PAPER

Efficient Beam Pruning for Speech Recognition with a Reward Considering the Potential to Reach Various Words on a Lexical Tree

Tsuneo KATO[†a)], **Kengo FUJITA**[†], ***and*** **Nobuyuki NISHIZAWA**[†], ***Members***

SUMMARY This paper presents efficient frame-synchronous beam pruning for HMM-based automatic speech recognition. In the conventional beam pruning, a few hypotheses that have greater potential to reach various words on a lexical tree are likely to be pruned out by a number of hypotheses that have limited potential, since all hypotheses are treated equally without considering this potential. To make the beam pruning less restrictive for hypotheses with greater potential and vice versa, the proposed method adds to the likelihood of each hypothesis a tentative reward as a monotonically increasing function of the number of reachable words from the HMM state where the hypothesis stays in a lexical tree. The reward is designed not to collapse the ASR probabilistic framework. The proposed method reduced 84% of the processing time for a grammar-based 10k-word short sentence recognition task. For a language-model-based dictation task, it also resulted in an additional 23% reduction in processing time from the beam pruning with the language model look-ahead technique.

key words: *pruning, frame synchronous beam search, lexical tree*

1. Introduction

Automatic speech recognition (ASR) engines always demand fast search algorithms. Server-based ASR engines for internet search or dictation need the fast search to expand their domain and vocabulary with larger language models and more accurate acoustic models. Embedded ASR engines for command and control (C&C) or local search on mobile devices need the fast search as well to get the recognition result as soon as possible with efficient use of limited CPU power and memory. Standard HMM-based ASR engines enhance their search efficiency in two ways. Firstly, the search space is hierarchically structured in a word-level network which represents the acceptable sentences, and an HMM-state-level network which represents the words composing the sentences as a lexical tree [1]. Secondly, the hypotheses used to search for the best path on the lexical tree are effectively reduced by frame-by-frame pruning without deteriorating word accuracy. The basic pruning techniques are beam width pruning [2] and histogram pruning [3]. The idea of these pruning techniques is to retain the most promising hypotheses for further searches and exclude the rest with reference to their likelihoods. These techniques basically function well by setting proper thresholds. However, the required number of hypotheses to bring out the best word accuracy is still excessive as the vocabulary size increases.

Various methods have been proposed to shorten the processing time. Two-pass search algorithms which rescore the lattice output of the first pass are effective ways to introduce a detailed *N-gram* language model [4], [5]. The language model look-ahead technique [6]–[8] significantly reduces the number of hypotheses required for the maximal word accuracy by incorporating a language model as early as possible into search on the lexical tree. Furthermore, stepwise addition of the linguistic probabilities along the HMM state sequence [9], [10] enhanced the efficiency. However, these powerful techniques are not applicable in grammar-based tasks because they do not use linguistic probabilities.

Efficiency of the beam pruning has also been improved. Word-end pruning, which sets another beam width pruning for hypotheses at the terminal states of words, is effective in suppressing an explosive increase in hypotheses caused by transitions from a word end to various words [8]. The efficiency of the beam pruning was enhanced by introducing a posterior probability-based confidence measure [11]. Recently, a machine-learning-based optimization framework combining multiple criteria with more detailed features has been proposed [12]. However, the geometric property of the lexical tree has not been sufficiently exploited in beam search. Equal-depth pruning [13], which partitions the lexical tree into multiple sections based on depth levels and executes a beam width search at each section, is a method that focuses on the geometric property of the lexical tree. However, the performance is supposed to be unstable when the number of hypotheses is severely reduced, because the basis, i.e. the top likelihood in each section for every frame is chosen from a limited number of hypotheses in the section.

We propose improved beam pruning which takes the geometric property of the lexical tree into consideration. The property is that a hypothesis staying at a state close to the root of a lexical tree has greater potential to produce various word hypotheses than one close to a leaf, and that pruning of a hypothesis close to the root generally has a greater adverse impact on the accuracy of the resultant recognized word sequence than that of a hypothesis close to a leaf. The proposed method thus makes the hypotheses close to the root less likely to be pruned out by easing the pruning condition for the hypotheses close to the root. Unlike the language model look-ahead, the proposed method is applicable to the grammar-based tasks. The proposed method is applicable in combination with the language model look-ahead and/or the word end pruning as well.

The remainder of this paper is organized as follows. An analysis on the distribution of the number of reachable

Manuscript received August 20, 2010.
Manuscript revised December 25, 2010.

