Proceedings of the Institute of Acoustics

SUB-WORD UNIT CLASSIFICATION USING A MULTI-LAYER PERCEPTRON

S.G. Smyth

British Telecom Research Laboratories, Martlesham Heath, Suffolk, U.K.

1. INTRODUCTION

Most artificial neural network (ANN) techniques do not address the problem of temporal variation in speech since the networks employed generally require fixed dimensional input. The system discussed in this paper uses the ANN, in the form of a multi-layer perceptron (MLP), to map fixed size short fragments of signal (of the order of a few frames—up to 90ms) on to symbols representing sub-word units, which are then integrated using dynamic programming (DP) alignment. This segmental MLP approach may be compared with hidden Markov model (HMM) techniques [3], with the MLP providing the local probability scores and the DP alignment embodying the state sequence processing.

The MLP is trained to perform the frame classification task and the DP method is used to combine legal sub-word unit sequences. The DP alignment can also be used to resegment the data, as in [2], and further training commenced on the new segmentation.

In this paper, the system has been applied to the task of speaker independent isolated word recognition, for which it attained a performance comparable to that of HMMs. The vocabulary used was the British English alphabet, from the BT Connectionist Project data set as described in [6].

The structure of the paper is as follows. First the sub-word units will be described. Section 3 explains the processing performed by the MLP and the DP task. The process of resegmentation is introduced in the following section. Experimental results are presented in section 5, followed by a discussion and suggestions for future work.

2. SUB-WORD UNITS

Each endpointed utterance was divided into a number of sub-word units. The data was initially linearly segmented, but automatic relabelling permitted a more reasonable segmentation which improved accuracy. It would have been preferable to use a linguistically defined segmentation [2] so that different pronunciations of words could be represented rather than assuming that each utterance of a word consisted of the same symbols, but a labelled database was not available.

The experiments were carried out with three units per word—a fairly arbitrary choice, but simple to implement. Further experiments were carried out with more units, and a variable number per word—however, these did not perform as well as the three unit per word experiments, as will be discussed in section 6.

2.1 Distinct
Initially each word consisted of three distinct units—i.e., a sub-word unit appearing in one word did not appear in any others. Since there were 26 words and three symbols per word, this resulted in a fairly large number of symbols (78), of which a significant number are very similar; for example, the final units in the 'E'-set ('BCDEGPTV').

2.2 Tied
From the data described in section 2.1, a sub-word unit confusion matrix was identified and the most confusable ones amalgamated—a couple of iterations of this process reduced the symbol set to 45 units, resulting in a much smaller MLP and no loss of performance.

2.3 Larger sets
Some experiments were tried with as many as seven units per word. Each word was defined as seven units and symbol confusion matrices produced to condense this to a more manageable number. However, the results were poor and reasons for this are discussed in section 6.

3. SYSTEM DESCRIPTION

The MLP was a single hidden layer fully interconnected network, trained using error back propagation [5].

Its task was to map a frame of speech (with context information) on to the identifier of the sub-word unit occupied by that frame. The MLP had (*size of frame* × *number of context frames*) inputs, a variable number of hidden nodes and *number of units* outputs, organised as a 1-in-n coding scheme. With this output coding, the MLP can be viewed as approximating an HMM local probability distribution (strictly a likelihood distribution) for the identity of the frame [1].

Naturally, this process was not very precise: symbol recognition rates of 50% are typical for a very well trained network. Rather than peak picking to identify symbol sequences, DP in the form of Viterbi alignment was used to find the maximum likelihood for each word in the vocabulary. The current implementation allowed only transitions to the next sub-word unit or remaining in the current one, with equal probability.

A block diagram of the full system is presented in figure 1.

4. SEGMENTATION

4.1 Initial segmentation
For these experiments, the initial segmentation was performed linearly, i.e., the first third of an utterance was labelled as the first sub-word unit, and so on. Some other experiments involved training an HMM on the data and using this to segment the data. However, there was no appreciable improvement in performance for the large amount of extra processing time, so this technique was abandoned in favour of the simple linear process.

4.2 Machine defined segmentation
Once the system had been trained, the DP process provided a full path as well as the maximum likelihood. The symbol transition information from the path was used to relabel the data frames. The MLP was then retrained on the new segmentation, and the process repeated. This is an example of segmental k-means training [4].

