THE 1995 ABBOT LVCSR SYSTEM FOR MULTIPLE UNKNOWN MICROPHONES

Dan Kershaw† *Tony Robinson*† *Steve Renals*‡

† Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.
Tel: [+44] 1223 332754
email : djk,ajr@eng.cam.ac.uk

‡ Department of Computer Science,
University of Sheffield, Sheffield S1 4DP, UK.
Tel: [+44] 114 282 5575
email : s.renals@dcs.shef.ac.uk

ABSTRACT

ABBOT is the hybrid connectionist-hidden Markov model large-vocabulary speech recognition system developed at Cambridge University. In this system, a recurrent network maps each acoustic vector to an estimate of the posterior probabilities of the phone classes, which are used as observation probabilities within an HMM. This paper describes the system which participated in the November 1995 ARPA Hub-3 Multiple Unknown Microphones (MUM) evaluation of continuous speech recognition systems, under the guise of the CU-CON system. The emphasis of the paper is on the changes made to the 1994 ABBOT system, specifically to accomodate the H3 task. This includes improved acoustic modelling using limited word-internal context-dependent models, training on the Wall Street Journal secondary channel database, and using the linear input network for speaker and environmental adaptation. Experimental results are reported for various test and development sets from the November 1994 and 1995 ARPA benchmark tests.

1. INTRODUCTION

The hybrid 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 [1]. ABBOT is a large-vocabulary speech recognition system based on the hybrid approach and utilizes a recurrent network for acoustic modelling. The major advantage of this approach is that the recurrent network is compact and able to capture temporal acoustic context. Consequently, the basic ABBOT system is able to achieve very good performance using only 54 context-independent phone models.

The ABBOT system participated in the 1994 ARPA continuous speech recognition (CSR) evaluations [8]. This paper reports recent improvements in the ABBOT system and the extra necessary features included for the multiple unknown microphones (MUM) evaluation. The acoustic modelling used for the 1995 evaluations is presented in the following section. This section describes training a new set of models on the secondary channel Wall Street Journal SI84 Corpus, and improved acoustic modelling using word-internal context-dependent phones. Section 3 describes the linear input network (LIN) technique for speaker and channel adaptation. Section 4 reports on the performance of the ABBOT system on various ARPA CSR development and evaluation tasks.

2. ACOUSTIC MODEL

This section describes the acoustic modelling process used in the ABBOT system. This includes a brief description of the front-end, structure of the observation model (i.e., the recurrent network), and the training process used for estimating the parameters of the connectionist component. It also describes the phonetic context-dependent modelling which augments the standard recurent network model.

2.1. Acoustic Feature Representation

Two sets of acoustic features have been used in the past by the ABBOT system: MEL+, a 20 channel mel-scaled filter bank with three voicing features [10], and PLP, 12th order cepstral coefficients derived using perceptual linear prediction and log energy [3].

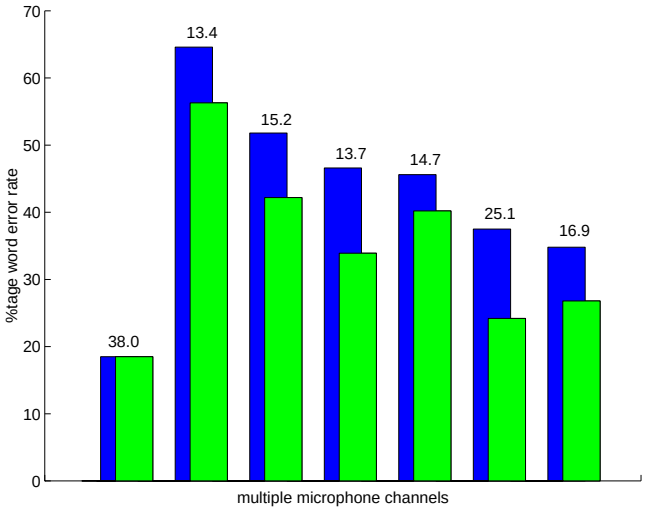

Figure 1: A comparison of the performance of MEL+ (in black) and PLP (in grey) RNN models, for speakers 700 and 703. The SNR figures above each bar are provided by the NIST "wavemd" tool.

This year the MEL+ representation has been dropped due to its poor channel robustness as demonstrated in Figure 1, where performance degrades by 24% when compared with PLP. Also to increase the robustness of the system to environmental conditions, the statistics of each feature channel were normalized to zero mean with unit variance over each utterance. Note that the 1995 ABBOT system was trained solely on the SI84 corpus.

