An Approach to Continuous Speech Recognition Based on Layered Self-Adjusting Decoding Graph

Qiru Zhou and Wu Chou

Bell Laboratories
Lucent Technologies
600 Mountain Avenue
Murray Hill, New Jersey, NJ 07974, U.S.A.
Email: wuchou@research.att.com

ABSTRACT

In this paper, an approach of continuous speech recognition based on layered self-adjusting decoding graph is described. It utilizes a scaffolding layer to support fast network expansion and releasing. A two level hashing structure is also described. It introduces self-adjusting capability in dynamic decoding on general re-entrant decoding network. In stack decoding, the scaffolding layer in the proposed approach enables the decoder to look several layers into the future so that long span inter-word context dependency can be exactly preserved. Experimental results indicate that highly efficient decoding can be achieved with a significant savings on recognition resources.

1. INTRODUCTION

The use of long span language models and the increase in recognition vocabulary size have pushed decoding and search methods in speech recognition to another level of complexity. In general, the best first based stack decoder often makes it easy to incorporate long span language models. But it requires good search heuristics to estimate the least upper bound of the path score in order to avoid search errors. On the other hand, the breadth first based beam search does not require heuristics, and the search can also be made frame-synchronous. However, one of drawbacks of the breadth first search is the amount of resources needed to support the large number of active nodes in the search beam. Integrating these two basic search schemes in speech recognition is one of major developments in decoder design.

Search algorithms which mix these two basic search schemes in decoding cannot be simply categorized as being a stack decoder or a beam search decoder. We call this new class of decoder, *quasi-stack* decoder. In quasi-stack decoder, it can be a stack based decoder but using beam search in local expansions, such as the envelop search based multi-stack decoder[2]. It can also be a beam search decoder and is frame-synchronous. But it also utilizes some stack decoding methods in determining arc expansion at word junctions, such as the one-pass dynamic decoder[3]. Quasi-stack decoder is quite successful in making it possible to integrate long span language models and detailed acoustic models in various tasks with manageable decoding resources. It also raises many interesting issues in search. In one-pass dynamic decoder, it is based on a non re-entrant decoding network. A non re-entrant decoding network is a network without looping back re-entrant arcs. However, in most applications, the decoding network is often compactly represented by a general re-entrant decoding networks. Mapping a re-entrant decoding network to a non re-entrant decoding work will lead to a network expansion. For the envelop search based multi-stack decoder, one issue is how to incorporate inter-word context dependency in a left to right search process. The inter-word context dependent model units often depend on the context occuring in the future substacks.

In this paper, an approach to continuous speech recognition based on layered self-adjusting decoding graph is described. It removes the non re-entrant constraint in one-pass dynamic decoding, and one-pass dynamic decoding is extended to general re-entrant network. In envelop search based multi-stack decoding, the proposed approach makes it possible to look several layers into the future in determining the correct context dependency at the current substack, which significantly improves the search envelop.

2. LAYERED SELF-ADJUSTING DECODING GRAPH

One observation made on the decoding network used in speech recognition is that the decoding network is very clumsy due to the integration of various knowledge

sources in the decoding process. This not only makes it difficult to expand the network dynamically but also makes it expensive to locally rebuild the needed pieces in order to sustain the search.

The approach of layered self-adjusting decoding graph proposed in this paper is based on a dynamically constructed scaffolding (skeleton) layer which serves as an anchor point layer to support fast network expansion and reconstruction. Fig 1. illustrates the decoding network structure used in this approach. The self-adjusting decoding graph is constructed by slicing the decoding network horizontally according to different knowledge sources and by slicing the decoding network vertically for each time instant. In this approach, the decoding graph can be sliced into three layers: word net layer, phone net layer and the DP (dynamic programming) layer. The word net layer keeps the word level knowledge sources such as grammar constraints and specifications from the language model. Phone net layer keeps the phonetic level knowledge sources such as phonetic lexicon graph, etc. The DP net layer is based on the integration of all knowledge sources and is the place where dynamic programming is performed.

