CONNECTIONIST EVIDENCE COMBINATION IN AUTOMATIC SPEECH RECOGNITION

Dave Abberley and Phil Green

Speech and Hearing Research Group, Dept. of Computer Science, University of Sheffield

1. INTRODUCTION

Hidden Markov Models (HMMs) have become the predominant solution to the continuous speech recognition problem due to their high performance in terms of recognition accuracy and computational efficiency. The power of HMMs arises from their ability to model both the dynamic and acoustic properties of a speech signal in such a way that a speech utterance can be decoded efficiently by the Viterbi algorithm to produce a recognition of the utterance. The ability of HMMs to model the speech signal is limited, however, by the fundamental assumptions which make them efficient. For example, the conditional independence assumption means that HMMs treat successive speech frames as uncorrelated with each other so that HMMs can't take advantage of any correlations which are present.

Neural networks (NNs) have emerged as a powerful tool for estimating posterior class probabilities in static pattern recognition tasks [1], but have problems with dynamic signals such as speech. Consequently, there is much interest in constructing hybrid HMM/NN systems which combine the best aspects of both paradigms. For example, one technique is to use a neural network classifier to estimate the HMM emission probabilities which improves discrimination between classes (e.g. [2]).

This paper demonstrates a different approach to improving the performance of a set of HMMs, continuing one of the main themes of the SYLK project [3]. The philosophy behind the approach is to use speech knowledge to augment the discrimination ability of a set of HMMs through a series of Refinement Tests. A Refinement Test is designed using phonetic knowledge to resolve potential confusions which arise from the HMM estimates of segment probabilities. Refinement Tests are free to perform whatever further processing of the speech signal is necessary to express this additional knowledge. In this work the Refinement Tests have been based on MLP classifiers and we report on two ways of combining Refinement Test measures with HMM segment probabilities. Our early results show that it is possible to improve on a set of HMMs in this way and that this is a promising avenue of research.

2. EVIDENCE COMBINATION

Work in the field of pattern classification has shown that combining sets of classifiers can produce combined performance better than that of the best individual classifier. Xu *et al* [4] proposed several schemes and demonstrated how combination improved performance on an optical character recognition task. In addition, they outlined three levels at which classifiers could be combined, depending on the information produced by the classifiers:

1. Abstract level. The individual classifiers produce a hard decision only, i.e. the 'winning' class label. The classifiers are combined by a voting scheme. This method is particularly useful when the classifiers produce different types of measurement (e.g. probabilities, distances, etc.).
2. Rank level. The individual classifiers produce a ranking of the classes under consideration. These rankings are combined to produce an overall ranking.
3. Measurement level. The individual classifiers produce a measurement value for each class (e.g. posterior probabilities, distances, etc.). These measurements are combined to produce an overall mea-

surement which can then be used to make the classification decision. The measurements produced by the classifiers may have to be converted to the same scale (e.g. distances converted to probabilities) before they can be combined.

We are interested in combination at the measurement level because we need information which can aid in the segmentation of the speech signal as well as its classification.

2.1. Combination of Neural Networks

Several techniques have been suggested for linearly combining networks at the measurement level. For example, Perrone's Generalized Ensemble Method (GEM) [5] produces a linear combination of the outputs of a set of neural networks that guarantees combined performance which is at least as good, in the mean squared error sense, as the best individual network. This does not guarantee, however, that classification performance will not decrease. Perrone also suggested that simply averaging the outputs of a set of networks can improve performance. He called this the Basic Ensemble Method (BEM) and it has been used successfully [2] although there is no guarantee of this. For a classification problem where each output unit estimates a posterior class probability, each of these class probabilities are combined independently to produce a new estimate. It should be noted that these combination methods can be used to combine any type of classifier and not just neural network classifiers. In Section 6 we report on the use of these techniques to combine HMM and MLP classifiers.

2.2. Evidence Combination in Speech Recognition

Evidence combination schemes have been employed in a variety of speech recognition systems. For example Zavaliagkos *et al* [6] used a linear combination scheme to combine HMM and segmental NN processing of the speech signal whereas Hochberg *et al* [2] average the outputs of multiple recurrent networks trained on different representations of the acoustic signal in their integrated HMM-NN system.

