Phone Deactivation Pruning in Large Vocabulary Continuous Speech Recognition

Steve Renals, *Member, IEEE*

Abstract— **In this letter, we introduce a new pruning strategy for large vocabulary continuous speech recognition based on direct estimates of local posterior phone probabilities. This approach is well suited to hybrid connectionist/hidden Markov model systems. Experiments on the *Wall Street Journal* task using a 20 000 word vocabulary and a trigram language model have demonstrated that phone deactivation pruning can increase the speed of recognition-time search by up to a factor of 10, with a relative increase in error rate of less than 2%.**

I. Introduction

THE dominant approach to speech recognition is statistical, typically using hidden Markov models (HMM's). Utterance models are hierarchically composed of word models, which in turn are built from basic subword HMM's, such as (context-dependent) phone models. The parameters of each HMM are usually iteratively estimated by maximizing the likelihood of the model given the acoustic data.

An alternative statistical approach uses connectionist probability estimators within the HMM framework [2], [6], [7]. These hybrid connectionist/HMM systems use connectionist networks to compute direct local estimates of posterior phone probabilities. These local posterior probabilities may be converted to scaled likelihoods and integrated into the HMM framework as (scaled) estimates of HMM output likelihoods [6]. Hence, at recognition time these scaled likelihoods are used in the decoding, in preference to the direct posterior probability estimates.

Application of this hybrid approach has produced high performance large vocabulary continuous speech recognizers, such as the ABBOT system [5], which use one or two orders of magnitude fewer parameters than similarly performing HMM systems. This saving in acoustic model complexity results from the fact that hybrid systems are able to accurately model the acoustics of speech using a small number of context-independent phone models, rather than requiring several thousand context-dependent phone models.

The principal contribution of this letter is a demonstration that the hybrid approach can offer substantial increases in recognition search efficiency in addition to the modeling advantages referred to above. A new pruning strategy, *phone deactivation pruning*, is introduced which makes direct use of the local posterior phone probability estimates computed by a connectionist network. Phone deactivation pruning can be used in conjunction with existing likelihood-based approaches. This pruning algorithm has been integrated into the search component of the ABBOT system and we report on a set of experiments to test its effectiveness. The experimental work was carried out using the ARPA *Wall Street Journal* (WSJ) database. Results are evaluated in terms of both word error rate and computational resource requirements.

Manuscript received January 24, 1995. This work was supported by an EPSRC Postdoctoral Fellowship held at Cambridge University and ESPRIT BRA 6487, WERNICKE. The associate editor coordinating the review of this paper and approving it for publication was Dr. L. Niles.

The author is with the Speech and Hearing Group, Department of Computer Science, University of Sheffield, Sheffield, UK.

Publisher Item Identifier S 1070–9908(96)01414–9

II. Search

As speech recognition systems begin to address unlimited vocabulary tasks, the efficiency and accuracy of the recognition-time search procedure increases in importance. The goal of the search procedure is to locate the most probable string of words for a spoken utterance given acoustic and language models. Because of the size of the system's vocabulary (typically 20 000 words or more), the size of the search space is very large. Evaluation of this space is made more complex when long-span language models (e.g., trigrams) are used. If the search is to be manageable, then approximations must be made to reduce the effective size of the search space. The placing of such restrictions on the search space is often referred to as *pruning* and may be regarded as removing words or partial utterance hypotheses from consideration at a particular time without computing the complete probabilities for those hypotheses.

In HMM systems the search space is usually evaluated by computing likelihood estimates of the acoustic data having been generated by a particular utterance model. Pruning strategies are generally likelihood-based, involving the definition of a *likelihood envelope* (or *beam*) Δ, around the likelihood L of the most probable partial hypothesis at time t. Only hypotheses whose likelihood falls within the envelope (i.e., those hypotheses with a likelihood $L' \geq L - \Delta$) remain in consideration. All other hypotheses at time t are pruned. In the simplest form of likelihood-based pruning, the envelope is defined statically at all times. An adaptive likelihood envelope may also be defined by imposing a fixed upper limit on the number of hypotheses to be extended and evaluated at any given time. Efficient pruning can be enhanced by structuring the lexicon as a tree, in which the nodes correspond to phone models, and every path from the root to a leaf (or a node at a word end) corresponds to a pronunciation of a word in the dictionary [4].

