6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

REDUCING TIME-SYNCHRONOUS BEAM SEARCH EFFORT USING STAGE BASED LOOK-AHEAD AND LANGUAGE MODEL RANK BASED PRUNING

Jian Wu and Fang Zheng

Center of Speech Technology, State Key Lab. of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University, P.R.China
jwu@sp.cs.tsinghua.edu.cn http://sp.cs.tsinghua.edu.cn

ABSTRACT

In this paper, we present an efficient look-ahead technique based on both the Language Model (LM) Look-Ahead and the Acoustic Model (AM) Look-Ahead, for the time-synchronous beam search in the large vocabulary speech recognition. In this so-call stage based look-ahead (SLA) technique, two predicting processes with different hypothesis evaluating criteria are organized by stages according to the different requirements for pruning the unlikely surviving hypotheses. Furthermore, in order to reduce the efforts for distributing the LM over the lexical tree more effectively, the LM Rank based Pruning (LMRP) is integrated with the extension of each new phoneme node. The recognition experiments performed on the 50k-word Mandarin Dictation task (Easytalk2000) show that a reduction by 10 percents in the search effort in comparison with the standard word-conditioned search using LM look-ahead only, and a reduction of 25 percents in the word error rates in comparison with the search algorithm without any look-ahead can be achieved.

1. INTRODUCTION

It is well known that establishing an effective hypothesis search algorithm plays an important role in the construction of the Large Vocabulary Continuous Speech Recognition systems. In order to make the system output the reasonable results in only few times of the real time, the trade-off between the quick response and the precise algorithm should be considered. In the present work, a lot of efforts have been made for this purpose and reached good performance, one of which is the word-conditioned time-synchronous beam search [1].

The word-conditioned time-synchronous beam search is one kind of the one-pass search strategies based on the copies of the tree-organized pronunciation lexicon. By this so-call lexical tree, it can incorporate an exact or approximate n-gram language model into the hypothesis pruning process easily and thus enhance the prediction effect of the language model, which is called the Language Model Look-Ahead (LMLA). Another mostly used technique in the word-conditioned beam search is the Phoneme Look-Ahead (PLA) [2], which prunes the active phoneme nodes according to the approximate acoustic probability estimation. By using the PLA, a lot of unlikely surviving hypotheses are pruned before being extended. Both of these two techniques can reduce unnecessary computation dramatically without too much loss in accuracy and can be easily combined together to achieve more robust performance [3].

In this paper, we propose another efficient Look-Ahead technique, namely SLA (Stage-based Look-Ahead), which is composed of a tri-gram Language Model Look-Ahead process and a HMM State Look-Ahead (HSLA) process. The idea of HSLA is the extension of the PLA. However, it does not need to check the acoustic frames of average phoneme duration in advance to determine which phoneme node can be extended as the probably correct one. Furthermore, in the context of the SLA, the shifting between the process of LMLA and HSLA can be controlled freely for the "stage" here means not only the state of the underlying HMM but also the position of the node in the lexical tree.

Secondly, we will describe a novel pruning strategy, namely LMRP, which is based on the rank of both the factored uni-gram probability of the next extended phoneme and the accumulated score of the current hypothesis. This approach ensures that, the more "excellent" one hypothesis behaves in the history, the more surviving chances it will get in the future.

The organization of this paper is as follows. In Section 2, the techniques of LMLA and PLA are reviewed. In Section 3, the HSLA is introduced briefly and the algorithm of SLA is described in detail. The LMRP strategy is presented in Section 4. In Section 5, the experiment results based on the 50k-word Mandarin Dictation task are given.

2. REVIEW OF LOOK-AHEAD TECHNIQUES

In this section, we mainly focus on the one-pass search methods seeking the best path of the possible hypotheses through the lexical tree. We will introduce the basic algorithm of the word-conditioned synchronous search, and then the idea of LMLA and PLA.

