Proceedings of the Institute of Acoustics

SPEAKER INDEPENDENT RECOGNITION OF THE DIGITS OVER THE TELEPHONE NETWORK

John Talintyre and Simon Ringland

British Telecom Research Laboratories, Martlesham Heath, Suffolk, IP5 7RE, England

1. INTRODUCTION

Speech recognition can enable the provision of many useful automatic telephone network based services. A large proportion of these services can be provided by speaker independent (SI), limited vocabulary, isolated word (IW) recognisers. A vocabulary of particular relevance to many applications is that of the digits. This paper investigates some important aspects of SI recognition of isolated digits over the telephone network.

The collection and processing of training data is both time consuming and costly. It is therefore vital to minimise the amount collected, whilst not compromising performance. The choice of classification technique may therefore be influenced not just by accuracy, but also by training data requirements. The impact of the amount of training data is investigated for what are arguably the most successful classification algorithms: Dynamic Time Warp (DTW), Vector Quantised and Continuous Density Hidden Markov Models (VQHMMs and CDHMMs) and Multi-Layer Perceptrons (MLPs).

Speech detection (endpointing) is a central process in any IW speech recogniser. Our experience has shown that reliable endpointing is essential for accurate recognition. This is particularly difficult in the telephony environment where both environmental and line conditions can be far from ideal. The interaction between endpointer and classifier will affect the system's resilience to endpointer errors. To investigate these matters, the endpointing ability of a connected recogniser is compared with that of a rule/threshold based system and the two are judged against the performance achieved using manually placed endpoints for the above classifiers.

2. SPEECH CORPUS

2.1 Database Collection

The database used in the experiments was a subset of approximately 700 talkers taken from a nationwide telephone collection which was recorded over a noisy network and hopefully represents worst case conditions. The participating talkers, the great majority of whom were unfamiliar with speech recognition technology, called without supervision from any phone. Each one was prompted to speak each of the digits ('one' to 'nine', 'zero', 'nought' and 'oh') once.

All the recordings were verified, an utterance only being rejected if it was considered that a human could not recognise it reliably, given the knowledge that it should be a digit. Every utterance was manually endpointed, the endpoints being chosen to make a playback of the digit sound correct (e.g. if the initial fricative in a 'six' was buried in noise, the start point was chosen to make the word sound as natural as possible) in the hope that this would lead to better modelling of the digits.

The database was split into sets with a nominal 100 talkers in each, referred to as sets A, B, C, etc. Sets A to E contained training material (a total of 6,118 utterances), while the F and G sets were used for testing (a total of 2,518 testing utterances).

2.2 Signal Parameterisation

The feature set chosen consisted of MFCCs [1] and values derived from them. One benefit of the chosen parameterisation is that it is independent of the absolute gain of the line. For a system which does not work over the telephone this might be seen as a disadvantage. However, on the telephone network where the received signal level varies enormously from connection to connection, gain independence is invaluable. Ideally one might wish to use an energy measure which is normalised to the level of each call. However, this introduces a number of problems. Apart from the difficulty of identifying the normalisation factor, recognition must be delayed until the factor has been identified.

3. DESCRIPTION OF CLASSIFIERS

This section contains a brief description of each of the classifiers used in the experiments. Further detail may be found in the many references.

3.1 Dynamic time warping

[2] is a very good introduction to using DTW for speaker independent speech recognition. The modified k-means algorithm [3] was used to train the DTW recogniser; it clusters the training data from each class into a pre-defined number of clusters. A number of reference templates were then formed by averaging the templates in each cluster. These reference templates became the models for the DTW speech recognition system.

3.2 Hidden Markov models

HMMs have proved to be one of the most successful techniques for speech recognition. HMMs are well described in many papers, such as [4], [5] and [6]. There has been some debate [4] as to the comparative merits of continuous density hidden Markov models (CDHMMs) and vector quantised hidden Markov models (VQHMMs); therefore experiments were carried out using both.

All the models had a left-to-right topology allowing no skips, with 8 emitting states and non-emitting start and end states. The features used were reasonably uncorrelated, allowing the use of multiple codebooks for the VQHMMs and diagonal covariance matrices to describe the Gaussian mixture pdf's for the CDHMMs.