[†]The authors are with the KDDI R&D Laboratories Inc., Fujimino-shi, 356–8502 Japan.
a) E-mail: tkato@kddilabs.jp

DOI: 10.1587/transinf.E94.D.1253

words on a lexical tree, and the proposed method are described in Sect. 2. Experiments on the processing time (real time factor: RTF) and accuracy (word error rate: WER) in three recognition tasks are reported in Sect. 3. The relation between the proposed method and conventional pruning methods are discussed in Sect. 4. Conclusions are given in Sect. 5.

2. Beam Pruning with a Reward Considering the Potential to Reach Various Words on a Lexical Tree

2.1 Distribution of HMM States in Terms of the Number of Reachable Words on a Lexical Tree

The lexical tree is formed by merging the common partial HMM state sequences from the beginning of words between word entries in the lexicon. An example of the lexical tree is shown in Fig. 1. The potential to produce various word hypotheses depends on the HMM states in the lexical tree. A hypothesis at a state close to the root has great potential, whereas a hypothesis at a state close to a leaf has limited potential. This potential is quantifiable by the number of reachable words. As easily seen from Fig. 1, a lexical tree comprises a small number of states with great potential, and a vast number of those with limited potential.

We composed a lexical tree for the 10k-word railway station name task in Sect. 3, and investigated the distribution of HMM states in terms of the number of reachable words. Figure 2 shows the histogram. The vertical axis is on a logarithmic scale. Figure 3 shows the pie chart of the same distribution. The HMM states reaching a single word, two words, three and four occupy 71%, 19%, 3.2% and 1.8%, respectively. The number of HMM states decreases rapidly from those of a single reachable word to those of more reachable words. On the other hand, a few HMM states close to the root have hundreds or thousands reachable words. An HMM state next to the root has the maximal number of 1,738 reachable words in this case.

Naturally, the hypotheses on the lexical tree comprise a small number of those with a great number of reachable words, and a vast number of those with a few reachable words. As mentioned above, a pruned hypothesis with more reachable words impinges more than a pruned one with fewer reachable words on the word accuracy of the resultant word sequence. Therefore, we ease the pruning condition for the few hypotheses with a great number of reachable words, and tighten it for the vast hypotheses with few reachable words.

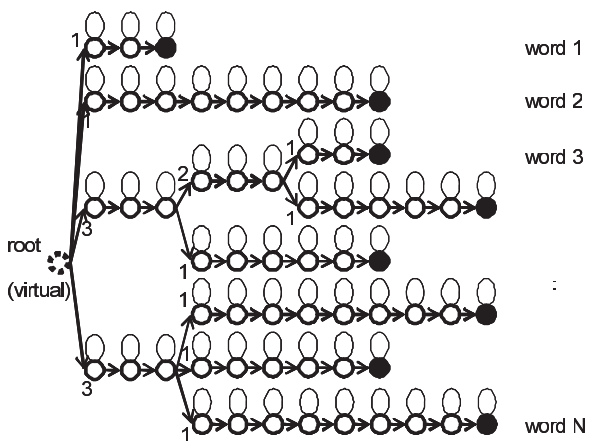


Fig. 1 Example of a lexical tree.
The numbers represent #reachable words from the HMM states.

2.2 Beam Pruning with a Reward as a Function of the Number of Reachable Words on a Lexical Tree

As a hypothesis advances on a path from root to leaf, the number of reachable words decreases monotonically, and is narrowed down to one after the hypothesis passes the last branching state in the lexical tree. Leveraging this property, a reward as a monotonically increasing function of the number of reachable words is tentatively added to the likelihood of the hypothesis for pruning. In addition, the value of the monotonically increasing function is set to be zero when the number of the reachable words is one ($f(1) = 0$). Consequently, as a hypothesis advances on the path, the reward starting from a positive value is renewed to a smaller value whenever the hypothesis passes a branching state, and then ends up at zero after the hypothesis passes the last branching state. The reward eases the pruning for the hypotheses closer to the root, while tightening it for the hypotheses closer to the leaves. Furthermore, this does not collapse the ASR probabilistic framework because the reward is always zero

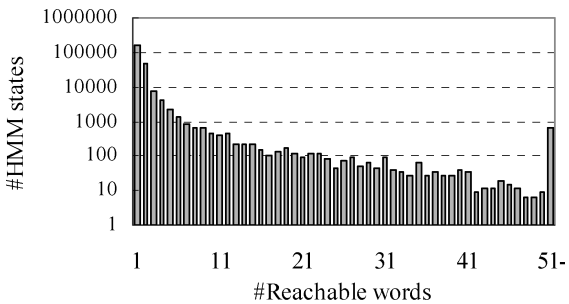


Fig. 2 Histogram of the HMM states in terms of the number of reachable words for the lexical tree of the railway station name task.