The word net and phone net are independent from the acoustic model and do not carry any time information. The use of cross-word tri-phone models and other detailed acoustic modeling techniques in search will only increase the size and the complexity of the DP net. But it has no effect on the upper two layers. Such a horizontal slicing of the decoding network has a very nice property that more than 90% memory usage is at the lower DP layer. The upper two layers form a very thin and light scaffolding structure upon which very fast dynamic DP net expansion and releasing can be supported. The layered self-adjusting decoding graph is both constructed dynamically and released frame synchronously. Dynamically grown the scaffolding structure in this approach is relatively easy, because it is a very thin slice corresponding to a focused knowledge source, such as word level or phonetic level information. The real issue is the dynamic expanding and releasing of the DP net according to various search strategies.

3. ONE-PASS DYNAMIC DECODING IN RE-ENTRANT NETWORK

The basic idea used in the proposed approach is to make the decoding network (in particular the DP net) into a self-adjusting graph which is self-adjusted to serve the need of one particular time frame. If viewed from the time axis, this corresponds to a vertical slice of the decoding network, $Net\_active(t)$, as illustrated by

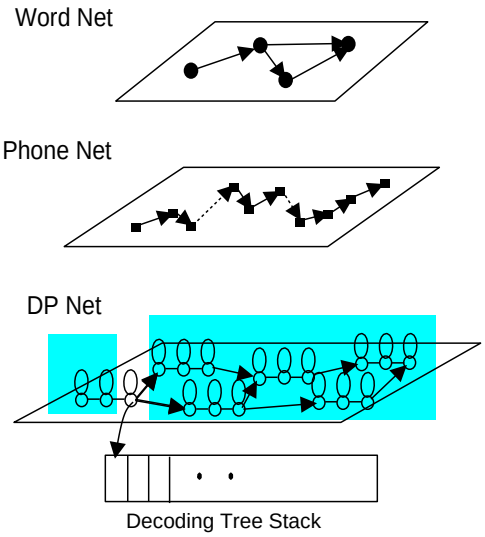

Figure 1: A layered decoding network architecture.

the unmarked area of DP net in Fig 1. $Net\_active(t)$ is the active portion of the decoding network at the time instant that the self-adjusting dynamic decoding graph represents. It is an approximation of the minimum and sufficient local decoding network to cover the need of active nodes at the current time frame. It is determined mainly by the acoustic model matching. It is relatively independent of the depth of the decoding grammar and the duration of the speech input. In layered self-adjusting decoding graph,

$$\begin{aligned} \max(size(Net)) &\approx \max(size(Net\_active(t))) \\ &\leq Net\_static. \end{aligned} \quad (1)$$

The actual search process is performed at the DP layer. Active path is represented as a node in the corresponding decoding (search) tree of the search graph[6]. Each node in the decoding tree carries with itself all necessary links, scores and trace-back informations. The self-adjusting capabilities are obtained by introducing a two level fast dynamic hashing structure in the layered decoding graph. It hashes the remaining fragments in DP layer after network pruning to the scaffolding layer. The exact pieces needed to support next input frames can be, therefore, identified and constructed.

When the decoding network is non re-entrant, the network expansion is performed forward and never go back. Therefore, locally the network is not punctured by dynamic network releasing. This situation is no longer true in re-entrant decoding network. It is often that the used network is reentered which is already partially or even totally released. This can result in the

so called "wanton expansions", where the network generation engine can not locate the remaining pieces of the used network and generate duplicated structures. This, in the past, has limited the one-pass dynamic decoding to only non re-entrant decoding network.

In the proposed approach, the two level dynamic hashing structure is done at the phone nodes of the scaffolding layer. At the phone node, it hashes the related nodes existing in the DP net to its node table and it hashes the related arcs (HMM instances) existing in the DP net in its arc table. Beam search is performed for all nodes in the decoding tree. Network pruning procedure is then followed to release nodes and arcs which are not active. The network expansion engine will grow the needed pieces based on the hashing information and self-adjusting the shape of the dynamic decoding graph. As a consequence, it results in a minimum and sufficient decoding graph covering the needs of supporting the remaining active nodes for the next time instant.