3. HMM TRAINING

Before embarking on the evidence combination experiments, we trained a set of HMMs which were needed for two reasons: 1) to give us a baseline level of performance for comparison purposes, and 2) to provide us with a set of probabilities for use in our evidence combination experiments.

3.1. Data

The data consisted of all utterances by male speakers on the prototype TIMIT database except the shibboleth 'sa' utterances which were discarded because they introduce a bias in favour of phones in certain contexts, as described by Lee [7]. The data was split into three datasets: a training set of 792 utterances, a cross-validation set of 760 utterances, and a test set of 768 utterances. Thus each dataset contained roughly one third of the data. The utterances were divided among the datasets so that speaker independence was maintained. The TIMIT labels were converted into the 39-phone set devised by Lee [7].

3.2. HMM Training

The HMMs used were 3-state left-to-right monophone models. The feature vector consisted of 12 mel-frequency cepstral coefficients (mfccs) plus energy together with their deltas and accelerations to give a 39-dimensional feature vector. No language model of any type was used in these experiments, i.e. no bigram, grammar scale factor or fixed transition probabilities. This was so that we could concentrate on the acoustic performance of the models in isolation from the effect of the language model. The HMMs were trained on the training and cross-validation sets described in Section 3.1 using HTK [8]. Performance improved as the number of mixtures per state was increased. Training was halted at 20 mixtures as adding extra mixtures was only producing small performance improvements by this stage. The performance of the 20 mixture set was 70.2 %Correct and 63.0 %Accurate.

4. FORMANT FREQUENCY CHARACTERISATION

For a given time segment and for each formant (F1, F2 and F3), formant frequency Refinement Tests are based on measurements of:

Where did the formant come from?, i.e. what was its frequency before the transition into the segment,
What value did the formant reach within the segment?,
Where did it go to?

We trace back to the start of the relevant formant movement, which may be different for each formant: we do not take measurements at a fixed place. Formant frequency tracks derived by Crowe's algorithm [9] are 'characterisationed' using a technique developed in previous work [3], [10] and this semi-symbolic representation is parsed to obtain the above description of formant behaviour. In brief, the procedure is as follows:

1. The median-smoothed formant track is broken into 'time-fragments' delineated by discontinuities. The remaining steps are performed separately for each fragment.
2. Curvature extrema in the track are used to identify potential segmentation points.
3. A development of an algorithm by Bridle and Sedgwick finds a piecewise-linear approximation to the trace with the following properties:
   Breaks are allowed only at curvature extrema.
   Each line segment is the least-squares approximation to the data it describes.
   The minimum possible number of line segments is used such that each segment approximates the data within a given tolerance.
4. The line-segment-sequence is then parsed for 'targets' - peaks and dips. The definition of 'target' is quite general, allowing 0 or more line segments in its attack, steady-state and decay. A grammar and parser allows any shape-descriptions to be defined.
5. The measurements are taken from the targets entering, centred in and leaving the segment.

5. MLP CLASSIFICATION EXPERIMENTS

A series of experiments were conducted using the formant frequencies from the formant characterisation software as input to an MLP classifier. A nine dimensional feature vector containing the formant frequencies (F1, F2 and F3) was created for each segment under consideration together with its neighbouring segments. The segments were obtained from the TIMIT label files. The raw formant frequencies were normalized linearly so that they lay approximately in the range [0,1] using the maximum and minimum frequency values found in the training set. The softmax output layer [1] consisted of 39 output units, one for each class. The output coding was 1 for the correct class and 0 for all others as suggested by Bridle [1] (amongst others) so that the outputs estimate the posterior class probabilities, assuming that the net has been correctly specified and trained. The nets were trained using gradient descent with a learning rate of 0.01(no momentum) and a relative entropy error function [1] using the training set described in Section 3.1. The set of weights which produced the best performance on the cross-validation set was used to evaluate the network on the test set. The best test set performance obtained was 31.9% for a network with 80 hidden units. Nets with 20 and 160 hidden units were also tried but they performed similarly.

The most likely reason for the poor classification performance is that the formant data is not sufficient to discriminate between the entire phone set. It was not considered worth expending a great deal of effort to try and maximize the performance of the nets, so it is possible that the performance could be improved with further training, varying learning rates, etc.

