TRANSCRIBING BROADCAST NEWS WITH THE 1997 ABBOT SYSTEM

Gary Cook and Tony Robinson

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK.
email: {gdc,ajr}@eng.cam.ac.uk

ABSTRACT

Recent DARPA CSR evaluations have focused on the transcription of broadcast news from both television and radio programmes [17]. This is a challenging task because the data includes a variety of speaking styles and channel conditions. This paper describes the development of a connectionist-hidden Markov model (HMM) system, and the enhancements designed to improve performance on broadcast news data. Both multilayer perceptron (MLP) and recurrent neural network acoustic models have been investigated. We asses the effect of using gender-dependent acoustic models, and the impact on performance of varying both the number of parameters and the amount of training data used for acoustic modelling. The use of context-dependent phone models is described, and the effect of the number of context classes is investigated. We also describe a method for incorporating syllable boundary information during search. Results are reported on the 1997 DARPA Hub-4 development test set.

1. INTRODUCTION

Television and radio news programmes typically contain a wide variety of speech. Speaking styles range from planned speech from native speakers of American English, to spontaneous speech from non-native speakers. Channel conditions include clean speech recorded in a studio, speech in the presence of background music or noise, and speech recorded over telephone channels. This variety of both speaking styles and channel conditions makes the transcription of broadcast news an extremely demanding task, even for state-of-the-art systems.

This paper describes experiments aimed at improving the performance of the ABBOT system on broadcast news data. ABBOT is a large vocabulary connectionist-HMM continuous speech recognition system developed at Cambridge University Engineering Department [16]. The connectionist-HMM approach uses an underlying hidden Markov process to model the time-varying nature of the speech signal and a connectionist system to estimate the observation likelihoods within the hidden Markov model framework [10].

The layout of this paper is as follows. We first describe the DARPA Hub-4 broadcast news data that has been used for all the experiments reported in this paper. The use of MLP acoustic models is then described, and results are reported for both gender-independent and gender-dependent systems. Section 4 describes the use of recurrent neural network models. We examine the effect of both model size (in terms of the number of model parameters) and the amount of training data on recognition performance. The use of context-dependent (CD) acoustic models is then described, and results reported for systems using varying numbers of CD models.

Section 5 describes a technique for incorporating syllable boundary information during the decoding procedure. We describe the method used to determine the syllable boundary points in the training data, and how this is used to train a syllable onset detector. The syllable boundary information has been incorporated in the decoding procedure without the need to modify the decoder, and the method by which this is achieved is described. Finally we discuss further techniques for improving the performance of the ABBOT system on broadcast news.

2. THE DARPA HUB-4 DATA

The Linguistic Data Consortium (LDC) and NIST have provided both acoustic and language model training data to sites that participate in the Hub-4 broadcast news benchmark tests. The acoustic training data consists of approximately 104 hours of data recorded from a variety of television and radio programmes [4]. The acoustic data is manually segmented into homogeneous regions termed "evaluation focus conditions". This was done to support the 1996 "partitioned evaluation" (PE) paradigm [2]. These focus conditions are as follows:

F0: Baseline broadcast speech

F1: Spontaneous broadcast speech

F2: Speech over telephone channels

F3: Speech in the presence of background music

F4: Speech under degraded acoustical conditions

F5: Speech from non-native speakers

Segments that do not fall within the specification for the focus conditions presented above are labelled FX. More details of the focus conditions can be found in [5]. The development test data is also manually segmented into these focus conditions, and these segment boundaries have been used for all the experiments reported in this paper. All results are for an episode of NPR Marketplace recorded on 12 July 1996 (this episode is denoted as k960712 in the Hub-4 development test data). This consists of 30 minutes of data containing 4413 words.

Language model training data is also available for the Hub-4 task. This covers the period from January 1992 to April 1996, and contains approximately 132 million words. The language model used for all the experiments reported in this paper also incorporated the 1995 Hub-4 language modelling data, which contains 108 million words and covers general North American business news. A trigram language model and a 65,532 word vocabulary were used for all the experiments.

3. MLP ACOUSTIC MODELLING

This section describes the use of MLP's as acoustic models. The models used are fully connected with a single hidden layer consisting of 4000 logistic sigmoid units, and an output layer with softmax units. A cross-entropy error criterion is used during training, and this ensures that the model outputs are estimates of the a posteriori probability of phone class given the acoustic data [14]. The input to the network consists of nine contiguous frames of 12^{th} order perceptual linear prediction (PLP) coefficients plus log energy. The networks are trained using back-propagation and gradient descent. The gradient descent learning rate is adapted during training based on the cross-validation error. Learning proceeds with the initial (empirically set) learning rate. When the decrease in cross-validation error falls below a threshold the learning rate is reduced by a factor of two. This continues after each iteration. When the decrease in cross-validation error again falls below a threshold the learning rate is set to zero and training is stopped [9].

