TWO-LEVEL WORD RECOGNITION USING THE MULTI-LAYER PERCEPTRON

IAN HOWARD

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

1. INTRODUCTION

This paper describes an experiment carried out on a multiple speaker digit database using the multi-layer perceptron (Rumelhart et al, 1985). The emphasis of this work is the incorporation of a priori knowledge of the recognition task and this is achieved by employing a two-level analysis of the speech: Firstly a bank of feature detectors emphasise phonetically significant regions in the speech and secondly these act as the input to word detectors that are only connected to the features necessary to detect the appropriate word. Each of these two levels were trained separately which reduced overall training times. The performance of the system on the digit database is shown to be better than a reference Hidden Markov Model recogniser.

2. LEVELS OF MODELLING OF SPEECH

Word modelling. The task of a system that performs automatic speech recognition of isolated words is to, when presented with a word from its vocabulary, generate a discrete symbol that indicates its identity. Here we only consider the problems associated with a 10 digit isolated word database. The simplest approach to isolated word recognition problem is to suitably represent the speech over a time-window long enough to contain the word and then to directly apply a pattern recognition technique to this pattern vector (Peeling & Moore, 1987). The advantage of such a direct mapping from a parametric representation of the input speech to a symbol to represent word identity is that training of such a system only requires the input data be labelled with the word identity. By employing a direct input to output mapping, the system designer is oblivious of how the mapping is carried out. If the designer has a poor understanding of how this analysis should be performed, then this may be a good thing. However, this requires that the necessary mapping be learnt during the training phase of operation of the system without any a priori limitations on what the transformation should be. The work in this paper makes use of the multi-layer perceptron and it will be appreciated a fully interconnected MLP with a hidden layer constitutes a very general purpose pattern recognition technique. As in any estimation problem, it is desirable to constrain the task as much as possible because a reduction in their degrees of freedom will also reduce the necessary training times and quantity of training data that is required. The approach used here is to break the word recognition problem down into simpler sub-problems, and this involved using sub-word models in the analysis of the speech.

Sub-word models. The reason it is possible to contemplate using sub-word models for speech stems largely from the fact that the speech signal is hierarchically organized (Borden & Harris, 1980). Consequently, to analyze speech into units smaller than word units is a natural thing to do. There are many good reasons why one would wish to do this. If speech could be represented by a relatively small number of simpler primitive patterns, these could then be used to describe the larger units in speech, such as words. Training will also be made easier by performing a hierarchical analysis of the speech into a set of primitive sub-patterns. If small set of sub-patterns can be detected in the speech, then the occurrences probability of a given sub-pattern will be much higher than that of a given word. Consequently, less data will be required to define the sub-patterns than is required to define each word in the vocabulary. In addition, a sub-pattern is, by definition, simpler than the original pattern. Therefore, one would expect that the transformation of the speech signal required to detect a sub-pattern will be of lower complexity than would be required to detect a word directly and so the sub-patterns detectors should be easier to train than direct word detectors.

3. ACOUSTIC PHONETIC REPRESENTATION OF SPEECH

In order to incorporate a priori knowledge into a speech recognition system by analyzing the speech into a set of primitive sub-patterns, it is necessary to have some idea what these sub-patterns should be. The requirement for the feature detectors are that they should maintain the necessary contrasts in the input signal, constitute a simpler representation of it and also be reliably detectable. There exists the broad field of phonetics which provides a source of information concerning the important sounds in speech necessary to discriminate between different words. One such set of units would be to make use of phone units, which are the basic sounds of speech. However, the intermediate representation used here is based on acoustic-phonetic qualities of the speech, because there are easier to detect.

Advantages of acoustic-phonetic features. A phonetic description of speech can correspond to many acoustic realization of the utterance. As a consequence of this, a phonetic description provides a means to normalize speaker differences at the acoustic level. The "complexity" of the phonetic representation of speech is lower than the acoustic representation of speech. This observation is verified by the fact that it is possible to get very high word recognition results using a linear classifier operating on the phonetic features whereas a non-linear classifier is needed if the acoustic data (vocoder data) is used directly as an input (Howard & Huckvale, 1987; Peeling & Moore, 1987).