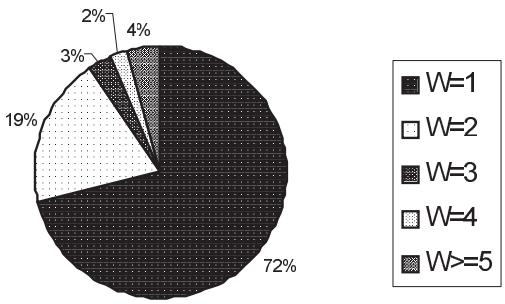


Fig. 3 Pie chart of the HMM states in terms of the number of reachable words for the lexical tree of the railway station name task.
W denotes #reachable words.

at the leaf HMM states unless homonyms or other words with that word as their prefix exist in the lexicon. Note, however, that the probabilistic framework is still preserved by discarding the reward in adding the word hypothesis into a lattice when homonyms or other words with that word as their prefix exist.

In a case of grammar-based recognition without linguistic probabilities, the score $S(h)$ for pruning of a hypothesis h is given by

$$S(h) = L_a(h) + R(W(h)) \tag{1}$$

where $L_a(h)$, $W(h)$ and $R(W)$ denote the accumulated acoustic likelihood, the number of reachable words of the hypothesis h in the lexical tree and the reward as a function of the number of reachable words, respectively. Strictly speaking, the reachable words depend on the grammatical context. However, $W(h)$ was precomputed approximately by just counting the number of reachable words on a lexical tree without considering the grammatical context here for the sake of simplicity.

In a case of recognition based on a probabilistic language model, the score $S(h)$ is given by

$$S(h) = L_a(h) + w_{lm}\{L_l + L_{la}(h)\} + R(W(h)) \tag{2}$$

where the additional parameters L_l, $L_{la}(h)$ and w_{lm} denote the accumulated linguistic likelihood from the word at the beginning to the previous word, the likelihood of language model look-ahead for the hypothesis h and the language model weight, respectively.

Considering the "long-tailed" distribution of the HMM states shown in Fig. 2, we assume two types of monotonically increasing functions which fulfill $R(1) = 0$, here. One is A) a logarithmic function as:

$$R(W) = a_{log}\left[log(W - b_{log}) - log(1 - b_{log})\right] \tag{3}$$

where a_{log} and b_{log} are constants to be optimized under the conditions $a_{log} > 0$ and $0 < b_{log} < 1$.

The other is B) an asymptotic exponential function converging on a value a_{exp} as:

$$R(W) = a_{exp}\left[1 - \exp\left\{-\frac{(W-1)}{b_{exp}}\right\}\right] \tag{4}$$

where a_{exp} and b_{exp} are constants to be optimized under the conditions $a_{exp} > 0$ and $b_{exp} > 0$. Outlines of the functions are shown in Fig. 4.

The pruning employed in this paper is standard beam width and histogram pruning. For beam width pruning, the maximum of $S(h)$ among all hypotheses is selected as the basis S_{max} every frame.

$$S_{max} = \max_{h \in H} S(h) \tag{5}$$

where H denotes the set of hypotheses at the frame. Then, all the hypotheses are determined to be retained or discarded according to whether the score $S(h)$ falls within beam width f_{GB} from the basis S_{max} or not. Hypotheses which fulfill the following inequation are retained.

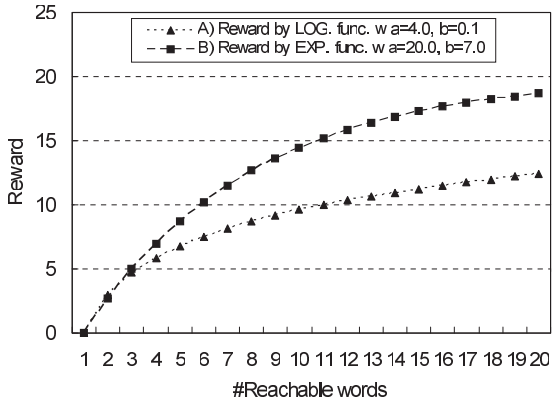


Fig. 4 Two monotonically increasing functions for the reward: A logarithmic function and an asymptotic exponential function. The constant values are optimized for a grammar-based task as follows: $a_{log} = 4.0$, $b_{log} = 0.1$, $a_{exp} = 20.0$ and $b_{exp} = 7.0$.

$$S(h) \geq S_{max} - f_{GB} \tag{6}$$

The histogram pruning is to limit the number of retained hypotheses under a predefined number N_{max}. To dispense with computationally expensive hypotheses sorting by their likelihoods, all the hypotheses are classified into ranges of a histogram once, and the hypotheses from the upper ranges are retained until the total number of retained hypotheses exceeds N_{max}. The beam width pruning and histogram pruning are used in combination.

