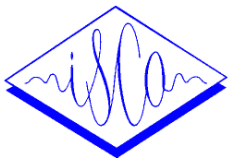

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION USING A HYBRID CONNECTIONIST-HMM SYSTEM

M. M. Hochberg *S. J. Renals* *A. J. Robinson* *D. J. Kershaw*
Cambridge University Engineering Department
Cambridge CB2 1PZ, England

ABSTRACT

ABBOT is a hybrid connectionist-hidden Markov model (HMM) system for large vocabulary speech recognition which participated in the November 1993 ARPA Wall Street Journal benchmark tests. This system uses a recurrent network to estimate the acoustic observation probabilities within the HMM framework. Since the 1993 benchmark tests, a number of improvements have been made to the ABBOT system. These improvements have been gained through better phone-duration modeling and connectionist model combination. In addition, ABBOT has been extended to handle large vocabulary tasks with a trigram language model. Fast decoding is obtained using a pruning strategy particularly well-suited for the hybrid approach. This paper describes the recent modifications to the system and experimental results are reported for various test and development sets from the November 1993 ARPA evaluations.

1. INTRODUCTION

ABBOT is the hybrid connectionist-hidden Markov model large-vocabulary speech recognition system developed at Cambridge University. The hybrid approach has proven to be very successful and the ABBOT system participated in the November 1993 ARPA evaluation of spoken language systems[1]. A major advantage of this approach is the good performance achieved using context- and gender-independent acoustic modeling, requiring many fewer parameters than comparable hidden Markov model (HMM) systems.

This paper describes recent improvements made to the ABBOT November 1993 system. The following section provides a basic description of the hybrid system. The next section describes the tasks on which the improvements to ABBOT are evaluated. Section 4 presents modifications and extensions made to the phone duration modeling of the system. In the 1993 tests, ABBOT employed model combination techniques to improve the acoustic modeling capabilities and Section 5 describes recent improvements. A major extension to the ABBOT system has been the capability to handle large vocabularies and a trigram language model. Fast pruning of the search space with minimal search errors is achieved by use of the connectionist component as a fast-match engine as well as the acoustic model. Section 6 describes the updated decoder. The paper concludes with some general discussion of the approach and presents the system performance on the 1993 benchmark tests.

[1] For the evaluation and associated publications, the ABBOT system is denoted as CU-CON.

2. SYSTEM DESCRIPTION

The basic framework of the ABBOT system is similar to the one described in [2] except that a recurrent network is used for the connectionist component. A more complete description of the basic approach can be found in [9] and a description of the 1993 evaluation system can be found in [5].

As in HMMs, the hybrid approach uses an underlying hidden Markov process to model the time-varying nature of the speech signal. The Markov process is determined in a hierarchical fashion, e.g., the language model is a Markov process on the words and the words are a Markov process on the phones.

A recurrent network is used as the acoustic model within the HMM framework. At each 16 msec frame, the input acoustic vector is mapped by the network to an output vector, $\mathbf{y}(t)$. The output vector represents an estimate of the posterior probability of each of the phone classes, i.e.,

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

where $q_i(t)$ is phone i at time t and $\mathbf{u}_1^t$ is the input from time 1 to t. Note that there is a single recurrent network for the system and the recurrent network generates *all* the phone probabilities in parallel.

Decoding with the hybrid connectionist-HMM approach is equivalent to conventional HMM decoding with the recurrent network modeling the observations. The main issue to consider is that – different from HMM acoustic modeling – the network estimates posterior phone probabilities. Because the decoding process makes use of the likelihood of the acoustic data, the network outputs are mapped to scaled likelihoods by

$$\Pr(\mathbf{u}(t)|q_i(t)) \sim \frac{y_i(t)}{\Pr(q_i)}. \tag{2}$$

Here, $\Pr(q_i)$ is estimated from the training data.

3. RECOGNITION TASKS

The tasks used to evaluate the improvements are taken from the November 1993 WSJ evaluation and development tests [7]. The spoke 5 and spoke 6 (Sennheiser microphone) development tests were used to tune parameters while evaluating the system changes. Final evaluation of the modifications used the hub 2 benchmark task. These are 5,000 word, closed vocabulary tasks. For the experiments described in this paper, the ABBOT system used the official bigram and trigram language models supplied by MIT-Lincoln Laboratories [8] and the pronunciation dictionary supplied by Dragon Systems [8]. For the 1993 evaluations, ABBOT used a 79 phone symbol set and employed context- and gender-independent acoustic models.