In an attempt to enhance performance, the nets were combined by the ensemble methods devised by Perrone but this only resulted in an improvement of 0.5%. Two possible explanations for this modest improvement are:

1. The nets trained are too similar, i.e. they represent the same local minimum in function space and there is not much to be gained by combining them.
2. The individual nets are already performing quite well on the data and are extracting most of the discrimination information.

The first explanation is supported by analysis of the confusion matrices which shows that, for each MLP, most of the correctly recognized phones are found in a small subset of the phones (which includes the silence phone - even though it should not have a distinctive formant signature!).

6. EVIDENCE COMBINATION EXPERIMENTS

In this section we report on two schemes which demonstrate that it is possible to improve the perfomance of a set of HMMs by integrating their probability estimates with extra evidence obtained from Refinement Tests. Section 4 shows how our formant characterisation software can estimate formant frequencies for a speech segment. There is still the problem of how to identify a segment to apply our formant-based Refinement Tests to. For the purpose of this demonstration, we decided to avoid this problem by using the segments defined by the TIMIT label files. This enables us to make direct comparisons of the performance before and after evidence combination without having to account for the different segmentations which may be produced by the different methods. Obviously, our system is not a complete phone recogniser, and cannot be compared with such systems, it is merely intended to show the potential of evidence combination.

For each segment defined by the label files, the set of HMM probabilities was calculated using the forward algorithm. This gave us a set of 39 probabilities (one value for each model) for each segment in the database. The forward algorithm estimates the probability of the observation given the model. This was converted into the posterior class probability by assuming that all phones are equally likely. Implementations of the forward algorithm normally estimate log probabilities to avoid underflow and these were converted back into actual probability values. We found that the HMMs were producing 'hard' probability estimates for many segments, i.e. the largest probability value was close to one and the other probabilities very small. This happened even when the HMMs were selecting the wrong phone. We were concerned that such hard estimates could cause problems during the evidence combination phase as strong counter evidence would be required to overturn them. Consequently, we attempted to soften these estimates by dividing each log probability value by the length (in frames) of the segment before converting them back into probabilities, as suggested by McInnes *et al* [11].

Approximately 8% of the segments were shorter than three frames in duration. The shortest path through the three-state left-to-right HMMs we were using is three frames. Thus there were some segments for which the HMMs could not produce any probability estimates. These segments have been excluded from the results presented below. For some phones, the majority of examples had to be excluded for this reason (e.g. over 80% of the examples of phone /b/ were short segments).

Obviously, it was only possible to perform evidence combination on those segments for which some formant data was available. If the formant characterisation software was unable to calculate the values of the formant features for the preceding or following segments, the formant target values were used in their place. On the rare occasions when it was impossible to calculate the formant target values for the centre segment, that segment was excluded. No attempt was made to calculate formant features for the first and last segments of each utterance as these are normally silence in the TIMIT database. The HMM probabilities alone were used to make the decision for those segments where formant features were unavailable.

6.1. Evidence Combination Results

Two different schemes have been tried so far: one using Perrone's ensemble methods and the other using an MLP to perform the evidence combination. The results are summarized in Table 1. (Since we know the correct segmentation, we have presented our results in terms of %Correct only as there are no insertions or deletions to account for.)

6.1.1. Ensemble Methods. The probabilities estimated by the HMMs and the MLP classifier were combined linearly via the BEM and GEM techniques (see Section 2.1). Figure 1 contains a block diagram illustrating the procedure. For those segments where the formant data was unavailable, the HMM probabilities alone were used to make a decision. Both methods produced an overall performance improvement of 0.6%.

6.1.2. One-Stage MLP Evidence Combination. This was achieved by feeding formant values and HMM probabilities for a given segment into an MLP and training it to reproduce the correct class label (see Figure 2). The formant values were normalized in the same way as for the MLP classifier. The MLP had 48 inputs (9 formant values plus 39 duration-normalized HMM probabilities) and 39 outputs (one for each class). Other training details are the same as for the MLP classifier in Section 5. With 20 hidden units, one-stage MLP combination achieved an improvement of 2.4% over the HMMs acting alone.

