Proceedings of the Institute of Acoustics

CONNECTIONIST ACOUSTIC MODELLING IN THE ABBOT SYSTEM

G. D. Cook, A. J. Robinson, and J. deM. Christie

Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK.

1. INTRODUCTION

This paper describes acoustic modelling in the ABBOT system. ABBOT is a large vocabulary, speaker independent, connectionist-HMM automatic speech recognition system developed at Cambridge University Engineering Department. The connectionist-hidden Markov model 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 (HMM) framework. The connectionist acoustic model in the ABBOT system is a recurrent neural network. The major advantage of this approach is that the recurrent network acts as a non-parametric model that is able to capture temporal acoustic context. Consequently, the basic abbot system is able to achieve very good performance using single pass decoding and context-independent phone models.

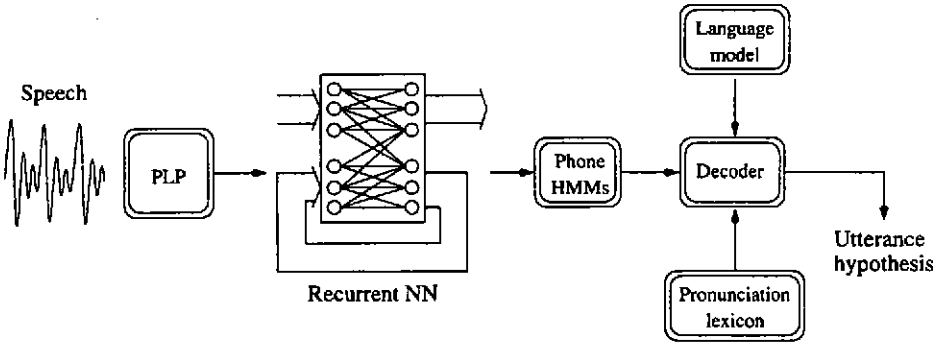

Figure 1: The ABBOT Connectionist-HMM Speech Recognition System

The basic components of the abbot system are shown in Figure 1. The acoustic waveform is segmented into 32 millisecond frames every 16 milliseconds, to which Perceptive Linear Predictive (PLP) coding is then applied. The resulting acoustic features are then normalised to unit mean and zero variance to increase the robustness to convolutional noise, and fed to the recurrent network, which maps each frame of acoustic features to a local estimate of the posterior probabilities of each phone class. This is described in more detail in the next section. The phone-class posteriors are then scaled to yield likelihoods, and fed to the decoder. The phone models used by the decoder are repeated single state HMMs, used to impose a minimum phone duration. The pronunciation lexicon provides the mapping between phone sequences and words, allowing multiple pronunciations for any word. Language modelling is implemented in the form of an n-gram model, with the baseline ABBOT system using a trigram model. The decoder merges information from the acoustic model, language model, phone models, and pronunciation lexicon, and produces the maximum likelihood word sequence.

The layout of the rest of the paper is as follows. The recurrent neural network acoustic modelling paradigm is described in Section 2. This includes a description of the model architecture and the training procedure used. Next we present results for a baseline ABBOT system on British English. This section includes a description of the training and test data. A method for combining multiple acoustic models is also described. Section 4 then introduces context-dependent acoustic modelling, and a compact method for estimating context-dependent phone probabilities is then described. Finally we describe ongoing work aimed at producing a system suitable for transcription of broadcast news television and radio programmes.

2. ACOUSTIC MODELLING

The basic acoustic modelling system [1] is illustrated in Figure 2. For each input 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, $\mathbf{x}(t+1)$. Sigmoid and softmax nonlinearities are applied to the state and output nodes, respectively. The output vector represents an estimate of the posterior probability of each of the phone classes, i.e.,

$$y_i(t) \simeq \Pr(q_i(t)|\mathbf{u}_1^{t+4}) \qquad (1)$$

where $q_i(t)$ is state i at time t and $\mathbf{u}_1^t = \{\mathbf{u}(1), \ldots, \mathbf{u}(t)\}$ is the input from time 1 to t. The output is delayed by four frames to account for forward acoustic context. The state vector provides the mechanism for modelling acoustic context and the dynamics of the acoustic signal. There is one output node per phone and the recurrent network generates all the frame-by-frame phone posterior probabilities in parallel.

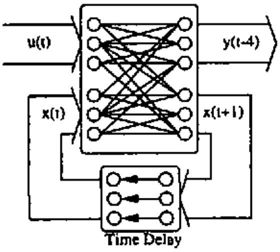

Figure 2: The recurrent network used for phone probability estimation.