Note that an excessive reward runs the risk of pruning out the correct hypothesis at a HMM state close to a leaf. However, the magnitude of the reward is controllable by the constants of functions.

2.3 Combination with Word End Pruning

The proposed method is to be evaluated in comparison to word end pruning (WEP) [8], and in combination with WEP. To suppress an explosive increase in hypotheses caused by cross-word transitions from word ends to various successive words, WEP applies another beam width pruning for hypotheses at the terminal states of words. (Note that the terminal states are indicated as filled circles in Fig. 1). The maximal likelihood among the hypotheses at the terminal states is found as:

$$S_{max\_leaves} = \max_{h \in H_{leaves}} S(h) \tag{7}$$

where H_{leaves} denotes the set of hypotheses at the terminal states. Then, only the hypotheses which fulfill the following inequation are retained.

$$S(h) \geq S_{max\_leaves} - f_{WEP} \tag{8}$$

Here, the beam width f_{WEP} is set tighter than the global beam width f_{GB}.

3. Experiments

3.1 Evaluation Tasks, Test Sets and Experimental Setup

The proposed method was evaluated by three recognition tasks: an isolated word recognition task, a grammar-based short sentence task without linguistic probabilities, and a dictation task based on a probabilistic language model. The isolated word recognition task is of 10k-word railway station names in Japanese. The short sentence task is of a formulaic train connection inquiry. The grammar accepts the pattern, "From ⟨a departure station⟩ to ⟨an arrival station⟩" in Japanese. The dictation task is a general mail dictation in Japanese on a 30k-word trigram language model.

Test sets of the tasks were collected using a recorder on cellphones in various noise environments. The noise environments were 30 places where people often use cellphones, including railway terminal stations, suburban railway stations, station square, offices, roadsides and shopping malls. The test set of the isolated word recognition task was 957 utterances made by 50 male and 50 female speakers. The test set of the train connection task was 500 utterances of the same speakers. The test set of the mail dictation task was 389 utterances of typical sentences from business mails. This test set was collected in a silent environment.

Two reward functions A) and B) were first compared with the basic beam width and histogram pruning without word end pruning (WEP). Then, the proposed functions were evaluated in combination with WEP.

The experimental conditions were as follows. A total of 38 dimensional acoustic features composed of the standard acoustic features of ETSI ES201108 [14] with CMS and their first and second derivatives excluding power were extracted from speech sampled at 8.0 kHz. Acoustic models were speaker-independent tied-state triphone models. In the isolated word recognition and the short sentence tasks, context-free grammars (CFGs) without linguistic probabilities on word entries were used with a one-pass frame-synchronous beam search. In the mail dictation task, a trigram language model was used with a one-pass frame-synchronous beam search.

The constants a_{log}, b_{log}, a_{exp} and b_{exp} of the proposed functions were optimized with a development set of the same size as the test set to minimize WER under a tight pruning condition. The pruning condition set the beam width f_{GB} at 140, and the maximal number of retained hypotheses N_{max} at 500. The processing time was measured on a PC with an Intel Pentium 4 3.0 GHz processor.

Sensitivity of the efficiency to these constants was evaluated as follows. Because RTF is nearly linear to the threshold N_{max} for histogram pruning, WER with respect to the change of a and b was evaluated at several RTF values controlled by N_{max}. Specifically, the beam width f_{GB} was fixed at 140, N_{max} is a variable to set RTF at a target value, either of a or b is another variable and the other is fixed at a constant.

Equal-depth pruning [13] partitions the lexical tree based on depth levels into multiple sections where the beam width pruning is executed separately. Equal-depth pruning has a parameter which specifies the degree of partitioning. We define this parameter "unit depth" with which the lexical tree is partitioned into sections. The performance is tuned by setting the proper unit depth. Though the original equal-depth pruning partitions the lexical tree finely with a small unit depth, we conducted our investigation with fine partitioning with a small unit depth through to rough partitioning such as bisection or trisection. The optimal value of the unit depth was obtained from RTF-WER curves with the beam width as a parameter for the development set.

3.2 Result of an Isolated Word Recognition Task

Figure 5 shows the averaged real time factor (RTF) and word error rates (WER) for the isolated word recognition task of the 10k-word railway station names. Five lines represent the basic pruning i.e. the beam width and histogram pruning without a reward, the beam width and histogram pruning with WEP, A) beam width and histogram pruning with the reward given by a logarithmic function, B) beam width and histogram pruning with the reward given by an asymptotic exponential function and the equal-depth pruning. The parameter of each line is the strength of the pruning. To be exact, the threshold N_{max} for the histogram pruning was shifted with the beam width f_{GB} fixed at 140. The looser the pruning, the lower the WER value, but the longer the RTF.

