RECENT IMPROVEMENTS TO THE ABBOT LARGE VOCABULARY CSR SYSTEM

M. M. Hochberg[†]*, S. J. Renals*[‡]*, A. J. Robinson*[†]*, and G. D. Cook*[†]

[†]Cambridge University Engineering Department, Cambridge CB2 1PZ, England
[‡]Department of Computer Science, University of Sheffield, Sheffield S1 4DP, England

ABSTRACT

ABBOT is the hybrid connectionist-hidden Markov model (HMM) large-vocabulary continuous speech recognition (CSR) system developed at Cambridge University. This system uses a recurrent network to estimate the acoustic observation probabilities within an HMM framework. A major advantage of this approach is that good performance is achieved using context-independent acoustic models and requiring many fewer parameters than comparable HMM systems. This paper presents substantial performance improvements gained from new approaches to connectionist model combination and phone-duration modeling. Additional capability has also been achieved by extending the decoder to handle larger vocabulary tasks (20,000 words and greater) with a trigram language model. This paper describes the recent modifications to the system and experimental results are reported for various test and development sets from the November 1992, 1993, and 1994 ARPA evaluations of spoken language systems.

1. INTRODUCTION

This paper describes recent improvements made to the ABBOT November 1993 system [1][1]. The improvements include new approaches to connectionist model merging and a duration model specifically optimized for the recurrent network hybrid approach. A 10% reduction in word-error rate (on average) has been observed for each of these modifications. The ABBOT system has also been extended to handle larger vocabularies (20,000 words and greater) using a back-off trigram language model. Results show that, while using significantly fewer acoustic modeling parameters, ABBOT recognition performance is comparable with state-of-the-art large vocabulary CSR systems.

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 [3] and a description of the baseline ABBOT system can be found in [1].

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 $\mathbf{u}(t)$ 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., $y_i(t) \simeq \Pr(q_i(t)|\mathbf{u}_1^{t+4})$ where $q_i(t)$ is phone i at time t and $\mathbf{u}_1^t$ is the input from time 1 to t. Because the HMM decoding process makes use of the likelihood of the acoustic data, the network outputs are mapped to scaled likelihoods by $\Pr(\mathbf{u}(t)|q_i(t)) \sim y_i(t) / \Pr(q_i)$. Here, $\Pr(q_i)$ is estimated by the relative frequency of the phone in the training data.

The recurrent network provides a number of benefits over standard HMM acoustic models (e.g., mixture Gaussian densities). The internal recurrent nodes are able to model acoustic context and the use of the network structure results in a nonparametric model of the acoustic features. These properties allow the system to operate with context-independent phone models. In addition, only a single state per phone is required. The compact representation of the phone model has many desirable properties, such as fast decoding.

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

2. MODEL MERGING

The recognition performance of ABBOT is strongly linked 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. Significant improvements have been observed by simply averaging the network outputs, i.e., setting

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

where $y_i^{(k)}(t)$ is the estimate of the kth model [1]. Previous work has reported merging results where the constituent models were all derived from the same speech waveform, but utilized different spectral representations (e.g., filter-banks, PLP-derived cepstra) and different time ordering of the input vectors (i.e., forward and backward). Although this merging has been successful, the approach is somewhat ad-hoc. This section presents two alternative methods for combining the outputs of multiple recurrent networks.

2.1. Merging Based on Multiple Front-Ends

A more principled approach to merging models from different front-ends is based on using the Kullback-Leibler information as a distance-like measure on multinomial distributions. Consider the following criterion

$$E(p) = \sum_{k=1}^{K} D(p||y^{(k)}) \tag{2}$$

where

$$D(p||q) \equiv \sum_i p_i \log \frac{p_i}{q_i} \tag{3}$$

is the Kullback-Leibler information. Minimization of E with respect to the distribution p can be interpreted as choosing the distribution which minimizes the average (across models) Kullback-Leibler information. Solving the minimization in (2) results in the log-domain merge of the network outputs, i.e.,

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

where B is a normalization constant such that $\mathbf{y}$ is a probability distribution.

2.2. Merge Based on Talker Clustering

Recent work on merging networks trained on different talkers has been motivated by two factors. The primary goal was to fully utilize the great amount of training data available. Due to memory and time limitations, it was difficult to directly train a recurrent network using the full SI-284 training corpus. The approach taken was to use multiple networks trained from subsets of the data and merge the outputs. The second motivating factor for this approach was to reduce the effects of inter-speaker variability. Inter-speaker variability is a major problem because parametric representations of speech are highly speaker dependent. To minimize this effect, multiple connectionist models are each trained on a subset of the training data. The subsets are formed by clustering the utterances so as to minimize the variability within each subset.

