SPEAKER IDENTIFICATION BASED ON VOWEL SOUNDS USING NEURAL NETWORKS

P. D. Templeton* and B. J. Guillemin**

* Department of Mathematical Sciences
University of Bath

** Department of Electrical & Electronic Engineering
University of Auckland

ABSTRACT - This paper presents the results of an experiment which applies a neural network approach to the problem of speaker identification. We restrict ourselves to the analysis of the 11 non-diphthongal English vowels. The neural networks were trained on a set of cepstral coefficients derived from an LPC analysis. Results are presented which show that this approach compares favourably with more traditional methods.

INTRODUCTION

Neural networks have been applied widely to the problem of speech recognition. However, to our knowledge they have not been used for speaker recognition.

Recent work by the Speech Processing Group at the University of Auckland has focused on vowel sounds as a basis for speaker identification (Miles & Guillemin, 1989; Miles, 1989). This is of particular interest in forensic applications where the amount of acoustic information for analysis is limited, and often severely corrupted by noise and transmission effects. Since vowels occur frequently and with high energy they are an obvious choice in these situations. Our previous work based the identification on the vowel formant frequencies, and used an Inverse-Variance-Weighted distance measure for template comparison. A major drawback with this approach is that the use of this weighting function is an *a priori* decision, and one which may well not be optimal.

As an extension to this work therefore, we have investigated the use of neural networks in this context. Of particular attraction is their ability to form optimally weighted mappings of their input parameters, thus obviating the need for the selection of an *a priori* weighting function.

Atal (1976) has compared a number of different spectral representations of the speech signal and found that the LPC cepstral coefficients, when used with a Mahalanobis distance measure, gave the best speaker recognition performance. More recently (Soong and Rosenberg, 1988) have investigated the use of weighted and unweighted cepstral coefficients in speaker recognition, and have drawn similar conclusions. They have also, interestingly, concluded that the instantaneous spectral features of the speech signal carry more speaker relevant information than

transitional spectral features. We therefore chose to use the cepstral coefficients, derived from an LPC analysis of the speech waveform, as the input parameter set for the neural networks.

The neural networks operate on the unweighted cepstral coefficients, converging automatically to an optimal solution. The convergence method is the back-propagation learning algorithm of the Multi-Layer Perceptron (MLP) (Rumelhart and McClelland, 1986), which is a neural network model well-suited to pattern recognition. This model has been widely applied to speech recognition with much success (Elman and Zipser, 1988; Waibel et al, 1989). Another method that has achieved good performance in speech recognition is the Learning Vector Quantisation algorithm of the Kohonen Self-Organising Feature Map. This has also recently been applied to speaker identification (Bennani et al., 1990). In the opinion of the authors, though, this method is a vector quantiser rather then a neural network.

The purpose of the experiments described in this paper is to see how well the MLP performs in terms of speaker identification. We compare these results with our earlier work and show that the performance of the MLP based identification is somewhat superior. It should be pointed out, though, that we cannot make a definitive comparison because the same data set has not been used in each case.

EXPERIMENT DETAILS

Data Collection

Our test population comprised 9 male subjects of New Zealand origin. Recordings were made of the 11 non-diphthongal English vowels (represented by the phonetic symbols /i/, /I/, /ɛ/, /æ/, /a/, /ɔ/, /ɜ/, /u/, /ʌ/, /o/ and /U/) embedded in 4 consonantal contexts: h/V/d, b/V/d, d/V/d and g/V/d. The reason for using different contexts is to achieve context independence of the analysis by accommodating possible coarticulatory modifications to the vowels. The complete set of 44 utterances were recorded 8 times for each subject over 2 sessions separated by a number of weeks.

Data Processing

These recordings were low-pass filtered to 4.5kHz and sampled at 10kHz. A 512 sample segment was extracted from the stationary vowel portion of each utterance. Each segment was pre-emphasised by a first order digital network with a transfer function,

$$H(z) = 1 - z^{-1}$$

These were then Hamming windowed prior to performing a standard 14th-order LPC autocorrelation analysis. From the LPC coefficients a set of 14 cepstral coefficients were recursively calculated.

Soong and Rosenberg (1988) used 8 cepstral coefficients in their analysis. They point out that the higher order cepstral coefficients are as important as the low order coefficients in

their ability to discriminate between speakers. For this reason we decided to use a slightly larger number of coefficients (14 in our case), though the greater the number of inputs to the neural network, the longer it will take to train.

Training the Neural Networks

The Multi-Layer Perceptron has a layered feed-forward architecture and is trained in a supervised manner using labelled training patterns. For these experiments we used an MLP simulator written in the C programming language that was developed at Bath University. The simulations were performed on a UNIX-based digital computer.

The MLP used in our experiments had 14 real valued inputs (one for each cepstral coefficient), 9 units in a single hidden layer, and 9 output units (one for each subject).

We used 11 networks, one for each of the vowels in our data set. Half the data for each vowel was used for training the networks (4 contexts x 4 rounds x 9 subjects = 144 training patterns for each vowel) and the other half for subsequent testing (144 test patterns per vowel).