The training approach is based on Viterbi training. Each frame of training data is assigned a phone label based on an utterance orthography and the current model. The recurrent network is then trained – using the back-propagation-through-time algorithm [2] – to map the input acoustic vector sequence to the phone label sequence. The labels are then reassigned and the process iterates. Initial alignments for the ABBOT system were derived from a recurrent network trained on the TIMIT database.

3. SYSTEM EVALUATION

This section describes the performance of the baseline ABBOT system on British English. We first describe the training and test data used, and then present word error rate results.

3.1 Data Sets

The data used for the evaluation of the ABBOT system is that used for the SQALE (speech quality assessment for linguistic engineering) project [3]. The aim of the project was to develop an evaluation paradigm for European multi-lingual speech recognition development. Training and test data, lexica, and language models were prescribed to allow accurate comparison of different systems. British English was one of the languages addressed by the SQALE project, and it is this data that has been used to evaluate ABBOT.

The training data for British English is the WSJCAM0 corpus [4]. This consists of 13.4 hours of data from 92 native speakers. The data was recorded in a quiet room with a noise-cancelling close-talk microphone (Sennheiser HMD414-6). The recorded sentences were all taken from the Wall Street Journal text corpus [5]. The total number of spoken words is approximately 131,000. The number of distinct words in the training data is 9084.

The results reported are for the SQALE evaluation test data. This consists of 10 sentences from each of 20 different speakers. The sentences were chosen to cover a range of perplexities, sentence lengths, and out-of-vocabulary (OOV) rates. A more detailed description of test data selection can be found in [6]. The standard SQALE 20k word vocabulary was used, which gives an OOV rate of 2.5%. The system utilises a trigram language model trained on 37.2 million words from the Wall Street Journal text corpus.

3.2 Combining Multiple Models

The ABBOT system utilises recurrent networks trained on forward-in-time and backward-in-time input sequences of PLP feature vectors. The recurrent network builds up a representation of the past acoustic context which implies the ordering of the input data is important. A significant performance improvement is achieved by merging multiple recurrent networks trained on these different input representations [7]. The most successful merging technique merges the network outputs in the log domain, i.e.,

$$\log y_i(t) = \frac{1}{K}\sum_{k=1}^{K} \log y_i^{(k)}(t) - Z \tag{2}$$

where Z is a constant to insure that $\mathbf{y}$ is a valid probability distribution.

Model	Substitutions	Insertions	Deletions	Word Error Rate
FORWARD	14.8%	4.0%	2.9%	21.6%
BACKWARD	14.3%	4.1%	2.8%	21.2%
COMBINED	12.1%	3.1%	2.6%	17.9%

Table 1: Error rates of the baseline ABBOT system on the SQALE evaluation test data.

The results for the baseline ABBOT system are shown in Table 1. As can be seen, the combination of acoustic models trained on forward-in-time and backward-in-time features gives a considerable (17%) reduction in word error rate. In addition to combining models trained on data presented forward and backward-in-time it is also possible to combine models which use different features representations. To this end we have trained models that use features derived from a 20 channel mel-scaled filter bank, plus energy, degree of voicing, and pitch [8]: these features are denoted MEL+. The combination of forward and backward-in-time PLP and MEL+ models gives a word error rate of 16.4%. This represents a 22.6% reduction in word error rate compared to the best single acoustic model.

4. CONTEXT-DEPENDENT ACOUSTIC MODELLING

Although the ABBOT system achieves good results using context-independent monophones, most state-of-the-art systems use some form of context-dependent acoustic modelling. Context-dependent modelling is designed to model the short-term contextual influence of co-articulation. This is achieved by creating models for all sufficiently differing phonetic contexts for which enough acoustic data exists to reliably train the models. The use of context-dependent modelling leads to large reduction in error rates for Gaussian mixture based HMM systems, e.g. [9]. This section describes a method for introducing context into the ABBOT.

4.1 Context-Dependent Architecture

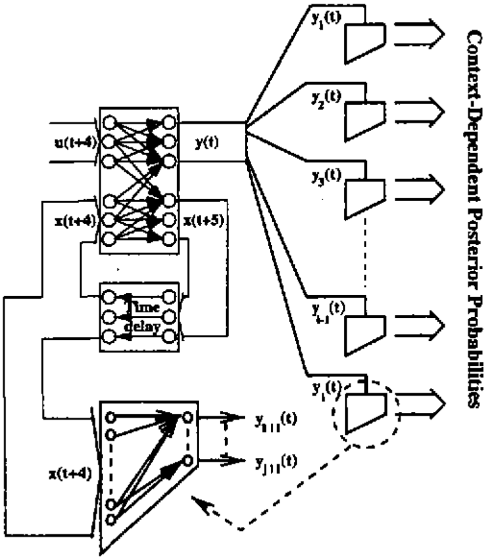

Figure 3: The phonetic context-dependent recurrent neural network modular system.

