"PARCOR" PARAMETERS AS FEATURES APPLIED TO AN ARTIFICIAL NEURAL NETWORK WORD RECOGNIZER

M. Saseetharan and K. E. Forward
Department of Electrical & Electronic Engineering
School of Information Technology & Electrical Engineering
The University of Melbourne

ABSTRACT - "PARCOR" parameters were extracted using linear predictive coding (LPC) of speech data. The fact that parameters extracted from a stable filter have a magnitude of less than unity, was used to confirm the stability of the filter. These parameters were time normalised and used as the input to the three-layer perceptron. Arbitrary non-linear decision surfaces were developed using an error back-propagation algorithm known as generalised delta rule (GDR) on a three-layer artificial neural network (ANN) of simple computing units. As a recognition task, a simulated perceptron of 140 inputs was trained to an accuracy of 0.1 rms with ten repetitions of twenty isolated words. Recognition was tested with sixteen repetitions of the same twenty isolated words spoken by the same person and an accuracy of 87.5% was achieved.

INTRODUCTION

Scientists and engineers have been trying to devise an automatic speech recognizer for sixty years. This will enable friendly interfacing to machines, such as voice controlled computers, type-writers, television sets, digit recognition of cellular phones etc. These machines can find applications in business environments, such as airline reservations and office automation.

In a typical word recognizer, pre-procesing is done to compress the raw speech data while retaining the information so that speech signals can be reconstructed from the extracted features. The existing speech recognition systems often perform the equivalent of a short-time spectral analysis on the signal, as the differences between speech sounds are more clearly and consistently represented in the frequency domain than in the time domain. In the frequency domain, the concentration of energy of different vowels will be different at different frequencies. Similarly, fricatives and nasals of speech data have different energy distribution over the frequency spectrum. In addition, repeated utterances of the same speaker often differ considerably in the time domain, although they are similar in the frequency domain.

Other researchers [1,2,4] have performed speech recognition experiments of isolated words using mel scale coefficients, cepstral coefficients and FFT coefficients to extract the features. The authors chose "PARCOR" parameters computed using LPC modelling of speech data. Experiments on initial speech recognition of isolated words have been carried out [11]. This paper reports further work in this area. In the next section, data compression techniques are reviewed briefly and this is followed by an introduction to the "PARCOR" coefficients. Subsequently, the choice of "PARCOR" parameters is justified. Then the three-layer artificial neural network is briefly described. Finally, the simulation results are detailed and followed by the conclusion.

DATA COMPRESSION

Data compression can be carried out through a number of techniques; for example; (a) a sound spectrograph, (b) filter bank analysis, (c) Fourier transformation analysis, (d) homomorphic analysis, (e) linear predictive coding (LPC) analysis. Since the advent of digital computers for speech processing, Fourier transformation techniques such as discrete Fourier transform (DFT) and fast Fourier transform (FFT) have been used. However, in speech processing, the spectrum produced by such techniques contains the harmonics of the fundamental frequency of the vocal tract which is not needed for speech recognition. Homomorphic analysis contains only the formant structure of the vocal tract transfer function in the frequency domain. This gives a smooth spectrum and also yields a set of coefficients which have more information per parameter than that of FFT (or DFT).

Homomorphic analysis, however, needs the computation of three FFT's (or DFT's) for each frame of signal under analysis. LPC on the other hand is a computationally efficient, robust, reliable technique which provides a good representation of the speech data [6,8]. This technique is used quite widely. The information content per parameter in LPC modelling is as good as homomorphic analysis.

PARTIAL CORRELATION ("PARCOR")

The raw speech data can be characterized as a linear time varying signal which can be represented as the impulse response of a linear time varying filter. If we choose an all pole LPC model for such a filter the impulse response H(z) with p number of poles is given by

$$H(z) = \frac{G}{1 + \sum_{j=1}^{p} a_j z^{-j}} \qquad (1)$$

where G is a gain factor and a_j is the LPC coefficient.

