ANALYSIS OF A NEURAL NETWORK MODEL
FOR SPEECH RECOGNITION

Steve Renals and Jonathan Dalby

ABSTRACT
An analysis of a feed-forward neural network model performing a vowel recognition task is reported. Two experiments analysed in detail involved training on speech processed using linear prediction and the discrete FFT. A phonetic analysis of the resultant weight matrices is offered; additionally we analyse the network structure using pairwise correlations and principal components analysis. We suggest methods of network initialisation obtained by inverting these analyses.

INTRODUCTION
Neural network models have been used in several speech recognition tasks in the past few years [1]. Perhaps the most popular algorithm has been the back-propagation of error algorithm [2] applied to layered, feed-forward networks. However most research has focussed on performance issues, with little work being directed toward questions such as what high order statistics are being computed by the network or what features are being extracted by the hidden units in a feed-forward network. Here the computations being performed by a back-propagation network are investigated, and the internal representations learned by the network are analysed phonetically.

The problem under investigation was that of vowel labelling. A back-propagation network was trained to label vowel tokens that had been hand-segmented from continuous speech. The speech data was analysed using two different methods: (a) linear predictive analysis from which cepstral coefficients were derived; (b) discrete FFT analysis, followed by a generalised centroid formant tracker producing frequency, bandwidth and amplitude information.

After training the networks, the resultant weight matrices were analysed, as was the networks internal response to both training and test data. The features derived from the network weight matrix are compared to similar features derived by performing a principal components analysis (PCA) of the training data (a linear process). Additionally we suggest a method of initialising networks prior to training using the method of PCA applied to the training data.

VOWEL LABELLING EXPERIMENTS

Speech Data
Phonetic labelling experiments were performed on hand-segmented vowel tokens taken from a phonemically dense speech database of 98 sentences [3] uttered twice each by two male speakers, giving a total of 392 sentences. Each set of 98 sentences contained approximately 750 vowel tokens of 20 classes (12 monophthongs and 8 diphthongs). One set of vowel tokens for each speaker was designated a training set, the other a test set. The input speech signals were sampled at 16 kHz, before one of two methods of analysis was employed: (a) a 20th order linear predictive analysis from which 20 cepstral coefficients were obtained; (b) a Fourier transform analysis to produce a spectrogram which was then tracked for three formants using a generalised centroid formant tracker, producing frequency, bandwidth and amplitude information for each formant.

For both methods of analysis a vowel token was regarded as being split into three parts of equal duration. For the cepstral analysis, each token was represented by a feature vector consisting of the median cepstral coefficients for each third of the token (60 coefficients in all) plus 12 coefficients representing a coarse coding of the duration in milliseconds (see below). For the formant analysis, 15 real valued coefficients were extracted: median formant frequencies (Bark-scaled) for each formant in each third of the token (9 in all), median bandwidths (Hz) for each formant in the central third of the token (3 parameters), two amplitude ratios from the central third of the token and the duration of the token (ms). This 15 element vector was then coarse coded over 218 input units.

Coarse Coding
A Gaussian coarse coding technique was used, enabling an arbitrary real number to be distributed over a group of input units.

To determine the clamped input values, $y_i$ when coarse coding a real value $X$ onto $N$ input units use the distribution:

$$y_i = \exp \left( -\frac{(X-m_i)^2}{2\sigma} \right) \quad 0 \leq i < N$$
was used, where $m_i$ is the pre-determined reference value for that unit and $\sigma$ is a variance.

Formant frequency values were distributed over 16 units:

$m_0 = 1$ Bark, $m_1 = 2$ Bark, ..., $m_{15} = 16$ Bark, $\sigma = 1$.

Bandwidth values were distributed over 10 units:

$m_0 = 0$ Hz, $m_1 = 50$ Hz, ..., $m_9 = 450$ Hz, $\sigma = 2000$.

Amplitude ratios $(A(F2)/A(F1)$ and $(A(F3)/A(F2))$ were distributed over 13 units:

$m_0 = 0$, $m_1 = 80$, ..., $m_{12} = 960$ Bark, $\sigma = 6000$.