Training of large recurrent neural networks is computationally very expensive, and so a method has been developed for estimating context-dependent phone probabilities without the need to retrain acoustic models.

By using the definition of conditional probability, the factorisation of conditional context-class probabilities is used to implement phonetic context-dependency in the acoustic model [10]. The joint posterior probability of context class j and phone class i is given by,

$$y_{ij}(t) = y_i(t) y_{j|i}(t), \tag{3}$$

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

$$y_{j|i}(t) \simeq \Pr(c_j(t)|\mathbf{u}_1^{t+4}, q_i(t)), \tag{4}$$

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 [11].

Figure 3 shows the context-dependent system in operation. The outputs on the right hand side of this figure are the context-dependent posterior probabilities as estimated by Equation 3. Viterbi segmentation is used to align the training data. Each context network is trained on a non-overlapping subset of the state vectors generated from all the Viterbi aligned training data. The context networks are trained using a gradient-based procedure.

4.2 Choosing the Context Classes

Given an architecture for context-dependent modelling it is necessary to determine which context classes are to be modelled. One choice available is to model all classes seen in the training data for which sufficient data exists. This can be refined by clustering of all the classes found in the training data. However this still presents two problems: the method doesn't necessarily pick the best context classes for discrimination purposes, and with large vocabulary speech recognition it is likely that context classes will occur in the test data that do not occur in the training data. The method for overcoming these problems is a decision-tree based approach to cluster the context classes. This guarantees a full coverage of all phones in any context, and the context classes are chosen using the acoustic evidence available. The tree clustering framework also allows for the building of a small number of context-dependent phones, keeping any new context-dependent connectionist system architecture compact.

Phonetic decision trees are grown in the following manner. Each terminal in the tree has some data associated with it. A particular question is asked at a terminal node which would initially split the data into two child nodes (one for an answer of yes, and one for an answer of no). The goodness of split scoring criterion is then calculated for this preliminary split. All the questions are asked at all the terminal nodes making preliminary splits, and the terminal with the best splitting score is split, while the question asked to achieve this best splitting score is stored at the terminal that is split. The splitting continues until the best overall goodness of split score falls below some threshold. The splitting score is determined by building a Gaussian model from the data in the terminal node and each of the two child nodes. The gain in log-likelihood due to splitting the data can then be calculated.

4.3 Context-Dependent System Evaluation

The context-dependent ABBOT system has been evaluated on the same data from the SQALE project as the context-independent system. Phonetic decision trees were grown using the context-independent alignments, resulting in a system with 465 context-dependent phone classes. As before both forward and backward-in-time PLP models are used. The results can be seen in Table 2.

Model	Substitutions	Insertions	Deletions	Word Error Rate
FORWARD	13.3%	3.9%	2.3%	19.6%
BACKWARD	12.2%	5.8%	2.0%	20.1%
COMBINED	10.1%	3.3%	2.3%	15.8%

Table 2: Error rates of the context-dependent ABBOT system on the SQALE evaluation test data.

As with the context-independent system combining multiple models reduces the word error rate considerably. Comparing the results in Tables 1 and 2 it can be seen that introducing context-dependency into the acoustic modelling reduces the word error rate by 11.7%. It is also possible to combine MEL+ context-dependent models with the PLP models, and this further reduces the error rate to 13.8%.

5. RECOGNITION OF BRITISH BROADCAST NEWS

One application of the ABBOT system is to perform near-real time recognition of British Broadcast News. One use for such a system is as a component in an Information Retrieval system. Such a system is being built by the THISL (Thematic Indexing of Spoken Language) project [1]. This section briefly describes the first large vocabulary recogniser for the THISL information retrieval system. The chronos decoder was employed throughout [12].

5.1 Test set

The test set in this section was two half hour radio shows which had been accurately transcribed (the 6pm news from BBC Radio 4 on 7th January and 28th January 1998). No compensation was made for "reasonable" speech recognition mistakes such as variable word hyphenation or contractions and this typically reduces the word error rate by a few percent. The vocabulary and language model were the same as used in the 1997 CU-CON DARPA evaluation [13].

5.2 Improved acoustic models

The baseline acoustic models were those trained on the WSJCAM0 database. This consists of planned speech in an acoustically spotless environment. This is in strong contrast to the often unplanned and noisy nature of broadcast speech. The THISL project is in the process of collecting a corpus of British broadcast news. Table 3 shows the improvement in the baseline system by using matched acoustic data.

