IPA: IMPROVED PHONE MODELLING WITH RECURRENT NEURAL NETWORKS

Tony Robinson *Mike Hochberg* *Steve Renals*

Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
UK

ABSTRACT

This paper describes phone modelling improvements to the hybrid connectionist-hidden Markov model speech recognition system developed at Cambridge University. These improvements are applied to phone recognition from the TIMIT task and word recognition from the Wall Street Journal (WSJ) task. A recurrent net is used to map acoustic vectors to posterior probabilities of phone classes. The maximum likelihood phone or word string is then extracted using Markov models. The paper describes three improvements: connectionist model merging; explicit presentation of acoustic context; and improved duration modelling. The first is shown to provide a significant improvement in the TIMIT phone recognition rate and all three provide an improvement in the WSJ word recognition rate.

1. INTRODUCTION

Many state-of-the-art speech recognition systems model phones or phone-like units as the basic linguistic component. This is necessary because the vocabulary size of current systems make full word modelling very difficult, if not impossible. Word recognition is performed by constructing word models from the basic phone models. Because the word models and, subsequently, word recognition performance depend on the phone models, it is very important to develop good phone models [1]. This paper describes phone modelling improvements made to a hybrid connectionist-hidden Markov model (HMM) speech recognition system. The improved system is evaluated on the TIMIT phone recognition and Wall Street Journal (WSJ) word recognition tasks.

The speech recognition system uses a recurrent network to map each frame of acoustic vectors to posterior phone probabilities. The network outputs are used as estimates of the observation probabilities within an HMM framework. The maximum likelihood phone or word string is then extracted using standard decoding techniques. A more complete description of the basic system can be found in [2, 3, 4].

This approach to speech recognition has proven to be very successful. Recent modifications to the original baseline system have resulted in still better recognition performance. The TIMIT results obtained with the updated system are the best reported to date. The WSJ results recently submitted for the November 93 evaluation are competitive with state-of-the-art HMM speech recognition systems. This paper reports on three improvements made to the original system. The improvements include:

- The incorporation of explicit context in the input representation to the network.
- The merging of different networks to improve the estimate of the posterior phone probabilities.
- The use of phone duration models to augment the connectionist framework.

2. RECOGNITION TASKS

Whilst there are applications for phone recognition alone, the main reason for developing a good acoustic modelling system is to improve word recognition accuracy. In [1], phone recognition performance has been shown to be an indicator of word recognition performance. However, a high performance word recognition system requires accurate estimates of the probabilities of all phones, while a phone recognition system can achieve good performance through accurate classification. Hence the system is evaluated on both the TIMIT phone recognition task and the WSJ word recognition task. (The full recognition systems for the TIMIT and WSJ tasks are referred to as IPA and ABBOT, respectively.)

2.1. TIMIT

TIMIT is one of the standard speech corpora for the evaluation of phone recognition systems. It is divided into 462 training speakers and 168 test speakers. Each speaker utters two calibration sentences and eight sentences that are used in these evaluations, giving a training set of 3696 sentences and 1344 test sentences. A summary of current results on this task can be found in [4]. The task chosen for this paper was to recognise the full set of 61 symbols, although there are good grounds for omitting the sentence initial and sentence final silences.

It should be noted that the TIMIT task has been investigated for several years, so there is a possibility that recognition systems are becoming indirectly tuned to the test set. However, TIMIT still represents a very worthwhile task. The error rates are such that considerable progress is required before systems are substantially over-tuned to the task.

2.2. WSJ

The Wall Street Journal is the current NIST/ARPA large vocabulary recognition task. The training data used was the short-term speakers from the WSJ0 corpus consisting of 84 speakers uttering a total of 7,200 sentences. Results for the hybrid system were obtained for the November 1992 (si_evl5.nvp) and November 1993 (h2_c1) NIST/ARPA evaluation test sets. These tests have a closed 5,000 word, non-verbalized pronunciation vocabulary. The results were obtained using the standard bigram language model developed at MIT Lincoln Laboratory [5] and the pronunciation

0-7803-1775-0/94 $3.00

lexicon provided by Dragon Systems [5]. There are 330 utterances from eight speakers in the November 92 test set and 215 utterances from ten speakers in the November 93 test set. The November 93 results were submitted to NIST as part of the official evaluation. At the time of writing, the adjudication period was still in force so the results should be treated as preliminary.

