IEE Symposium, London, 1988

ACOUSTIC-PHONETIC ATTRIBUTE DETERMINATION USING MULTI-LAYER PERCEPTRONS

I.S. Howard and M.A. Huckvale*

Introduction

Speech is a signal that exhibits considerable variability and possesses a complex underlying structure. In order to determine phonetic correlates of the speech signal it is important to take this structure into account. However to specify the structure in advance of any analysis at the necessary level of detail would be a very difficult task. One would have to ensure that any fixed structure imposed was appropriate. Computational techniques from the parallel distributed processing paradigm appear useful for this type of problem because they have potential for acquiring internal structure automatically by means of a training procedure.

The multi-layer perceptron[1] provides a means of performing complex pattern recognition and feature detection tasks. It is capable of implementing non-linear transformations which may be found by means of an iterative training procedure, known at the generalized delta rule. Such a technique is therefore of particular interest in determining elementary phonetic attributes of the speech signal, as first results suggest[2,3].

This paper describes the application of an MLP network to the feature-labelling of simple speech material. The output of the network is a feature vector for each input time-window. The network can be viewed as either a pre-processing stage for a phonetic recognition system, or simply as a non-linear data reduction of the signal for input to pattern-matching recognisers.

A difficulty for such a system is the determination of the appropriate time-window over which to "view" the signal. The nature of speech is such that there exist constraints in the signal that extend over different time-windows. The size of the window depends on the "level" of abstraction of the aspect of the speech. For example, the effects of articulator dynamics result in constraints over shorter windows than morphology or syntax.

For the feature labelling task, the use of a wide time window will help in establishing a "context" for the frame under scrutiny. However, the larger the window, the more different input vectors are possible, and hence the greater the size of the training set required. With a limited training set it becomes difficult to determine the decision boundary. We believe the solution to this problem is to build a classifier in a number of layers, with each layer having access to a larger time-window. Thus the current work uses a small window on the signal (30ms) but is eventually to be part of a classification system that operates at a higher level of abstraction.

Feature detector construction

Fig 1 shows the architecture of the feature-detection system used in the experiments. The input speech signal is encoded into 19-channel log energies using a vocoder[4]. The input window of the detector is over 3 10ms frames of the vocoder output, with the top 50dB of the energies mapped to the scale 0 to 1. The output of the detector consists of six feature values: silence, frication, vocalic, nasality, front-back vowel quality and open-close vowel quality. The first four features are binary, the last two are multi-valued.

The transformations that may be performed by a multi-layer perceptron are determined by its hidden units. With no hidden units, it performs as a linear classifier. With one layer of hidden units, the decision boundary can be curved or a closed linear region. With two layers of hidden units, the decision boundary may be of arbitrary complexity. In the feature detection system we have chosen a single layer of 12 hidden units. It is hoped this number provides a "bottleneck" in the communication between the inputs and the outputs, forcing generalisations of the input data without requiring large numbers of training cycles during the learning phase.

Speech corpus and method

The speech data used in these experiments consisted of 20 repetitions of the digits spoken by a single male speaker in an anechoic room. The data was low-pass filtered and digitised at 10kHz and passed through a 19-

*Department of Phonetics and Linguistics, University College London, London WC1E 6BT.

channel vocoder. One instance of each digit was hand-annotated, and repetitions of each digit were annotated automatically from the reference by a dynamic-programming procedure[5]. A simple look-up table was used to transform the phonemic annotations to a frame feature matrix for training. Ten repetitions of the digits were used for training and ten for testing. The MLP training algorithms were run for 20 passes. The algorithms were all written in C and ran under Unix on a Masscomp MC5500 series computer.

Evaluation of Results

The results produced are in the form of the feature waveforms and also, in the case of binary features, in the form of equal hit/false alarm rates. The latter measure is derived from the *receiver operating characteristic* of a particular feature detector. This involves choosing a threshold such that the number of correct classifications (hits) equals the number of false classifications (false alarms).

Experiment 1

In the first experiment, the feature detector was trained directly on the anechoic speech to establish a baseline performance. Fig 2 shows typical feature detector waveforms for the digit "5" in the test set. Notice that the silence, frication and voicing detectors operate as expected, and that the diphthong is tracked by the vowel-quality detectors from an open half-front position to a close front position. Fig 3 shows the digit "1". Notice the activation of the nasal detector, and the differential transition rates of the two vowel quality detectors during /w/. Good results were obtained for all the digits with the exception of the voiced fricative in seven which was almost entirely ignored by the frication detector. The performance of the binary feature detectors on the anechoic data was as follows:

Feature	Hit-rate (at equal error)
SILENCE	93.6%
FRICATION	86.4%
VOCALIC	95.5%
NASALITY	92.3%

The performance of the frication detector is low due to the mis-labelling of "7". Errors are almost entirely at the borders of detected regions of the signal.