2.1. Basic Algorithm of Word-Conditioned Time-Synchronous Search

The idea of traditional word-conditioned search is to establish a new copy of the lexical tree at the end of each possible word pair. During the search process within the lexical tree, only the acoustic score of the passed phonemes from the root are accumulated and combined with the history score before entering this tree. The tri-gram scores are considered only at the end frame of one word, which means that the language model will have no any influence on the pruning of hypotheses before the words are identified. The following notation can be defined to express the idea:

Let $Q_{uv,n}(t,s)$ denote the quantitative score of the hypothesis which stays in the state *s* of the node *n* in the lexical tree with the history word pair (u,v) at current time *t*.

Therefore, for the hypothesis that still lies within a certain word, the recursion formulation for accumulating score is:

$$Q_{uv,n}(t+1,s)=\max_{s}\{q_n(x_{t+1},s|\boldsymbol{s})\cdot Q_{uv,n}(t,\boldsymbol{s})\}, \tag{1}$$

where $q_n(x_t,s|\boldsymbol{s})$ is the corresponding transformation or emission probability of HMM.

For those hypotheses that reach the leaves of the lexical tree, which represent the word boundaries, the tri-gram probabilities should be integrated with the accumulated scores and then the new copies of the lexical tree are duplicated and assigned to the best hypothesis paths with the initializing scores as Eq. (2) .

$$Q_{vw,root}(t,0)=\max_{u}\{p(w|u,v)\cdot Q_{uv,S_w=n}(t,\boldsymbol{q}(n))\}, \tag{2}$$

where $root$ means the root of the lexical tree, $p(w|u,v)$ means the conditional tri-gram probability, $S(w)$ means the last phoneme node of the word w, and $\boldsymbol{q}(n)$ denotes the last state of node n.

By using these two quantities, we have a dynamic programming recursion to evaluate all the hypotheses not only in the word interior but also at the word boundaries.

2.2. Language Model Look-Ahead

The purpose of LMLA is to incorporate language model probabilities into the early stage of the synchronous search within the lexical tree. It can be achieved by calculating a factored LM probability for each phoneme through all the leave nodes that can be reached from this phoneme. The basic operation of the factorization could be SUM or MAX over all the language model probabilities associated with the corresponding leave nodes. For example, assuming that m is the parent node of the node n, the factored probability of the node n can be defined as:

$$h_{uv}(n|m)=\max_{w\in\Pi(n)}p(w|u,v)\Big/\max_{w\in\Pi(m)}p(w|u,v), \tag{3}$$

where $P\ (n)$ represents the set of the leaves that can be reached from the phoneme node n.

After the factored probabilities are calculated, they can be incorporated into the quantities of hypotheses while phoneme transferences occur:

$$Q_{uv,n}(t+1,0)=h_{uv}(n|m)\cdot q_n(x_{t+1},0|\boldsymbol{q}(m))\cdot Q_{uv}(t,\boldsymbol{q}(m)). \tag{4}$$

Actually, the factored probabilities can be calculated on demand by a dynamic program procedure and then cached in the memory to reduce the chances of redundant computation.

2.3. Phoneme Look-Ahead

When a phoneme node is started in the search process to form a new hypothesis, an approximate probability, which is referred to as phoneme look-ahead score, can be estimated and then used to check whether it is probable to survive the following pruning process. In the so-call PLA strategy, a few frames of acoustic feature vectors, which cover about an average phoneme duration, should be looked ahead to estimate the predicting score. Furthermore, since the look-ahead score of the phoneme is computed by performing a time alignment procedure, the acoustic model used in predicting could not be the same one in the following steps. Usually, the CI phoneme models with a small number of component densities instead of the CD tri-phone models are adopted [3][4].