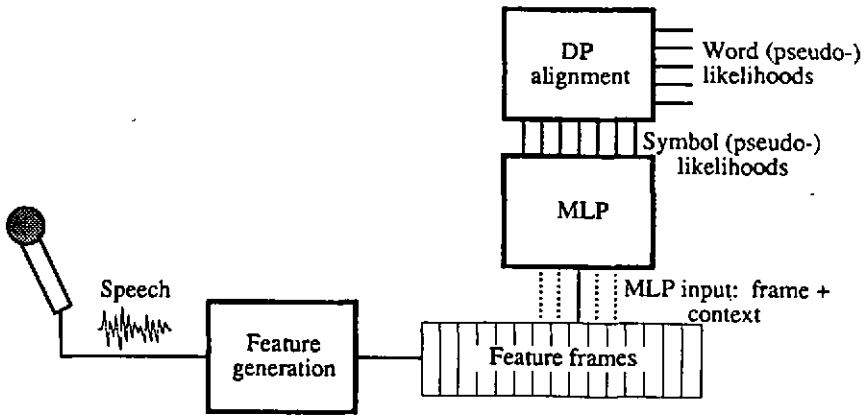

Figure 1: Segmental MLP recognition system.

5. EXPERIMENTS

5.1 Database

The data set contained three utterances of each letter of the British alphabet from 104 talkers, collected at a 20kHz sampling rate in a low noise environment. The speech was endpointed and manually checked (with bad utterances removed) before further processing. The database was split into 52 training and testing speakers, balanced with respect to sex and age group resulting in a training set of 3,999 utterances and a test set of 3,977 utterances.

The speech waveform was converted into Mel frequency cepstral coefficients (MFCCs) with a frame length of 25.6ms. There was a total of 17 features for each frame.

The data available as input to the MLP were the features over a window of several frames. The output of the MLP was a 1-in-n vector with the high output being the desired segment identifier.

5.2 Results

An initial experiment was carried out with the MLP having only a single frame, with no context information, to establish a baseline score. In further experiments, the context window was expanded to seven frames (three on each side), representing a total of about 90ms of speech.

From the results of the experiments with totally distinct sub-word units, symbol confusions were identified and the symbol set was reduced to 67 sub-word units. A second iteration allowed this number to drop to 45 units. No further significant symbol confusions were discovered at this stage.

Each configuration was evaluated with a range of hidden layer sizes and several random starts. A summary of the results appears in table 1.

At this stage, the data was resegmented and the MLP retrained, giving the results in table 2, all based on the best performing 45 sub-word unit, 7 frame, 70 hidden node system.

Number of sub-word units	Context window size	Number of hidden nodes	Frame accuracy Training	Frame accuracy Test	Word accuracy Training	Word accuracy Test
78	1	20	18.26%	17.22%	—	—
78	5	70	31.86%	29.68%	78.17%	74.64%
78	7	70	35.29%	32.64%	83.72%	80.62%
67	7	70	42.53%	39.40%	74.62%	72.72%
45	7	70	55.88%	53.28%	83.97%	83.11%

Table 1: Results before resegmentation.

Resegmentation	Frame accuracy Training	Frame accuracy Test	Word accuracy Training	Word accuracy Test
Initial segmentation	55.88%	53.28%	83.97%	83.11%
First resegmentation	57.87%	55.11%	86.87%	85.37%
Second resegmentation	68.11%	65.28%	89.07%	86.53%
Third resegmentation	69.81%	66.99%	88.50%	85.93%

Table 2: Results after resegmentation.

As can be seen, the change in performance after the first resegmentation is minimal. In fact, after the second resegmentation, word accuracy dropped slightly while frame accuracy was still increasing.

6. DISCUSSION AND COMPARISON WITH OTHER SYSTEMS

The results above suggest that the window size could have been increased further, but then the context would have been a significant portion of a word (average length approximately 40–50 frames).

The transition frames are poor examples of either of the adjoining sub-word units and so could have been eliminated from the training set. (The transition frames form unreliable training data but provide valid context information, so they can still appear as context frames for other data). However, doing so reduced the training set to such an extent that the resulting MLP did not generalise very well. A faster frame rate would provide more frames so that the transitions could safely be ignored.