Experiment 2

In order to get some idea of the performance of the features in the presence of background noise, the feature detectors were then run (for both training and test modes) on speech contaminated to 3 dB SNR with uniform density white noise. The SNR was defined in terms of the windows of the signals, of length 500mS, that contained the maximum power.

Feature	Hit-rate (at equal error)
SILENCE	70.9%
FRICATION	61.8%
VOCALIC	89.1%
NASALITY	68.9%

It can be seen that the delection of silence and frication have been affected the most by the white noise. This is to be expected, since it is difficult to distinguish between the background noise and the presence of frication. The detection of vocalic features was the least affected.

Experiment 3

In order to give some indication of performance in more natural noisy conditions, the feature detectors were run on the speech was contaminated to 0 dB SNR with "canteen" noise (environmental noise from the college refectory at lunch-time). The SNR was defined as before.

Feature	Hit-rate (at equal error)
SILENCE	81.4%
FRICATION	78.8%
VOCALIC	87.3%
NASALITY	69.8%

The detection of all features are again somewhat worse than in experiment 1, but better than in experiment 2. The exception is the vocalic feature, which gave poorer performance than in experiment 2. The most

likely explanation for this is that because the canteen noise contains the background conversations of other speakers, it interferes more with the vocalic aspect of the speech than the white noise.

Conclusions

The results show that it is possible to train a MLP to extract sensible acoustic-phonetic features. It is felt that these features may be more appropriate than grosser measures of signal quality for applications such as speech recognition. One possible reason for their usefulness is that they may give a lower data-rate representation without discarding useful information. The absolute performance of the network above should only be gauged as part of a recognition system. This work would benefit from the use of more training data, particularly from more than one speaker. It would also be worthwhile to investigate what micro-features of the speech data the hidden units were detecting.

Acknowledgements

This work was supported by MRC studentship RS-85-2 and by an SERC fellowship.

References

1 Rumelhart, D.E., Hinton, G.E., Williams, R.J., "Learning Internal Representations by Error Propagation", in Rumelhart & McClelland (eds), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press, 1986, Vol 1, Ch 8.

2 Howard, I.S., Huckvale, M., "The application of adaptive constraint satisfaction networks to acoustic phonetic attribute determination", Proc. European Conference on Speech Technology, Edinburgh, 1987.

3 Peeling, S.M., Bridle, J.S., "Experiments with a learning network for a simple phonetic recognition task", Proc. I.O.A. Conference, Windemere, 1986.

4 Rabiner,L.,R., Schafer,R.W., Digital processing of speech signals, Prentice-Hall, (1978).

5. Chamberlain, R.M., Bridle, J.S., "ZIP: a dynamic programming algorithm for time-aligning two indefinitely-long utterances", IEEE ICASSP 1983, 816.