3.STAGE BASED LOOK AHEAD IN THE WORD-CONDITIONED SYNCHRONOUS SEARCH

Although the combination of the LMLA and the PLA can take both advantages of these two different techniques, it is still difficult to reach a good balance between the speed and accuracy. That is to say, if we prefer to get less error in the procedure of PLA, we have to compute the alignment scores more frequently and precisely. On the contrary, if we compute the phoneme look-ahead scores by a synchronous algorithm incrementally instead of the time alignment, a lot of search efforts will be reduced. Furthermore, if the score can be reused for further using, the detail model can also be used for predicting. The stage based look-ahead technique is accordingly proposed in this section to improve the search efficiency from the point of the above aspects.

3.1 HMM State Look-Ahead

According to the statistical experimental results, a lot of impossible candidate phonemes could be detected before reaching the last states of corresponding HMM during the time synchronous viterbi search. Under this assumption, it only needs to look ahead a few HMM states for the determination of the surviving phonemes.

Let $\tilde{q}(n,\boldsymbol{g};t_1,t_2)$ denote the look-ahead score of the phoneme n, which starts at time t_1 and ends at time t_2 and state *g*. Then the look-ahead score can be computed by a synchronous equation as Eq. (5) instead of the time alignment in PLA.

$$\tilde{q}(n,\boldsymbol{g};t_1,t_2)=\max_{s}\tilde{q}(n,\boldsymbol{s};t_1,t_2-1)\cdot q_n(x_{t_2},\boldsymbol{g}|\boldsymbol{s}) \tag{5}$$

It should be noted that, at each frame, the maximal score of the hypotheses ending at state *g* could be calculated and treated as the basis of the state pruning procedure shown by the following equations:

$$\hat{q}(\boldsymbol{g};t_2)=\max_{n,t_1}\tilde{q}(n,\boldsymbol{g};t_1,t_2) \tag{6}$$

and

$$\hat{q}_b(\boldsymbol{g};t_2)=bw(\boldsymbol{g})\cdot\hat{q}(\boldsymbol{g};t_2). \tag{7}$$

$bw(\boldsymbol{g})$ means the threshold of beam width at state *g* and the phoneme with the scores less than $\hat{q}_b(\boldsymbol{g};t_2)$ are removed from the candidate set.

Since these phonemes have different beginning time and then different duration time, the look-ahead scores should be normalized before using Eq. (6) and Eq. (7) according to the number of crossed frames:

$$\tilde{q}_{norm}(n,g\,;t_1,t_2) = \tilde{q}(n,g\,;t_1,t_2)^{1/(t_2-t_1)}\,. \tag{8}$$

Furthermore, the look-ahead score can also be combined with the history score $Q_{uv}(t_1,q(m))$ as the second pruning tier to check the remaining nodes more carefully.

Although the pruning equations of HSLA and PLA are quite similar, there are still many differences between them, especially in the variant basic units for confidence evaluation. Firstly, in the PLA, the quantity of evaluation depends on the whole phoneme. In the HSLA, the quantity depends on the passing HMM states, which are only part of the phoneme. Secondly, the PLA compares all the phonemes that begin at the same time, while the HSLA compares those candidate phonemes that have the same ending state. Thirdly, the PLA can only be performed at the start of the phoneme, while the HSLA can be started at any state. The first property of HSLA avoids a great number of computations of the time alignment. The latter two ensure that the HSLA can be performed in a simple synchronous way and thus the searching does not need to rewind back to the beginning of that phoneme.

3.2 Stage based Look-Ahead Technique

In the word-conditioned search algorithm utilizing traditional look-ahead techniques, the factored language model probabilities could only be incorporated into the accumulated score at the beginning or the ending of each tree nodes. Actually, the factored LM score can also be seen as a linguistic penalty on the state transformation during search. Therefore, it could be added at any time if required and appropriate. In addition, as mentioned before, the HSLA can be executed in the synchronous framework, so it can also be combined with the Language Model Look-Ahead easily.