The initial linear segmentation is actually not too unfavourable as is illustrated in table 2. The first resegmentation imparts a sizeable improvement while extra work offers little benefit. The improvement in performance with the first resegmentation is the effect of the system 'correcting' the clustering from the linear division. Subsequent relabelling moved the transition boundaries only very slightly, and the drop in word accuracy after the third resegmentation is probably due to corruption of the MLP's sub-word unit models by the transition frames.

Some further experiments were run with a much larger symbol set—seven sub-word units per word—the intention being to reduce the total number of symbols through identifying confusions to a more manageable size whilst still retaining distinct units where they were required. These experiments gave very poor results, probably due to the initial linear segmentation giving poor clustering and having much less training data per sub-word unit. Again, a faster frame rate would help.

One problem with the process of identifying confusable units is that the confusion matrix shows symbols which may be combined, but cannot indicate symbols which should be split, due to several distinct pronunciations of a unit—this is where a linguistically based initial data labelling would be useful.

Table 3 shows results from other recognition systems, illustrating that the performance of the segmental MLP method is comparable to dynamic time warping (DTW) and HMM systems. (It should be mentioned that there is not really enough data to train a discrete symbol, or *vector quantized* (VQ), HMM, and a continuous density HMM has achieved about 90% test set accuracy on this database, but with a modified front end.)

Classifier	Test set accuracy
DTW	84.32%
VQ HMM	82.38%
Segmental MLP	86.53%

Table 3: Comparison with other systems.

In HMM terms, the MLP is providing 'local probabilities' and the DP process is traversing states. The MLP makes use of context information which in unavailable to the frame level HMM processing. Other work [2] has taken the analogy further and included HMM state transition probabilities within the DP task. However, the HMM state traversal is flawed in that the state duration is restricted to an exponential distribution—this is also true of the equiprobability process implemented here. It was therefore felt that the extra work involved in modelling the state transitions more fully would not have been justified.

7. FUTURE EXTENSIONS

This section presents some ideas for future work with the segmental MLP system.

7.1 Faster frame rate

A faster frame rate would allow (a combination of) more sub-word units, larger context windows and removal of transition frames from the training set. With more sub-word units, each unit is modelling a smaller segment of speech, thus the MLP's task would be simplified. Removing transition frames would also prevent the system from trying to represent the spurious data in these frames.

7.2 Durational modelling
The Viterbi process is a very simplistic temporal model. The next step in these experiments is to introduce durational modelling [3]. This will not make very much difference to the endpointed isolated word results, but will improve performance on connected and non-endpointed speech.

7.3 Silence modelling
Extra sub-word units will be created to model background noise or silence data—the MLP will have extra output nodes for this data, treating frames of noise in the same way as the other sub-word units, though the alignment process will allow a modified durational distribution since inter-word gaps can be of any length.

7.4 Connected speech
With some added complexity, the DP process may be extended to cope with simple grammars. This could be used to implement a simple connected recognition scheme.

8. CONCLUSION

A technique for speaker independent isolated word recognition combining MLP and DP has been described, and results obtained. The MLP is used for frame recognition, and DP combines sub-word units.

The results are comparable to HMM performance on the same problem, and the process has much in common with HMM systems.

Limitations have been discussed, particularly the need for a faster frame rate.

The method can be extended to connected recognition, and work is underway to investigate this.

9. REFERENCES

[1] H BOURLARD & C J WELLEKENS, 'Links Between Markov Models and Multilayer Perceptrons', to appear in *IEEE Trans Pattern Analysis and Machine Intelligence*
[2] N MORGAN & H BOURLARD, 'Continuous Speech Recognition Using Multilayer Perceptron with Hidden Markov Models', *Proc ICASSP*, p413 (1990)
[3] L R RABINER, 'A Tutorial of Hidden Markov Models and Selected Applications in Speech Recognition', *Proc IEEE*, 77(2) p257 (1989)
[4] L R RABINER, J G WILPON & B H JUANG, 'A segmental k-means training procedure for connected word recognition', *AT&T Tech J*, 65(3) p21 (1986)
[5] D E RUMELHART, G E HINTON & R J WILLIAMS, 'Learning internal representations by error propagation', *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* eds Rumelhart & McClelland, MIT Press (1986)
[6] P C WOODLAND & S G SMYTH, 'A Neural Network Speech Recogniser for Directory Access Applications', *Proc Voice Systems Worldwide*, p196 (1990)