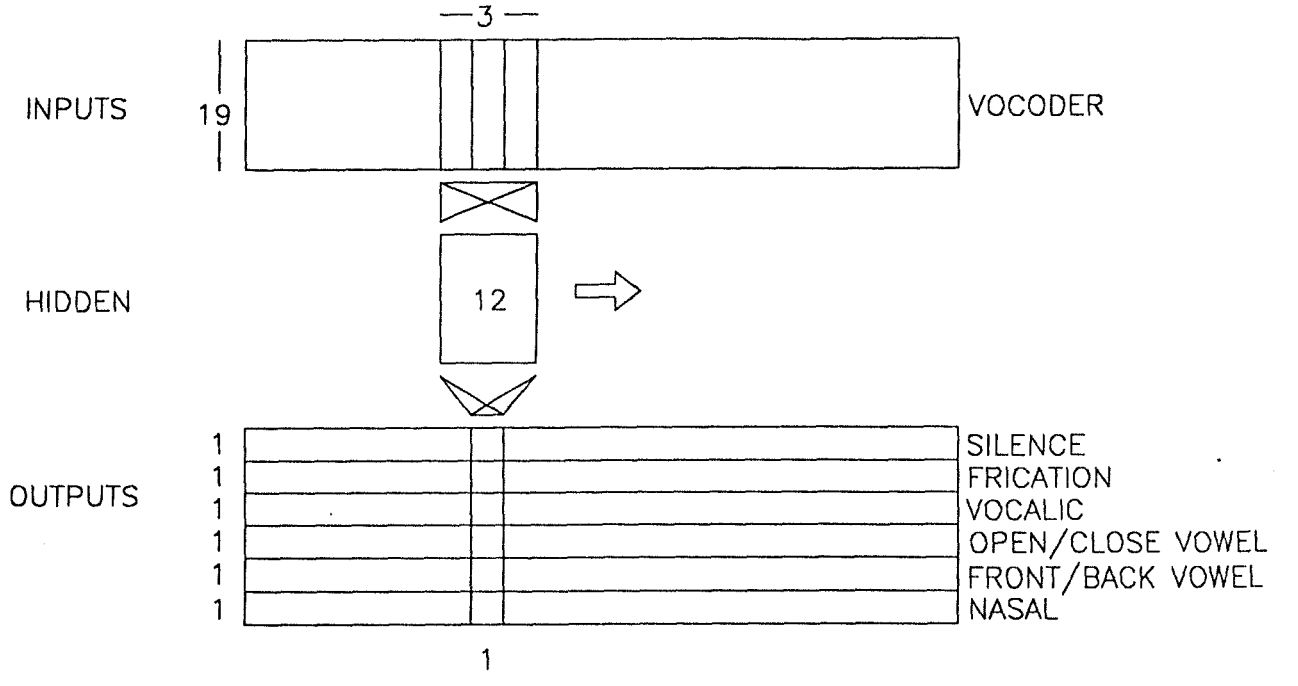

Figure 1. Architecture of multi-layer perceptron feature detectors.