3. DEFINITION BASELINE SYSTEM

The basic acoustic modelling system is illustrated in figure 1 and fully described in [4]. At each 16ms frame, an acoustic vector, $\mathbf{u}(t)$, is presented at the input to the network along with the current state, $\mathbf{x}(t)$. These two vectors are passed through a standard single layer, feed-forward network to give the output vector, $\mathbf{y}(t-4)$, and the next state vector, $\mathbf{x}(t+1)$. The output vector represents the posterior probability of each of the phone classes and is delayed by four frames to allow for forward acoustic context. The state vector provides the mechanism for modelling context and the dynamics of the acoustic signal.

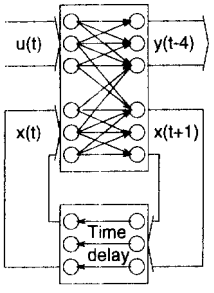

Figure 1. The recurrent net used for phone probability estimation.

This network is trained by back-propagation through time (details may be found in [4]). The dynamics of the training algorithm are quite sensitive to the interval over which the gradient vector is accumulated before changing the weights in the network. Experience indicates that better parameter values are found by decreasing this interval. If, however, the interval is made too small, the network fails to converge. The results in this paper are for the smallest update interval value that was found to be stable, although sometimes it was not possible to make an exhaustive search of all possible intervals.

4. ACOUSTIC FEATURE REPRESENTATION

The baseline system accepts one frame of acoustic features at every time step. An evaluation of different acoustic features was performed and two approaches gave approximately equal performance [3]. The two types of acoustic features used in this paper are:

MEL+ a 20 channel mel-scaled filter bank with voicing features [2],

PLP 12th order perceptual linear prediction [6].

Most HMM-based systems augment the feature vector with the first and sometimes second order differences / temporal derivatives in order to better incorporate acoustic context. It is possible for the recurrent net structure to internally compute time derivatives of the acoustic vectors and to represent short-term context within the dynamics of recurrent states. However, there are two advantages in explicitly presenting contextual information as additional features:

- Storage is expensive in a fully connected recurrent network. The number of parameters in the model scales as the square of the dimension of the state space. Hence it makes sense to use the internal state for slowly varying features and present the short term context explicitly.
- The derivatives of the acoustic vectors are known to correlate well with the phone classes. Hence by building in this prior information, it is expected that the required mapping will be easier to extract using the network.

Two approaches were used to incorporate temporal information in the acoustic vector:

- Incorporation of channel "deltas" calculated by linear regression over five frames.
- Use of a five frame window of acoustic vectors that are presented to the recurrent net.

The results of these experiments can be seen in table 1. The multi-frame TIMIT result was not available but related experiments suggest this figure would be similar to the other listed results. As can be seen, incorporation of these changes had little impact on the TIMIT results, but made a substantial improvement on the WSJ task. Due to the substantial effort required with WSJ, no temporal derivatives results are available for the Nov92 and Nov93 tests.

Input Representation	Error %		
	TIMIT	Nov92	Nov93
Single Frame	30.7	12.3	17.5
Multiframe	n/a	11.0	16.5
Derivatives	30.5	n/a	n/a

Table 1. The effect of different input representations on recognition performance for the TIMIT, WSJ November 92 si_evl5.nvp, and WSJ November 93 h2_c1 tasks.

5. CONNECTIONIST MODEL COMBINATION

Connectionist model combination refers to the process of merging the outputs of two or more networks. Recent work in merging decision trees [7] and connectionist networks [8] has indicated that combining a set of models will produce a higher performance system compared with simply choosing the best single model. In the experiments performed here, a linear combination of the network outputs is performed to generate the merged output. In each experiment, all models were weighted equally. This is optimal in the case where the single networks perform equally well in isolation and the errors are independent of the acoustic information. More research into different weighting schemes and non-linear combinations of the networks is planned for the future.

The original motivation for model merging with the hybrid system came from analysis of the recurrent network. Unlike a standard HMM, the recurrent net structure is time asymmetric. Training a network to recognise forward in time will result in different dynamics than training to recognise backwards in time. As different information is available to both processes, it seems reasonable that better modelling can be achieved by combining both information sources.

As shown in table 2, a network was trained on the time reversed acoustic vectors for the TIMIT task and found to yield the same recognition rates as the standard system. However, the distribution of errors is different and combining the phone probabilities from the two systems reduces the number of errors made by 8%.