3.3 Multi-layer perceptron

An MLP [7] is a feedforward artificial neural network that is trained to perform a particular mapping from a certain set of input patterns to a pattern on its output nodes. When a particular class is present on the inputs, the output node corresponding to the class of the input should respond with a high output value, while all the other nodes should respond with a low value. The network learns to respond in this way to input data by adjusting all the weight and bias values in the network, using an error back propagation algorithm. The MLP requires a fixed dimensional input, this was achieved in these experiments by compressing each utterance to 20 frames using linear interpolation. All the MLPs had 1 hidden layer of 70 nodes.

4. RESULTS FOR DIFFERENT TRAINING SET SIZES

Recognition results are presented in Figures 1 and 2 for the four classifiers using manually endpointed data. For each training set 3 MLPs were trained using different random starts; the results quoted are for the network that gave the best test set result.

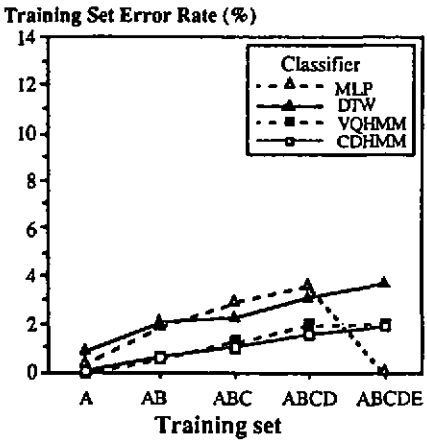

Figure 1 - Training set error rate as a function of the training set size

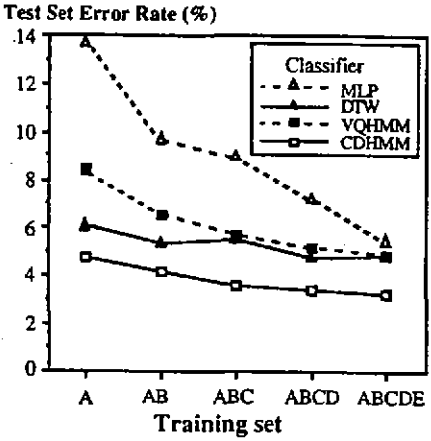

Figure 2 - Test set error rate as a function of the training set size

Some general observations can be made from Figures 1 and 2. Generally speaking, as the training set size increased the test set error rate decreased and the training set error increased. The classifier with the lowest test set error rate was always the CDHMM, the one with the highest error rate was always the MLP. The DTW classifier never did worse than the VQHMM classifier, but with the larger training sets (ABC and up) the difference in error rates is small (<0.4%).

With the exception of the MLP classifier, the training algorithms work by building a *model* of the training data which is then used in the classification task. During recognition, the class with the model(s) that best matches the unknown utterance is chosen as the recognised class. For this approach to work, the trained models must generalise well to unseen data. This leads to two basic requirements: the modelling assumptions must be good and the models must be well trained. The MLP is different in that it is trained in a discriminatory manner on all classes simultaneously, whereas with the other techniques each class is trained separately. Whilst the MLP does more than model each class, the same criteria apply to its success as a recognition technique. To satisfy the criterion that the models must be well trained, and hence as representative as possible of unseen data, requires a representative training set that is large enough to obtain good estimates of all the free parameters in the models. Therefore, we would expect a larger training set to lead to a lower test set error rate. The error rate on the training set should go up as the training set size increases, because the task is becoming more difficult — the models may better represent unseen data, but with the number of free parameters in the model remaining constant, they are unable to model the training set as accurately.

The observed results (Figures 1 & 2) were largely as expected. The anomalous MLP result for the ABCDE training set is a consequence of the nature of the back propagation training algorithm — the final network is heavily dependent on its initial parameter values, which in this case were fortuitously very good. It is noticeable in Figure 2 that as the training set size increases, the DTW error rate does not improve as well as the other classifiers.

5. DESCRIPTION OF DIFFERENT ENDPOINTERS

5.1 The rule/threshold based endpointer (RE)

The threshold based endpointer uses a number of simple rules operating on two separate measures. A set of endpoints is generated from each of the measures, which are then combined, accepting the area of overlap between the two as the endpointed utterance. The first measure is log energy, while the second is based on the distance between the current frame and a noise vector. Two thresholds are set, both of which must be crossed before a start point is accepted. Similarly, the measure must fall below both of the thresholds before the end point is accepted. A number of other rules are used to determine whether the utterance is too short or long, and to add fixed offsets to the chosen endpoints. One feature of this endpointer, important for a fast response, is that it operates in a frame synchronous manner, that is, an utterance is processed as it is spoken.