4. DURATION MODELING

Although a single acoustic model is used for each phone, a duration model can be employed to enforce constraints in the decoding process. A number of different duration models were investigated in [10], but little variation in performance has been observed. This section describes two new approaches to duration modeling which have resulted in improved performance.

4.1. Phone-Deletion Penalty

In the November 1993 evaluations, a simple left-to-right Markov chain with no skip states was used to model the duration for each phone. The number of states in each phone model was set to one half the average duration (in frames) and all transition probabilities were set 0.5 as shown in Figure 1a. The goal was to approximate a Poisson distribution (i.e., duration variance equals the mean) with a minimum duration constraint. However,

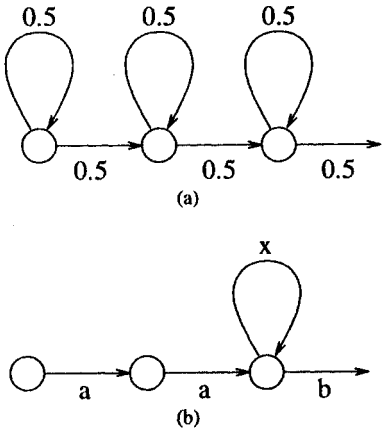

Figure 1. Topologies for (a) pseudo-Poisson and (b) general Viterbi decoder duration models.

the Poisson distribution is approximated only if all paths through the duration model are considered. Viterbi decoding actually led to a duration model as is shown in Figure 1b where $a = x = b = 0.5$. Since all the transition probabilities are 0.5 at every time step, this model only enforces minimum duration as a constraint during decoding. All phone sequences satisfying the minimum duration constraint have the same duration score.

A duration model which applies a phone-deletion penalty can be expressed in the form of the model in Figure 1b. If $a = x$ and b and x are the same for all phone models, then the likelihood ratio[2] of having two phones versus one phone over any given segment of the speech signal is given by b/x. For $b/x > 1$, the model penalizes phone deletions (or, equivalently, encourages phone insertions). A search over different values found that $b/x = 3$ (with $x = 0.5$) resulted in the best performance. Table 1 shows the results of using this duration model versus the original minimum duration model. As the table clearly shows, penalizing the phone deletions results in substantial improvement for spoke 5 and a reasonable improvement for spoke 6.

Model	Error Rate %	
	spoke 5	spoke 6
Minimum Duration	15.4	11.5
Phone-Deletion Penalty	12.7	11.0

Table 1. Performance of the minimum duration and phone-deletion penalty models on the WSJ spoke 5 and spoke 6 development tests.

4.2. Context-Dependent Modeling

It has been observed that the mean duration of some phones vary substantially within different contexts [3]. Analysis of the WSJ0 training corpus confirms this and Table 2 indicates that phone-duration variability can be reduced by employing context-dependent phone-duration models. In the table, the context-dependent phones were computed by clustering co-occurences of phones based on left or right context (biphones) or left and right context (triphones). A "back-off" to a lesser context was used if too few models (in this case, less than 200) exist for a particular context.

Context Model	Number of Models	Std. Dev., msec
Independent	79	33.31
Biphones	4680	26.14
Triphones	7430	25.53

Table 2. Average standard deviations over the training set for varying amounts of context.

To implement the duration models, Markov chains were built using statistics gathered from the data and incorporated into the pronunciation lexicon. Various different models were built in order to evaluate duration context dependency. Results in Table 3 show that improvement can be achieved by modeling context. The gamma duration model type is a Markov chain fit to gamma

[2]Actually, this is not the ratio of true likelihoods since the models are not probability distributions, but can be considered this in the context of Viterbi decoding.

distribution [3]. The work reported here is preliminary, but the table shows that there is promise for the context-dependent approach.