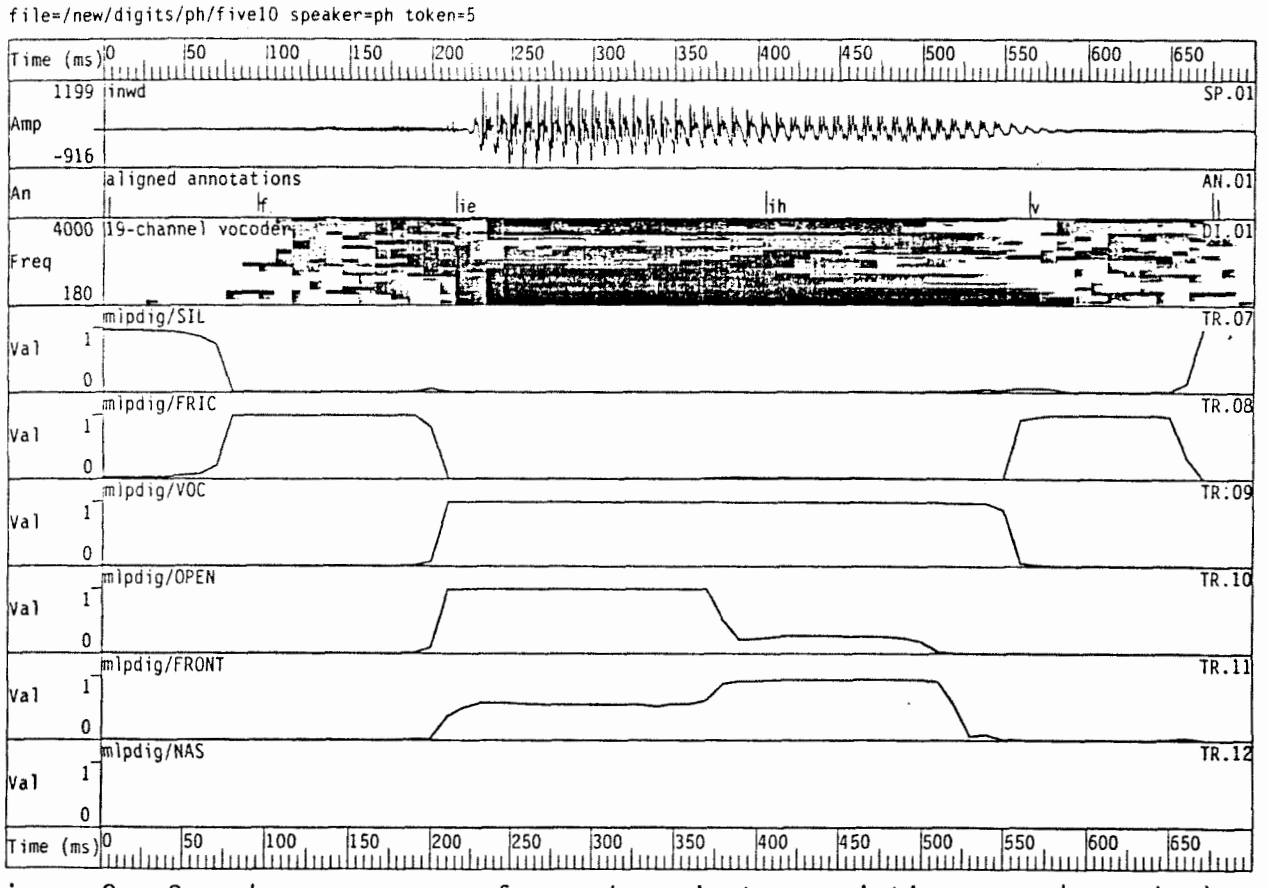

Figure 2. Speech pressure waveform, phonemic transcription, vocoder output and feature waveforms for the digit "five".

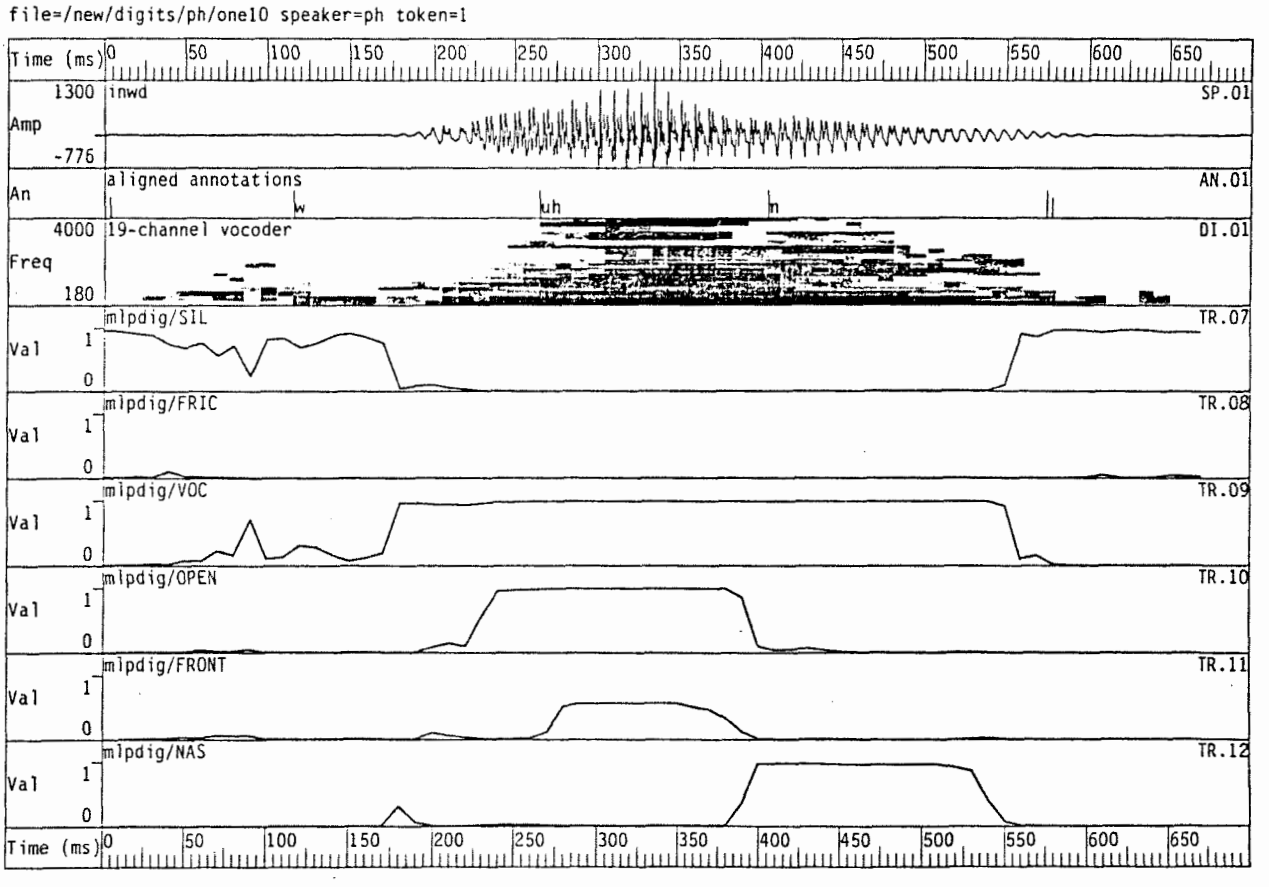

Figure 3. Speech pressure waveform, phonemic transcription, vocoder output and feature waveforms for the digit "one".