2.2. Recurrent Network Structure

The basic acoustic modelling system is described in [11]. For each input frame, an acoustic vector, $\mathbf{u}(t)$, is presented at the input to the network. Within the recurrent structure, the state vector provides the mechanism for modelling of past context and the dynamics of the acoustic signal. The output vector represents an estimate of the posterior probability of each of the phone classes given the acoustic input and the model parameters. The output is delayed by four frames to account for future acoustic context.

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 [11] – 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.

2.3. Secondary Channel Microphone

New recurrent network models were trained on the secondary channel microphone data available on the Wall Street Journal training corpus. The conditions of this training set are more closely matched to those of the multiple unknown microphones than the clean speech. For decoding, two sets of models are used; if the signal-to-noise ratio, as calculated by the NIST "wavemd" tool is above a certain threshold (as determined by the model's performance on a development set), then the clean models are used. Otherwise the secondary channel models are used. On an H3 development set, with a random selection of a "secondary microphone" for each speaker, using secondary channel models, as opposed to clean models, resulted in a reduction in word error rate of 27.4%.

2.4. Context-Dependent Modelling

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 [1]. The joint posterior probability of context class j and phone class i is given by,

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

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

$$y_{j|i}(t) \simeq \Pr(c_j(t) \mid \mathbf{x}(t+4), q_i(t)), \tag{2}$$

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 [4].

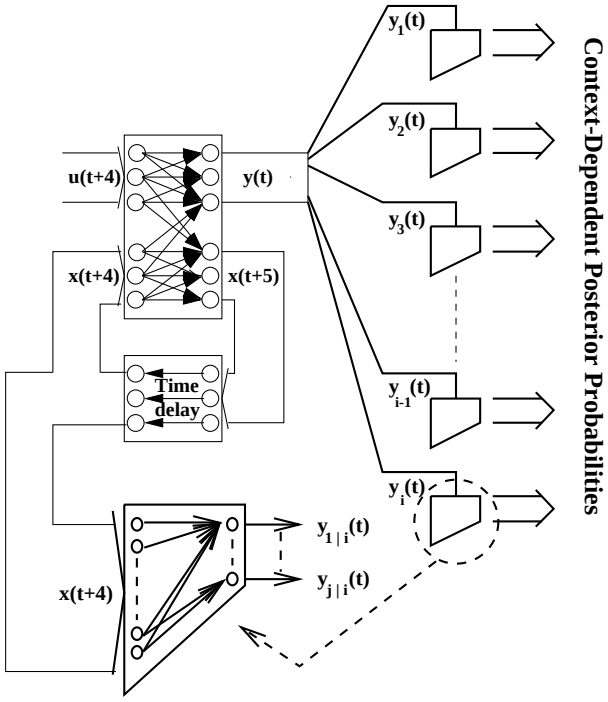

Figure 2: The phonetic context-dependent recurrent neural network modular system.

Figure 2 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 1.

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. 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). The decision trees are also used to relabel the pronunciation lexicon.

3. THE LINEAR INPUT NETWORK

The linear input network (LIN) has been successfully applied to connectionist-HMM hybrid systems for both supervised [6] and unsupervised [7] speaker adaptation. A linear mapping is created to transform the acoustic vector. During recognition, this transformed vector is fed as input into the speaker-independent RNN, as in Figure 3.

A gradient decent technique is used to train the LIN for a new speaker. The input is propagated forward (through the LIN and RNN) to the output layer of the RNN. At this point the error is back-propagated through the RNN, to provide gradient information

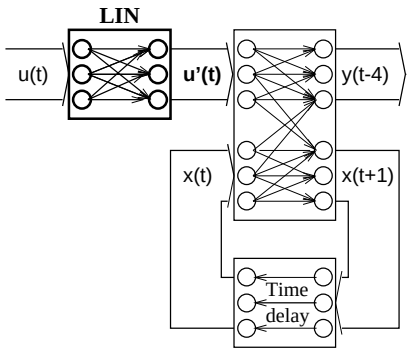

Figure 3: The linear input network "bolts on" to the recurrent network, performing a linear transformation of the acoustic feature space.

to the linear input layer. Only the LIN's parameters are updated; the RNN's are kept frozen.

For the MUM evaluations, unsupervised block adaptation is performed over each speaker session – i.e. a Viterbi alignment using the current model is carried out on decoded utterance hypotheses of a session, to label the acoustic frames. The LIN is now trained in a supervised fashion. Decoding is then performed with the "adapted" recurrent network. This process is iterated until there is little or no change in the session hypotheses from one adaptation pass to the next [5].