Table 1: Effect of Evidence Combination on Classification Performance

Classifier	Combination Method	%Correct
MLP	None	31.9
HMM	None	70.6
HMM + MLP	BEM	71.2
HMM + MLP	GEM	71.2
HMM + MLP	MLP	73.0

7. DISCUSSION

The results show that it is possible to improve on HMM discrimination performance by adding extra information, at least within the framework of our experiment. Further work is required to discover how effective this approach will prove. There are several avenues to pursue. Firstly, what data can be used in a Refinement Test? We have demonstrated the use of Refinement Tests based on formant values, but we can use our knowledge of speech to obtain other sets of features (including other sets of formant features, e.g. bandwidth, energy) which will help discriminate between phones. We believe that the most effective sets of features will contain information that the HMMs cannot use very effectively. Formant features fall into this category because they are present over a number of consecutive frames which the HMMs treat as uncorrelated.

A second issue is how best to use the extra evidence. In the experiments reported here, the extra evidence was used to produce classifiers which attempt to discriminate between the entire phone set. It may prove more fruitful to attempt discriminations between subsets of the phone set (e.g. a vowel classifier) or just to concentrate on individual confusions (e.g. a classifier which discriminates between /s/ and /z/). This raises the question of which segments should be presented to such a classifier: if the correct label for a segment is /m/, the information provided by an /s/-/z/ classifier will not be very helpful! It may be possible to find which Refinement Tests are appropriate to a given segment from the ranking of phones produced by the HMMs. Although the probability values produced by the HMMs are unreliable, the ranking of phones based on these values appear to be more sensible: for example, the correct phone is amongst the top five on 97% of occasions.

Thirdly, there are other evidence combination schemes which can be tried and which may be more effective than the ones reported on in this paper.

We have also trained a set of HMMs which use formant data in their feature vector. Their performance was comparable with that of the standard HMMs. They will be used in a control experiment which will demonstrate whether or not it is better to use the formant data directly as part of a HMM feature vector, or indirectly via Refinement Tests.

When we have optimized the performance of the Refinement Tests and evidence combination scheme, we can then attempt to integrate them into a full speech recognition system. The better discrimination achieved by using extra evidence should not only reduce the number of substitutions but could also help produce a better segmentation of the utterance which will help with insertions and deletions as well.

In summary, we have demonstrated that it is possible to improve on the discrimination performance of a set of HMMs by using extra evidence from a Refinement Test. Further work is needed to establish how effective this approach will be but we believe our preliminary results are very encouraging.

8. ACKNOWLEDGEMENTS

Thanks are due to Dorota Iskra and Leo-Geert van den Berg of the Department of Phonetics, University of Utrecht, Netherlands for training the HMMs used in these experiments. Dave Abberley was supported by a SERC CASE award and sponsored by the DRA, Malvern.

9. REFERENCES

[1] J S BRIDLE, 'Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition', NATO ASI Series, Vol. F68, Neurocomputing, eds. F Fogelman Soulie and J Herault, Springer-Verlag, 1990.

[2] M M HOCHBERG, S J RENALS, A J ROBINSON & D J KERSHAW, 'Large Vocabulary Continuous Speech Recognition using a Hybrid Connectionist-HMM System', Proc ICSLP-94, Vol. 3, pp1499-1502, 1994.

[3] P D GREEN, N R KEW & L A BOUCHER 'Experiments with the SYLK Speech Recognition System', Proc Inst Acoustics 14 (6), pp25-32, 1992.

[4] L XU, A KRZYZAK & C Y SUEN, 'Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition', IEEE Trans. on Systems, Man and Cybernetics, 22(3) pp418-435, May/June 1992

[5] M P PERRONE, 'Improving Regression Estimation: Averaging Methods for Variance Reduction with Extensions to General Convex Measure Optimization', PhD Thesis, Dept. of Physics, Brown University, May 1993.

[6] G ZAVALIAGKOS, S AUSTIN, J MAKHOUL & R SCHWARTZ, 'A Hybrid Continuous Speech Recognition System using Segmental Neural Nets with Hidden Markov Models', Int. J. of Pattern Recognition and Artificial Intelligence, 7(4), pp949-963, August 1993.

[7] K-F LEE, 'Automatic Speech Recognition: the Development of the SPHINX System', Kluwer Academic Publishers, 1989.