Let S be the sample covariance matrix from an utterance and Σ_j be the covariance matrix of cluster j. The distance measure used for clustering the utterances is an estimate of the log-likelihood of S given Σ_j and is given by

$$l(S; \Sigma_j) = \frac{Nn}{2} \left\{ \log(|\Sigma_j^{-1} S|)^{1/n} - \frac{1}{n} \mathrm{tr}(\Sigma_j^{-1} S) \right\}. \tag{5}$$

Here, N is the number of samples in the utterance, n is the dimensionality of the feature vector and tr indicates the trace of a matrix. The use of (5) is based on the assumption that the feature vectors are independent with identical, zero mean, normal distributions. Research by Gish *et al* [4] on speaker identification has shown that (5) provides good discrimination between different talkers.

The clustering algorithm is a hierarchical divisive procedure based on the LBG algorithm for vector quantization [5]. It starts with a single cluster consisting of all the patterns (utterance covariances). The data is randomly split into two disjoint clusters and the cluster covariances are re-computed. Using (5), each pattern is assigned to a cluster. The cluster covariances are re-computed after all of the patterns have been assigned to a cluster. This process continues until each cluster is stable and there is no movement of patterns between clusters. The cluster consisting of the largest number of patterns is then randomly split into two clusters, and the process continues as before. This continues until the desired number of clusters have been created.

For reasons related to training of the recurrent networks, it is desirable to have the same number of tokens in each cluster. This is accomplished by assigning a scale factor β_j to each cluster log-likelihood score. This subset-size normalization is applied after completion of the clustering algorithm. The clustering procedure is then re-run using fixed cluster covariances and only re-assigning the utterance labels. This is an iterative procedure where β_j is defined as

$$\beta_j^m = \left\{ 1 + \left(\frac{n_j - N}{N} \right) \epsilon \right\} \beta_j^{m-1} \tag{6}$$

and where n_j is the number of patterns in cluster j, N is the desired number of patterns per cluster, m is the iteration, and ϵ is a small constant.

Each cluster is defined in terms of its covariance Σ_j, its weight β_j and a list of utterances U_j which generate the covariance. The U_j are used as training data for cluster dependent models. Thus, for each subset of the data, a recurrent network is trained to estimate phone probabilities. When an utterance is to be decoded, the covariance of the acoustic feature vectors S is computed. The posterior probability of the cluster model ω_j given the data X is then estimated by

$$P(\omega_j | X, \alpha) \sim l(S; \Sigma_j)^\alpha \beta_j^\alpha \tag{7}$$

where α is a heuristically determined tuning parameter. The outputs of the recurrent networks are then merged using

$$y_i(t) = \sum_{k=1}^{K} P(\omega_j | X, \alpha)\, y_i^{(k)}(t). \tag{8}$$

3. DURATION MODELING

In previous versions of the ABBOT system, 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 to 0.5. The goal was to approximate a Poisson distribution (i.e., duration variance equals the mean) with a minimum duration constraint. However, the Poisson distribution

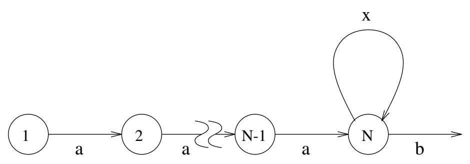

Figure 1: Topology for a general Viterbi decoder duration model.

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 1 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 1. If $a = x$ and b and x are constant for all phone models, then the likelihood ratio[2]

[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.

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 $b / x = 1.5$ (with $x = 0.5$) resulted in the best performance.

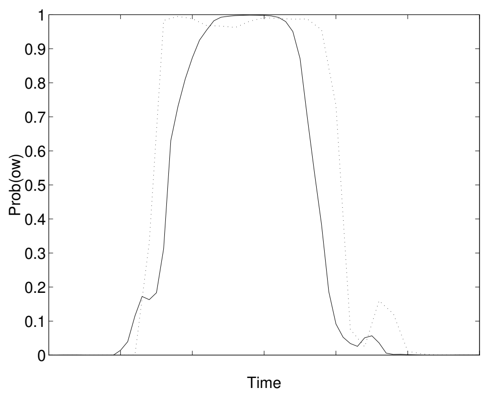

Figure 2: Comparison of network outputs for the acoustic input at normal (dotted) and slow (solid) rates for the phone `ow`.

