AUTOMATIC VOICE RECOGNITION USING TRADITIONAL AND ARTIFICIAL NEURAL NETWORK APPROACHES

Final Report

NASA/ASEE Summer Faculty Fellowship Program-1988

Johnson Space Center

Prepared By:	Nazeih M. Botros, Ph.D.
Academic Rank:	Assistant Professor
University & Department:	Southern Illinois University Department of Electrical Engineering Carbondale, Illinois 62901

NASA / JSC

Directorate:	Engineering
Division:	Tracking and Communications
JSC Colleague:	William Jordan
Date Submitted:	August 19, 1988
Contract Number:	NGT 44-005-803

ABSTRACT

The main objective of this research is to develop an algorithm for isolated-word recognition. This research is focused on digital signal analysis rather than linguistic analysis of speech. Features extraction is carried out by applying a Linear Predictive Coding (LPC) algorithm with order of 10. Continuous-word and speaker independent recognition will be considered in future study after accomplishing this isolated word research

To implement and test the proposed algorithm a microcomputer-based data acquisition system has been designed and constructed. The system digitizes the voice signal, after passing through a 100 c/s-3.8 Kc/s band pass filter, with a sample rate of 8 KHz and stores the digitized data into a 64Kx10-dynamic random access memory (DRAM) buffer. A squelch circuit consists mainly of comparators (74ls85) detects the beginning of the spoken word. The end of the word is detected by a software algorithm based on comparing the speech energy with a precalculated threshold. A flag signals the end of sampling and the data is transferred from the buffer to an IBM-PC where it is segmented into frames each, 30 millisecond long, and the LPC coefficients are calculated.

To examine the similarity between the reference and the training sets, two approaches are explored. The first is implementing traditional pattern recognition techniques where a dynamic time warping algorithm is applied to align the two sets and calculate the probability of matching by measuring the Euclidean distance between the two sets. The second is implementing a backpropagation artificial neural net model with three layers as the pattern classifier. The adaptation rule implemented in this network is the generalized least mean square (LMS) rule.

The first approach has been accomplished. A vocabulary of 50 words was selected and tested, the accuracy of the algorithm was found to be around 85%. The second approach is in progress at the present time. The topology of the backpropagation model consists of three layers: input, hidden, and output. The actual output of each node is calculated using a sigmoid nonlinearity function of the inner product of the weight and the input; the weights are adapted by using the formula $W_{ij}^{new} = W_{ij}^{old} + u\ E_j X_i$ where u is the gain factor, E_j is the error, and X_i is the input. The network is being simulated on a PC.

ORIGINAL PAGE IS
OF POOR QUALITY

INTRODUCTION

For more than a decade the United States Government, foreign countries especially Japan, private corporations, and universities have been engaged in extensive research on human-machine interaction by voice. The benefits of this interaction is especially noteworthy in situations when the individual is engaged in such hands/eyes-busy task, or in low light or darkness, or when tactile contact is impractical/impossible. These benefits make voice control a very effective tool for space-related tasks. Some of the voice control applications that have been studied in NASA-JSC are: VCS Flight experiments, payload bay cameras, EVA heads up display, mission control center display units, and voice command robot. A special benefit of voice control is in zero gravity condition where voice is a very suitable tool in controlling space vehicle equipment.

Automatic speech recognition is carried out mostly by extracting features from the speech signal and storing them in reference templates in the computer. These features carry the signature of the speech signal. These reference templates contain the features of a phoneme, word, or a sentence, depending on the structure of the recognizer. If a voice interaction with the computer takes place, the computer extracts features from this voice signal and compares it with the reference templates. If a match is found, the computer executes a programmable task such as moving the camera up or down.

Several digital signal processing algorithms are available for speech feature extraction. The efficiency of the current algorithms is limited by: hardware restriction, execution time, and easiness of use. Some of these algorithms are: Linear Predictive Coding (LPC), Short-time Fourier Analysis, and Cepstrum analysis. Among these algorithms, LPC is the most widely used since it is easy to use, has short execution time, and do not require large memory storage. However, this algorithm has several limitations due to the assumptions upon which it is based upon.