When we have the time domain representation of the signal the problem is to compute the LPC coefficients a_j. The three main formulations, through which the a_j's are computed, are covariance method, autocorrelation method and lattice method. Among the three formulations, the autocorrelation method is computationally efficient and guarantees a stable synthesis filter. In the autocorrelation formulation, the signal outside the frame of analysis is assumed to be zero. In order to reduce the error caused by this assumption, each and every frame of signals under analysis is multiplied by a Hamming window or similar window to smooth the signal to zero or near zero, at the boundaries of the frame of analysis.

Makhoul [6] has shown that a set of autocorrelation functions R(i) of the speech signal can be organized to give a Toeplitz matrix, which is a symmetric matrix in which elements along any diagonal are identical. The Levinson-Durbin recursive procedure uses the redundancy in the Toeplitz matrix to form an efficient solution to the LPC coefficients [6] by minimizing the error as follows

$$E_0 = R(0) \qquad (2)$$

$$a_0 = 0 \qquad (3)$$

$$k_i = - [R(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)] \; / \; E_{i-1} \qquad (4)$$

$$a_i^{(i)} = k_i \qquad (5)$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i \, a_{i-j}^{(i-1)} \qquad 1 <= j <= i-1 \qquad (6)$$

$$E_i = (1 - k_i^2) \, E_{i-1} \qquad (7)$$

where E_i is the minimum squared error and k_i is the reflection coefficient. In addition, from equation (7), $| k_i | <= 1$. This condition on k_i guarantees a stable LPC synthesis filter H(z) [8]. Equations (4) to (7) are solved recursively for i = 1, 2, .., p. The LPC coefficients a_j's are given by

$$a_j = a_j^{(i)} \qquad 1 <= j <= p \qquad (8)$$

The procedure outlined above can be used to calculate a_j. The intermediate parameter k_i can be more easily calculated using another method of Roux and Gueguen[9]. For this reason when the object of the computation is data reduction, the k_i's are often computed since they contain similar

information to a_j's. This can be demonstrated by showing that the spectra derived using a_j and k_i are identical. The partial correlation parameters are simply the negatives of the reflection coefficients and usually the term partial correlation is contracted to "parcor" and hence they are known as the "PARCOR" coefficients.

CHOICE OF THE "PARCOR" PARAMETERS

In the previous section, we saw that the filter corresponding to the autocorrelation formulation of LPC is guaranteed to be stable, however, in reality computation on a finite word length can cause the Toeplitz matrix to become ill-conditioned. This would result in the magnitude of k_j being greater than unity. In order to overcome this problem a fixed point computation of "PARCOR" parameters was implemented through intermediate variables, different from those shown above, after the method of Roux and Gueguen [9].

A number of feature extractors have been used to reduce the number of samples required to represent speech at the input to an ANN used as a recognizer. We have chosen "PARCOR" parameters for this purpose because they are (1) computationally efficient, (2) contain more information per parameter than that of FFT or DFT, and (3) do not need to be amplitude normalised as they range from +1 to -1. They thus have the potential to permit a real-time implementation of a speech recognition system.

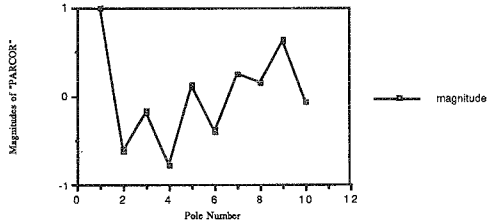

Figure 1. Typical plot of a frame of "PARCOR" parameter with ten poles.

ARTIFICIAL NEURAL NETWORK

Artificial neurons like biological neurons are interconnected networks which are massively parallel. This gives them a degree of fault tolerance and robustness which is quite unlike that of traditional sequential computers of the Von Neumann type. This robustness means that several links can be broken or several weights can be incorrect and the computation will still produce the correct result.

The structure of the formal neuron based on the elementary neural net modelling [4] in the middle 1970s is shown in Figure 2.

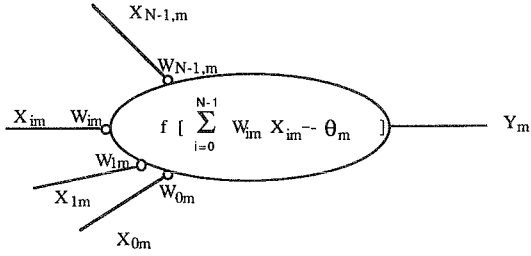

Figure 2. The Formal Neuron.