The networks were trained using the standard back-propagation of error algorithm (Rumelhart and McClelland, 1986) . A network is initialised to a random internal state. Training involves presenting a pattern from the training set at the inputs simultaneously with the desired pattern at the outputs. The error between the actual and desired outputs is calculated automatically and propagated back through the network. The strengths or weights of the internal connections in the network are then modified such that this error is minimised. For any particular training pattern the desired output pattern would have a high signal (> 0.75 x full scale) on the output corresponding to the correct subject and a low signal (< 0.25 x full scale) on all the other outputs. Intrinsic to this algorithm is a learning rate which can take a value between 0 and 1 and affects the speed of convergence to a solution. If this parameter is set too high an optimal solution may not be found or the state of the network may oscillate indefinitely. Previous work at Bath University has shown that a learning rate of 0.5 is a good compromise.

In our experiments the networks converged to a stable state after about 1000 presentations of the training data set, which on our computer (an HP 9000 Series 300) took typically one hour for each network. In some cases the networks failed to learn the correct association for one or two of the training patterns.

After training, each of the networks was tested by applying the test patterns in turn to the inputs, causing an output pattern to be generated. The output with the highest level then identified one of the speakers. This method is valid because we wished to perform speaker identification from within a known set. In the case of speaker verification it would be necessary to ensure that the highest output was above a pre-determined threshold and all others were small by comparison.

RESULTS AND DISCUSSION

The results for each of the 11 vowels and 9 subjects are summarised in Table 1. The vowel or phonemic class is listed down the left side, and the speakers or output classes are listed along the top of the table. Each entry represents the percentage of the test patterns that the network successfully associated with the correct speaker.

VOWEL	Example word	1 mm	2 ph	3 ag	4 tr	5 ga	6 mn	7 mt	8 pm	9 fe
i+	heed	88	38	94	75	75	88	100	100	100
I	hid	94	19	94	63	50	81	100	100	50
ɛ*+	head	94	63	88	56	75	88	100	100	100
æ*	had	56	56	100	31	75	69	75	100	100
a+	hard	100	63	81	88	75	63	100	25	100
ɔ	hoard	75	6	50	50	100	69	100	75	100
ɜ*	heard	56	25	56	75	25	50	100	75	75
u*+	who'd	75	31	100	88	100	81	75	100	100
ʌ*+	hud	69	56	100	94	75	38	100	50	75
o	hod	94	19	25	44	75	50	100	50	75
U+	hood	100	38	88	69	100	69	100	100	100
Average Rate		82	37	80	66	75	68	95	80	86

Table 1 : % Identification rates for each vowel for each of the nine subjects

*optimum vowel set using formant frequencies alone
+vowels that achieved average identification rates of > 75%

Unfortunately some of our subjects were not available for the second recording session. For this reason subjects 5, 7, 8, and 9 had only 44 patterns in the test set (i.e., 4 for each vowel). The results for subjects 1, 2, 3, 4 and 6 are statistically more valid having 16 patterns for each vowel.

Speaker identification rates averaged over all the vowels vary between 37% and 95% across the speakers. Performance within a particular vowel category varies between 6% for the worst subject and 100% for the best.

At the outset these results might seem poor by comparison with other speaker identification experiments, but it should be emphasised that in this case the identification process has been restricted to an analysis of vowel sounds alone, which is a severe limitation. Further, the figures given relate to identification based upon a single vowel sound. In practice one would base the identification on a number of vowel sounds, which

will significantly improve the overall identification rates.

It is interesting that the optimum vowel set from these experiments does not accord with that deduced from experiments based upon formant frequency (Miles, 1989). This implies that the cepstral representation contains other information in the speech signal that is important in discriminating between subjects. The average identification rate over all vowels and subjects for these experiments was 73.7%, whereas for the experiments using formant frequencies it was 66.9%. One might conclude that the cepstral coefficients contain more speaker related information than the formant frequencies.

FUTURE WORK

Using the same data set we intend to run a speaker identification experiment using the Kohohen Self-Organising Feature Map with the LVQ2 training algorithm to see how performance compares with the MLP model. We also intend to investigate identification accuracies based upon groups of vowels rather than on single vowels.

CONCLUSION

The main purpose of these experiments has been to determine the degree to which individual vowels can be used to distinguish between speakers. In our opinion the neural network approach has given us a more objective method of assessing this compared with more traditional techniques.

REFERENCES

Atal, B. (1976) "Automatic Recognition of Speakers From Their Voices", Proc. IEEE, 64, 460-475.

Bennani, Y., Fogelman-Soulie, F. & Gallinari, P. (1990) "Text-Dependent Speaker Identification Using Learning Vector Quantization", Int. Neural Network Conf. Paris, France, Vol 2, 1087-1090.

Elman, J.L. & Zipser, D. (1988) "Learning the Hidden Structures of Speech", J. Acoust. Soc. Am. 83, 1615-1626.

Miles, M.J. (1989) "Speaker Recognition Based upon an Analysis of Vowel Sounds and its Application to Forensic Work", Masters Dissertation, University of Auckland, New Zealand.

Miles, M.J. & Guillemin, B.J. (1989) "Speaker Recognition Based on an Analysis of Vowel Sounds", IREECON Int. Digest of Papers, Melbourne, Australia, 120-123.

Rumelhart, D.E. & McClelland, J.L. (1986) *Parallel Distributed Processing; Explorations in the Microstructure of Cognition*, Vol. 1, (M.I.T. Press: Cambridge, MA).

Soong, F.K. & Rosenburg, A.E. (1988) "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", IEEE Trans. Acoust., Speech & Signal Processing, ASSP-36, 871-879.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. & Lang, K.J. (1989) "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans. Acoust., Speech and Signal Processing, ASSP-37, 328-339.