Current speech recognition technology is not sufficiently advanced to achieve high performance on continuous spoken input with large vocabularies and/or arbitrary speakers. A major obstacle in achieving such high performance is the limited capability of the traditional pattern recognition (classifier) algorithms that are currently implemented. Due to this limited capability, and considering the fact that humans have a fascinating capability of recognizing the spoken words, researchers have started to explore the possibility of implementing human-like models, or what is known as artificial neural networks, as the pattern classifiers.

Neural net models have the greatest potential in area such as speech and image recognition where many hypotheses are pursued in parallel, high computation rates are required, ability to learn is desired, and the current best systems are far from equaling human performance. Most neural net algorithms adapt connection weights in time to improve performance based on current results. Adaptation or learning is a major focus of neural net research. The ability to adapt and continue learning is essential in area such as speech recognition where training data is limited and new talkers, new words, new dialects, new phrases, and new environments are continuously encountered.

LINEAR PREDICTIVE CODING (LPC)

Feature extraction in our research is carried out through a Linear Predictive Coding algorithm. The signal corresponding to a spoken word is segmented into frames each 30 milliseconds long and the algorithm is applied to replace each frame with 10 coefficients. In the following, we briefly review the LPC algorithm. Details of this algorithm can be found elsewhere [1-9]. This algorithm is built on the fact that there is a high correlation between adjacent samples of the speech signal in the time domain. This fact means that an nth sample of speech signal can be predicted from previous samples. The correlation can be put in a linear relationship as:

$$\hat{Y}_n = a_1Y_{n-1}+a_2Y_{n-2}+...+a_pY_{n-p} \qquad ....(1)$$

where p is the order of analysis. Usually p ranges from 8 to 12. $\hat{y}_n$ is the predicted value of speech at time n and a's are the linear predictive coefficients. The prediction error E_n that resulted from the above linear relationship is:

$$E_n = \sum_{i=0}^{p} a_iY_{n-i}, \qquad a_o = 1. \qquad ....(2)$$

To find the predictive coefficients which give least mean square error, the above equation is squared, partially differentiated with respect to a's, and time average term by term. The result is p equations in p unknowns as shown below:

$$\begin{bmatrix} r_0 & r_1 & r_2 & \ldots, & r_{p-1} \\ r_1 & r_0 & r_1 & \ldots, & r_{p-2} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ r_{p-1} & r_{p-2} & r_{p-3} & \ldots, & r_0 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \cdot \\ \cdot \\ \cdot \\ r_p \end{bmatrix} \quad \ldots\ldots\ldots(3)$$

with

$$\begin{aligned} r_0 &= \overline{Y_n Y_n}, \\ r_j &= \overline{Y_n Y_{n+j}} = \overline{Y_{n-j} Y_n} \end{aligned} \quad \ldots\ldots\ldots\ldots(4)$$

where r_j is a correlation coefficient of waveform $\{y_n\}$ and $r_{-j} = r_j$ by the assumptions of stationary state of y_n The coëfficients a_i's exist only if the matrix in equation 3 is a positive definite. To ensure that this condition is satisfied, y_n is multiplexed by a time window W_n. This multiplexing makes y_n exist in a finite interval from 0 to N-1, where N is the interval of the Window; a stable solution for equation 3 is always obtained. Accordingly, r_j can be written as:

$$r_j = \frac{1}{N} \sum_{n=0}^{N-j-1} Y_n Y_{n+j} W_n W_{n+j} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

In our study, a Hamming window is implemented. Calculation of the correlation coefficients by window multiplexing is called the correlation method. a_i's correspond to the resonance frequencies of the signal, and if p, the order of the analysis is selected correctly, these a_i's represent the formants, frequencies at which peaks of the power spectrum of the speech signal occur. A block diagram representing an algorithm for voice recognition based on LPC analysis is shown in Figure 1.