The fact that the phone-deletion penalty duration model provides better performance than more sophisticated approaches can be attributed to the dynamical modeling capability of the recurrent network. The recurrent network has an implicit, internal representation of phone duration and the acoustic dynamics associated with the phone. An example of this is shown in figure 2. The figure shows the output of a well-trained recurrent network (estimates of the posterior probabilities) for the ARPA-bet phone `ow` on a portion of the speech waveform from the WSJ corpus. The dotted lines represent the network output for the acoustic input at a normal rate. The solid lines represent the network output for the same acoustic input at half the input rate. This slow signal was generated by interpolating between the input vectors of the normal signal. The time axis has been normalized so that the outputs of the network would be (nearly) identical if the network were independent of the duration of the phone. Figure 2 shows that doubling the duration of the vowel `ow` is not reflected in the output and the network dynamics limit the duration of the vowel.

4. LEXICAL REPRESENTATION

All ABBOT system results reported prior to this paper have been obtained using the Dragon Systems pronunciation lexicon [6]. In more recent work, a pronunciation lexicon supplied by LIMSI-CRNS has been employed. This lexicon employs 45 phones. Before using this dictionary, a set of rules provided by the International Computer Science Institute (ICSI) were employed to expand the LIMSI phone set to include flaps and closures [7]. This resulted in a total of 54 phones in the lexicon.

The LIMSI pronunciations were sufficient only for the 1992 and 1993 data. To extend the lexicon to 65,533 words required generation of additional pronunciations. The 1993 LIMSI dictionary was extended with those words that could easily be derived from the existing entries (by the addition of suffices or the merging of two words). This resulted in about 35,000 entries. The remaining entries were derived by ICSI from the CMU, COMLEX, TIMIT, OGI Numbers, BEEP and a TTS system using a probabilistic mapping technique to unify the phone sets and provide multiple pronunciations (see [7] for details).

5. TASKS

Enhanced capabilities relating to more complicated tasks have been added to the ABBOT system. In particular, the original Viterbi beam search decoder has been replaced with a time-asynchronous decoder which can handle 65,533 word vocabularies using a back-off trigram language model. This single-pass decoder makes direct use of the network posterior phone probability estimates to substantially reduce the search space and 20,000 word trigram tasks are decoded in near real-time with minimal degradation in performance [1]. A complete description of the NOWAY decoder and it's application to these large vocabulary tasks can be found in [8].

6. RESULTS

This section presents the experimental results used to evaluate the modifications to the ABBOT system. Results are reported for a number of different tasks from the ARPA evaluations:

`s5dev93`: the Nov. 1993 spoke 5 dev. test;
`s6dev93`: the Nov. 1993 spoke 6 dev. test;
`20keval92`: the Nov. 1992 20K word NVP eval. test;
`h1eval93`: the Nov. 1993 H1 20K word eval. test;
`h1dev94`: the Nov. 1994 H1 unlimited vocabulary dev. test;
`h1eval94`: the Nov. 1994 H1 unlimited vocabulary eval. test.

In the above, `s5dev93` and `s6dev93` are evaluated using a 5,000 word lexicon and bigram language model, `20keval92` and `h1eval93` are evaluated using a 20,000 word lexicon and trigram language model, and `h1eval94` is evaluated using both a 20,000 word and 65,533 word lexicon and trigram language model.

Table 1 shows the results for merging the recurrent networks using different acoustic front-ends (the networks were trained on forward and backward in time filter-bank and PLP-cepstra spectral representations for a total of four networks). As can be clearly seen from the table, the log-domain merge provides the better results.

Model Type	Error Rate %	
	`s5dev93`	`s6dev93`
LINEAR MERGE	15.4	11.4
LOG MERGE	13.6	11.0

Table 1: Word recognition results for linear and log-domain model combination for different acoustic front-ends. Models trained with SI-84 data.