Focus Condition	Gender Ind. Model	Gender Dep. Models
F0	24.0	25.3
F1	37.8	42.3
F2	38.2	43.6
F3	40.4	44.2
F4	38.5	41.8
F5	34.9	42.7
FX	65.0	66.3
OVERALL	32.7	35.5

Table 1: Word error rates by focus conditions for both gender independent and gender dependent MLP acoustic models.

We looked at both gender-independent and gender-dependent acoustic modelling using MLPs. The mark up of the acoustic training data includes gender tags, and these were used to produce training sets for male and female speakers. The selection of the gender at recognition time was based on the log likelihood of the decoded utterances. All the test data was decoded using both the male and female acoustic models, and the decoded utterance with the highest log likelihood selected to form the final system output. The results are are shown in Table 1, and as can be seen the gender independent system performs better than the gender dependent system. This may be due to the relatively small (30%) proportion of training data from female speakers.

4. RECURRENT NEURAL NETWORK ACOUSTIC MODELLING

In this section we report results for both context-independent and context-dependent recurrent neural network (RNN) acoustic models. The RNN architecture provides a mechanism for modelling acoustic context and the dynamics of the acoustic signal. Training uses backpropagation-through-time and an adaptive step size algorithm for weight updates. A detailed description of the RNN architecture and training algorithm is given in [15].

The first set of experiments examine the effect of both the size of the acoustic model and the amount of training data. Table 2 shows results for a model with 256 state units (83700 parameters) trained on 35 hours of data (denoted **Model 1**), and a model with 384 state units (174324 parameters) trained on 60 hours of data (denoted **Model 2**). It can be seen that increasing the model size and the training data results in an 8.2% relative reduction in word error rate.

Focus Condition	Model 1	Model 2
F0	25.4	22.5
F1	41.8	38.4
F2	38.2	43.6
F3	44.7	39.2
F4	38.2	32.1
F5	31.8	33.3
FX	61.8	63.4
OVERALL	34.3	31.5

Table 2: Word error rates by focus conditions for RNN acoustic models with different numbers of parameters.

Comparing the results from Tables 1 and 2 shows that there is little difference in performance between the gender-independent MLP system, and a system using an RNN acoustic model (**Model 2**). Indeed, the performance difference between the two systems is not significant at $p < 0.05$[1]. However, the MLP acoustic model has four times the number of parameters of the RNN model.

4.1. Context-Dependent Acoustic Modelling

This section describes the use of word-internal context-dependent phone models. The method used to implement CD phone models is based on the factorisation of conditional context-class probabilities [7, 8]. The joint a posteriori probability of context class j and phone class i is given by

$$y_{ij}(t) = y_i(t) y_{j|i}(t), \tag{1}$$

where $y_i(t)$ is estimated by the recurrent network. Single-layer networks or "modules" are used to estimate the conditional context-class posterior,

$$y_{j|i}(t) \simeq \Pr(c_j(t)|q_i(t)), \tag{2}$$

where $c_j(t)$ is the context class for phone class $q_i(t)$. The input to each module is the internal state (similar to the hidden layer of an MLP) of the recurrent network, since it is assumed that the state vector contains all the relevant contextual information necessary to discriminate between different context classes of the same monophone. The context classes for each context module are determined by using a decision tree based approach. This allows for sufficient statistics for training and keeps the system compact (allowing fast context training).

Word error rates are shown in Table 3 for systems with different numbers of context-dependent phone models. It can be seen that the number of context-dependent models has only a small effect on recognition performance. The differences between each of the context-dependent systems are not significant at $p < 0.05$. However, introducing context-dependent models provides a significant (at $p < 0.05$) improvement over a context-independent system.

[1] Significance tests were performed using the two-tailed matched pairs method described in [3]

Focus Condition	CI System	Number of CD phone models			
		589	697	792	1002
F0	22.5	20.1	19.9	20.5	21.2
F1	38.4	34.6	33.7	35.5	34.5
F2	43.6	45.5	40.0	39.1	43.6
F3	39.2	32.2	31.4	28.8	31.2
F4	32.1	30.9	31.2	29.7	29.4
F5	33.3	35.4	34.4	34.9	37.5
FX	63.4	63.8	60.6	61.0	63.4
OVERALL	31.5	28.9	28.2	28.5	29.2

Table 3: Word error rates by focus conditions for different numbers of context-dependent phone models.

5. INCORPORATING SYLLABLE BOUNDARY INFORMATION

This section reports experiments aimed at improving recognition accuracy by incorporating syllable boundary information during search. Previous research on detecting syllable boundaries and using this information to improve recognition accuracy has been reported [18, 6]. In this work we use the method of Wu *et al* [18].