There are three structures in the layered self-adjusting decoding graph that are managed dynamically for the DP net. These are dynamic programming arcs, dynamic programming nodes and the hashing tables. The dynamic programming arcs are released once they are not an active part in the search beam. The dynamic programming nodes are released when they no longer terminate an active dynamic programming arc. The dynamic hashing tables, which are used to hold the record of existing arcs and nodes under the scaffolding layer, are released when they become empty.

When reentering a used decoding network, the dynamic network generation engine first examines the node hash table at the scaffolding phone node to re-establish the missing nodes and reuse the existing nodes which are leftovers from the previous time instant. Then it examines the arc hashing table at the scaffolding phone node to re-establish the missing arc connections and reuse the existing ones. The use of fast dynamic hashing is important because linear search can incur quite a cost in recovering a punctured network.

Under the layered self-adjusting decoding graph, one-pass dynamic decoding is made possible on punctured decoding network, and therefore, it no longer requires the decoding network to be non re-entrant. Moreover, one key issue in one-pass dynamic decoding is how to control the network growth. Unlike the conventional beam search, the one-pass dynamic decoding is a quasi-stack decoding scheme. Active word end nodes are sorted and linked into various dominance chains. These chains in fact form a stack, and each dominance chain is an entry in the stack. Only the most likely word-end on the chain is allowed to create successor nodes, and each expanding word-end node has its own copy of its successor network. This can lead to a tremendous growth in decoding network if not controlled properly. The proposed layered self-adjusting decoding graph provides enhanced control on the network growth and can also support a much deeper network pruning due to its unique scaffolding layers.

4. LAYERED SELF-ADJUSTING DECODING GRAPH FOR MULTI-STACK DECODER

The envelop based multi-stack decoding is according to the forward A^* principle[4]. It is based on the estimated likelihood score of making a partial theory into a complete theory

$$f_i(t) = g_i(t) + h^*(t), \tag{2}$$

where $f_i(t)$ is the estimated log-likelihood of completing a sentence with the partial theory i ending at time t, $g_i(t)$ is the log-likelihood of partial theory i, and $h^*(t)$ is the log-likelihood of the best extension of any theory from time t to the end of the data. Define

$$f^*(t) = g^*(t) + h^*(t), \tag{3}$$

for the best theory with a word transition at time t, $g^*(t) = lubg_i(t)$ is the least score upper bound for all partial theory ending at time t. The envelop of the stack is

$$\hat{f}_i(t) = f_i(t) - f^*(t) = g_i(t) - g^*(t), \tag{4}$$

which is built around the least score upper bound on the partial theories at time t. In multi-stack decoding, the envelop depth is specified by a depth constant Δ. Partial theory i will be dropped from the stack if $\hat{f}_i(t) > \Delta$.

In multi-stack decoding, active partial theories are sorted into substacks corresponding to their end-time. Partial theories in the same substack are also sorted according to their likelihood scores and the current least upper bound on partial theory scores $g^*(t)$ is determined. In this approach, $g^*(t)$ determines the search envelop, and a good estimate of $g^*(t)$ is crucial. In applying inter-word context dependent model units in decoding, partial theories ending at time t become conditioned on the possible succeeding word sequences which can complete the partial theories at time i. Words in succeeding word sequences are described in future substacks, and this makes it difficult to incorporate correct context dependency in envelop based stack decoding.

In layered self-adjusting decoding graph, the scaffolding layer provides an extra dimension in search.

It allows the stack decoder to look context dependencies of the future through the scaffolding layer without the need to expand the future stacks. In this way, exact inter-word context dependencies can be determined during each expansion of partial theories in the stack. The context dependency lookup can go down several levels into the future, and this makes it possible for incorporating inter-word context dependencies even for one-phone words which cross two word junctions.