5.2 The Viterbi search based end-pointer (VE)

The VE uses connected recognition techniques to perform utterance segmentation. Noise/silence models are used in combination with the usual vocabulary models and a simple recognition syntax, equivalent to the one-level level building strategy used in [8]. To generate the endpoints, a VQHMM recognition pass was run using the syntax shown in Figure 3. The Viterbi segmentation was then extracted, providing endpoints for the vocabulary segment of each utterance.

5.2.1 Noise/silence models. Two noise/silence models were used. The first was a simple 1-state model trained on all the data in the training set which lay outside the manual endpoints. For the second an 8-state 'insertion' model was used. Its training data was identified by first performing an unconstrained syntax recognition pass on the training set and then isolating all the segments which had been identified as insertion errors by a DP algorithm [9].

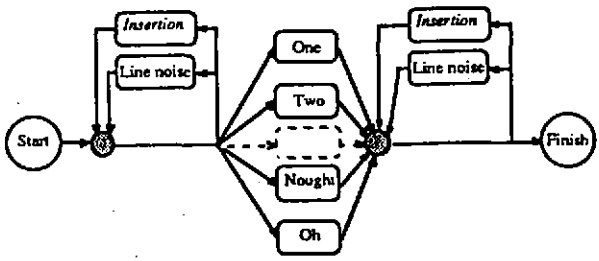

Figure 3 - Restricted syntax used for segmentation

6. RESULTS FOR THE DIFFERENT ENDPOINTERS

The histograms in Figures 4 & 5 show how the RE placed its start and end points relative to the manually placed ones, while Figures 6 & 7 show the corresponding histograms for the VE. The extreme bins of the histograms (at ±20 frames) include any utterances with endpoints which lay outside the ±20 frame range.

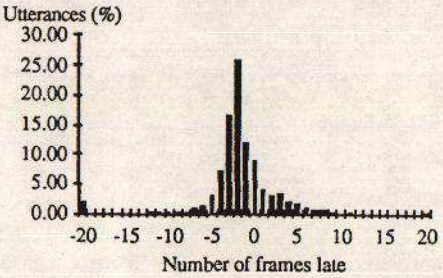

Figure 4 - RE start-points relative to manual start points

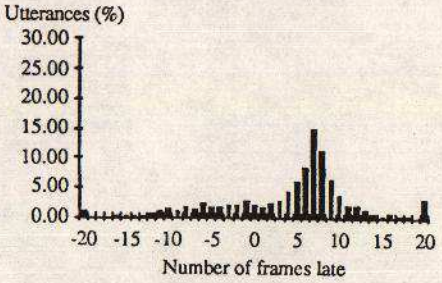

Figure 5 - RE end points relative to manual end points

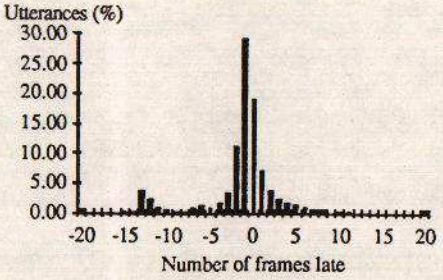

Figure 6 - VE start points relative to manual start points

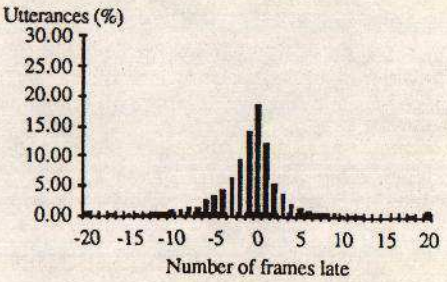

Figure 7 - VE end points relative to manual end points

It is noticeable from these diagrams that the two endpointers perform similarly in detecting speech onset, while differing markedly in their ability to detect the end of speech. If one examines the proportion of utterances where each endpointer marked the end of speech as being within 5 frames of the manual endpoint, the VE achieved 82% whilst the RE only managed 30%. The extremely low figure for the RE was almost totally due to end points that were later than the manual end points.

A striking feature of the RE plots is the skewed nature of the distributions, particularly for end of speech detection. For both the start and end points, the tails on the plots show that the RE was cutting into the speech in spite of being biassed to start early and finish late.