Model	correct	insert.	subst.	delet.	errors
forward	72.7%	3.4%	21.1%	6.2%	30.7%
backward	72.9%	3.7%	21.0%	6.1%	30.8%
combined	74.2%	2.6%	19.2%	6.6%	28.4%

Table 2. Conectionist model merging for the TIMIT task.

The same approach was also taken for the WSJ tasks. In addition, networks trained with different input features were also merged with forward and backward representations. As in the TIMIT case, the distribution of errors are different between the different acoustic feature representations. Table 3 shows the merging results on the WSJ tests. Here both the MEL+ and PLP acoustic vectors are used. In this case the word error drops by 16% for three times as many parameters.

Model Type	Word Error %	
	Nov92	Nov93
Forward MEL+	10.7	16.5
Forward PLP	10.8	16.7
Backward PLP	11.2	15.5
f-MEL+ & f-PLP	8.9	14.8
f-MEL+ & b-PLP	8.2	14.0
f-PLP & b-PLP	8.6	13.7
f-MEL+ & f-PLP & b-PLP	7.9	13.8

Table 3. Model merging results for the WSJ November 92 si_ev15 nvp and WSJ November 93 h2_c1 tasks.

6. DURATION MODELLING

The original hybrid system employed a single state in the Markov chain for each of the modeled phones. Although this has been effective for previous phone classification experiments, the resulting model of phone duration is restricted to a simple, geometric distribution. By expanding the underlying phone model from a single state to multiple states with tied-output probability estimates, it is possible to obtain more sophisticated distributions on phone duration. A simple form of this approach has been employed in a word recognition system where pseudo-poisson and shifted exponential distributions have been used to model phone duration.

It is also possible to incorporate a more sophisticated duration distribution into the phone models. Work by Crystal and House has shown that the duration distribution of a variety of phonetic segments can be closely modeled with an HMM [9]. In this HMM, the observation distributions are known and only the Markov process parameters need to be estimated. Phone duration is modeled by replacing the single state with the Markov chain from the HMM. The acoustic model of the phone is then tied across all the states. A diagram of the process of mapping the single-state phone model to the more complex duration model is shown in figure 2. By using a first-order Markov chain with a left-to-right transition structure (as shown in the figure), the resulting duration models will have a gamma-like distribution. The process of generating a duration distribution can be described as follows:

1. Generate the duration distribution for each phone from labeled data,
2. Train an HMM to model the duration distribution,
3. Insert the phone duration HMM into the hybrid system.

This process is used to model a variety of distributions. The distributions evaluated for this paper include:

Single State: Baseline duration model.

Shifted Exponential: Similar to single state duration model except a minimum duration is enforced.

Psuedo-Poisson: This is a left-to-right Markov chain with no skips where the number of states is 1/2 the average duration and all transition probabilities are 0.5.

Poisson: A Poisson distribution is approximated by a Markov chain estimated as an HMM.

Gamma: A gamma distribution is approximated by a Markov chain estimated as an HMM. The parameters of the gamma are estimated using the method of moments.

3 Gamma: A three parameter gamma distribution is approximated by a Markov chain estimated as an HMM. The shift parameter is estimated as the minimum duration.

These phone duration distributions were evaluated on the TIMIT and Wall Street Journal November 93 Hub 2 tasks and the results are shown in Table 4.

Duration Model	Note	Error Rate %	
		TIMIT	WSJ
Single State		28.4	18.3
Shifted Exponential (1)		28.4	15.0
Shifted Exponential (2)	i	28.2	n/a
Psuedo-Poisson		28.3	13.8
Poisson		28.2	14.8
Gamma (1)		28.2	14.9
Gamma (2)	ii	28.8	13.9
3 Gamma (1)		28.3	13.7
3 Gamma (2)	i	28.4	n/a
3 Gamma (3)	ii	28.9	13.8
3 Gamma (4)	i, ii	29.0	n/a

Table 4. Different phone duration models evaluated on the TIMIT and WSJ November 93 Hub 2 tasks. TIMIT results are phone error rates and WSJ results are word error rates. The notes indicate the following: (i) 10% of training data with duration less than minimum duration and (ii) variance scaled by 0.5.

The results in the table indicate that duration modelling has a relatively minor effect on phone recognition rates. The differences over phone recognition performance is minor. In fact, most of the difference are probably attributable to differences in the insertion-deletion ratios. These can be "twiddled" (e.g., scale the language model) to pick-up 0.2%. On the other hand, phone duration modelling has a substantial effect on word recognition accuracy.

