S10.17

A Connectionist Approach To Speech Recognition Using Peripheral Auditory Modelling.

Mark Terry*, Stephen Renals, Richard Rohwer, Jonathan Harrington.

Center for Speech Technology Research,
University of Edinburgh,
Edinburgh EH1 1HN, UK.

* currently with US West Advanced Technologies,
6200 S. Quebec St., Englewood, Colorado 80111.

ABSTRACT

In this study, a prototype isolated word recogniser was constructed, with an auditory-based analysis component and a pattern classification module based on a parallel distributed processing paradigm [1]. The auditory model [2, 3] used was a band-pass non-linear (BPNL) configuration which incorporates the effects of lateral suppression. Pattern classification was performed by a layered, feed-forward neural network [4], consisting of an array of input nodes representing the binary features output by the auditory model, a set of hidden nodes and an array of output nodes representing the word to be recognised. A suitable internal representation was learned by the method of back-propagation of errors by gradient descent, using the generalised delta rule. This prototype recogniser was trained to recognise English digits spoken by male and female speakers. Recognition rates for the digit set, (zero to ten) were better than 80%.

INTRODUCTION

Automatic speech recognition systems require an initial analysis which adequately represents perceptually important features. A decision-making component should be able to learn on available training data, and to perform a fast efficient evaluation of the incoming unknown utterance, coping with inter-speaker and intra-speaker variability. A combination of interest is to use a model of the peripheral auditory system as an analysis module together with a neural net decision making component.

Recent work on auditory modeling has indicated that the auditory representation of the speech signal is perceptually more relevant (2,5) reflects better the phonetic events in speech (6,7) is more robust to inter-speaker variability (8) and is more robust to noise (9) in comparison with more traditional spectral analysis techniques, such as linear prediction and the discrete Fourier transform. Although modeling the higher levels of the auditory system is at present unrealistic it would seem that a connectionist paradigm offers more plausibility than rule based artificial intelligence decision making techniques.

Our paper describes the ASR system which used an auditory model as its front end and a neural net based decision component. This recogniser is applied to the speaker independent isolated digit automatic speech recognition.

CH2561-9/88/0000-0699 $1.00 © 1988 IEEE

AUDITORY MODEL

The auditory model used in this study was a band-pass nonlinear model (BPNL), consisting of 64 channels regularly spaced on a Bark scale from 1 to 20 Bark. Each channel was comprised of a compressive nonlinearity positioned between an initial band-pass triangular shaped filter with asymmetric slopes (90 dB/oct and -200 dB/oct) and a more sharply tuned second band-pass filter (200 dB/oct and -200 dB/oct). Such a model is known to exhibit some of the effects of lateral suppression [10], which may serve to emphasize the formant structure of voiced speech. Each filter in the channel was constructed as 256 tap FIR digital filter. The output from the second filter was then subjected to a temporal analysis to obtain an interval histogram of the periodicities present in the speech signal [11]. The zero crossing were estimated via linear interpolation of successive positive and negative speech samples from the the output of each channel. The interval histogram was computed successively over a 5 millisecond time window. The histogram was then frequency compensated (-6 dB/oct) to account for increasing number of zero crossings with frequency in the fixed 5 ms interval. Typical results of this processing are displayed as an auditory spectrogram(see Figure 1).

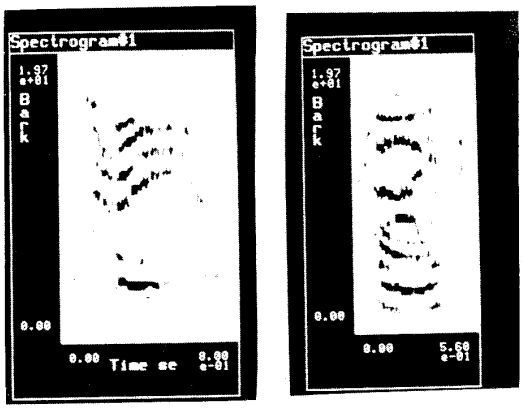

Figure 1. Auditory spectrograms for the digits three (male speaker), and the digit five (female speaker). Note that the lower harmonics and upper formants are resolved.