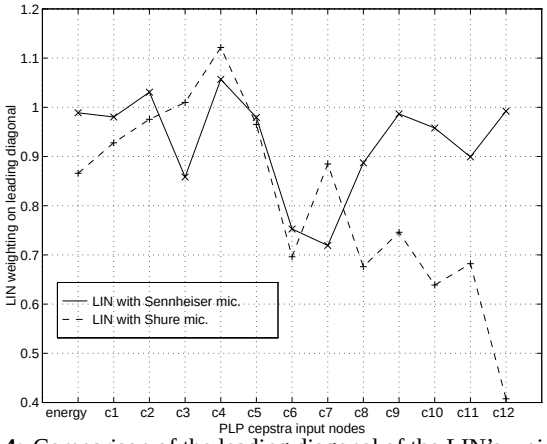

Figure 4: Comparison of the leading diagonal of the LIN's weight matrix for clean and secondary channel conditions.

The transformation matrix (bias weights excluded) tends to be close to an identity matrix, as might be expected. Figure 4 shows the diagonal elements of a LIN adapting a "clean" RNN to a new speaker (71j) using the Sennheiser microphone compared with those of a LIN adapting the same RNN to the same speaker, but with a Shure microphone. For the leading diagonal, the first six cepstra transforms are similar (apart from channel 3), indicating that these could be responsible for a "general picture of the speech signal"; the other higher order cepstra transforms are quite markedly different. Clearly the LIN is trying to adapt to a different acoustic channel input, where the transformation is likely to be responsible for translating the acoustic channel response of the Shure microphone back towards that of the Sennheiser. The difference in the energy channel is probably due to the microphone position and background noise.

4. RESULTS

This section reports decoding results for the ARPA 1995 H3 multiple unknown microphones Task. Decoding is performed by the ABBOT decoder, NOWAY [9], which uses a modified stack decoding algorithm, with a pruning strategy that is well matched to the hybrid connectionist-HMM approach. The various tests and their conditions are denoted as follows:-

H3:P0-DT A MUM development test set comprising 20 speakers and 7 different secondary microphones. The 1994 20k trigram language model was used. There was no adaptation.

H3:C0-DT The development contrast test set. Conditions are the same as H3:P0-DT, but with the Sennheiser microphone. The standard 60k trigram was used here. There was no adaptation.

H3:P0 The 1995 H3 MUM unlimited vocabulary test. The standard 60k trigram language model was used, along with unsupervised block adaptation.

H3:C0 The 1995 H3 unlimited vocabulary contrast. Conditions are the same as H3:P0, but with the Sennheiser microphone.

H3:C1A The 1995 H3 MUM unlimited vocabulary contrast. No adaptation is allowed, and standard 60k trigram must be used.

H3:C1B The 1995 H3 MUM unlimited vocabulary Sennheiser microphone contrast. No adaptation is allowed, and standard 60k trigram must be used.

The 1995 ABBOT system used the 60,000 word vocabulary and standard trigram generated by CMU throughout the evaluation. The pronunciation lexicon was derived primarily from a lexicon supplied by LIMSI-CNRS and expanded to cover the 60,000 word vocabulary [5].

4.1. Context-Dependent Results

Table 1 shows a comparison of the standard ABBOT system and the ABBOT system with the limited context-dependent word-internal phone modelling as discussed in section 2.5. Results are reported for various other tasks in [4]. Note, there are only 527 context-dependent phones. As can be seen, the context-dependent phone modelling attains a consistent reduction in word error rate.

Test Set	CI	CD	Red$^{n}_{o}$ WER
H3:P0-DT	24.6	21.9	11.0
H3:C0-DT	14.4	12.3	14.6
H3:C1B	15.3	13.6	11.1

Table 1: Performance of the context-dependent system on the 1995 ARPA tasks compared with the baseline RNN (CI). Results for H3:C1B are not run with the official CU-CON evaluation system.

4.2. The 1995 H3 MUM Results

The results shown in this section make full use of all the techniques described in previous sections: training on matched conditions, context-dependent modelling and adaptation using the linear input network. Table 2 shows the reduction in error for each iteration of unsupervised LIN adaptation, on both the H3:P0 (MUM) and H3:C0 (Sennheiser microphone only) tasks. For the P0 condition the error rate is reduced by 18.2% after 3 iterations of adaptation. For the C0 condition the error rate is reduced by 10.7%.