For talker-cluster merging, the SI-284 training set was clustered into five subsets of approximately 7,000 utterances. A recurrent network was trained for each talker cluster. The *optimal* value for the α parameter of (8) was determined empirically. The effect of varying α is shown in figure 3. From (7), note that $\alpha = \infty$ implies only using the most probable model and $\alpha = 0$ implies uniform averaging of the models. After choosing the α, decoding experiments were performed and the results are shown in table 2. Here, the baseline result refers to a cluster-independent system trained on the SI-84 data set. The improvement moving from the SI-84 to the SI-284 training set is less than what traditional HMM systems have reported (e.g., [9]) and further research is planned.

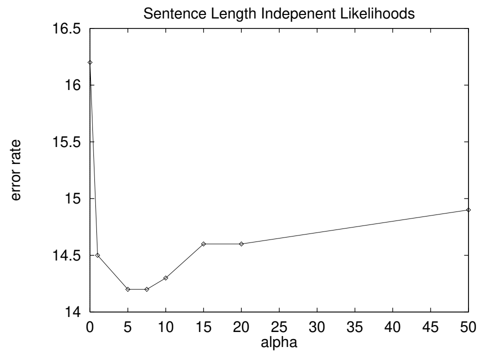

Figure 3: Effect of posterior probability scaling on error rate.

	Error Rate, %		
Test	Baseline	Cluster	Improvement
`s5dev93`	17.3	14.2	17.9
`h1dev94`	12.2	11.2	8.2

Table 2: Results for talker-dependent cluster merging. Baseline refers to a SI-84 system and cluster indicates merged cluster dependent models trained on SI-284.

Table 3 shows the results of using the phone-deletion penalty duration model versus the original minimum duration model. As the table clearly shows, penalizing the phone deletions results in substantial improvement across the four test sets.

	Error Rate %	
Task	Min. Duration	Phone-Deletion
`s5dev93`	15.4	12.7
`s6dev93`	11.5	11.0
`20keval92`	12.6	12.0
`h1eval93`	18.2	17.1

Table 3: Performance of the minimum duration and phone-deletion penalty models. The `h1eval93` has not been through the adjudication processing. All systems were trained using SI-84 database.

As mentioned above, the ABBOT system has been extended to handle very large vocabularies. The ABBOT system achieved (unofficially) a 14.9% word error rate on the `h1eval94` task using a 20,000 word vocabulary and trigram language model (H1:C1 system). This system employed both front-end and talker-cluster merging. By extending the vocabulary to 65,533 words, this error rate was reduced to (unofficially) 13.0% (H1:P0 system). This improvement directly reflects the difference in the percentage of out-of-vocabulary words for the two lexicons.

7. SUMMARY

This paper has described a number of new features of the ABBOT system which have resulted in a substantial reduction in error rate. The log-domain model combination is both a more principled method of merging phone-probability estimates and a more effective method than simple linear averaging. The talker-cluster approach is also a viable method for reducing the error rate, but there is still research to be performed in the area to fully take advantage of the massive amounts of data now available. The phone-deletion penalty duration model has been found to complement the dynamics of the recurrent network. The complete system has achieved very good performance on the standard benchmark tests of large vocabulary systems. This represents a significant achievement for a context-independent CSR system with relatively few acoustic parameters.

8. ACKNOWLEDGMENTS

This work was partially funded by ESPRIT project 6487, WERNICKE. AJR and SJR were supported by EPSRC fellowships. We acknowledge: MIT Lincoln Laboratory for providing language models for the 1992 and 1993 tasks; Dragon Systems, LIMSI and ICSI for providing pronunciation lexicons; and Rob Schechtman for help generating the 64K word-list and language models.

9. REFERENCES

[1] M. M. Hochberg, S. J. Renals, A. J. Robinson, and D. J. Kershaw, "Large vocabulary continuous speech recognition using a hybrid connectionist-HMM system," in *Proc. ICSLP*, 1994.

[2] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[3] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Transactions on Neural Networks*, vol. 5, pp. 298–305, Mar. 1994.

[4] H. Gish, W. Kransner, W. Russell, and J. Wolf, "Methods and experiments for text-indepedent speaker recognition over telephone channels," in *Proc. ICASSP*, 1986.

[5] Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. COM-28, pp. 84 – 95, Jan. 1980.

[6] 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*, Feb. 1992.

[7] G. Tajchman, D. Jurafsky, and E. Fosler, "Learning phonological rule probabilities from speech corpora with exploratory computational phonology." Submitted to ACL-95.

[8] S. Renals and M. Hochberg, "Efficient search using posterior phone probability estimates," in *Proc. ICASSP*, 1995.

[9] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using HTK," in *Proc. ICASSP*, 1994.