PATTERN CLASSIFICATION

Pattern classification was performed by a layered, feed-forward connectionist system[1,4]. This system consisted of a network of processing units (or nodes) linked together by weighted connections. After training the knowledge of the system is contained in the weighted connections. There were three sorts of nodes: *input, output* and *hidden.* The input layer of the network consisted of the array of nodes. As described later, the number of nodes depended on the representation of signal presented to the network. The array of the output nodes represented the word to be recognised. The number of output nodes corresponded to the number of words under investigation, with recognition being indicated by the activation of the relevant node and the deactivation of the remaining output nodes. The remaining hidden units contained the system's internal representation of the problem. This was a feed-forward network; that is connections were forbidden between units in the same layer or from a higher layer to a lower layer; all other connections were allowed. A suitable internal representation to carry out the desired classification was learned using the method of back-propagation of errors by gradient descent, using the generalised delta rule [4].

EXPERIMENTS

The speech material consisted of the digit set, zero to ten, spoken by two male and two female native British talkers. The speakers had varying accents(southern English, northern English and southern Scottish). The speech was sampled at 16 kHz by a 12 bit A/D converter. The speech was recorded by a condenser microphone in a computer terminal room. Five tokens of each digit were obtained. Endpoints were determined manually. About 20 ms of background noise was intentionally left on each side of the utterance.

Each word token was processed by the auditory model and the results stored on file for future processing. In early experiments attempts were made to present a small matrix of features as the input layer to the neural network. The features used were crude voicing and frication measures based on energy content in low and high frequency channels together with estimates of the first three formants abstracted from the auditory spectrograph. In these experiments we typically used about 50 input nodes in the neural net. Although the network learned to classify subsets of the digits {one,three,and seven}, {two,four, and six} in these tests, there was difficulty in both generalisation to unseen tokens and to extending the number of digits in the training set. We concluded that the feature extraction stage was unreliable and that possibly the features were inappropriate for the task. We then decided to allow the network access to the basic output of the auditory model. In this case the simplest representation consisted of an auditory spectrogram, which was quantised into 20 sections along the Bark axis and into 10 sections along the time axis, giving a total input array of 200 nodes. Various networks were tried: with between 3 and 10 hidden units, and both strictly layered networks, and networks with direct input-output connections. We found that strictly layered networks learned at a slower rate and were more likely to become stuck in local minima than those networks with layer-skipping connections. After some experimentation, a hidden layer consisting of 10 hidden units was chosen for subsequent recognition experiments. The configuration of the network is shown in Figure 2.

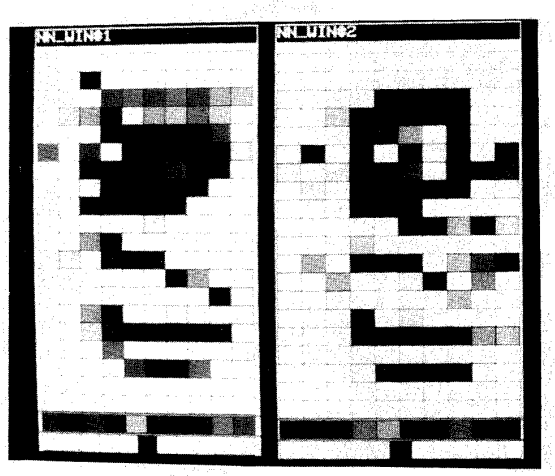

Figure 2. States of the neural network when presented with a training token (left picture) and test token (right picture). The tokens were the digit five (female speaker). The input layer consists of the auditory spectrogram quantised into 20 intervals on the Bark Scale and ten time slices across the whole word. The state of the hidden layer of ten units is displayed in the second row from the bottom. The state of the output layer of eleven units, each of which represents one of the digits, is shown in the bottom row.