The proposed method A) reached WER below 18% at RTF 0.18, while the basic pruning reached the same WER at RTF around 0.2. The proposed method A) was not worse than the basic pruning. The proposed method B) was as effective as the method A). Because cross-word transitions do not cause an explosive increase in hypotheses in isolated word recognition tasks, the fact that the WEP had no effect is a reasonable outcome. The results of the proposed methods in combination with WEP were identical to those without WEP, though they are not shown here. The optimized values of the constants were $a_{log} = 4.0$, $b_{log} = 0.1$ for the method A), $a_{exp} = 20.0$ and $b_{exp} = 7.0$ for the method B). The reward functions with the optimized constant values are

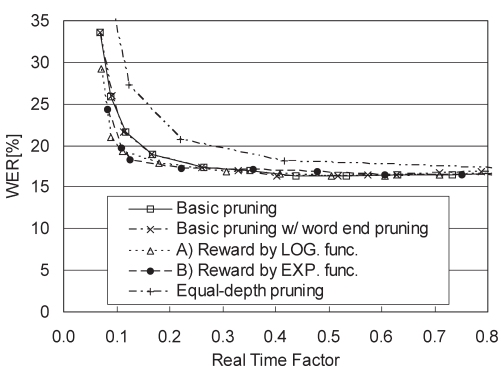


Fig. 5 RTF and WER for the isolated word recognition task.

shown in Fig. 4.

The equal-depth pruning was worse than the others even after the unit depth was optimized. The optimal value of the unit depth was 20, which means partitioning into 6 sections.

3.3 Results of a Grammar-Based Short Sentence Task

Figure 6 shows the RTF and WER for the grammar-based task of the formulaic train connection inquiries. The WER was calculated based on 1,000 departure and arrival station names in 500 utterances. Five lines represent the same as shown in Fig. 5. While the WER of the basic pruning gradually approached the minimal value, those of the proposed methods A) and B) fell below the minimal value of the basic pruning 20.7% at RTF 0.27, which meant approximately an 84% reduction from the minimal value at RTF 1.68. The threshold of histogram pruning N_{max} required for WER of 20.7% was actually reduced from 6,500 to 1,000, which was an 84% reduction as well. The proposed methods A) and B) reached a 1.0% lower minimal WER than the basic pruning. The optimized values of the constants were the same as in the case of the isolated word recognition as $a_{log} = 4.0$, $b_{log} = 0.1$ for the method A), $a_{exp} = 20.0$ and $b_{exp} = 7.0$ for the method B). WEP gave no improvement because the explosive increase in hypotheses was limited by the strong constraints of the grammar in this case.

Equal-depth pruning was also worse than the others for this task. The best performance of equal-depth pruning was achieved by bisection of the lexical tree, which corresponds to a unit depth of 60. The reason of the inferior performance is considered to be as follows. Equal-depth pruning has an effect of dispersing the hypotheses widely across the lexical tree. Originally, the basic beam width and histogram pruning naturally suppresses hypotheses close to the leaves for speech segments of former half of words in the isolated word recognition and the grammar-based short sentence tasks. However, equal-depth pruning makes more hypotheses in deep sections to be retained for these speech segments. This unnecessary retention of hypotheses deteriorates the search efficiency and RTF.

The sensitivity of WER to the two constants a_{exp} and b_{exp} of Eq. (4) for the method B) is shown in Fig. 7. Panel A) shows the sensitivity to a_{exp} at several RTFs 0.15, 0.25, 0.40 and 0.50 with b_{exp} fixed at 7.0, while Panel B) shows the sensitivity to b_{exp} at the same RTFs with a_{exp} fixed at 20.0. The beam width f_{GB} was fixed at 140, and the threshold N_{max} for the histogram pruning was tuned to set RTF at a target value. The basic pruning corresponds to $a_{exp} = 0$ in Panel A). Introduction of a small reward with a small a_{exp} value gave a significant WER reduction, but generally, WER changed slowly to the change of a_{exp}. The WER gradually decreased as a_{exp} increased at RTF 0.15, while the minimal WER did not change at RTF 0.40 in Panel A). The WER was insensitive to b_{exp} in Panel B).