Model Type	Error Rate, % CI	CD	Number of Models
Gamma	14.0	13.6	4451
Min. Dur.	15.4	14.7	281
P.D. Penalty	12.7	13.3	281

Table 3. Comparison of word error results for context-independent (CI) models and context-dependent (CD) phone duration models for spoke 5 using gamma, minimum duration (Min. Dur.), and phone-deletion penalty (P.D. Penalty) model types.

5. MODEL COMBINATION

The recognition performance of ABBOT is strongly tied to the acoustic modeling capability of the connectionist component. A very successful approach to improving the estimates of the phone probabilities has been to combine the outputs of multiple recurrent networks trained on different representations of the acoustic signal. Significant improvements on the WSJ [5] and TIMIT [10] tasks have been observed by simply averaging the network outputs, i.e., setting

$$y_i(t) = \frac{1}{K}\sum_{k=1}^{K} y_i^{(k)}(t) \tag{3}$$

where $y_i^{(k)}(t)$ is the estimate of the kth model. Recent work [4] has indicated that a better approach is to merge the network outputs in the log domain, i.e.,

$$\log y_i(t) = \frac{1}{K}\sum_{k=1}^{K} \log y_i^{(k)}(t). \tag{4}$$

With this approach, it is difficult to assign a probabilistic interpretation to the merged outputs. However, if the models are assumed to be independent, then the estimated joint likelihood of the different data is proportional to the product (or sum in the log-domain) of the network outputs.

In the experiments presented here, the parameters for each network are estimated on the same speech data, but processed with different front-ends. Two successful spectral representations have been found to be a 20 channel mel-scaled filter bank with voicing features and 12th order cepstral coefficients derived from perceptual linear prediction. The filter bank and cepstra are referred to in this paper as MEL+ and PLP, respectively. In addition, because the recurrent network is time asymmetric, training the network to classify forward in time will result in different dynamics than training to classify backwards in time. Based on the above considerations, four networks were constructed from the possible representations; FORWARD MEL+, BACKWARD MEL+, FORWARD PLP, and BACKWARD PLP. Table 4 shows the WSJ results for recurrent network merging. Each of the networks trained on different front-ends have similar performance, but the error rate of the merged system is reduced significantly. In addition, the log-domain merge provides the best results.

Model Type	Error Rate % spoke 5	spoke 6
FORWARD MEL+	17.3	15.0
FORWARD PLP	17.1	15.1
BACKWARD MEL+	17.8	15.5
BACKWARD PLP	16.9	14.4
AVERAGE	17.3	15.0
LINEAR MERGE	15.4	11.4
LOG MERGE	13.6	11.0

Table 4. Word recognition results for linear and log-domain model combination.

6. SEARCH

6.1. Tree-Based Lexicon

Like most medium to large vocabulary systems, word models in ABBOT are represented as sequences of sub-word models (e.g., phones, etc.). However, the original ABBOT decoder – referred to as Y0[3] – independently modeled each word in the vocabulary. This resulted in a large number of redundant computations and the search became very time consuming for large-vocabulary decoding. One method of reducing the number of redundant computations is to employ a tree-based lexicon [6]. On a 5,000 word task using a bigram language model there was little speed difference between the tree- and linear-based lexicon. However, for a 20,000 word task (also with a bigram language model), a factor of 10 improvement was seen in the computation time for the tree-structured lexicon.

6.2. No Way? Why Not?

Two approaches were taken for performing the search over the possible word sequences. The first approach was a time-synchronous, tree-based lexicon, extension of the original Y0 decoder. In this approach, a new tree is started at every frame. As discussed in [6], because each tree represents all the words in the vocabulary, it is necessary to apply the bigram or trigram language score for the current word when exiting the word. Although the possibility exists for having too many copies of the lexicon, in practice it was found that pruning kept the number of active trees and active nodes in the trees to very reasonable levels. This approach has only implemented a bigram language model capability.

The second approach taken in the decoding was to implement a quasi-time-asynchronous, single pass decoder. This decoder – referred to as **noway** – employs techniques related to stack decoding [1]. Like Y0, **noway** also used a tree-based lexicon.

Both of the above approaches provided similar computational performance on the 5,000 and

[3]Y0 was developed jointly with the Realization Group at the International Computer Science Institute.

