CUED/F-INFENG/TR 22, Cambridge University Engineering Department.

Omohundro, S.M. 1987. Efficient Algorithms with Neural Network Behavior. *Complex Systems* **1**, 273–347.

Pallett, D.S. 1986. A PCM/VCR Speech Database Exchange Format. *In:* Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, Tokyo, Japan, 317–320.

Parsons, T. 1986. *Voice and Speech Processing*. New York: McGraw-Hill.

Paul, D.B. 1987. A Speaker-Stress Resistant HMM Isolated Word Recognizer. *ICASSP 87*, 713–716.

Peeling, S.M. and R.K. Moore. 1987. *Experiments in Isolated Digit Recognition Using the Multi-Layer Perceptron*. Technical Report 4073, Royal Speech and Radar Establishment, Malvern, Worcester, Great Britain.

Peterson, Gordon E. and Harold L. Barney. 1952. Control Methods Used in a Study of Vowels. *The Journal of the Acoustical Society of America* **24:2**, 175–84.

Pineda, F.J. 1987. Generalization of Back-Propagation to Recurrent Neural Networks. *Physical Review Letters* **59**, 2229–2232.

Poggio, T. and C. Koch. 1987. Synapses that Compute Motion. *Scientific American* **256**, 46–52.

Prager, R.W. and F. Fallside. 1987. A Comparison of the Boltzmann Machine and the Back Propagation Network as Recognizers of Static Speech Patterns. *Computer Speech and Language* **2**, 179–183.

Prager, R.W., T.D. Harrison, and F. Fallside. 1986. Boltzmann Machines for Speech Recognition. *Computer Speech and Language* **1**, 2–27.

Price, P., W.M. Fisher, J. Bernstein, D.S. Pallett. 1988. The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition. *In:* Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, New York **1**, 651–654.

Rabiner, L.R. and B.H. Juang. 1986. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine* **3:1**, 4–16.

Rabiner, Lawrence R. and Ronald W. Schafer. 1978. *Digital Processing of Speech*. New Jersey: Prentice-Hall.

Raffel, J., J. Mann, R. Berger, A. Soares, and S. Gilbert. 1987. A Generic Architecture for Wafer-Scale Neuromorphic Systems. *In:* 1st International Conference on Neural Networks, IEEE.

Robinson, A.J. and F. Fallside. 1988a. A Dynamic Connectionist Model for Phoneme Recognition. *nEuro '88*, Paris, France.

———. 1988b. Static and Dynamic Error Propagation Networks with Application to Speech Coding. *In:* Neural Information Processing Systems, ed. D. Anderson, 632–641. New York: American Institute of Physics.

Rohwer, R. and B. Forrest. 1987. Training Time-Dependencies in Neural Networks. *In:* 1st International Conference on Neural Networks, IEEE, II–701.

Rossen, M.L., L.T. Niles, G.N. Tajchman, M.A. Bush, J.A. Anderson, and S.E. Blumstein. 1988. A Connectionist Model for Consonant-vowel Syllable Recognition. *In:* Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, New York, NY, 59–66.

Rumelhart, D.E., G.E. Hinton, and R.J. Williams. 1986a. Interactive Processes in Speech Perception: The TRACE Model. *In:* Parallel Distributed Processing: Vol. 2, Psychological and Biological Models, eds. D.E. Rumelhart and J.L. McClelland. Cambridge, MA: MIT Press.

———. 1986b. Learning Internal Representations by Error Propagation. *In:* Parallel Distributed Processing: Vol. 1, Foundations. Cambridge, MA: MIT Press.

Sakoe, H. and K. Iso. 1987. *Dynamic Neural Network — A New Speech Recognition*

Model Based on Dynamic Programming and Neural Network. IEICE Technical Report 87, NEC Corporation.

Seneff, S. 1988. A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing. *Journal of Phonetics* **16**, 55–76.

Servan-Schreiber, D., A. Cleeremans, and J.L. McClellan. 1988. *Encoding Sequential Structure in Simple Recurrent Networks.* Technical Report CMU-CS-88-183, Carnegie Mellon University.

Shamma, S. 1988. The Acoustic Features of Speech Sounds in a Model of Auditory Processing: Vowels and Voiceless Fricatives. *Journal of Phonetics* **16**, 77–91.

Shore, J.E. and D.K. Burton. 1983. Discrete Utterance Speech Recognition Without Time Alignment. *IEEE Transactions on Information Theory* **IT-29**, 473–491.

Sompolinsky, H. and I. Kanter. 1986. Temporal Association in Asymmetrical Neural Networks. *Physical Review Letters* **57**, 2861–2864.

Stornetta, W.S., T. Hogg, and B.A. Huberman. 1988. A Dynamical Approach to Temporal Pattern Processing. *In:* Neural Information Processing Systems, ed. D. Anderson, 750–759. New York: American Institute of Physics.

Tank, D. and J.J. Hopfield. 1987. Concentrating Information in Time: Analog Neural Networks with Applications to Speech Recognition Problems. *In:* 1st International Conference on Neural Networks, IEEE.

Tattersall, G.D., P.W. Linford, and R. Linggard. 1988. Neural Arrays for Speech Recognition. *British Telecommunications Technology Journal* **6**, 140–163.

Unnikrishnan, K.P., J.J. Hopfield, and D.W. Tank. 1988. Learning Time-delayed Connections in a Speech Recognition Circuit. Neural Networks for Computing Conference, Snowbird, Utah.

Waibel, Alex, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. 1987. *Phoneme Recognition Using Time-Delay Neural Networks.* Technical Report TR-1-006, ATR Interpreting Telephony Research Laboratories, Japan. Scheduled to appear in March 1989 issue of *IEEE Transactions on Acoustics Speech and Signal Processing.*

Waibel, Alex, H. Sawai, and K. Shikano. 1988. *Modularity and Scaling in Large Phonemic Neural Nets.* Technical Report TR-I-0034, ATR Interpreting Telephony Research Laboratories, Japan.

Watrous, R.L. 1988. *Speech Recognition Using Connectionist Networks.* Ph.D thesis, University of Pennsylvania.

———. 1986. *Learning Algorithms for Connectionist Networks: Applied Gradient Methods of Nonlinear Optimization.* Technical Report MS-CIS-87-51, Linc Lab 72, University of Pennsylvania.

Watrous, R.L. and Lokendra Shastri. 1987. Learning Phonetic Features using Connectionist Networks: An Experiment in Speech Recognition. *In:* 1st International Conference on Neural Networks, IEEE, IV–381.

Wong, M.K. and H.W. Chen. 1986. Toward a Massively Parallel System for Word Recognition. *In:* Proceedings IEEE International Conference on Acoustics Speech and Signal Processing, 37.4.1–37.4.4.

Received 10 November; accepted 14 November 1988.