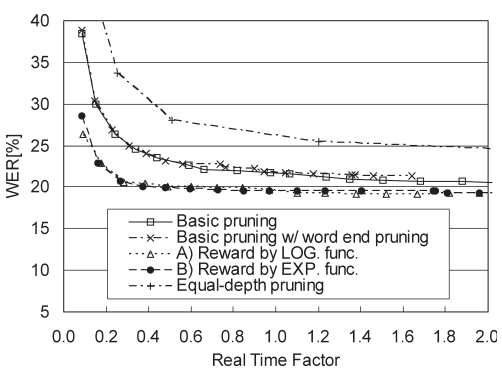
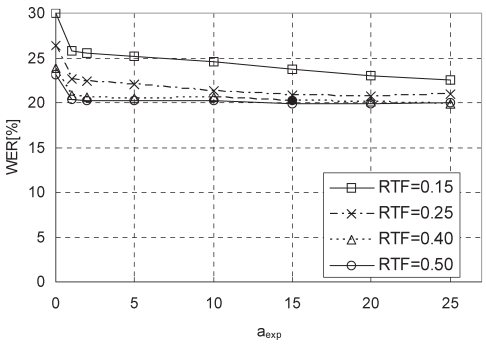


Fig. 6 RTF and WER for the grammar-based recognition task.

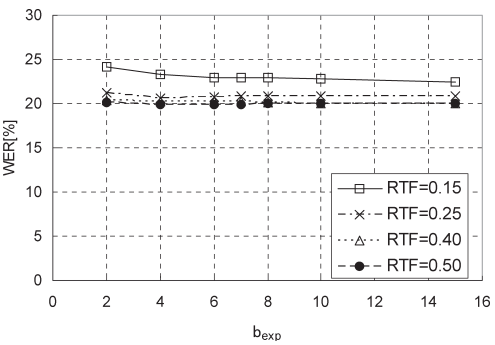
3.4 Results of a Language-Model-Based Dictation Task

Figure 8 shows the RTF and WER for the 30k-word mail dictation task. In this dictation task, the basic pruning includes the language model look-ahead technique [6]–[8]. This technique actually reduces the processing time to less than 1/10 from that without the look-ahead technique. All the other lines also include the look-ahead technique as their baseline.

While the basic pruning reached a WER of 20.0% with RTF 0.90, the proposed method B) reached the WER value with RTF 0.55, which was a 39% reduction. The proposed method A) reached the WER of 20.0% with RTF around



A) WER sensitivity to a_{exp} at RTFs of 0.15, 0.25, 0.40 and 0.50. b_{exp} was set at 7.0.



B) WER sensitivity to b_{exp} at RTFs of 0.15, 0.25, 0.40 and 0.50. a_{exp} was set at 20.0.

Fig. 7 WER sensitivity to parameters a_{exp} and b_{exp} at several RTFs for the method B) in the grammar-based recognition task.

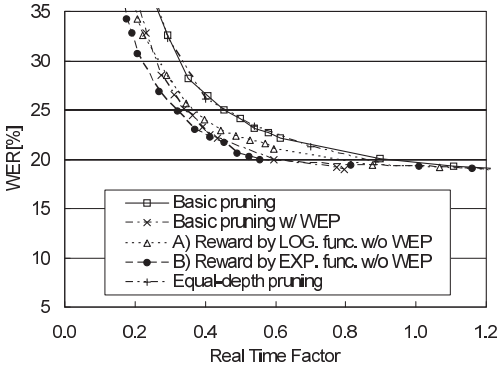


Fig. 8 RTF and WER of respective pruning techniques for the 30k-word language-model-based dictation task.

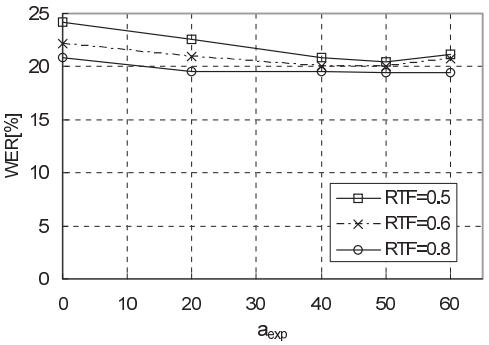
0.8. The optimized values of the constants were $a_{log} = 4.3$, $b_{log} = 0.9$ for the method A), $a_{exp} = 46.0$ and $b_{exp} = 0.4$ for the method B). WEP reached the WER of 20.0% with RTF 0.61, which was a little less effective than the proposed method B). The optimal threshold f_{WEP} was 80.0.

The equal-depth pruning was not worse than the baseline, but made no improvement over the basic pruning. The optimal value of the unit depth was 3 for this task. In this dictation task with less linguistic constraints than the isolated word recognition task or the grammar-based short sentence task, the hypotheses are basically distributed widely across the lexical tree in the basic pruning. Therefore, equal-depth pruning, which has an effect of dispersing hypotheses, did not change the search efficiency.

