EFFICIENT SEARCH STRATEGY IN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION USING PROSODIC BOUNDARY INFORMATION

Shi-wook Lee, Keikichi Hirose** and Nobuaki Minematsu**

* Department of Information and Communication Engineering, School of Engineering
** Department of Frontier Informatics, School of Frontier Sciences
The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{lsw, hirose, mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT

Prosodic-syntactic boundary as an information source can be used to improve the performance of Large Vocabulary Continuous Speech Recognition (LVCSR) in both efficiency and accuracy. This paper presents a study of two effective methods to exploit prosodic boundary information in a multi-pass decoder. In this paper, we address the effect of a language model on setting pruning beam width and how to control the Cross-word Context Dependent (CCD) models by prosodic boundary information. In the first pass decoding, dynamic beam search strategy regarding inner-word and cross-word paths is proposed to reduce search space efficiently, and then cross-word context dependent models are optimized using prosodic boundary information in the second pass decoding. The recognition experiments, which were carried out on the Japanese Newspaper Article Sentences (JNAS) 20k word task using a multi-pass decoder, demonstrated that the proposed method led to significant reduction in the search space with accuracy improvement.

1. INTRODUCTION

Recent advances in speech recognition and computing power have brought highly accurate recognition systems close to reality. However, continuous speech recognition with a very large vocabulary still requires a large amount of computation. Recently, dynamic beam search strategies have been studied to achieve high performance in recognition efficiency. S. Renals and M. Hochberg[1] proposed the phone deactivation pruning with language model interaction and achieved extremely successful improvement in the search efficiency of LVCSR systems. V. Steinbiss[2] proposed histogram pruning and language model(LM) look-ahead method, which incorporated LM probability as early in the search as possible. In view of these approaches, we proposed a new approach to control pruning beam width[3].

In continuous speech recognition, word boundaries of input speech are uncertain before the search process. Thus, the linguistic state cannot be determined in 1-pass search. With prosodic information, speakers tend to group words into a phrase whose boundaries are marked by pauses or break indices, and many phonological rules are constrained to operate only within a phrase, usually termed a *prosodic phrase*. As a knowledge source, the prosodic phrases can be applied to constrain the search space for a speech recognition system. In the application, one could use either an explicit [4] or a Multi-Layer Perceptron(MLP) [5] decision as to what type of prosodic phrase boundary occurs at a particular point in time, or a set of hypothesized phrasings for the utterance. However, since the phrase detection algorithms are not sufficiently reliable yet, there has been a little success reported for the pre-word recognition approach.

The pronunciation of a word depends greatly on the coarticulation effects of the connected words in continuous speech. To reduce the acoustic mismatch at cross-word point, CCD models have been used in continuous speech recognition and resulted in high accuracy. However, if a pause is not placed on word transition, non-coarticulated word transition is difficult to be dealt with. Since most phonological rules are constrained to operate only within a prosodic phrase, it is beneficial to generate CCD models especially designed for cross-word transitions at prosodic boundaries. In our approach, CCD models are controlled by prosodic boundary information.

In section 2, we briefly describe an algorithm for the detection of prosodic boundaries of continuous speech. The proposed dynamic beam search strategy and optimization technique of applying CCD models are described in section 3 and 4 respectively. The results of the experiments conducted to evaluate the performance of the proposed methods are shown in Section 5 and Section 6 concludes the paper.