5. EXPERIMENTAL RESULTS

The approach of layered self-adjusting decoding graph were applied to three search schemes, namely beam search, one-pass dynamic decoding and envelop search based multi-stack decoding. One advantage in the proposed approach is that various search schemes can be built in one framework using the same scaffolding structure. Switching from one decoding scheme to another is seamless, even though each decoding scheme has its unique search strategy and properties. Beam search based decoding is often more stable and convenient in dealing with complicated grammar constraints. A beam search decoder based on layered self-adjusting decoding graph can reduce the decoding resources significantly[5]. One-pass dynamic decoding and multi-stack decoding are quasi-stack decoding schemes and more suitable for long span N-gram based language model.

In our experiments, we found that one-pass dynamic decoding is more robust with respect to the quality of search heuristics. This is because it is mainly a beam search decoder. On the other hand, good search heuristics are very important for multi-stack decoding. One-pass dynamic decoding is a frame-synchronous decoding scheme but needs good control of the network expansion. It is much easier to control network expansion in multi-stack decoder because network expansion is separated in time and in various substacks. However, multi-stack decoder is mainly a stack decoder and it is not frame-synchronous. It's processing block length should be greater than the maximum allowed duration of words (including compound words) in the vocabulary.

We tested the one-pass dynamic decoding on a 7K vocabulary tasks. More than 80% words in the vocabulary were long compound word phrases. A phone tree constructed from this vocabulary had an average 15.2 phones per word, which is roughly the size of a phone tree for 20K regular word vocabulary (if the average phone per word is five). Table 1 compares the peak memory usage between a layered self-adjusting decoding graph based one-pass dynamic decoder and a conventional static beam search decoder. The one-pass dynamic decoder based on the proposed approach was quite efficient. It ran the task real time on an 150MHz SGI R4400 machine without using special fast match method. The peak DP memory usage was reduced by a factor of five.

	DP	Total Usage
Baseline	30MB	41MB
Layered graph	6MB	16MB

Table 1: Comparisons between different approaches

6. SUMMARY

In this paper, an approach of continuous speech recognition based on layered self-adjusting decoding graph is described. It utilizes a scaffolding layer to support fast network expansion and releasing. A two level hashing structure is also described. It introduces self-adjusting capability for dynamic decoding on general re-entrant decoding network. In stack decoding, the scaffolding layer in the proposed approach enables the decoder to look several levels into the future so that long span inter-word context dependency can be exactly preserved. Experimental results indicate that highly efficient decoding can be achieved with a significant savings on recognition resources.

REFERENCES

[1] Chin-Hui Lee and Lawrence R. Rabiner "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition", *IEEE Trans. on Acoustic and Signal Processing*, vol 37, No. 11, November 1989.

[2] S. Renals and M. Hochberg "Efficient Evaluation of the LVCSR Search Space Using the Noway Decoder", *Proc. ICASSP-96, vol 1, pp149–153*

[3] J.J. Odell, V. Valtchev, P.C. Woodland and S.J. Young "A One Pass Decoder Design for Large Vocabulary Recognition", *DARPA Work-Shop on Continuous Speech Recognition, March 94*, pp 380–385.

[4] D. Paul "An Efficient A^* Stack Decoder Algorithm for Continuous Speech Recognition with a Stochastic Language Model", *Proc. ICASSP-92, vol. 1, pp25–29*

[5] E. Burhke, W. Chou and Q. Zhou "A *Wave* Decoder for Continuous Speech Recognition", *Proc. ICSLP-96.*

[6] N.J. Nelsson "Principles of Artificial Intelligence", *McGraw Hill*, 1980.

[7] H. Ney, "Architecture and Search Strategies for Large-Vocabulary Continuous Speech Recognition", *pp 59 – 84* NATO-ASI BUBION 93.

[8] J. Pearl "Heuristics", *Addison and Weiley 1984*