Training Data	Forward	Backward	Combined
WSJCAM0	54.8%	55.2%	-
10 hours	42.1%	-	-
20 hours, pass 1	35.7%	34.6%	32.4%
20 hours, pass 2	33.8%	35.1%	31.1%

Table 3: Retraining acoustic models

[1] http://www.dcs.shef.ac.uk/research/groups/spandh/projects/thisl/

5.3 Incorporating sentence boundaries

Our language models use a special symbol, <s>, to indicate the end of a sentence. In earlier work we expect to recognise one sentence at a time. However, in this task we are performing recognition on half an hour of audio at once. It was found that if the symbol <s> was given the acoustic realisation of at least one frame of silence, and the decoder modified so that <s> was treated like any other word, about half of the sentence boundaries were detected. Moreover, as shown in Table 4 this resulted in a modest decrease in the word error rate.

Language Model	Substitutions	Deletions	Insertions	Word Error Rate
sentence internal LM	22.3%	4.8%	4.0%	31.1%
cross-sentence LM	21.9%	4.8%	3.9%	30.6%

Table 4: Retraining acoustic models

5.4 Incorporating segmentation

The above results normalised the acoustic vectors to zero mean and unit variance over the whole half hour show then decoded the whole show. However, the show contains structure such as speaker changes and other changes of acoustic condition. A new means of performing acoustic segmentation has been developed and here it was used to provide boundaries over which the acoustic vectors were normalised. Table 5 shows that this results in a significant reduction in the word error rate.

Segmentation	Substitutions	Deletions	Insertion	Word Error Rate
no normalisation	21.9%	4.8%	3.9%	30.6%
with normalisation	20.5%	3.9%	4.3%	28.8%

Table 5: Acoustic normalisation

The work on building an information retrieval system for British English broadcast news is ongoing. The current word error rate is under 30% and this is comparable with the errors rates seen on the American English TREC evaluation data.

6. CONCLUSIONS

This paper has described the ABBOT large vocabulary continuous speech recognition system. We have described the basic acoustic modelling paradigm used, and shown how this can be extended to include context-dependent models. We have evaluated the performance of the system on read British English. It has been shown that combining multiple acoustic models, and introducing context-dependency into the system can reduce the word error rate considerably. We have also briefly described work at producing a system for the transcription of broadcast news.

7. ACKNOWLEDGEMENTS

This work has been partially supported by the ESPRIT Project 20077 SPRACH. The method for incorporating context into the recurrent neural network acoustic model was introduced by Dan Kershaw [14].

8. REFERENCES

[1] 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.

[2] P.J. Werbos. Backpropagation Through Time: What Does It Mean and How to Do It. In *IEEE*, volume 78, pages 1550–60, October 1990.

[3] H.J.M. Steeneken and D.A. van Leeuwen. Multi-Lingual Assessment of Speaker Independent Large Vocabulary Speech Recognition Systems: The SQALE Project. In *EuroSpeech*, volume 2, pages 1271–1274, Madrid, September 1995.

[4] J. Fransen, D. Pye, A.J. Robinson, P.C. Woodland, and S.J. Young. WSJCAM0 Corpus and Recording Description. Technical Report CUED/F-INFENG/TR.192, Cambridge University Engineering Department, September 1994.

[5] D.B. Paul and J.M. Baker. The Design of the Wall Street Journal-based CSR Corpus. In *Proceedings of the fifth DARPA Speech and Natural Language Workshop*, pages 357–362. Morgan Kaufmann Publishers Inc., 1992.

[6] S.J. Young, M. Adda-Dekker, X. Aubert, C. Dugast, J.-L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, A.J. Robinson, H.J.M. Steeneken, and P.C. Woodland. Multlingual Large Vocabulary Speech Recognition: The European SQALE project. *Computer Speech and Language*, 11:73–89, 1997.

[7] M.M. Hochberg, G.D. Cook, S.J. Renals, and A.J. Robinson. Connectionist Model Combination for Large Vocabulary Speech Recognition. In *Neural Networks for Signal Processing*, volume IV, pages 269–278, 1994.

[8] A.J. 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.

[9] P.C. Woodland and Young S.J. The HTK Tied-State Continuous Speech Recogniser. In *EuroSpeech*, volume 3, pages 2207–2210, 1993.

[10] H. Bourlard and N. Morgan. Continuous Speech Recognition by Connectionist Statistical Methods. *IEEE Transactions on Neural Networks*, 4(6):893–909, November 1993.

[11] D.J. Kershaw, M.M. Hochberg, and A.J. Robinson. Incorporating Context-Dependent Classes in a Hybrid Recurrent Network-HMM Speech Recognition System. F-INFENG TR217, Cambridge University Engineering Department, May 1995.

[12] Tony Robinson and James Christie. Time-first search for large vocabulary speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.

[13] G.D. Cook and A.J. Robinson. The 1997 ABBOT System for the Transcription of Broadcast News. *DARPA Broadcast News Transcription and Understanding Workshop*, pages 49–54, February 1998.

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