Neural networks also have the potential to demonstrate a degree of computational robustness when the inputs have a tendency to vary. Thus, speech, which shows variation from speaker to speaker, due to accent, gender, age, emotion, etc., is more likely to be recognized by an ANN than by a parametric technique. Adaptation also gives a degree of robustness by compensating for small variations in characteristics of processing elements.

A three-layer perceptron is sufficient to make arbitrary complex decision surfaces [5], hence it was chosen to be used as a classifier based on the arguments by Lippmann [5]. This choice was further facilitated by the discovery of generalised delta rule [5,10] which enables the learning of a multi-layer perceptron. Lippmann has also developed arguments to optimise the number of neurons to be used in a three-layer perceptron [5].

SIMULATION RESULTS

These experiments use Texas Instruments Sixteen Speaker Isolated Word Database. The available twenty words are ***yes, no, erase, rubout, repeat, go, enter, help, stop, start, one, two, three, four, five, six, seven, eight, nine*** and ***zero*** which are sampled at 12kHz. For every speaker, the Database consists of ten repetitions of twenty words for training and sixteen repetitions of the same twenty words for testing.

LPC analysis was carried out on the segmented speech data frame by frame by shifting the frame along the input at an overlapping of 12.5% between frames. Each frame consisted of Hamming windowed 33.3ms data (400 samples). Windowed samples were processed and ten "PARCOR" coefficients were extracted from the training words. The extracted "PARCOR" coefficients were then time normalised using the following set of equations [7].

$$P_T(m) = (1-s)P_a(n) + sP_a\ (n+1) \qquad (9)$$

where m=1,2...N_T, and

$$n = \text{modulus}[(m-1)\ (N-1)\ /\ (N_T - 1) + 1] \qquad (10)$$

$$s = (m-1)\ (N-1)\ /\ (N_T - 1) + 1 - n \qquad (11)$$

In the equations (9), (10), and (11), N represents the actual number of frame lengths, and N_T represents the modified number of frame length after time normalising. P_a is the actual "PARCOR" coefficient and P_T is the time normalised "PARCOR" coefficient. Time normalisation was carried out to give fourteen frames of "PARCOR" parameters from each word. Therefore, 140 time normalised "PARCOR" parameters for each word formed the input to the three-layer perceptron.

The three-layer perceptron was simulated using a program written in C. The generalised delta rule was applied until the error at the output of the three-layer perceptron corresponding to the the training "PARCOR" parameters was reduced to 0.1 rms value. The learning rate was set to 0.1 and the gain for the momentum term was set to 0.4. Considerable work is being done in choosing the value for learning rate and gain term and also strategies to improve training.

Then, the system was tested with time normalised "PARCOR" parameters, which were extracted from another sixteen repetitions of the same twenty words from the Database, which are assigned for testing. These words are denoted by Wd(i), e.g. Wd(1) = ***yes***, Wd(2) = ***no***, etc. Suppose F[Wd(i), Wd(j)] is used to denote the number of times the Wd(i) is recognized as Wd(j). The ideal case is when F[Wd(i), Wd(j)]=0 for i≠j, and F[Wd(i), Wd(j)]=M for i=j, where M is the total number of testings of recognition. The following table summarises the result of the recognition test with M=16.

F[yes, yes]	= 15;	F[yes, rubout]	= 1;		
F[no, no]	= 13;	F[no, go]	= 2;	F[no, zero]	= 1;
F[erase, erase]	= 16;				
F[rubout, rubout]	= 14;	F[rubout, go]	= 1;	F[rubout, five]	= 1;
F[repeat, repeat]	= 14;	F[repeat, eight]	= 1;	F[repeat, nine]	= 1;