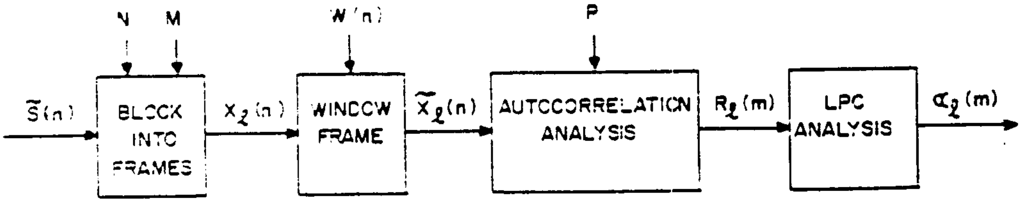

FIGURE 1. CALCULATION OF THE LPC COEFFICIENTS [5]

EXPERIMENTAL DESIGN

To calculate the LPC coefficients and test the performance of the proposed algorithm, a microprocessor-based data acquisition system has been designed and constructed. Figure 2 shows a block diagram of the system. The system is designed to lay the foundation for further expansion and enhancement for more sophisticated microprocessor-based speech identification/ recognition research. The system receives the voice signal through a microphone coupled with an audio amplifier. The voice signal passes through an active multiple feedback bandpass filter with a 3db bandwidth of 3.5 KHz approximately. The output signal of the filter is applied to the analog-to-digital converter where it is digitized with a sampling rate of 8 KHz. A squelch circuit is constructed to detect the beginning of the utterance and accordingly activates a temporary storage buffer to store the digitized data. The buffer consists of DRAMS with maximum capacity of 64 Kbyte. A hardware flag (the output bit of a flip flop) signals the end of sampling and the data is transferred to the microprocessor where digital signal analysis and pattern recognition algorithms are applied. The digital-to-analog converter, power amplifier and loud speaker are used to verify the storage. If the output of this circuit matches the original signal, then the storage is successful. The system has been tested successfully by the aid of a function generator. Details of the hardware of the

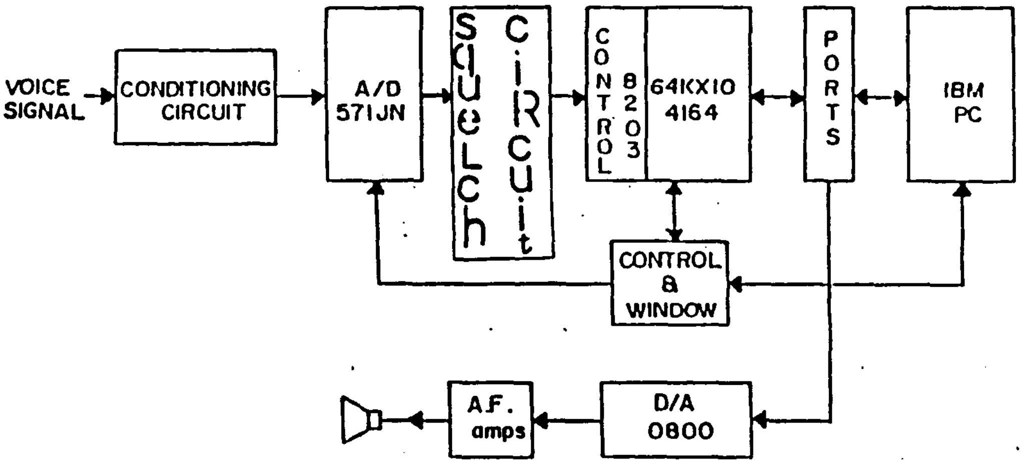

FIGURE 2. A BLOCK DIAGRAM OF THE DATA ACQUISITION SYSTEM.

system can be found elsewhere [9].

PATTERN RECOGNITION

To recognize the spoken word, an algorithm that compares between the reference and the training patterns to see whether they match or not should be developed. Two approaches will be discussed here. The first is a traditional pattern classifier approach where training pattern is aligned in time with the reference pattern and the Euclidean distance between the two patterns is caloulated and taken as the probability of matching. The second is based on implementing a backpropagation artificial neural network as the pattern classifier. In the following, we discuss briefly the two approaches.