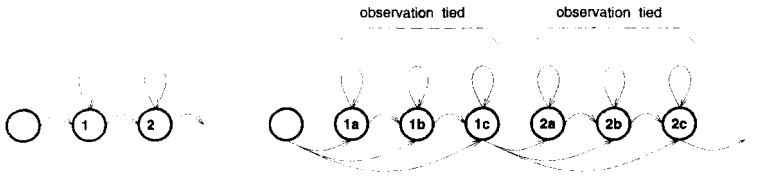

Figure 2. Mapping of single-state phone model to multi-state duration model.

7. CONCLUSION

This paper has presented several improvements to phone modelling with applications to both phone recognition and large vocabulary recognition. The improvements are summarised as follows:

- Adding explicit contextual information to the input representation reduced the error rate by 10% on the WSJ task. The effect on phone recognition results were not so dramatic.
- Combining the outputs of models with different input representations (i.e., time-reversed, different feature extraction method) improved the TIMIT and the WSJ error rates by 8% to 25%.
- Augmenting the recurrent network with more sophisticated duration models provides a little improvement on the TIMIT task but 25% fewer errors on the WSJ task.

One of the main observations resulting from this work is the effect of phone modelling on phone recognition and word recognition results. The relative improvements seen on TIMIT and WSJ have highlighted the differences between the two tasks. In the case of combining models, both the TIMIT and WSJ tests realised substantial reductions in error rate through improved phone classification of the individual frames. Because the TIMIT task is nearly a phone classification task, the improved frame classification results in better phone recognition. The model merging also results in a smoothing of the frame probabilities which helps with the word recognition task.

In the case of context and duration modelling, the focus of the improvements was on the capturing of phone information over time. This should certainly improve the word modelling capability because it reduces the likelihood of phone sequences with arbitrary durations and contexts. On the other hand, neither should really help much with the frame classification problem and this is reflected in the relatively minor improvements on the TIMIT task.

As a final point, the hybrid connectionist-HMM approach continues to show great promise as an alternative to traditional HMM-based methods. The TIMIT results are the best reported results known to the authors. The Wall Street Journal results, although not the best, are competitive with state-of-the-art speech recognition systems. Even when three models are merged the system still only has 330,000 parameters, which is far fewer than HMM systems with similar performance.

ACKNOWLEDGEMENTS

This work was partially funded by ESPRIT project 6487, WERNICKE. Two of the authors (T.R. and S.R.) are supported by SERC fellowships. The authors would like to acknowledge MIT Lincoln Laboratory for providing the grammar and Dragon Systems for providing the pronunciation lexicon for the WSJ task.

REFERENCES

[1] L. F. Lamel and J. L. Gauvain. High performance speaker-independent phone recognition using cdhmm. In *Proceedings of 3rd European Conference on Speech Communication and Technology*, volume Volume 1, pages 121–124, Berlin, September 1993.

[2] Tony Robinson. Several improvements to a recurrent error propagation network phone recognition system. Technical Report CUED/F-INFENG/TR.82, Cambridge University Engineering Department, September 1991.

[3] Tony Robinson, Luis Almeida, Jean-Marc Boite, Hervé Bourlard, Frank Fallside, Mike Hochberg, Dan Kershaw, Phil Kohn, Yochai Konig, Steve Renals, Marco Saerens, Joao Paulo Neto, Nelson Morgan, and Chuck Wooters. A neural network based, speaker independent, large vocabulary, continuous speech recognition system: The WERNICKE project. In *Proceedings of the European Conference on Speech Technology*, 1993.

[4] Tony Robinson. The application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, March 1994.

[5] Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proc. Fifth DARPA Speech and Natural Language Workshop*, pages 357–362, Harriman, New York, February 1992. DARPA, Morgan Kaufman Publishers, Inc.

[6] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87:1738–1752, 1990.

[7] Wray Buntine. Learning classification trees. In D. J. Hand, editor, *Artificial Intelligence Frontiers in Statistics III*, pages 182–201. Chapman & Hall, 1993.

[8] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[9] Thomas H. Crystal and Arthur S. House. Segmental durations in connected-speech signals: Current results. *J. Acoust. Soc. Am.*, 83(4):1553–1573, April 1988.