Adaptation for H3 Multiple Unkown Microphones					
# Passes of	Sub^n.	Del^n.	Ins^n.	WER	SER
0	17.0	3.6	3.6	**24.2***	89.0
1	14.2	3.2	3.0	20.4	85.0
2	13.8	3.1	3.0	19.9	83.3
3	13.7	3.2	2.9	**19.8**†	83.7

Adaptation for the H3 Sennheiser Microphone					
# Passes of	Sub^n.	Del^n.	Ins^n.	WER	SER
0	9.7	2.1	2.3	**14.0***	76.0
1	8.5	1.9	2.2	12.5	73.3
2	8.4	1.9	2.2	**12.5**†	73.3

Table 2: Decoding performance on the H3 tasks, at each adaptation pass. The stars represent the official entries for H3:C1A and H3:C1B respectively, while the daggers represent the official entries for H3:P0 and H3:C1 tasks.

5. SUMMARY

There are a few general conclusions which can be drawn about the 1995 ABBOT system.

- The linear input network is a very effective procedure for adapting to both speaker and channel conditions for connectionist systems. In addition, it requires very few extra parameters.
- Context-dependent modelling provides a consistent improvement in performance, even with only 527 word-internal context-dependent phone classes in this system.
- A more closely matched training conditions gives superior performance.

The 1995 ABBOT system, even though extended with context-dependent modelling and the linear input network, still has an order of magnitude fewer parameters than conventional HMM systems. The discriminative nature of this system coupled together with the phone deactivation pruning strategy employed in the decoder leads to real-time performance with only a small increase in search errors. The official, adjudicated results for the ABBOT system on the 1995 H3:P0 and H3:C0 tasks were 19.8% and 12.5%, respectively.

Recent improvements to the ABBOT system include training of the recurrent networks for effective use of the SI284 training corpus [2], and local speaker-adaptation approaches [12], while application of state-based context-dependent phone modelling is planned for the near future.

6. ACKNOWLEDGMENTS

This work was partially funded by ESPIRIT project 6487 WERNICKE. The authors would like to acknowledge MIT Lincoln Laboratory and CMU for providing language models and associated tools, LIMSI-CNRS and ICSI for providing the pronunciation lexicons. A special acknowledgement also goes to Mike Hochberg for all his invaluable help.

7. REFERENCES

1. H. Bourlard and N. Morgan. Continuous Speech Recognition by Connectionist Statistical Methods. *IEEE Transactions on Neural Networks*, 4(6):893–909, November 1993.
2. G.D. Cook and A.J. Robinson. Boosting the Performance of Connectionist Large-Vocabulary Speech Recognition. In *ICSLP '96*, October 1996.
3. H. Hermansky and N. Morgan. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–89, October 1994.
4. D.J. Kershaw, M.M. Hochberg, and A.J. Robinson. Context-Dependent Classes in a Hybrid Recurrent Network-HMM Speech Recognition System. To appear in Neural Information Processing Systems 8, 1996.
5. D.J. Kershaw, A.J. Robinson, S.J.Renals, and M.M. Hochberg. The 1995 ABBOT LVCSR System. In *The ARPA Speech Recognition Workshop*, Arden House, Harriman, New York, February 1996.
6. J. Neto, L. Almeida, M.M. Hochberg, C. Martins, L. Nunes, S.J. Renals, and A.J. Robinson. Speaker-Adaptation for Hybrid HMM-ANN Continuous Speech Recognition System. In *Eurospeech*, pages 2171–4, September 1995.
7. J.P. Neto, C.A. Martins, and L.B. Almeida. Unsupervised Speaker-Adaptation For Hybrid HMM-MLP Continuous Speech Recognition System. In *IEEE Speech Recognition Workshop*, pages 187–8, December 1995.
8. D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. Lund, A. Martin, and M. Pryzbocki. 1994 Benchmark Tests for the ARPA Spoken Language Program. In *Proceedings of the Spoken Language Technology Workshop*. Morgan Kaufmann, 1995.
9. S. Renals and M. Hochberg. Efficient search using posterior phone probability estimates. In *Proc. ICASSP*, volume 1, pages 596–599, Detroit, 1995.
10. 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.
11. A.J. Robinson. An Application of Recurrent Nets to Phone Probability Estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, March 1994.
12. S.R. Waterhouse, D.J. Kershaw, and A.J. Robinson. Smoothed Local Adaptation of Connectionist Systems. In *ICSLP '96*, October 1996.