Durations were distributed over 18 units:

$m_0 = 0$ ms, $m_1 = 16$ ms, ..., $m_{17} = 272$ ms, $\sigma = 250$.

In the cepstrally-based feature vector, the duration was coded over 12 units. All values that were coarse coded over $n$ units were bounded to lie between $m_0$ and $m_n$.

Experiments

Several experiments were performed using this input data. We trained back-propagation networks on both types of feature vector, using the back-propagation algorithm on a two layer (one hidden layer) network. The vowels were divided into 20 classes (12 monophthongs and 8 diphthongs) coded on the output layer using a "1 out of $n$" output coding. The number of hidden units was varied (from 2 to 52) and generalisation statistics collected. For both types of feature vector it was found that the generalisation performance increased markedly as hidden units were added until there were 18 - 20 hidden units, beyond which addition of hidden units gave only marginal performance increases. Additionally it was found that the cepstrally-based inputs produced slightly higher recognition rates compared with the formant-based inputs. These results are reported in detail elsewhere [4].

NETWORK ANALYSIS

The hidden layer(s) of a back-propagation network may be regarded as extracting higher-level features from the input data. Two networks were analysed: both featured a single hidden layer with 24 hidden units. One was trained with the cepstral data, the other with the formant data.

Formant Input

The weights into a given hidden unit may be displayed graphically (fig 1). To aid interpretation the weights have been divided into portions corresponding to the input coarse codings. A more compact representation, displaying the frequency components of a set of input to hidden weights is shown in figs 2a-f. Since the formant frequency values are all coarse coded on the same scale, the 3 formant components may be summed within each time frame, and a "spectrogram" representation may be employed. Figure 3 illustrates what classes each hidden unit responds to when presented with the training set. This explicitly demonstrates which feature detectors have been developed in the network.

The spectrograms of the hidden unit weights were examined by a phonetician trained in interpreting speech spectrograms with a view to arriving at an informal estimate of the extent to which they are interpretable in a similar manner. If only the middle portions of these plots are taken into account (there was no control in the training for place of articulation of the adjacent segments) some appear to bear a clear resemblance to canonical vowel target frequency schema. Node 0 for example shows two clear and well separated regions of activation with strong inhibition between them. These activation regions correspond to a relatively low F1 and a relatively high F2 (fig 1, fig 2a) typical of high front vowels such as /i:/, /i/ or /ei/. Figure 3 shows that this node is indeed responding most strongly to high front vowels and to diphthongs with a high front component. Node 4 (fig 2b) shows an activation pattern similar to that of node 0, though the F2 activation area is not quite so high. The responses for this node (fig 3) are not so clearly interpretable, however. Perhaps the most that can be said is that it disfavours the low back rounded vowels /oo/ and /o/ (and the centralising diphthongs /e@/ and /u@/).

The spectrogram of the weights for node 8 (fig 2c) shows a pattern that is typical of low back vowels which have low F2 and high F1 values, with the two formant edges meeting at about 1000 Hz for male talkers of average vocal tract lengths. The outputs for this node (fig 3) show that it tends to respond more strongly to vowels of this class and to diphthongs which have a low back component than to others, though it seems to 'like' the low front vowel /a/ as much or more than the low back round /oo/ and /o/ and it is also responding to the non high front centralising diphthong /e@/. The middle portion of the spectrogram of node 3 (fig 2d) also resembles the typical F1/F2 configuration for low and back vowels though its activations are not as compact as those of node 8. The responses (fig 3) show that it is selecting the low back vowels /oo/ and /o/ strongly but it is also responding strongly to the high back centralising diphthong /u@/ and, anomalously, to /i:/, /i/ and /ei/.