5.1. Detecting Syllable Boundaries

The broadcast news training data does not include syllable boundary or phonetic alignment information. An automatic procedure for determining syllable boundaries is therefore required. The method used in this work is based on deriving syllable boundaries from phonetic alignments. The first step in determining the syllable boundaries is to produce pronunciations with tagged syllable boundaries. Syllable tagged pronunciations are required for every word in the training data. This was done automatically using the NIST software *tsylb2*[2]. The first phone of each syllable is tagged as an onset phone. Viterbi forced alignment is then used to determine phone alignments for the training data. These can be used in conjunction with the syllable tagged lexicon to derive the syllable onsets.

A single hidden layer, fully connected MLP with 500 hidden units was trained to estimate the probability that a given frame is a syllable onset. The input to this MLP consists if 9 contiguous frames of perceptual linear prediction (PLP) features computed over a 32ms window every 16ms. For the purposes of training, the syllable onsets were represented as a series of four frames, with the initial frame corresponding to the actual onset derived from the phonetic alignments.

A simple numeric threshold applied to the probability estimates generated by the neural network determined the identification of any frame as a syllable onset. This method correctly detected 92% of the onsets derived from the phonetic alignments. However, this method also detected syllable onsets where there were none in 30% of frames outside the four-frame window defined for training. This effect can be seen in Figure 1 which shows an example of the neural network output. The width of the onsets detected tends to be much wider than the four-frame window used during training.

[2]The actual syllabification of the lexicon was done by Eric Fosler, of the International Computer Science Institute.

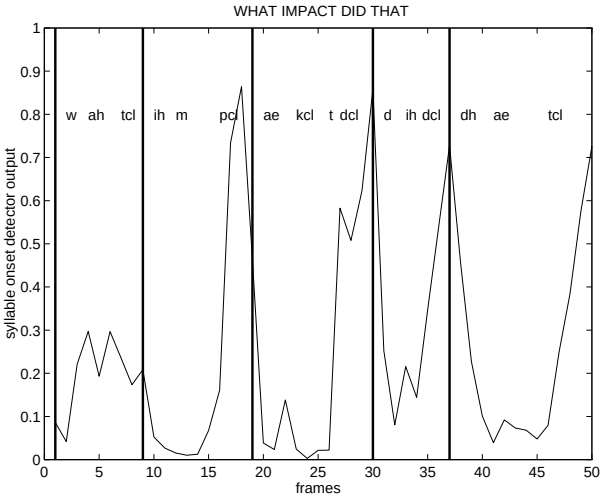

Figure 1: Example of the output of the syllable onset detector for the utterance "what impact did that". The vertical lines are the syllable onsets as derived from Viterbi aligned phone labels.

5.2. Syllable based Decoding

The NOWAY [12, 13] stack decoder was used to incorporate syllable boundary information in the decoding process. The context-independent phones may occur both at a syllable onset, or not directly after the syllable onset. This can be seen in the example pronunciation shown below in which the schwa (ax) occurs both at the beginning of the first syllable, and as the second phone of the last syllable. Phones that occur at syllable onsets are tagged with _on.

ABATEMENTS = { ax_on bcl b_on ey tcl m_on ax n tcl s }

Therefore two phone models are required for each context independent phone in the system, one model for when the phone occurs at a syllable onset, and one when it does not. The same acoustic model is used to generate the observation probabilities for the syllable onset phones and the standard (ie. not at syllable onsets) phones. This assumes that the realisation of any particular phone is not affected by whether or not it is the onset of a syllable. The observation probabilities of the onset phone models are set to zero when no onset is detected, and to those of the standard model when a syllable onset is detected. This effectively means that the decoder can only choose syllable onset phones when a syllable onset is detected, and thus allows the incorporation of syllable boundary information into a standard decoder.

The results for context-independent systems with and without syllable boundary information can be seen in Table 4. Incorporating syllable onset information has reduced the word error rate for each of the focus conditions, and resulted in an overall reduction in word error rate of 8.6% (which is significant at $p < 0.05$).

6. DISCUSSION

The experiments described in this paper have been performed as part of the development of the 1997 ABBOT system for the DARPA Hub-4 English Broadcast News Evaluation. All of the results reported in this paper are for systems using a single acoustic model.

Focus Condition	Standard CI system	CI + syllable onset system
F0	22.5	21.1
F1	38.4	32.1
F2	43.6	40.0
F3	39.2	37.1
F4	32.1	31.5
F5	33.3	32.3
FX	63.4	59.3
OVERALL	31.5	28.8

Table 4: Word error rates by focus conditions for a context-independent system, and a context-independent system incorporating syllable boundary information.

Previous work has shown that combining multiple acoustic models can lead to significant reductions in word error rate [1], and we plan to incorporate multiple acoustic models into the 1997 ABBOT system.