2. DETECTION OF PROSODIC BOUNDARIES OF CONTINUOUS SPEECH

Among various types of syntactic boundaries, phrase boundaries can be used as a very effective constraint in the recognition process, but their correct detection with prosodic features is considerably difficult when they are not accompanied by long pauses. In our previous study, a method was proposed for the detection of prosodic-syntactic boundaries, where not only major boundaries like clause and phrase boundaries, but also minor boundaries like *bunsetsu* boundaries can be detected[4]. Here, *bunsetsu* is a basic unit of Japanese syntax consisting of a content word with or without being followed by a few of function words. In the method, a rule-based analysis of macro- and microscopic(original) features of F_0 is conducted to find prosodic-syntactic boundaries. The F_0 contour is segmented at dips (minima) in the energy contour, and rules are applied to generate candidate syntactic boundaries at dips in the F_0

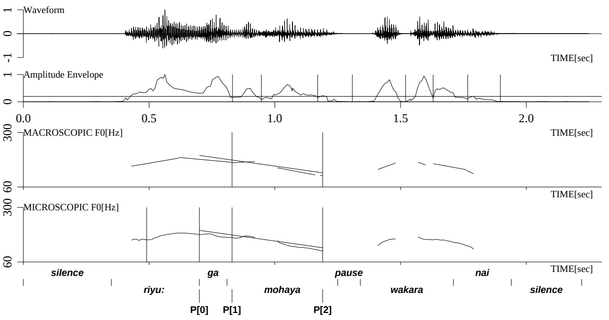

Figure 1: An example of the prosodic-syntactic boundaries for the utterance "*riyu:ga / mohaya / wakaranai.*" (The reason has become obscured.)

contour. A macroscopic F_0 contour is calculated as a piecewise linear contour from the original contour. Among the candidates detected from the macro- and microscopic F_0 contours, selection of prosodic-syntactic boundaries is conducted. Several thresholds are used in the rules, and these can be adjusted to control insertion errors (a typical problem of automatic boundary detection). The system detects over 80% of syntactic boundaries with an insertion error rate of 30%. Figure 1 shows the result of detecting prosodic boundaries.

3. DYNAMIC BEAM SEARCH STRATEGY

3.1. Incorporating language model probability

Since a number of words have the same initial sequence of phonemes, a pronunciation lexicon in LVCSR systems can be arranged as a tree structure. It plays an important role in the efficient search strategy. However, the problem is that the LM score is not incorporated before reaching the leaf, because the word identity is only known at the leaf of the tree while it is already known at the root in linear-structured lexicon. To overcome this problem, G. Antoniol, et. al. [6] have proposed that bigram probabilities $p(w_2|w_1)$ can be factorized in order to employ linguistic information as early as possible during the beam search. With this factorized n-gram language model, bigram probabilities can be incorporated more aggressively into acoustic-phonetic decoding in the computation of Viterbi score. As a result of this implementation, the search space and the computational cost of the recognition process are substantially reduced.

3.2. Dynamic control of beam width

Since the LM probabilities are factorized into tree nodes by an estimated LM probability of all possible successors, the language model scores near the root nodes of the tree-structured lexicon are different from actual LM scores, namely, the scores at the succeeding leaf nodes. As a result, it causes unexpected pruning near the root nodes with excessively aggressive pruning beam width, which eliminates the potentially best path. In other words, this factorized LM score is not guaranteed to coincide with the true LM score at its succeeding leaf nodes. In order to obtain reasonable recognition accuracy, the beam width has to be set wide enough in conventional beam search, because some of locally unlikely paths may form part of the globally most likely path and pruning them will result in search errors. However, as the search process goes closer to the end of the word and/or prosodic boundary, the globally most likely path achieves a relatively high rank in search space for two reasons. One is a linguistic reason that a greater degree of certainty about word identity is obtained near word ends than word starts. The other one is an acoustic reason that the acoustic certainty of a particular word is increased by matching upon more input frames. Consequently, the necessary requirement for search space at the end of words and/or the prosodic boundaries becomes smaller. As a constraint rule in the proposed method, the pruning beam width is dynamically controlled regarding this phenomenon; using a tighter inner-word beam width than cross-word beam width. Figure 2 shows an example of search space constrained by static pruning beam width.