III. Phone Deactivation Pruning

A new pruning strategy is introduced taking advantage of three features of the hybrid approach:

- Direct estimation of posterior probabilities $P(\text{phone}|\text{data})$ by the network, rather than likelihoods $P(\text{data}|\text{phone})$.
- Context-independent acoustic modeling leads to a small set of basic HMM's (typically 40–80), rather than several thousand context-dependent models.
- Network probability estimation enables the computation of all phone probability estimates at each frame, without imposing much additional computational cost.

For each frame of data the connectionist acoustic model produces a complete vector of context-independent posterior phone probabilities.

The phone posteriors may be regarded as a local estimate of the presence of a phone at a particular time frame. If the posterior probability estimate of a phone given a frame of acoustic data is below a threshold, then all words containing that phone at that time frame may be pruned (deactivated), i.e.

$$\texttt{if } P(\text{phone}_i|\text{data}) < \text{threshold } \texttt{then } P(\text{phone}_i|\text{data}) = 0.$$

In practice the log (scaled) likelihood is set to a large negative value.

We refer to this process as *phone deactivation pruning*. The posterior probability threshold used to make the pruning decision may be empirically determined using a development set, and is constant for all phones. The effect of varying this threshold on both recognition accuracy and CPU time is reported in Section IV.

Phone deactivation pruning takes advantage of the fact that our basic acoustic component estimates posterior probabilities rather than likelihoods. (Posteriors may be regarded as discriminative probabilities that do not incorporate an estimate of $P(\text{data})$). Direct estimation of posteriors saves summing over baseform HMM's, which would be required to carry out an equivalent approach in a likelihood-based system—a costly computation for a large context-dependent system. The channel-bank-based approach of Gopalakrishnan *et al.* [3] does attempt to use likelihoods to carry out a similar operation to phone deactivation pruning. However, this approach is somewhat more complex and requires phone-dependent thresholds. When used in a fast match, a factor of two speedup is achieved with a 5–10% increase in search error.

Furthermore, phone deactivation pruning can diminish the effect of inaccurate estimation of small posterior probabilities by connectionist estimators. It is known that connectionist networks estimate small probabilities poorly (where "small" is defined relative to the amount of training data) [1]. Phone deactivation pruning may be used with a threshold that is effectively at the limit of accurate probability estimation, thus avoiding the need to compare small, inaccurately estimated posterior probabilities.

IV. Experiments

A combined pruning strategy using both likelihoods and posterior probabilities has been implemented in the search component of ABBOT. The search algorithm used is based on stack decoding and is partially time-asynchronous. For efficiency the lexicon is tree-structured, with multiple pronunciations treated as separate lexical items.

TABLE I

DECODING PERFORMANCE ON THE *WALL STREET JOURNAL* TASK, USING A 20 000 WORD VOCABULARY AND A TRIGRAM LANGUAGE MODEL. WORD ERROR RATE, RELATIVE SEARCH ERROR (THE ERRORS DUE TO PRUNING THE MOST LIKELY HYPOTHESIS DURING THE SEARCH) DUE TO PHONE DEACTIVATION PRUNING AND CPU TIME (IN MULTIPLES OF REALTIME ON AN HP735) ARE GIVEN WITH RESPECT TO VARYING THE LIKELIHOOD ENVELOPE AND THE POSTERIOR-BASED PHONE DEACTIVATION PRUNING THRESHOLD. THE MAXIMUM NUMBER OF HYPOTHESES CONSIDERED AT ANY TIME WAS SET TO BE 31