Node 14 (fig 2e) is an example of a node which is difficult to interpret. At the mid point the (rather weak) activation appears to be with vowels with low F2 and low F1, characteristic of the high back /u@/ and /u/, though the response (fig 3) for this node is strongest for /a@/ and /e@/, the /u@/, /u/ response being about the same as for high front vowels. Node 23 (fig 2f) is of interest because while the spectrogram shows a strong bias for a rather low F2 value, there is almost no biasing in either direction in the low frequency area. Interpreting this as a vowel formant pattern is impossible. The output for this node (fig 3) shows quite clearly that it biased against high front vowels /i:/, /i/ and /e/ and diphthongs with high front components /e@/ and /u@/.

In summary, we can say that while there appears to be a tendency for the nodes to train to select (or deselect) for phonetically natural classes, the interpretation of them as such is far from unambiguous. Probably, much more training data should be seen by the network as the populations for
some of these vowels are quite sparse.

Initialisation

Initialisation of feed-forward networks is a problematic area. Three principal choices must be made: the number of hidden units, the number of connections and free weights and the initialisation of weights. Here we shall assume that the networks being studied are fully connected within a feed-forward scheme.

The variety of features being extracted by hidden units may be quantified by computing pairwise correlations between pairs of weight vectors leading into hidden units. The mean and maximum correlations are graphed for the cepstral input network with respect to the number of hidden units in figure 6. It is apparent that as the number of hidden units increases, so does the maximum correlation: i.e. one unit comes close to duplicating another. The mean correlation curve rises much less steeply, implying that most pairs of hidden units have low correlation coefficients. Computation of correlation coefficients may be a way to carry out a principled pruning of hidden units from a network during training.

Principal components analysis (PCA) may also prove useful in determining the number of hidden units to use in an network. Performing a PCA on the input data produces an eigenbasis and a set of corresponding eigenvalues. The magnitude of each eigenvalue corresponds to the proportion of variance in the input accounted for by that eigenvector. Hence, an examination of the eigenvalues of a PCA might give an indication of how many hidden units are required. It must be emphasised that PCA is a linear analysis and is equivalent to performing an identity mapping on the input data using a single hidden layer network [5]: here, the network has been performing a (non-linear) discriminant analysis [6] which is a different operation to PCA. However, if pairwise correlations of the principal components are taken with the input-to-hidden weight vectors in the 24 hidden node cepstral network, then it is found that the first 6 or 7 principal components are correlated with at least one hidden unit with correlation coefficients at least one standard deviation above the mean. This suggests that the basis vectors obtained from a PCA might be a good way to initialise the input weight vectors to a set of hidden units.

CONCLUSIONS

We have demonstrated that, to an extent, the internal representations produced by a feed-forward network correspond to phonetic categories and that the weight matrices may be interpreted in an acoustic phonetic manner. We have quantified the diversity of the hidden units by performing pairwise correlations on the set of weight vectors leading into hidden units and additionally, have correlated these weights with the principal components of the input data. This work also suggests methods of initialising the structure of a feed-forward network and the values of the weights by using conventional techniques such as PCA or closest-means. Additionally, it also suggests that speech-specific knowledge may also be used to initialise the weight matrix by inverting the process of weight matrix analysis.

ACKNOWLEDGEMENTS

We wish to thank Richard Rohwer and Andrew Sutherland for valuable discussions. One of the authors (SR) was supported by a UK Science and Engineering Research Council studentship. This work was supported by the Information Engineering Directorate/Science and Engineering Research Council as part of the IED/SERC Large Scale Integrated Speech Technology Demonstrator Project (SERC grant D/29640, D/29611, P/10309, P/10316) in collaboration with Marconi Speech and Information Systems and Loughborough University of Technology.

REFERENCES

Figure 1: Display of weight vector leading into hidden unit 0. The weights have been divided into groups corresponding to the input coarse coding.

Figures 2a-6: "Spectrograms" formant frequency components of weights leading into 6 hidden units.

Figure 3: Means (filled circles) and standard deviations (outer circles) over input patterns of hidden unit response by vowel class.

Figure 4: Plot of maximum correlation (upper line) and rms correlation (lower line) between input to hidden weight vectors for networks with varying numbers of hidden units. 5 simulations were run for each network size and pairwise correlations were computed for the network with median training performance for each network size.