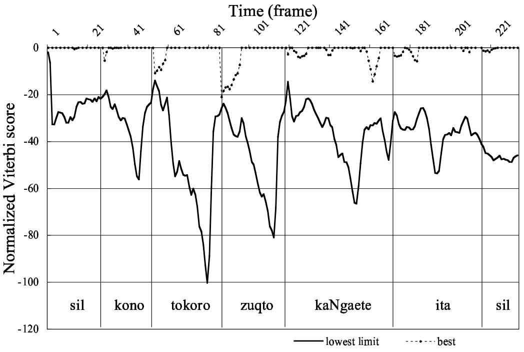

Figure 2: An example of search space constrained by static pruning beam width for sentence "*kono / tokoro / zuqto / kaNgaete / ita* " (I have been considering it.)

In the figure, the dots show the globally best path which is correctly recognized and the solid line is the lowest limit resulted from applying the static pruning beam width; keeping the number of succeeding nodes constant. Since the factorized LM probabilities in early portion of word is higher than the portion of word ends, incorporating LM probabilities at the word initial portion needs pruning beam width to be wide enough. Therefore, a lot of useless search space within a word can be observed with a given static pruning beam width which is set to obtain reasonable accuracy. Based on this observation of the effect of LM probability incorporation, the pruning beam width within prosodic phrase can be set more tightly than that of word initial portion. Furthermore, speakers tend to group words into a phrase, whose boundary is marked by prosodic boundaries. At the prosodic-syntactic boundary point, the correlation between the words may be less than that of the words within a phrase. Thus, the ambiguity at the prosodic-syntactic boundary is increased. Prosodic information can be employed in constraint rules, to set an optimal pruning beam width dynamically, and can preserve efficient search.

3.3. Updating LM probabilities

The other unavoidable consideration of restricting beam width according to the prosodic-syntactic boundary is the

LM score incorporation of the cross-word transition within a prosodic phrase and phone branch node in tree-structured lexicon. Furthermore, since the lexicon for Japanese LVCSR is based on morphemes, linguistic unit usually smaller than word, it causes drastic degradation of recognition accuracy. To cope with this problem, we utilize the number of hypothesis that reached lexical-entries end nodes and phone branch nodes as a supplementary factor. This method was proposed to determine word boundaries and to use the word boundaries to restrict search space. The basic idea is that many predecessor words are expected to have the same word ending portion in search procedure.

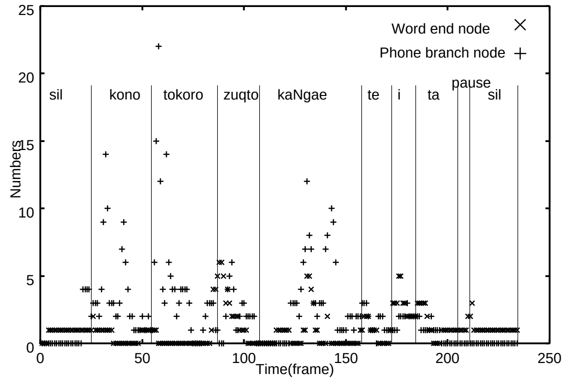

Figure 3: Number of word-end nodes and phone-branch nodes

However, in our approach, this observation is introduced as a supplementary factor, which enlarges beam width. It is considered that this supplementary factor plays a good role in removing search error not only from the effect of cross-word within prosodic-syntactic phrase and phone branch node, especially in Japanese whose lexicon is based on morphemes, but also from the deletion error in detection of prosodic-syntactic boundary. From these observations, the optimal beam width is dynamically computed from the following equations.

$$P_t(s) \quad < \quad P_t^{max}(s) - \hat{\lambda}(t) \tag{1}$$

$$\begin{aligned} \hat{\lambda}(t) \quad &= \quad \lambda(0) + \{\lambda_{var}(t) \times LM_w \\ &\times \quad (\frac{\sum word\_end(t) + \sum phone\_branch(t)}{\sum active(t)})\} \\ &\quad\ where, \lambda(0) \leq \lambda_{var}(t) \leq 1 \end{aligned} \tag{2}$$