In contrast with the RE, the VE histograms are approximately symmetric. The small bump seen at around -15 frames in the start point histogram might be explained by the assimilation of lip-smacks and similar noises into the start of the words.

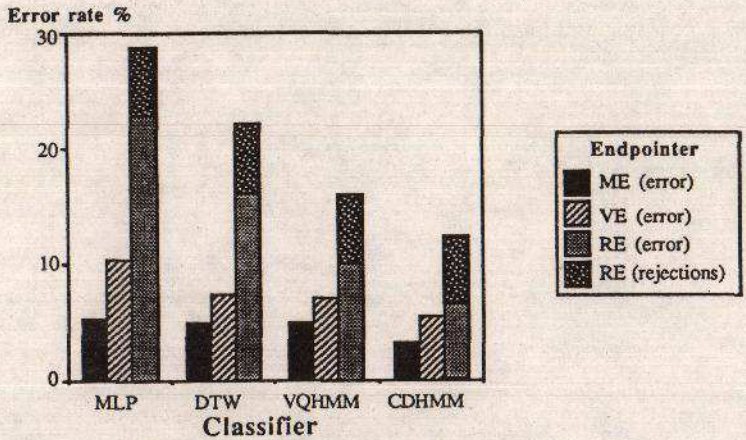

Figure 8 - Endpointer results for various classifiers

The recognition error rates for the endpointers are shown for each of the classifiers in Figure 8. The ordering of their performance is clear and consistent: The RE was always worse than the VE, which was in turn always worse than using manual endpoints. This is consistent with tests which showed that recognition error rate is closely correlated with endpointer error. With the manual endpoints, the only classifier to distinguish itself from the rest was the CDHMM. However, when moving to the VE and the RE a clear pattern emerged in the ability of the classifiers to tolerate endpointer error, the ranking being CDHMM, VQHMM, DTW then MLP (going from best to worst).

7. DISCUSSION & CONCLUSIONS

7.1 Statistical significance of results

It is important when comparing recogniser performance figures to check that any differences which are observed in recogniser performance are statistically significant. McNemar's test [10] was used to compare a number of the results and, where differences in the error rates were apparent, gave high levels of confidence that the recognisers were different. For example, comparison of the CDHMM and VQHMM classifiers trained on the ABCDE set (see Figure 2) showed a difference significant at >99.9% level, while changing the training set of the VQHMM from ABCD to ABCDE produced a difference significant at the 85% level.

7.2 Comparison with other work

Comparison of the results presented in this paper with other published results is difficult as only a few papers deal with the recognition of telephony quality digits. However, some observations can be made. Our results seem to fall between those on 'cleaner' data [11] and those on more realistically recorded data [12].

In [11] results are presented for SI IW recognition of a digit vocabulary (11 digits – 'nought' is not present) with a Linear Predictive Coding (LPC) based front end. Unfortunately, the nature of the endpointer used was not made clear. A 12 template DTW system gave SI error rates of between 1.2% and 1.6%. A CDHMM system gave errors between 0.5% and 1.6%. No rejections were indicated for either system. These results are substantially better than ours, the best of which are for CDHMMs with error rates of 3.2% for MEs and 5.5% for VEs. A major cause of our poorer results may be the nature of the database. Our database was recorded over a noisy analogue network (hopefully representing worst case conditions), and the calls were made from a wide variety of environments. On the other hand, the database in [11] was recorded over a 'local' connection with all the talkers using the same handset in a silence cabinet.

[12] presents results on what the authors term pseudo-isolated digits (extracted from a connected digit sequence and with a least 200ms of silence preceding and following either digit) collected by automatic call interception equipment. Three commercial IW recognisers gave results of between 5% and 17.3% digits incorrect with 19.1% to 12.5% rejections (the higher the rejection rate, the lower the number of errors) when tested on this data. The results presented in this paper are generally superior to those in [12], our best results our considerably better. We probably gained by having our training and test data recorded on the same equipment. However, tests we have performed on data collected over different equipment and lines have yielded very similar results to the ones presented here.

7.3 Choice of endpointer

Our results clearly show that the VE is superior to the RE irrespective of the choice of classifier. While the VE as described here does not allow frame synchronous recognition (the segmentation being performed prior to the classification), tests using the Viterbi search directly for recognition gave very similar results. The VE has a further advantage in that it has no heuristically determined parameters, unlike the RE case where the thresholds are set manually. Unsurprisingly, using manual endpoints outperforms both of the automatic techniques.