In order to take advantage of the both properties, a predicting technique named Stage based Look-Ahead is proposed, in which the search within the lexicon tree is divided into several stages and each of them has a unique tag: “A” or “L”. If one certain stage is tagged with “A”, then all the hypotheses that belong to the stage should be checked by HSLA frame by frame. If the tag is “L”, then the accumulated scores of all the hypotheses should be integrated with the relative factored LM scores at the beginning of the stage. Generally speaking, each phoneme can be assigned two stages: the stage “A” first and then the stage “L”. Hence the search within the lexical tree using SLA technique can also be treated as the exploring within a series of stage “A” and stage “L” in turn. The transferring time from the stage “A” to the stage “L” is not fixed. It could occur at any frame where HMM state transformation may occur. For example, we could define HMM state *M* as the dividing line between these two stages, then all the hypotheses of which the current occupancy states are less than *M* belong to stage “A” of related phonemes and the rest belong to stage “L”.

Figure 1 is an illustration of the pruning result among three hypotheses after adopting the SLA technique, supposed that the pruning is taken action at each encountering point of any two paths and only the best path (shown as the solid line) is preserving.

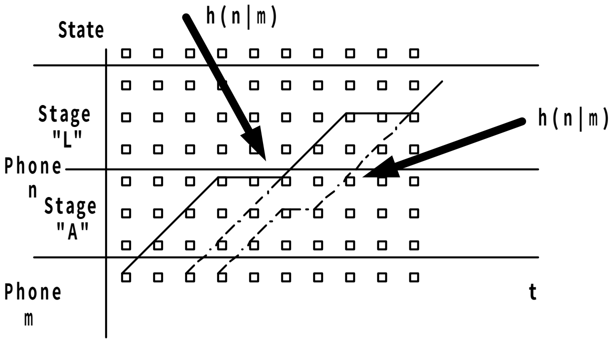

Figure 1: The illustration of the algorithm of the SLA

4. PRUNING STRATEGY BASED ON THE RANK OF LANGUAGE MODEL PROBABILITY

Although lots of efforts, such as calculate-on-demand, have been adopted for reducing the computation of factoring LM score, it still needs to spend much CPU time and other resources on looking for the tri-gram probabilities and factorization when starting a new lexical tree because hundreds of derived leaf nodes should be considered for working out the factored score of each node in the first layer of lexical tree. Obviously, one of the solutions is to make the number of look ahead scores to be factored as less as possible without too much increase of search errors. The PLA and HSLA are such approaches to reduce directly the phonemes to be extended. But they may lose effects or even introduce more search errors for the inaccuracy of the predicting assumption when the hypothesis surviving condition becomes stricter with the increase of the pruning threshold.

The Language Model Rank based Pruning is an improving version of such solutions, which will focus on not only how to prune more phonemes reasonably but also how to make use of the information of reliabilities provided by the hypotheses with different histories. It assumes that those paths that have higher scores are more credible and thus they should be given more chances to be surviving. In other words, the better one hypothesis performs in the history, the more phonemes it would be extended to after leaving the current phoneme. But for the path that has lower score, only a few phonemes that are most likely to be extended are considered.

In order to implement such a ‘prejudice’ policy, the hypotheses and the phonemes should be evaluated or ranked according to some certain decision rules. It is indubitable that the accumulated score is the most believable criteria for the hypothesis. As for the phoneme, both of the HMM state look-ahead score and the factored LM score are suitable. In the LMRP, only the factored uni-gram probabilities are adopted to determine which phoneme is likely to be retained because the state look-ahead score will be utilized in the HSLA procedure of the following stages. The Figure 2 shows the detail algorithm of the LMRP.