A. Dynamic Time Warping (DTW)

The dynamic time warping (DTW) algorithm finds the "optimal" (least cost) warping path w(n) which minimizes the accumulated distance, D, between training and reference patterns, subject to a set of path and endpoint constraints. The dynamic programming algorithm is based upon the fact that the optimal path to point (i,j) in the two dimensional matrix illustrated in Figure 3 must pass through either the point (i-1,j), or (i-1,j-1), or (i,j-1). The minimum accumulated distance to point (i,j) is then given by:

$$D(i,j)=Dist(i,j)+Min\{D(i-1,j),D(i-1,j-1),D(i,j-1)\} \quad \ldots\ldots\ldots\ldots\ldots(11)$$

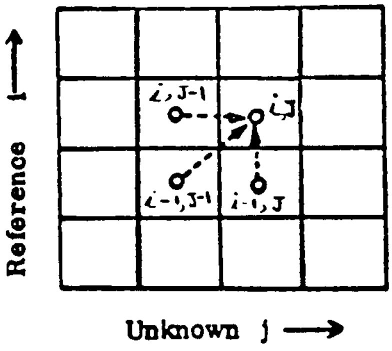

FIGURE 3. DYNAMIC TIME ALIGNMENT

where Dist (i,j) is the distance between the reference and training pattern at time j. The algorithm recursively computes this distance column by column to determine the minimum accumulated distance to the point (M,N), where M is the number of frames in the reference and N is the number of frames in the unknown (training). This path results in a time alignment in which the reference word has the maximum acoustic similarity with the input.

B. Artificial Neural Network [10-18]

Artificial neural network is a non-algorithmic information processing structure based on the architecture of our biological nervous system. The structure is composed of a massive number of processing elements operating in a predetermined parallel operation. The processing elements are connected by links with variable weight. The topology of each network determines the way each processor is connected to the other. The link can be excitory, inhibitory, or has no effect on the activity of the processing element. See Figure 4. The primary processing at each element consists of the calculation of weighted sums of the form $f(W_{ij},X_j)$ and weight changes of the form $W_{ij}^{new} = G(W_{ij}^{old},X_i,X_j...)$. the function f is usually a nonlinear function, W_{ij} is the weight of the link from element i to element j and X_i is the input to element i. See Figure 5. Figure 6 shows some of the most popular neural

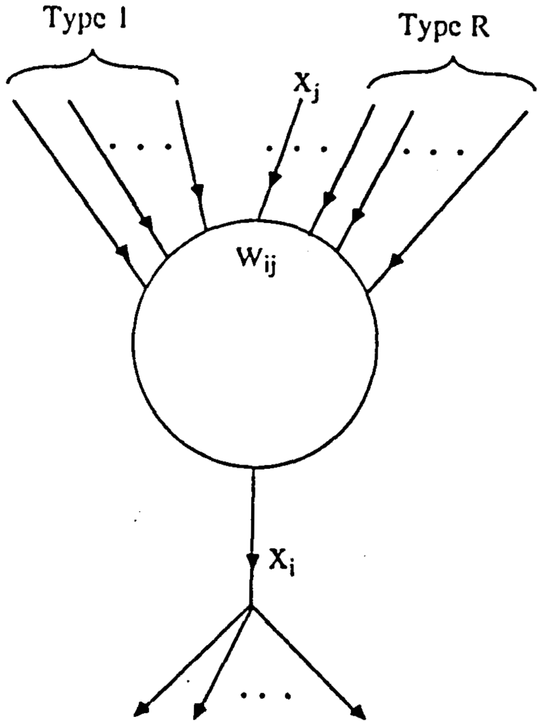

FIGURE 4. A PROCESSING ELEMENT.