20K Trigram, Trained on SI-84							
Pruning Parameters		si_dt_s5			Nov_92		
Envelope	Threshold	Time	Error/%	Search Error/%	Time	Error/%	Search Error/%
10	0.0	165.3	12.2	0.0	175.1	12.4	0.0
10	0.000075	16.1	12.1	-0.8	15.7	12.6	1.6
10	0.0005	4.3	12.2	0.0	3.9	12.9	4.0
10	0.003	1.4	14.3	17.2	1.3	14.9	20.2
8	0.0	46.8	12.5	0.0	50.4	12.6	0.0
8	0.000075	5.4	12.2	-2.4	4.9	12.8	1.6
8	0.0005	1.7	12.6	0.8	1.5	13.6	7.9
8	0.003	0.6	15.0	20.0	0.6	15.8	25.4

We have experimented with phone deactivation pruning using this system applied to the ARPA *Wall Street Journal* (WSJ) task. This is a very large vocabulary speaker-independent task, which uses read speech data. In the experiments reported here, ABBOT was used with a 20 000 word vocabulary[1] and a backed-off trigram language model[2]. The acoustic model used was a combination of recurrent networks with 78 phone classes (plus silence) trained on the WSJ0 short-term speaker data (SI-84).

These experiments were designed to evaluate phone deactivation pruning in terms of word error and computer (CPU) time. For the experiments, two parameters were varied: The phone deactivation threshold and the likelihood envelope.

The experiments were carried out using two data sets. The first, labeled si_dt_s5[3] contained 216 utterances from ten speakers. This was a 5000 word closed vocabulary set (the sentences being filtered from a larger open vocabulary set). In the experiments here, the 20 000 word dictionary and standard trigram language model were used for both sets. The second set, labeled Nov_92[4], contained 333 utterances from eight speakers. This used a 64 000 word vocabulary; about 1.9% of the words were out of vocabulary with respect to the 20 000 word dictionary. Comparative results with a varying envelope and phone deactivation threshold are presented in Table I.

These results indicate that phone deactivation pruning can improve the search time in this large vocabulary recognition task by up to a factor of 10, with less than 2% relative search error. Increasing the amount of posterior-based pruning results in increased relative search error (up to 25%), but can improve speed by a factor of over 100.

For the 20k trigram task we have found that "evaluation quality" decoding (i.e., minimal search error) can be obtained in 15× real time on a Hewlett-Packard HP735 workstation.

[1] Pronunciation dictionary provided by Dragon Systems.

[2] Provided by MIT Lincoln Laboratories.

[3] This was the ARPA 1993 spoke five development set, using Sennheiser microphone.

[4] This was the ARPA 1992 20k open NVP evaluation set.

We achieved realtime search on the same computer with an added 7% relative search error using an envelope of eight, a maximum of seven hypotheses at any time frame and a phone deactivation pruning threshold of 0.005. Further experimentation on more recent development and test sets released by ARPA conform to the speed/accuracy tradeoff, reported here.

V. Conclusion

In this letter we have introduced a novel pruning strategy for large vocabulary continuous speech recognition based on local posterior phone probability estimates. Phone deactivation pruning is of particular application to hybrid connectionist/HMM systems. Experiments on the *Wall Street Journal* database using a 20 000 word vocabulary and trigram grammar have demonstrated that search time may be reduced by up to a factor of ten with the introduction of little or no search error. A speed improvement of two orders of magnitude is possible with up to 25% relative search error.

Acknowledgment

The author would like to thank M. Hochberg and A. Robinson for many fruitful discussions. The author acknowledges MIT Lincoln Laboratory for language model provision and Dragon Systems for pronunciation dictionary provision. This work was partially carried out at Cambridge University Engineering Department.

References

[1] E. Barnard and E. C. Botha, "Back-propagation uses prior information efficiently," *IEEE Trans. Neural Networks*, vol. 4, pp. 794–802, 1993.

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

[3] P. S. Gopalakrishnan, D. Nahamoo, M. Padmanabhan, and M. A. Picheny, "A channel-bank-based phone detection strategy," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 161–164, Adelaide, Australia, 1994.

[4] R. Haeb-Umbach and H. Ney, "Improvements in beam search for 10 000-word continuous-speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 353–356, 1994.

[5] 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. Int. Conf. Spoken Language Processing*, pp. 1499–1502, Yokohama, Japan, 1994.

[6] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 161–175, 1994.

[7] A. J. Robinson, "The application of recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, vol. 5, pp. 298–305, 1994.