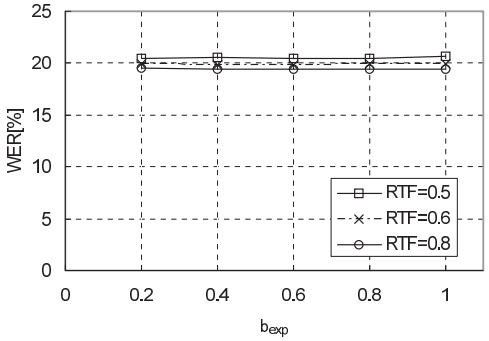
The sensitivity of WER to a_{exp} and b_{exp} for the method B) is shown in Fig. 9. Panel A) shows the sensitivity to a_{exp} at several RTFs 0.5, 0.6 and 0.8 with b_{exp} fixed at 0.4, while Panel B) shows the sensitivity to b_{exp} at the same RTFs with a_{exp} fixed at 46.0. Generally, WER changed slowly with the change of a_{exp}. The WER gradually decreased as the reward increased, reached the minimal value at $a_{exp} = 50$ and increased slowly again in Panel A). The minimal WER at RTF 0.8 was less sensitive to the change of a_{exp}. The WER was insensitive to b_{exp} in Panel B).

In Fig. 10, the proposed methods A) and B) were evaluated in combination with WEP. The RTF to achieve WER of 20% was reduced from 0.55 to 0.46 in case of the method B). Compared with the basic pruning with language model look-ahead and WEP, the proposed method B) achieved an additional 23% reduction.

Though the proposed methods improved the efficiency compared to the basic pruning with the language model look-ahead, the effect was weaker than that of the grammar-based short sentence recognition task. We consider this to be due to the similarity of the effects given on the likelihood of hypotheses by the language model look-ahead technique and the proposed reward. The look-ahead value also decreases monotonically as a hypothesis advances on a path from the root in the lexical tree, because the look-ahead value uses the maximal value of the linguistic likelihood among the reachable words. Moreover, stepwise renewal of the reward along HMM state sequence like the previous



A) WER sensitivity to a_{exp} at RTFs of 0.5, 0.6, and 0.8. b_{exp} was set at 0.4.



B) WER sensitivity to b_{exp} at RTFs of 0.5, 0.6, and 0.8. a_{exp} was set at 46.0.

Fig. 9 WER sensitivity to parameters a_{exp} and b_{exp} at several RTFs for the method B) in the 30k-word language-model-based dictation task.

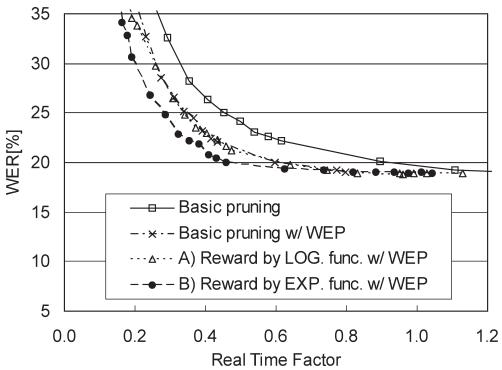


Fig. 10 RTF and WER of the proposed method in combination with word end pruning (WEP) for the language-model-based dictation task.

studies [9], [10] may further improve the efficiency.

4. Discussion

The proposed method is viewed as an extension of WEP. While WEP applies tighter beam width pruning to the hypotheses at the terminal states in the lexical tree, the proposed method gradually tighten the beam width as the hypotheses approach the terminal states. Thus, the proposed method retains hypotheses more efficiently than WEP. Note that the beam width gap between the root and leaves, which corresponds a_{exp} in case of the exponential function, was

different from the gap between the global beam width and the beam width for WEP.

Viewed from another perspective, the monotonically decreasing reward along the path on a lexical tree can be interpreted as a heuristic gain of the likelihood on the path from the current HMM state to a leaf state in the lexical tree.

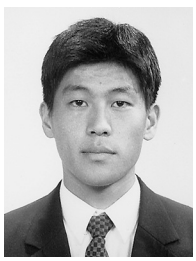
Comparing the two types of the reward functions, the exponential function performed better than the logarithmic function. In terms of the stability in optimization of the constants, the non-asymptotic logarithmic function may produce a prominent reward for a hypothesis having a significantly-great number of reachable words close to the root, and this prominent reward has a decisive effect upon determination of S_{max} and f_{GB}. The asymptotic exponential function is preferable in that the optimized constants are more stable to the maximal number of reachable words i.e. the vocabulary size. Regarding the optimized values, the constants of the two grammar-based tasks, the isolated word recognition and the short sentence task, were coincident. The optimized values of the constants between the grammar based tasks without linguistic probabilities and the language-model-based dictation task were much different. The greatest factor is considered to be the language model look-ahead.