20,000 word recognition tasks with the bigram language model. The **noway** decoder, however, was initially developed for the large vocabulary recognition task and provides a much more flexible language model interface. For this reason, most future work is expected to employ the **noway** decoder.

6.3. Posterior-Directed Path Pruning

Pruning of the search space is a very important feature of any large vocabulary speech recognition system. In both decoders, standard pruning techniques (e.g., beam search, fixed number of hypothesis, etc.) were implemented to reduce the computational requirements. One point to note about the standard pruning techniques is that it is desirable to set the tightest pruning thresholds which still result in very few search errors. To achieve this, it is important to propagate the language model score through the tree as much as possible. Use of the best possible bigram score at each node of a tree, resulted in a 90% reduction in the number of active nodes required for Y0.

One advantage of the hybrid connectionist-HMM approach is that the acoustic processing computes the posterior probabilities of phones given the current frame of acoustic data. These probabilities can be used directly for fast-match and/or fast-pruning of the search space. The fast pruning aspect of the approach was investigated for this paper. This approach effectively pruned all phones whose current, local probabilities were below some threshold (typically 10^{-5}). Using posterior-directed fast pruning, recognition time was reduced by 50% for both the 5,000 and 20,000 word tasks using a bigram language model. This speed-up was achieved with no increase in the number of search errors.

7. CONCLUSION

As described in the preceding sections, a number of substantial improvements have been made to the ABBOT system. To benchmark the improvements, ABBOT was evaluated on the November 1993 hub 2 test using the phone-deletion penalty duration model and log-domain model merging. For the bigram language model, the ABBOT system achieved an 11.1% word error rate. This represents an 18% improvement over the official 1993 results. This is a very big improvement considering that the system employs virtually the same acoustic models as used in the evaluation. ABBOT had a 8.8% word error on the hub 2 test using a trigram language model. This represents a 20% reduction in the error rate and is consistent with other comparisons between bigram and trigram language models on this task.

Further work is still planned for improving the system. In particular, improvement is certainly expected from training new acoustic models using the full WSJ0 + WSJ1 training set. In addition, continued investigation into context-dependent duration modeling and connectionist model merging is in progress.

8. ACKNOWLEDGEMENTS

This work was partially funded by ESPRIT project 6487, WERNICKE. A.J.R. and S.J.R. are supported by SERC fellowships. We would like to acknowledge MIT Lincoln Laboratory and Dragon Systems for providing the language model and pronunciation lexicon, respectively.

REFERENCES

[1] L. R. Bahl and F. Jelinek. Apparatus and method for determining a likely word sequence from labels generated by an acoustic processor. United States Patent, May 1988. Number 4,748670.

[2] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach.* The Kluwer International Series in Engineering and Computer Science. VLSI, Computer Architecture, and Digital Signal Processing. Kluwer Academic Publishers, Boston, Massachusetts, 1994.

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

[4] M. M. Hochberg, G. D. Cook, S. J. Renals, and A. J. Robinson. Connectionist model combination for large vocabulary speech recognition. In *Proc. of NNSP-94 Workshop*, 1994.

[5] M. M. Hochberg, S. J. Renals, and A. J. Robinson. ABBOT: The CUED hybrid connectionist-HMM large-vocabulary recognition system. In *Proc. of Spoken Language Systems Technology Workshop.* ARPA, Mar. 1994.

[6] H. Ney. Modeling and search in continuous speech recognition. In *Proceedings of 3rd European Conference on Speech Communication and Technology*, volume Volume 1, pages 491–498, Berlin, Sept. 1993.

[7] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. Lund, and M. Pryzbocki. 1993 benchmark tests for the ARPA spoken language program. In *Proc. of Human Language Technology Workshop.* ARPA, Mar. 1994. to appear.

[8] D. B. Paul and J. 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, Feb. 1992. DARPA, Morgan Kaufman Publishers, Inc.

[9] A. J. Robinson. An application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, Mar. 1994.

[10] T. Robinson, M. Hochberg, and S. Renals. IPA: Improved phone modelling with recurrent neural networks. In *1994 International Conference on Acoustics, Speech, and Signal Processing*, pages 37–40, Adelaide, Australia, Apr. 1994. IEEE. Volume 1.