4. THE MULTI-LAYER PERCEPTRON

The pattern recognition technique employed in this work for both operation at the acoustic-feature level and the word level was the multi-layer perceptron (Rumelhart et al, 1985). This technique has been shown effective for many speech pattern processing tasks (Boulard & Wellekens, 1987; Peeling et al, 1986; Howard & Huckvale, 1988) and better than many standard methods for speech pattern processing (Huang & Lippmann, 1987). In addition it provides a convenient formalism for constructing systems that are trained in parts (Waibel et al, 1988) and then combined together to provide a network that is organized hierarchically (Howard & Huckvale, 1989). The MLP used here was trained using the back-propagation rule and in some cases with an adaptive selection of the momentum term and learning rate (Chan & Fallside, 1987). In addition, the learning also made use of selective emphasis of the training data.

Pattern sorting. The data pattern vectors used to train the MLPs were sorted into representative groups such that each group had at least one of each kind of pattern class in it. The proportion of each pattern classes in each representative group reflected the overall proportions of the pattern classes in the training data. The advantage of using this procedure is that it is then possible to make MLP weight changes over a relatively small set of patterns, but ones which are a good reflection of the possible range of patterns in the data set.

Selective emphasis training of the MLP. A technique that has been found successful in speeding up training and in some cases also increasing performance involves selective emphasis of the weight taken of different pattern vectors. Thus if a pattern was correctly recognized, w.r.t a pre-defined threshold, it was de-emphasized whereas if it was wrongly recognized, more emphasis was placed on it. For the MLP, a high class threshold of 0.9 is employed and for a low class, a threshold of 0.1 is employed. These values were chosen because they constitute values that can be reached without too much difficulty (whereas 0.0 and 1.0 cannot be reached since the sigmoid non-linearity reaches these values only asymptotically). In addition, the emphasis value of 1.0 and the de-emphasis values of 0.0 were chosen. With these parameter values this procedure is the same strategy as used to train in NETtalk (Sejnowski & Rosenberg, 1986) although it was devised independently. In a large number of cases, the selective emphasis method reduced the training time and also gave rise to output that were better canonical representations of their targets than could be obtained with the same network with normal training. One reason for the speeding-up of the training using this method is due to the fact that when a pattern is considered to be correctly recognized, no back-propagation of error or updating of the weights needs to be carried out, and this can typically take up around 70% of the processing time during training.

5. THE RSRE DIGITS DATABASE

The database used was the RSRE 40 speaker ten digit database and the least consistent 20 speakers were used for all the recognition experiments. There were 15 speakers used for training, and the remaining 5 were used for testing. Initially 1500 digits were used for training and 500 for testing. In the final experiment, the test data set was expanded to 2000 digits. The data was supplied by RSRE with the digits detected, cut and processed by means of a 19-channel vocoder (Holmes, 1980). The data frames were quantized into 16 levels and sampled in 20ms frames. The input frames were typically in the range 0-50 and were scaled to the range 0.0-1.0 by multiplying the input elements by a factor of 0.02.

6. ACOUSTIC-PHONETIC FEATURE DETECTORS

Labelling of the speech data. The phonetic labels used were chosen to permit discrimination of the digit database and more phonetic labels would be needed to discriminate a larger vocabulary. One repetition of each digit for each speaker was labelled by hand and then an automatic technique was employed to transfer the annotations to the other repetitions by means of a dynamic programming algorithm, somewhat akin to DTW speech recognition (Chamberlain & Bridle, 1983). The advantage of such a scheme is that it is quick, easy (no more hand labelling is involved) and the labels generated in this way are self-consistent. The annotation labels were converted to features by means of a look-up table. The features present in each word are shown in table 2, with the exception of the word 'two', which did not contain 'FRIC', 'EE-IH' and 'ER'.

Training the MLP acoustic-phonetic feature detectors. The input patterns were generated by sweeping a window over 5 adjacent frames of the vocoder data and the corresponding output vector consisted of the 17 features as they were defined at the centre of the input window. In the new experiments, the patterns were then sorted into representative groups, which in this case consisted of 251 pattern vectors. Initially one multiple output MLP was then trained to perform the required transformation, using the generalized delta rule. In the last instance of the experiment 17 single-output MLPs were trained using the selective emphasis method. Training was generally carried out until the error on the training data suggested that no more learning would occur.