F[go, go]	= 15;	F[go, no]	= 1;		
F[enter, enter]	= 15;	F[enter, rubout]	= 1;		
F[help, help]	= 13;	F[help, two]	= 1;	F[help, nine]	= 2;
F[stop, stop]	= 14;	F[stop, seven]	= 2;		
F[start, start]	= 6;	F[start, stop]	= 9;	F[start, zero]	= 1;
F[one, one]	= 15;	F[one, rubout]	= 1;		
F[two, two]	= 15;	F[two, zero]	= 1;		
F[three, three]	= 15;	F[three, zero]	= 1;		
F[four, four]	= 15;	F[four, one]	= 1;		
F[five, five]	= 13;	F[five, one]	= 3;		
F[six, six]	= 16;				
F[seven, seven]	= 15;	F[seven, stop]	= 1;		
F[eight, eight]	= 14;	F[eight, no]	= 1;	F[eight, three]	= 1;
F[nine, nine]	= 13;	F[nine, rubout]	= 1;	F[nine, no]	= 2;
F[zero, zero]	= 15;	F[zero, no]	= 1;		

Table No 1. Recognition results.

This results in an overall accuracy of 87.5%.

CONCLUSION

In this paper, an experimental evaluation of the artificial neural network approach to isolated word recognition has been described. "PARCOR" parameters are extracted using LPC of speech data. The extracted parameters were time normalised to give fourteen frames and used as the input to a three-layer perceptron. The accuracy of the isolated word recognition has been experimentally determined to be 87.5%.

From Table No 1, it can be seen that some words are incorrectly recognized. This is due to the fact that an all pole LPC representation of speech data does not represent nasals and fricatives accurately compared to pole-zero LPC representation [6]; e.g. F[no, go] = 1 & F[start, stop] = 9. Furthermore, the process of time normalising reduces the number of frames to fourteen by discarding a large number of "PARCOR" parameters, thus introducing serious error, especially with words having large numbers of phonemes; e.g. F[yes, rubout] = 1.

In the proposed method, the extracted "PARCOR" parameters range from -1 to +1 exclusive, provided the LPC filter corresponding to the frame under analysis is stable. Hence, it has the advantage that the stability of the LPC filter corresponding to each frame can be readily confirmed. In addition, the "PARCOR" parameters are suitable as the input to the three-layer perceptron for the same reason. They can be computed effectively following the implementation of Roux and Gueguen [9]. Hence, a real-time implementation of this technique is possible[9]. Research is being continued in this area, particularly in training the three-layer perceptron, time normalisation and pole-zero modelling of the filter, rather than the all-pole modelling represented here.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Dr. A. B. Bradley, Department of Communication and Electrical Engineering, RMIT for the valuable discussions and Mr. J. Tierney, Lincoln Laboratory, MIT, USA, for providing with speech data and for being an adviser for this work during his time at the University of Melbourne. Thanks are also due to our colleague, Mr. V. Pang for permitting us to modify and use his version of the ANN simulator.

REFERENCES

[1] Burr, D. J., "Experiments on neural net recognition of spoken and written text", IEEE Transactions on Acoustics, Speech and Signal Processing, pp 1162-1168, July 1988.

[2] Kammerer, B., and Kupper, W., "Experiments for isolated-word recognition with single and multi-layer perceptrons", Neural Networks 1 (Supplement 1), pp 302, 1988. Abstracts of first annual meeting, Boston.

[3] Kohonen, T., "State of art in Computing", ICCN 1987, San Diego, Cal., June 21-24, 1987.
[4] Kohonen, T., "Self-Organization and Associative Memory", Springer-Verlag, Berlin, 1984.
[5] Lippmann, Richard P., "An Introduction to computing with Neural Nets", IEEE ASSP magazine, vol 3, NO 4, pp 4-22.
[6] Makhoul, John, "Linear Prediction: A Tutorial Review", Proc. of the IEEE, vol.63, No.4, April 1975.
[7] Myers, C., Rabiner, L., R., Rosenberg, A., E., "Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-28, No.6, Dec 1980.
[8] O'Shaughnessy, Douglas, Speech Communication, Addison-Wesley Publishing company.
[9] Roux Le, J., Gueguen, C., "A Fixed Point Computation of Partial Correlation Coefficients", IEEE Transactions on Acoustics, Speech and Signal processing, June 1977, pp 257-259.
[10] Rumelhart, D. E., Hinton, G. E., and Williams, R. J., Parallel Distributed Processing, Vol.1,.MIT Press, 1986, pp 318-362.
[11] Saseetharan, M., "Isolated word recognition using artificial neural networks", (ACNN'90) Proceedings of the First Australian conference on neural networks,pp 88, January 1990.