Where P_t^{max} is the likelihood[1] of the most likely path at time t, LM_w is the relative weight of the acoustic and language models, $\lambda_{var}(t)$ is decreasing factor from a prosodic boundary point to the next prosodic boundary point, and $\hat{\lambda}(t)$, the dynamic beam width at time t, is finally calculated by eq.(2). To avoid an extremely pruning case, which can be caused by deletion errors in automatic detection of prosodic-syntactic boundary, lower bound of $\hat{\lambda}(t)$,is set at 1.

4. CROSS-WORD ACOUSTIC MODELS

In continuous speech, the pronunciation of a word depends greatly on the coarticulation caused by its surrounding words. To reduce the acoustic mismatch at word boundaries and model the effect of phonetic context across word boundaries, CCD models are used in improving continuous speech recognition. Even though CCD models are reported very helpful in recognition accuracy, it increases complexity in search process because the last phone in the current word becomes dependent on the next word, which is unknown. To cope with this problem, CCD models are adapted to the further stage. In the current practical systems, CCD models are usually integrated in 2-pass decoding as high order knowledge source. The problem here is that CCD models are applied regardless of whether the coarticulation is actually occurred or not. The coarticulation effect within a word can be handled easily with multiple lexical entries of the same word. However, the coarticulation across words is not evident. Non-coarticulated transition can be found and have to be considered especially when a prosodic boundary exists. In our approach, we control CCD models according to prosodic boundaries and use them in the second pass decoding.

5. EXPERIMENTAL RESULTS

The recognition system is a multi-pass decoder, shown in figure 4. The language model and Japanese acoustic model which are provided by IPA[2] are used to conduct experimental evaluation[7]. The test-set sentences in this study are a portion of the JNAS[3] speech database. It consists of 50 sentences by 10 male speakers (5 sentences per speaker).

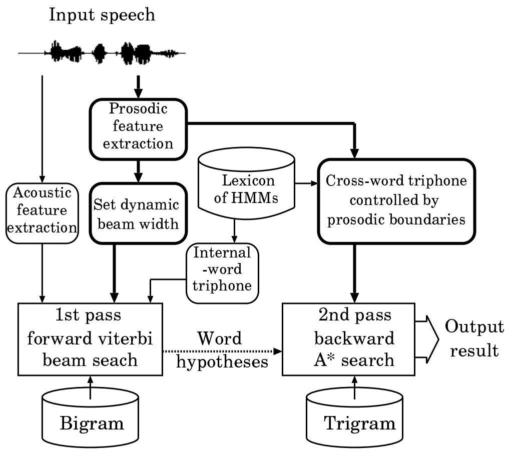
Input speech
Prosodic feature extraction
Acoustic feature extraction
Set dynamic beam width
Lexicon of HMMs
Cross-word triphone controlled by prosodic boundaries
Internal -word triphone
1st pass forward viterbi beam seach
Word hypotheses
2nd pass backward A* search
Output result
Bigram
Trigram

Figure 4: A multi-pass decoder for achieving efficient search using prosodic information

Table 1 shows the result of experimental evaluation of the baseline system. The realtime factors were measured on SUN 336MHz CPU.

[1] All likelihoods are assumed to be logarithmic.