[8] S J YOUNG, 'HTK Version 1.4: a Hidden Markov Model Toolkit', Cambridge University Engineering Department, 1992.

[9] A S CROWE, 'Generalised Centroids: a New Perspective on Peak-Picking and Formant Extraction', Proc. 7th Symposium of FASE (SPEECH '88), eds. W A Ainsworth and J N Holmes, pp683-690, IoA 1988.

[10] P D GREEN, G J BROWN, M P COOKE, M D CRAWFORD & A J H SIMONS, 'Bridging the Gap between Signals and Symbols in Speech Recognition', Advances in Speech, Hearing and Language Processing, ed. W A Ainsworth (JAI Press, 1990), pp149-191.

[11] F R MCINNES, Y ARIKI & A A WRENCH, 'Enhancement and Optimisation of a Speech Recognition Front End Based on Hidden Markov Models', Proc Eurospeech 89, Vol. 2, pp461-464, Paris 1989.

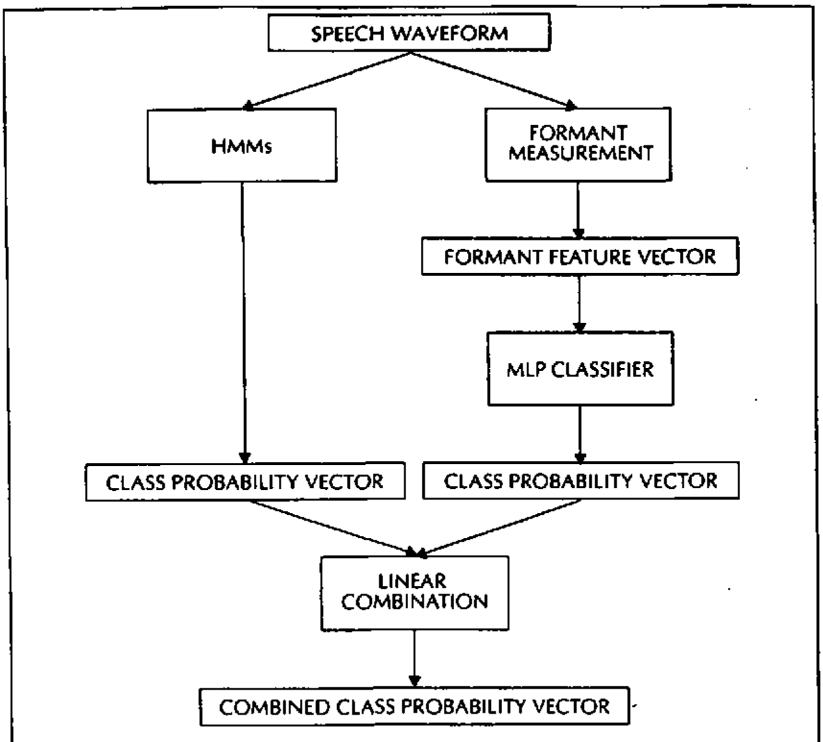

Figure 1: Linear Combination of Probabilities for a Speech Segment

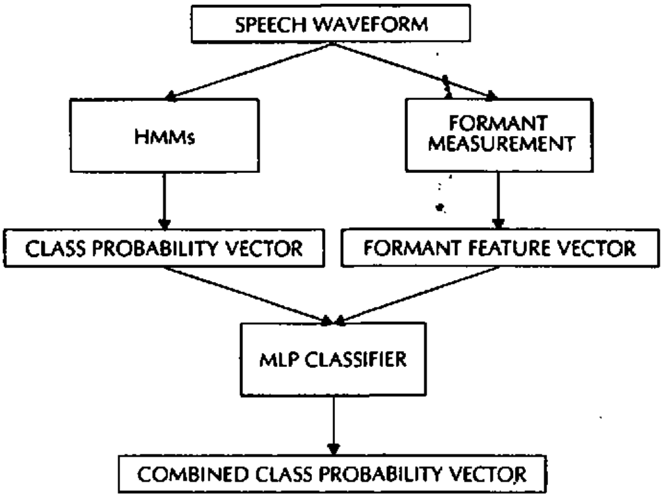

Figure 2: Evidence Combination by MLP for a Speech Segment