The transcription of broadcast news has highlighted a weakness in current techniques for large vocabulary speech recognition. Word error rates increase significantly when speech is of a spontaneous/conversational nature. This effect can be seen in the results of all the systems that participated in the 1996 Hub 4 evaluation [11]. It can be seen from the results in Table 4 that incorporating syllable boundary information has reduced the error rate on spontaneous speech (focus condition F1) by 16.4%. We plan to investigate this further, and to extend the syllable boundary work to a context-dependent system.

7. ACKNOWLEDGEMENTS

This work was partially funded by ESPRIT project 20077 SPRACH. The work on incorporating syllable boundary information is based on ideas from Steve Greenberg, Nelson Moragn, Su-Lin Wu, Mike Shire, and Eric Fosler from the International Computer Science Institute [18]. The authors wish to thank Eric Fosler and Su-Lin Wu for the assistance in implementing the syllable boundary experiments.

8. REFERENCES

[1] G.D. Cook, D.J. Kershaw, J.D.M. Christie, and A.J. Robinson. Transcription of Broadcast Television and Radio News: The 1996 Abbot System. *DARPA Speech Recogniton Workshop*, February 1997. Westdfields International Conference Center, Chantilly, Viginia.

[2] J.S. Garofolo, J.G. Fiscus, and W.M. Fisher. Design and Preparation of the 1996 Hub 4 Broadcast News Benchmark Test Corpora. *DARPA Speech Recognition Workshop*, February 1997. Westfields International Conference Center, Chantilly, Virginia.

[3] L. Gilliek and S.J. Cox. Some Statistical issues in the Comparison of Speech Recognition Algorithms. *International Conference on Acoustics, Speech and Signal Processing*, 1:532–535, 1989.

[4] D. Graff. The 1996 Broadcast News Speech and Language-Model Corpus. *DARPA Speech Recognition Workshop*, February 1997. Westfields International Conference Center, Chantilly, Virginia.

[5] Hub 4 Working Group. *Specification of the DARPA November 1996 Hub 4 Evaluation*.

[6] M.J. Hunt, M. Lennig, and P. Mermelstein. Experiments in Syllable-based Recognition of Continuous Speech. *International Conference on Acoustics, Speech, and Signal Processing*, 3:880–883, April 1980. Denver, Colorado.

[7] D.J. Kershaw. *Phonetic Context-Dependency in a Hybrid ANN/HMM Speech Recognition System*. PhD thesis, Cambridge University Engineering Department, September 1996.

[8] D.J. Kershaw, M.M. Hochberg, and A.J. Robinson. Context-Dependent Classes in a Hybrid Recurrent Network-HMM Speech Recognition System. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, Cambridge, MA 02142-1399, 1996.

[9] N. Morgan and H. Bourlard. Generalization and Parameter Estimation in Feedforward Nets: Some Experiments. In D.S. Touretzsky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan Kaufmann, 1990.

[10] Nelson Morgan and Hervé Bourlard. Continuous Speech Recognition. *IEEE Signal Processing Magazine*, 12(3):24–42, May 1995.

[11] D.S. Pallett, J.G. Fiscus, and M.A. Przybocki. 1996 Preliminary Broadcast News Benchmark Tests. *DARPA Speech Recognition Workshop*, February 1997. Westfields International Conference Center, Chantilly, Virginia.

[12] S. Renals and M. Hochberg. Decoder Technology for Connectionist Large Vocabulary Speech Recognition. Technical Report CS-95-17, Dept. of Computer Science, University of Sheffield, 1995.

[13] S. Renals and M. Hochberg. Efficient Evaluation of the LVCSR Search Space Using the NOWAY Decoder. *International Conference on Acoustics, Speech, and Signal Processing*, 1:149–152, 1996.

[14] M.D. Richard and R.P. Lippmann. Neural Network Classifiers Estimate Bayesian a posteriori Probabilities. *Neural Computation*, (3):461–483, 1991.

[15] A.J. Robinson. An Application of Recurrent Nets to Phone Probability Estimation. *IEEE transactions on Neural Networks*, 5(3), 1994.

[16] A.J. Robinson, M.M. Hochberg, and S.J. Renals. The Use of Recurrent Neural Networks in Continuous Speech Recognition. In C. H. Lee, K. K. Paliwal, and F. K. Soong, editors, *Automatic Speech and Speaker Recognition – Advanced Topics*, chapter 19. Kluwer Academic Publishers, 1995.

[17] R.M. Stern. Specification of the 1996 Hub 4 Broadcast News Evaluation. *DARPA Speech Recognition Workshop*, February 1997. Westfields International Conference Center, Chantilly, Virginia.

[18] S-L. Wu, M.L. Shire, S. Greenberg, and N.Morgan. Integrating Syllable Boundary Information into Speech Recognition. *International Conference on Acoustics, Speech, and Signal Processing*, 2:987–990, April 1997. Berlin.