	Active states per frame		Active nodes per frame		Active trees per frame		WER(%)	
	I	II	I	II	I	II	I	II
No Look Ahead	261.8	278.1	801.7	900.2	3.9	4.2	11.9	32.1
LMLA	251.7	274.3	756.3	823.2	3.2	3.8	11.0	23.1
HSLA+LMLA (M=3)	253.6	272.5	681.9	804.7	2.7	3.4	10.8	23.3
HSLA+LMLA+LMRP	252.4	271.6	672.0	785.3	2.7	3.3	10.8	23.4

Table 1 Comparison on the search efforts and word error

```
Off-line work:
    Calculate all the factored uni-gram scores of the lexical tree.
    Set up a rank-beamwidth lookup table: $RB=\{(r,bw(r))\}$ for $r=1..R+1$ and $0=bw(1)<bw(2)<..<bw(R)<bw(R+1)=1$.
    For any node m in the lexical tree
        Give each son of node m a rank tag that ranges from 1 to R according to the factored LM score. Make sure that:
        (A) the node with the highest score of all the sons belongs to the rank 1.
        (B) the rank value of the node with higher score is not greater than that with lower score.
On-line search:
    Sort all of the hypotheses and get the best score Sb.
    For all of those hypotheses
        If the score of the hypothesis lies between $bw(i)*Sb$ and $bw(i+1)*Sb$, then only the son phoneme with rank tag less than i can be considered for extension by the hypothesis.
```

Figure 2: Algorithm of the Language Model Rank based Pruning

5. EXPERIMENTAL RESULTS

All of the experiments are based on a 50k-Word Mandarin dictation system, namely Easytalk2000 [5], in which both the training and tests are carried on the corpus provided by the National 863 High-Tech Project. The training corpus includes about 41600 utterances from 80 male speakers. The test set I includes 1000 utterances from 2 male speakers and the test set II includes 300 utterances from 3 male speakers. Table 1 is the experimental results on these two test sets.

In the first experiments, we adjust the system with SLA to have the comparable search efforts as that of system without any look-ahead techniques. From the Table 1, we can find that the average word error rates can be reduced by approximate 25%.

In the second experiments, the problems of search effort are addressed. It should be noted that the number of active nodes is the most important influence facts on the search efforts since the complexity of factorization procedure is much greater than that of HMM state look ahead. From the last three columns of Table 1, we can find that in order to achieve a similar word error rates, the system adopted only LMLA should spend almost 6% CPU time more than SLA on the language model factorization. Furthermore, if the LMRP is used, the search efforts can even be reduced by about 10%.

6. CONCLUSIONS

In this paper, we present two techniques used in the one-pass synchronous search: Stage based Look-Ahead and Language Model Rank based Pruning. The first one is actually an extension of the idea of combining the PLA and the LMLA. But it works under a totally different framework, especially in the procedure of acoustic model look ahead. The LMRP is an effective rule based approach to incorporate the language model into pruning as early as possible. In our experiments, both techniques show their effects on reducing search efforts as expected.

7. REFERENCES

1. Ortmanns, S., Ney, H., Seide, F. and Lindam I., "A Comparison of Time Conditioned and Word Conditioned Search Technique for Large Vocabulary Speech Recognition," *Proc. Int. Conf. On Spoken Language Processing,* Philadelphia, PA, pp.2091-2094, October 1996.

2. Ney, H., Haeb-Umbach, R., Tran, B.H. and Oerder, M., "Improvements in Beam Search for 10000-word Continuous Speech Recognition," *Proc. IEEE Int. Conf. On Acoustics, Speech and Signal Processing,* San Francisco, CA, pp.13-16, March 1992.

3. Ortmanns, S., Ney, H., Eiden, A. and Coenen, N., "Look-Ahead Techniques for Improved Beam Search", *Proc. CRIM-FORWISS Workshop, Montreal,* pp. 10-22, October 1996.

4. Li, Z., Boulianne, G., Labute P., Barszca, M., Garudadri, H., and Kenny, P., "Bi-Directional Graph Search

Strategies for Speech Recognition", *Computer Speech and Language*, v10, pp.295-321, 1996.

5. Zheng, F., Song, Z.J., Xu, M.X. etc., "EasyTalk: A large-vocabulary speaker-independent Chinese dictation machine", *Proc. 6th European Conference on Speech Communication and Techniques*, EUROSPEECH' 99. Budapest, Hungary, 1999, v2, pp.819-822.