Evaluation of the performance of the phonetic-feature detectors. The performance of the feature detectors was found from the receiver operating characteristic (ROC) of each detector (Levine & Schefner, 1981). This is a plot of the number (or proportion) of correctly identified patterns against the number of false alarms, for a changing detection criterion. The point used to characterise the ROC curve is the equal-error point, which is the point where the percentage of misses equals the percentage of false alarms. In addition, the threshold value for which the equal-error point is found is also given in results table 1. It is important to note that this performance criterion compares the outputs from the feature detectors with the targets, and consequently any differences that occur at boundaries will show up as errors, whereas in actual fact they may be unimportant for the subsequent word recognition stage.

7. WORD RECOGNITION USING THE MLP

The isolated word recognition experiments carried out use the same basic approach as that adopted by Peeling & Moore, 1987. An input pattern vector was generated from an isolated utterance of a word by placing its acoustic-phonetic representation in the centre the pattern vector window of width 50 frames. The outer elements were then padded with the no speech present condition element values, which in this case corresponded to setting the silence feature output high and all the others feature outputs low. The vectors were then used to train either 10 single output individual word detectors MLPs. During recognition mode, the output node with the largest output value was found and the output word class assigned accordingly.

8. PREVIOUS EXPERIMENTAL RESULTS

Acoustic-phonetic feature performance. The network used in a previous experiment (Howard & Huckvale, 1989) had the configuration 95-34-17, that is 95 inputs, 34 hidden nodes and 17 outputs. The equal-error ROC results for this network appear in table 1. It was found that overall, good performance was achieved but some features gave bad performance with this network. Different training runs gave rise to different performances, especially for the lower performance features such as the 'UE' detector. The interpretation of this is that the network got stuck in local minima from which it could not escape. It can be seen that the performance for the 'UE' detector is rather poor, at only 60%

TABLE 1. Acoustic-phonetic feature performance, speaker-independent.

Network configuration		95-34-17	95-50-17		95-5-1	
Feature	Description	test	test	thresh	test	thresh
thresh						
SIL	(silence)	91.9%	94.9%	0.31	94.4%	0.32
FRIC	(frication)	90.7%	94.1%	0.25	92.4%	0.13
VOC	(voicing)	93.0%	95.9%	0.47	93.9%	0.40
NAS	(nasality)	84.8%	91.5%	0.10	88.2%	0.11
VFRIC	(voiced frication)	76.9%	85.5%	0.02	78.8%	0.05
S	(/s/ fricative)	93.4%	95.2%	0.24	94.4%	0.19
FTH	(/f/,/T/ fricative)	75.1%	89.1%	0.06	81.1%	0.10
?	(glottal stop)	88.8%	97.2%	0.03	26.4%	0.00
K-REL	(/K/-release)	86.2%	98.0%	0.06	82.4%	0.01
T-ASP	(/t/-aspiration)	60.8%	92.8%	0.01	87.6%	0.03
EE-IH	(/i/,/I/ vowel)	90.0%	92.3%	0.17	90.2%	0.18
EH	(/e/ vowel)	94.0%	96.1%	0.05	94.6%	0.11
UH	(/V/ vowel)	94.4%	96.5%	0.17	96.6%	0.03
ER	(schwa vowel)	77.3%	84.9%	0.08	81.8%	0.01
AW	(/O/ vowel)	98.9%	99.0%	0.16	99.3%	0.40
UE	(/u/,/U/ vowel)	60.6%	92.3%	0.09	83.4%	0.11
R	(/r/glide)	95.3%	97.5%	0.23	97.3%	0.30

Word recognition performance. The old MLP word recognition experiment used all the feature detector outputs over 50 frames in time. The acoustic-phonetic feature detectors were run on the training data, and these tracks were then used to train the word classifiers. The resulting confusion matrix appears in figure 1.