Various sets of isolated digits were presented to the network. The learning was generally encouraging; in the easily confusable set {eight, five, nine}, learning was complete after 1500 sweeps on a 48 pattern input (4 tokens of each digit by each speaker). A further 12 tokens ,

previously unseen by the network, were then presented and 11 out 12 were correctly recognised(i.e. the "correct" output node was at a greater value the 0.9 and the other output nodes were at values less than 0.1). Initial results for discrimination between the digit set (zero to ten) show that the classification problem using the for training tokens was learnt after 8000 training sweeps. Testing the network on the remaining tokens gave a recognition rate of 80%. The greatest confusion was between the digits one and four. However there was no significant difference in discriminating between male and female utterances of the same digit.

DISCUSSION AND CONCLUSION

The present auditory model is only one of many possible. The optimality of its parameters remains to be determined. The effectiveness of its particular processing stages(e.g. lateral suppression, periodicity analysis) is matter of current interest. We also plan to study effect of adaptation to determine whether this would enhance the representation of transients and rapid formant changes in the speech signal[2]. Presentation of the high bit-rate speech signal requires heavy processing by neural networks. It is our current belief that presenting the network with perceptually relevant information may reduce the amount of computation spent in training of the neural network. The current system does appear to have some potential for isolated word recognition. Further improvements of performance may come from improvements in the auditory model, better parameterisation of the auditory model output and improvements in the network architecture. Some efforts in this direction are presented at this conference [12].

Note on Implementation:

This recognition system was implemented on a vector-accelerated MassComp 5700 within the AUDLAB interactive speech processing system [13].

REFERENCES:

[1] D.E. Rumelhart and J.L. McClelland (eds.), "Parallel Distributed Processing: Explorations in the microstructure of cognition. Vol. 1: Foundations", (MIT, Ca.spidge MA, 1986).

[2] M.P. Cooke, "A computer model of peripheral auditory processing incorporating phase-locking, suppression and adaptation effects", Speech Communication, 5(3) (1986).

[3] A.M.P. Terry (1981), "Suppression effects in hearing", PhD Thesis, (University of Reading).

[4] D.E. Rumelhart, G.E. Hinton and R.J. Williams (1986), *"Learning representations by back-propagating errors"*, Nature, 323 (1986) pp. 533-536.

[5] B. Moore and B. Glasberg (1987), "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns", Hearing Research, 28, pp. 209-225.

[6] S. Seneff (1986), "A computational model for the peripheral auditory system: application to speech recognition research", Proc. ICASSP-86, pp. 1983-1986.

[7] M. Hunt (1986), "Speech recognition using a Cochlear Model", Proc. ICASSP-86, pp. 1979-1982.

[8] H. Hermansky (1987), "An efficient speaker-independent automatic speech recognition based on a simple simulation of human auditory processing." Proc. ICASSP-87, pp. 1159-1162.

[9] O. Ghitza (1986), "Auditory nerve representation as a front-end for speech recognition in a noisy enviroment", Computer Speech and Language, 1, pp. 109-130.

[10] R.R. Pfieffer, "A model for two-tone inhibition os single cochlear nerve fibers", J.Acoust.Soc.Am. Vol 48(6), (1970) pp. 1373-1378.

[11] R. Carlson and B. Granstrom (1982), "Towards an auditory spectrograph", in *The representation of speech in the peripheral auditory system*, Carlson and Granstrom Eds, Elsevier, Amsterdam, 1982, pp. 109 -114.

[12] R.Rohwer, S. Renals, and M. Terry (1988), Unstable Connectionist Networks in Speech Recognition, This Conference.

[13] M. Terry, S. Hiller, J. Laver and G. Duncan (1986), "The AUDLAB interactive speech analysis system", IEE Conf. publication 258, pp. 263-265.