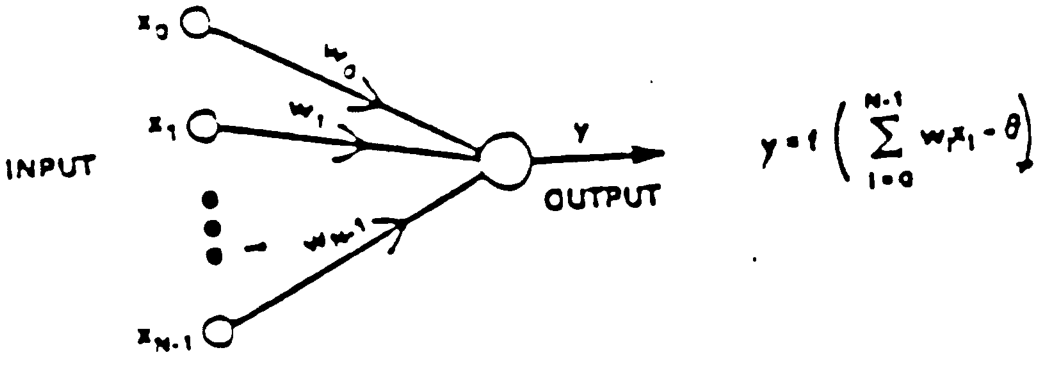

FIGURE 5. OUTPUT, INPUT, AND CONNECTION FOR A PROCESSING ELEMENT

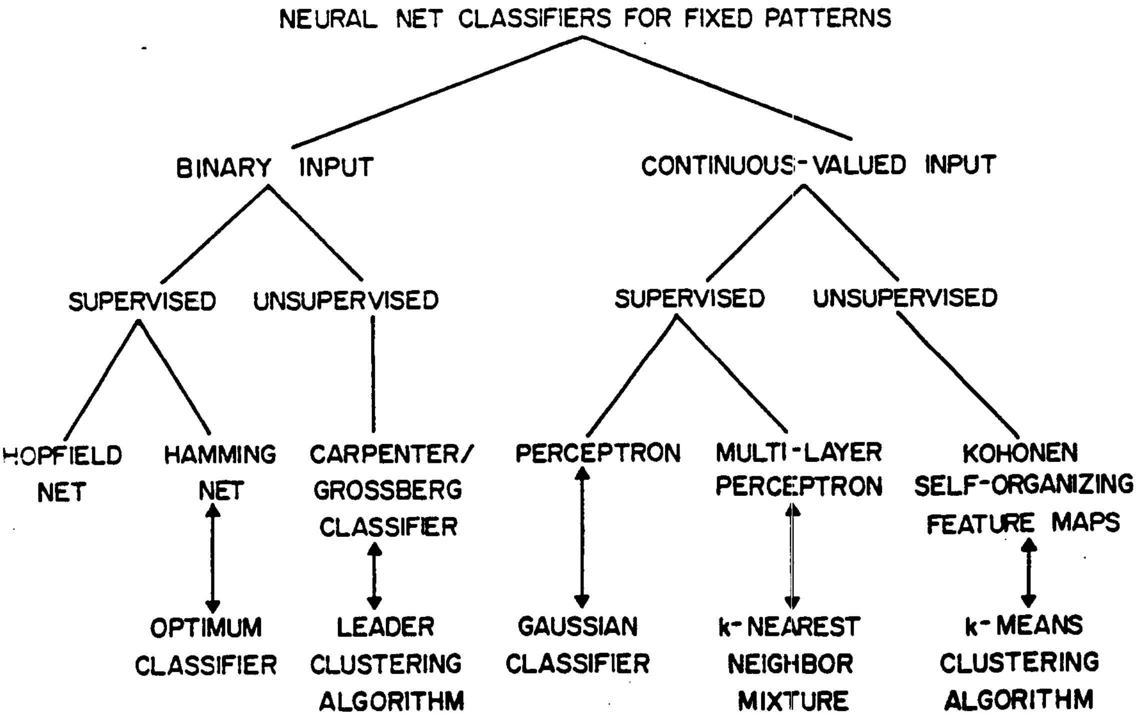
NEURAL NET CLASSIFIERS FOR FIXED PATTERNS
BINARY INPUT
CONTINUOUS-VALUED INPUT
SUPERVISED
UNSUPERVISED
SUPERVISED
UNSUPERVISED
HOPFIELD NET
HAMMING NET
CARPENTER/ GROSSBERG CLASSIFIER
PERCEPTRON
MULTI-LAYER PERCEPTRON
KOHONEN SELF-ORGANIZING FEATURE MAPS
OPTIMUM CLASSIFIER
LEADER CLUSTERING ALGORITHM
GAUSSIAN CLASSIFIER
k-NEAREST NEIGHBOR MIXTURE
k-MEANS CLUSTERING ALGORITHM