FIGURE 1. Confusion matrix for old MLP word recognition experiment.
Overall recognition rate = 96.6%

	1	2	3	4	5	6	7	8	9	0
1	48	0	0	0	0	0	0	0	1	1
2	0	39	1	1	0	0	0	0	0	9
3	0	0	50	0	0	0	0	0	0	0
4	0	0	0	50	0	0	0	0	0	0
5	0	0	0	0	50	0	0	0	0	0
6	0	0	0	0	0	50	0	0	0	0
7	0	0	0	0	0	0	50	0	0	0
8	0	2	0	0	0	1	0	47	0	0
9	1	0	0	2	0	0	0	0	47	0
0	0	0	0	0	0	0	0	0	0	50

In addition, the same training (part of it) data was used to train an 8-state Gaussian continuous distribution HMM recognizer (Russell & Cooke, 1986), using the vocoder data directly. Figure 2 shows the performance of the standard 8-state continuous distribution HMM on the same test digit set. There was no significant difference in overall performance between the HMM and

MLP algorithms. However, it can be seen from the confusion matrix in figure 1, that the errors for the MLP were largely due to the failure of the digit 'two' detector. This poor performance can be traced back to the low performance of the 'UE' detector which is necessary for the discrimination of 'two' and 'zero' in this case.

FIGURE 2. Confusion matrix for old HMM word recognition experiment on vocoder data. Overall recognition rate = 96.2%

	1	2	3	4	5	6	7	8	9	0
1	49	0	0	0	1	0	0	0	0	0
2	0	42	0	0	0	0	2	0	0	6
3	0	0	48	0	0	0	0	2	0	0
4	0	0	0	50	0	0	0	0	0	0
5	1	0	0	0	49	0	0	0	0	0
6	0	0	0	0	0	50	0	0	0	0
7	0	0	0	0	0	0	50	0	0	0
8	0	0	0	0	0	6	0	44	0	0
9	2	0	0	0	2	0	0	0	46	0
0	0	0	0	0	0	0	0	0	0	50

9. NEW SPEAKER-INDEPENDENT RECOGNITION EXPERIMENTS

The acoustic-phonetic feature detectors were retrained speaker-independently with a MLP network with more hidden units (50) than the original network (34). It was found that significantly higher speaker independent performance could be achieved with the larger network, which was also less prone to getting stuck in local minima during training. These results also appear in table 1.

Separate training of feature detectors, with selected emphasis method. In another experiment each feature detector was trained using a separate 95-5-1 network, rather than using one large network for all the features at the same time. This approach was adopted because there are difficulties encountered when training one large network. Firstly, if any feature is to be removed, changed or added, then the entire large network has to be retrained. Secondly, it has often been noticed that different output features learn at different rates. It was found that the separate feature detectors trained well, and the selected emphasis methods was used for the last part of their training. Some detectors showed lower performance than those in the large network, suggesting that more hidden nodes were probably required. Again results appear in table 1.

Partial interconnect of word detectors. An approach to reduce word recognizer size was to use partial interconnect on the word classifiers, so that a given word detector only received input from features known a priori to occur in that word. The partial interconnections used for the word detectors are shown in table 2. The previous number of weights in digit word recognizer was 8500, whereas the number of weights in partial interconnect word recognizer is 2850, which constitutes only 33.5% of the original number of weights with no measurable reduction in recognition performance. The confusion matrix obtained after joining all the networks together and using the multiple-output network as before appears in figure 3.

TABLE 2. List of partial interconnect features used in recognition.

one	VOC ,UE, UH, NAS, SIL
two	T-ASP, UE, SIL, VOC, FRIC, EE-IH, ER
three	F-TH, R, EE-IH, SIL, FRIC
four	F-TH, AW, SIL, VOC, FRIC
five	F-TH, UH, ER, EE-IH, VFRIC, SIL
six	EE-IH, SIL, KREL, FRIC, VOC
seven	S, EH, VFRIC, ER, NAS, SIL
eight	GLOT, EH, EE-IH, T-ASP, SIL
nine	AS, UH, EE-IH, SIL
zero	S, VFRIC, EE-IH, R, UE, EH, SIL, FRIC, VOC

FIGURE 3. Confusion matrix for MLP on partially-interconnected acoustic-phonetic features. Overall recognition rate = 99.2%