7.4 Amount of training data

On the question of how much training data is required, our answer can only be to use as much as possible, as none of the classifiers we tested appeared to have saturated even on our largest training set. For the HMM classifiers the expected smooth reduction in the test set error rate was observed as the training set size increased. For the MLP and DTW systems, however, the improvement was not so smooth. We believe that this can be attributed to the nature of the training algorithms. The HMM training algorithm guarantees to move towards a local minimum, and can be started with good initial parameter estimates. The gradient descent MLP algorithm will also move towards a local minimum, but choosing where to start is a problem. Our use of 3 random starts for the MLP training appears to have been insufficient to guarantee convergence towards a good minimum. The DTW clustering algorithm, meanwhile, does not even guarantee to produce a local minimum and unlike the MLP algorithm has no capacity to use multiple random starts.

7.5 Choice of classifier

The MLP's major deficiency in the recognition of speech is its inability to cope with temporal distortion. This was reflected in its very poor performance when presented with the additional distortions introduced by the RE. In order to determine what proportion of the MLP's errors was due to the compression of its input to 20 frames, the DTW recogniser was trained and tested on the compressed data; this resulted in a performance degradation of 0.9%. Irrespective of the training

set size or the endpointer employed, the CDHMM classifier always outperformed the other classifiers by a significant margin. Although the CDHMMs make more assumptions regarding the form of their output distributions than their VQ counterparts, it would appear that this factor was outweighed by the distortions introduced in the vector quantisation process. This belief is supported by the results of tests which showed a similar drop in accuracy when the input features to a DTW system were quantised. The CDHMMs may also have gained by being better able to generalise to unseen data than the VQ models.

8. REFERENCES

[1] G Chollet & C Gagnoulet, 'On the Evaluation of Speech Recognisers and Databases Using a Reference System', Proc. ICASSP 1982, Vol. 3, pp 2026-2029

[2] L R Rabiner, S E Levinson, A E Rosenberg & J G Wilpon, 'Speaker-Independent Recognition of Isolated Words Using Clustering Techniques', IEEE Trans. on ASSP (1979), Vol. 24, No. 4

[3] J G Wilpon & L R Rabiner, 'A modified k-means clustering algorithm for use in speaker independent isolated word recognition', IEEE Trans. ASSP (1985), Vol. 33, No. 3

[4] K-F Lee, 'Automatic Speech Recognition – The Development of the SPHINX System', Kluwer Academic Publishers (1989), London

[5] L R Rabiner, 'A tutorial on hidden Markov models and selected applications in speech recognition', Proc. IEEE, Vol. 77, No. 2

[6] S J Cox, 'Hidden Markov models for automatic speech recognition: theory and application', British Telecom Technology Journal (1988), Vol. 6 No. 2, pp 105-115

[7] D E Rumelhart, G E Hinton & R J Williams, 'Learning internal representations by error propagation', in *Parallel Distributed Processing*: Explorations in the Microstructure of Cognition, eds Rumelhart & McClelland, MIT Press (1986), pp 318-362

[8] J G Wilpon & L R Rabiner, 'Application of hidden Markov models to automatic speech endpoint detection', Computer Speech and Language (1987) Vol. 2, pp 321-341

[9] M J Hunt, 'Evaluating the Performance of Connected-Word Speech Recognition Systems', Proc. ICASSP 1988, Vol. 1, pp 457-460

[10] L Gillick, S J Cox, 'Some Statistical Issues in the Comparison of Speech Recognition Algorithms', Proc. ICASSP 1989, pp 532-535

[11] L R Rabiner & J G Wilpon, 'Some performance benchmarks for isolated word speech recognition systems', Computer Speech and Language (1987) Vol. 4, pp 343-357

[12] D Yashchin, S Basson, N Lauritzen, S Levas, A. Loring, J Rubin-Spitz, 'Performance of Speech Recognition Devices: Evaluating Speech Produced Over the Telephone Network', Proc. ICASSP 1990, pp 552-555

Acknowledgments

The authors particularly wish to thank Gavin Smyth (BTRL) for doing the MLP experiments and Frank Scahill (BTRL) for providing the RE software. Thanks also to our many colleagues at BTRL who assisted with this work, especially those people that helped with the database preparation.