FIGURE 6. CLASSIFICATION OF ARTIFICIAL NEURAL NETWORKS [10]

net models that can be used as classifiers with the classical algorithms that are most similar to the neural net models are listed along the bottom[10]. As shown in this figure, these nets are first divided according to whether the input is binary or continuous-value. Second, they are divided according to whether they need supervision during training or not. Since a multi-layer perceptron backpropagation model is implemented in our study, we discuss in the following section the topology and the learning rule of that model.

The Backpropagation Model [10,14,15]

Figure 7 shows the topology of the backpropagation

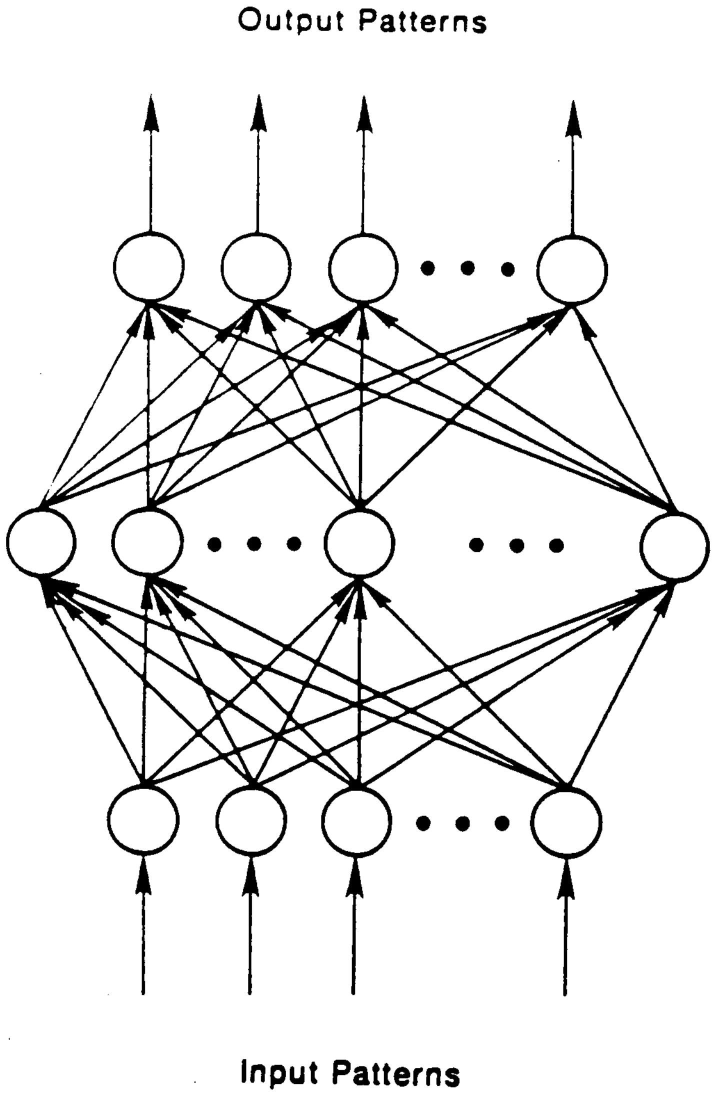

FIGURE 7. BACKPROPAGATION NETWORK.

model. The model consists of 3 layers (input, hidden, and output). each node in the input layer is connected to every node in the hidden layer also each node in the hidden layer is connected to every node in the output layer. The input of the model is a continuous valued vector x_o, $x_1, \ldots x_{N-1}$ representing either the LPC coefficients or the frequencies of the Formants which are calculated from the coefficients. Investigations, mainly experimental, will be carried out to see which input is most suitable for word recognition. The actual output y_o, $y_1, \ldots y_{M-1}$ is calculated as:

$y = f(\sum W_{ij}X_j - \Theta)$ where Θ is a predetermined threshold

and the function f is the sigmoid nonlinearity:

$$f(a) = \frac{1}{1 + e^{-(a-\Theta)}}$$

Training of the network is carried out by setting the output of the model to the desired output vector d_o, $d_1, \ldots d_{M-1}$. All elements of the desired output vector are set to zero except for that corresponding to the current input training word which is set to 1. The weights are adjusted recursively from the output nodes to the hidden nodes by the formula:

$$W_{ij}^{new} = W_{ij}^{old} + u\ \delta_j X'_i$$

where u is the gain factor, δ_j is the error, and X'_i is either the output of node i or is an input. If node j is an output node, then

$$\delta_j = y_j(1-y_j)(d_j-y_j),$$

where d_j is the desired output of node j and y_j is the actual output. If node j is an internal hidden node, then

$$\delta_j = x_j'(1-x_j') \sum_k \delta_k w_{jk},$$

where k is over all nodes in the layers above node j.

REFERENCES

1. S. Saito, and K. Nakato, 1985, Fundamental of Speech Signal Processing, Shuzo Saito and Kazuo Nakato, Academic Press, Inc.

2. F. Fallside, and W. Woods, 1985, Computer Speech Processing, Prentice Hall International (U.K.) Ltd.

3. F. Itakura, 1975, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. Acoust., Speech, Signal, Processing, Vol. ASSP-23, pp. 67-72.

4. L.R. Rabiner, 1978, "On Creating Reference Template for Speaker Independent Recognition of Isolated Word", IEEE Trans. Acoust., Speech, Signal, Processing, Vol. ASSP-26, pp. 34- 42.

5. L.R. Rabiner, and S.E. Levinson, 1981, "Isolated and Connected Word Recognition-Theory and Selected Application", IEEE Trans. on Communication, Vol. comm. 29, No. 5.

6. W.A. Lea, (1979) "Trends in Speech Recognition." Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

7. N.M. Botros, "Digital Signal Processing Algorithms for Automatic Voice Recognition." Final report, NASA/ASEE Summer Faculty Fellowship Program-1987.

8. T. Parsons, "Voice and Speech Processing", McGraw-Hill Book Company, New York.

9. N. Botros, P. Hsu, and K. Xu, "A microprocessor-based system for voice identification", Proceedings of the ISMM International Symposium on Software and Hardware Applications of Microcomputers, pp. 52-56, Fort Collins, Co. 1987.

10. R. Lippmann, " An Introduction to Computing With Neural Nets," IEEE-ASSP Magazine, 4-22, April 1987.

11. J.J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," Proc. Natl. Acad. Sci. USA, Vol. 79, 2554-2558, April 1982.

12. T. Kohonen, K. Masisara and T. Saramaki, "Phonotopic Maps -Insightful Representation of Phonological Features for Speech Representation," Proceedings IEEE 7th Inter. Conf. on Pattern Recognition, Montreal, Canada, 1984.

13. R.F. Lyon and E.P. Loeb, "Isolated Digital Recognition Experiments with a Cochlear Model," in Proceedings International Conference on Acoustics Speech and Signal Processing, ICASSP-87, Dallas, Texas, April 1987.

14. D. E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning International Representations by Error Propagation" in D.E. Rumelhart & J.L. McClelland (Ed.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations. MIT Press (1986).

15. D.E. Rumelhart, and J.L. McClelland, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press (1986).

16. T. Sejnowski and C.R. Rosenberg, "NETtalk: A Parallel Network That Learns to Read Aloud," Johns Hopkins Univ. Technical Report JHU/EECS-86/01, 1986.

17. D.W. Tank and J.J. Hopfield, "Simple 'Neural' Optimization Networks: An A/D Converter, Signal Decision Circuit, and a Linear Programming Circuit," IEEE Trans. Circuits Systems CAS-33, 533-541, 1986.

18. B. Widrow and S.D. Stearns, Adaptive Signal Processing, Prentice-Hall, New Jersey (1985).