[2] Information-technology Promotion Agency, Japan, 1999

[3] provided by Acoustical Society of Japan, 1997

Beam width	WAR (%)	# average active nodes/frame	x RT
20	67.35	33.57	2.0
25	78.23	71.87	2.5
30	79.88	149.06	3.2
35	86.04	282.81	4.4
40	88.91	502.63	5.9
45	90.14	829.02	7.9
50	90.35	1273.89	10.5
55	91.58	1826.49	13.5

Table 1: Experimental evaluation of the baseline system Real Time(RT)factor: 3.69sec

5.1. Dynamic control of pruning beam width in the first pass decoding

Typically, most of the computation in the multi-pass decoder is concentrated on the first pass. The proposed method sets pruning beam width dynamically according to automatically detected prosodic boundaries in the first pass from eq.(2). As can be seen from Table 2, compared to the static beam width search where the Word Accuracy Rate (WAR;$\%Correct - \%insertion$) is about 79%, the use of dynamic beam using prosodic boundary information reduced the average number of active nodes by about 60%, and the real-time factor is decreased by 31%. The recognition accuracy of 86.04% can be obtained in x3.1 RT using dynamic beam width, while the x4.4 RT would be required with the static beam width.

Maximum beam width	WAR (%)	# average active nodes/frame	x RT
40	78.64	59.44	2.2
50	86.04	139.10	3.1
60	87.06	302.04	4.5
70	89.53	594.12	6.6
80	90.76	1054.34	9.6
90	91.58	1661.17	12.7

Table 2: Experimental evaluation using dynamic beam width

5.2. Applying controlled CCD models to the second pass decoding

In another approach using prosodic-syntactic boundary information, we controlled acoustic models to improve accuracy in the second pass. The performance in table 3 shows context dependency of acoustic models. With controlled C-CD models according to prosodic boundaries, the WAR and Sentence Correct Rate(SCR) are better than CCD models without the consideration of prosodic boundary. Above all, the proposed method shows notable improvement in SCR, by 8%. It means that using prosodic-syntactic information is very helpful for estimating correct syntactic structure and improving accuracy in LVCSR.

Lexicon of HMMs	WAR(%)	SCR(%)	x RT
Internal-word triphones	86.04	52	9.2
CCD triphones	90.14	56	7.9
static beam width + CCD with PB	91.3	64	8.1
dynamic beam width + CCD with PB	89.53	62	6.5

Table 3: Experimental evaluation using CCD models controlled by prosodic boundaries

6. CONCLUSIONS

Two approaches using prosodic-syntactic boundary information in LVCSR, dynamic beam search strategy and C-CD models controlled by prosodic boundaries, have been described and evaluated on a 20k(JNAS) task. The experimental results showed that the proposed method leads to much smaller search space by about 60% in active nodes per frame than the static beam search. And with controlled CCD models by prosodic boundary information, we could achieve the higher accuracy, especially in SCR by 8%. From these results, prosodic information can be used very effectively for LVCSR systems, in both efficiency and accuracy. We are now planning to extract more sophisticated prosodic boundary information, which is harmonic with stochastic search process, and to adapt it to LVCSR systems.

7. REFERENCES

[1] S. Renals and M. Hochberg, "Efficient evaluation of the LVCSR search space using the NOWAY decoder", *Proc. of ICASSP'96*, pp.149-152, 1996

[2] V. Steinbiss, B.-H. Tran and H. Ney, "Improvements in Beam Search", *Proc. of ICSLP'94*, pp.2143-2146, 1994

[3] S.W. Lee and K. Hirose, "Dynamic beam search strategy using prosodic-syntactic information", *Proc. of ASRU'99*, pp.189-192, 1999

[4] K. Hirose, A. Sakurai and H. Konno, "Use of prosodic features in the recognition of continuous speech",*Proc. of ICSLP'94*, pp.1123-1126, 1994

[5] J. Buckow, et. al., "Dovetailing of Acoustics and Prosody in Spontaneous Speech Recognition",*Proc. of ICSLP'98*, pp.571-574, 1998

[6] G. Antoniol, F. Brugnara, M. Cettolo and M. Federico, "Language Model Representations for Beam-search Decoding", *Proc. of ICASSP'95*, pp.588-591, 1995

[7] T. Kawahara, et. al., "Sharable software repository for Japanese large vocabulary continuous speech recognition", *Proc. of ICSLP'98*, pp.3257-3260, 1998