5. Conclusions

To make the frame-synchronous beam search more efficient and reduce the processing time, we introduced a tentative reward considering the potential to reach various words on a lexical tree into the beam width and histogram pruning. Two types of the reward given by an asymptotic exponential function and a logarithmic function greatly reduced the number of hypotheses required to maintain the maximal word accuracy in grammar-based tasks. The reward given by the exponential function resulted in an 84% reduction in processing time for a grammar-based short sentence recognition task without losing accuracy. For a language-model-based dictation task, it showed an additional 23% reduction in processing time from the pruning with the language model look-ahead technique.

References

[1] R. Haeb-Umbach and H. Ney, "Improvements in beam search for 10,000-word continuous speech recognition," IEEE Trans. Speech Audio Process., vol.2, no.2, pp.353–356, 1994.
[2] H. Ney, D. Mergel, A. Noll, and A. Paeseler, "Data-driven search organization for continuous speech recognitions," IEEE Trans. Signal Process., vol.40, no.2, pp.272–281, 1992.
[3] V. Steinbiss, B.-H. Tran, and H. Ney, "Improvements in beam search," Proc. ICSLP 94, vol.1, pp.397–400, 1994.
[4] S. Ortmanns and H. Ney, "A word graph algorithm for large vocabulary continuous speech recognition," Comput. Speech Lang., vol.11, no.1, pp.43–72, 1997.
[5] J. Ogata and Y. Ariki, "An efficient lexical tree search for large vocabulary continuous speech recognition," Proc. ICSLP'00, vol.2, pp.967–970, 2000.
[6] S. Ortmanns, A. Eiden, and H. Ney, "Look-ahead techniques for fast beam search," Proc. ICASSP 97, pp.1783–1786, 1997.
[7] S. Ortmanns, A. Eiden, and H. Ney, "Improved lexical tree search for large vocabulary speech recognition," Proc. ICASSP 98, pp.817–820, 1998.
[8] S. Ortmanns and H. Ney, "Look-ahead techniques for fast beam search," Comput. Speech Lang., vol.14, pp.15–32, 2000.
[9] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga, and Y. Sagisaka, "Spontaneous dialogue speech recognition using cross-word context constrained word graphs," Proc. ICASSP 96, pp.145–148, 1996.
[10] J. Davenport, L. Nguyen, S. Matsoukas, R. Schwartz, and J. Makhoul, "Toward realtime transcription of broadcast news," Proc. Eurospeech 99, pp.651–654, 1999.
[11] S. Abdou and M.S. Scordilis, "Beam search pruning in speech recognition using a posterior probability-based confidence measure," Speech Commun., vol.42, pp.409–428, 2004.
[12] T. Hori, S. Watanabe, and A. Nakamura, "Search error risk minimization in Viterbi beam search for speech recognition," Proc. ICASSP 2010, pp.4934–4937, 2010.
[13] J. Pylkkonen, "New pruning criteria for efficient decoding," Proc. Interspeech 2005, pp.581–584, 2005.
[14] "ETSI ES 201108 speech processing, transmission and quality aspects; Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," ETSI Standard, 2000.



Tsuneo Kato received the B.E. and M.E. degrees from University of Tokyo in 1994 and 1996, respectively. He joined Kokusai Denshin Denwa Co. Ltd. in 1996. He is currently with KDDI R&D Laboratories Inc. He has been engaged in research and development of automatic speech recognition and speaker verification. He is a member of the Acoustical Society of Japan (ASJ), Information Processing Society of Japan (IPSJ) and IEEE.

Kengo Fujita received the B.E. and M.E. degrees in electrical and electronics engineering from Kobe University, Japan, in 2001 and 2003, respectively. Since 2003, he has been with KDDI R&D Laboratories Inc., Saitama, Japan. He has been engaged in the research and development of speech recognition and evaluation of telephone speech quality. He was a visiting researcher of acoustics and audio engineering at NHK Science & Technical Research Laboratories in 2007. He is a member of the Acoustical Society of Japan (ASJ).

Nobuyuki Nishizawa received the B.E. degree in electrical engineering, and the M.E. and Ph.D. degrees in information and communication engineering from the University of Tokyo in 1998, 2000 and 2003. He was a researcher in ATR Spoken Language Translation Research Laboratories from 2003 to 2006. Since 2006, he has been in KDDI R&D Laboratories. His research interests include speech synthesis. He is a member of the Acoustic Society of Japan (ASJ).