	1	2	3	4	5	6	7	8	9	0
1	50	0	0	0	0	0	0	0	0	0
2	0	50	0	0	0	0	0	0	0	0
3	0	0	50	0	0	0	0	0	0	0
4	0	0	0	49	0	0	0	1	0	0
5	0	0	0	0	50	0	0	0	0	0
6	0	0	0	0	0	49	0	0	0	1
7	0	0	0	0	0	0	50	0	0	0
8	0	0	0	0	0	0	0	50	0	0
9	0	0	0	0	0	0	0	0	50	0
0	0	2	0	0	0	0	0	0	0	48

This recognition system was also tested on 2000 digits from the test data set, which gives a more statistically reliable recognition performance. The confusion matrix for the 2000 digit test appears in figure 4. This is better than the original speaker-independent experiment and the increased performance is due to the fact that now the 'UE' phone detector was operating well and consequently the significant misclassifications of the digits 'two' and 'zero' no longer occurred.

10. SUMMARY AND CONCLUSIONS

The main result of this paper was to show that the use of an intermediate acoustic-phonetic representation in the analysis of speech is beneficial and the results obtained so far indicate that good recognition performance can be achieved using the MLP. By breaking up the analysis into a set of simpler sub-problems, the overall problem becomes easier to solve. The use of acoustic-phonetic features can be interpreted as a means of including a priori knowledge into a MLP network for the purposes of speech recognition. It was found useful to train each feature detector and each word detector in separate MLP networks, because the different features and word detectors train at different rates and such a scheme can accommodate this as well as speeding up training. It has also been found possible to reduce the number of weights required in the recognition MLP system by using partial interconnections in the word detectors only to features known a priori to be important for the given word.

FIGURE 4. Confusion matrix for MLP on partially connected acoustic-phonetic features. Overall recognition rate = 98.5%

	1	2	3	4	5	6	7	8	9	0
1	200	0	0	0	0	0	0	0	0	0
2	0	200	0	0	0	0	0	0	0	0
3	1	0	195	1	3	0	0	0	0	0
4	0	0	0	199	0	0	0	1	0	0
5	0	0	0	0	200	0	0	0	0	0
6	0	0	0	0	0	185	0	0	0	15
7	0	0	0	0	0	1	199	0	0	0
8	0	0	0	0	0	0	0	200	0	0
9	0	0	0	1	0	0	0	0	199	0
0	0	8	0	0	0	0	0	0	0	192

The recognition results achieved with the current database are too high to be able to reliably evaluate improvements to the recognition strategy and experiments on a large 700 word vocabulary are now being carried out in conjunction with Mark Huckvale.

11. ACKNOWLEDGEMENTS

The pattern processing system used for this work was written in conjunction with Mark Huckvale. This work was supported by MOD University research agreement 2047/137 with SRU at RSRE.

12. REFERENCES

G J BORDEN & K S HARRIS, (1980), Speech science primer, Baltimore, Williams & Wilkins.
H BOULARD & C J WELLEKENS, (1987), IEE 1st Ann. Confr. on Neural Networks, San Diego, California.
R M CHAMBERLAIN & J S BRIDLE, (1983), IEEE ICASSP 816.
L CHAN & F FALLSIDE, (1987), CUED/F-INFENG/TR.2.
J HOLMES, (1980), IEE Proc., 127, Part F, No. 1.
I S HOWARD & M A HUCKVALE, (1988), FASE88,Edinburgh.
I S HOWARD & M A HUCKVALE, (1989), 1st IEE confr. on ANNs, London.
W Y HUANG & R LIPPMANN, (1987), ICNN, San Diego, CA, 21-24 JUNE, 1987.
M LEVINE & J SCHEFNER, (1981), Fundamentals of sensation and perception, Addison-Wesley.
S M PEELING & R K MOORE, (1987), RSRE Memorandum 4073.
D E RUMELHART, G E HINTON & R J WILLIAMS (1985), in Rumelhart & McClelland, "Parallel distributed processing", MIT press.
M J RUSSELL & A E COOKE, (1986), Proc. IOA, 8, Part 7, p291-297.
T J SEJNOWSKI & C R ROSENBERG, (1986), The John Hopkins University EE & CS Technical Report, JHU/EECS-86/32 PP.
A WAIBEL, T HANAZAWA, G HINTON, K SHIKANO, & K LANG, (1988), IEEE Trans. ASSP, ASSP-37, p328.