音声言語情報処理　9－8
(1995. 12. 14)

単語グラフを用いた自由発話音声認識

清水徹　山本博史　松永昭一　匂坂芳典
ATR 音声翻訳通信研究所
〒619-02 京都府相楽郡精華町光台 2-2
Tel.: 0774-95-1345　e-mail: shimizu@itl.atr.co.jp

あらまし
本論文では、自由発話の大語彙音声認識を目的とした、音素環境に基づき制約された単語グラフを用いる連続音声認識手法を提案する。本認識手法では、単語グラフ生成時の計算コストを削減するため、“単語境界の文脈に依存した単語間遷移時刻の近似”と“木構造辞書における言語スコアの予測値の付与”の2つの近似手法を導入した。本手法の効果を、“travel planning corpus”を用いて評価した。この結果、単語間遷移時刻の近似により、単語仮説の数は25-40%減少し、この結果同一サイズの単語グラフ生成に要するCPU時間を約30-60%減少させることができた。また、木構造辞書の非終端ノードに与える言語スコアの予測値としてクラスバイグラムを用いた場合、予測値なしの場合に比較して単語誤り率が25-30%削減された。

キーワード　• 連続音声認識　• 探索手法　• 単語グラフ　• 木構造辞書

Spontaneous Dialogue Speech Recognition using Cross-Word context Constrained Word Graph

Tohru Shimizu　Hirofumi Yamamoto　Shoichi Matsunaga
Yoshinori Sagisaka
ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02
Tel. 0774-95-1345　e-mail: shimizu@itl.atr.co.jp

Abstract
This paper proposes a large vocabulary spontaneous dialogue speech recognizer using cross-word context constrained word graphs. In this method, two approximation methods “cross-word context approximation” and “lenient language score smearing” are introduced to reduce the computational cost for word graph generation. The experimental results using a “travel planning corpus” show that this recognition method achieves a word hypotheses reduction of 25-40% and cpu-time reduction of 30-60% compared to no approximation, and that the use of class bigram scores as the expected language score for each lexicon tree node decreases the 25-30% of word error rate compared to no approximation.

key words　• Continuous Speech Recognition　• Search method　• Word graph　• Lexicon tree

1 introduction

Reducing the number of word (sentence) hypotheses is an important issue for reducing the total amount of computational cost for a search.

We have already proposed a sentence hypotheses merging method on a time-synchronous continuous speech recognizer driven by a context-free grammar[1]. In this method, a huge computational cost reduction is achieved by merging hypotheses with the same phoneme history but with a different syntactic parse. However, the computational cost required for the spontaneous dialogue recognition is still quite large because a large number of sentence hypotheses have to be considered to cope with the spontaneous utterances which include many ungrammatical phenomena such as filled pauses, hesitations, and corrections.

To cope with this problem, or to cope with computational cost reduction of large vocabulary recognition, recently, recognition schemes using word graphs have been proposed [2][3][4][5]. In these schemes, the following word hypotheses reduction method has been successfully applied for word graph generation:

- **Word hypotheses pruning using "word pair approximation"**[4][6]
  Assume the word boundary between two succeeding words to be independent of further predecessor words.
- **Language model smearing on tree structured lexicon**[5][7]
  Another way to reduce the number of word hypotheses is to use a narrow beam width. However, as pruning errors are likely to occur when a language score is close to the pruning threshold, smear the language score over the tree.

However, these two methods have the following problems respectively.

- **Increasing the number of word hypotheses caused by the existence of many predecessor words**
  As a "word pair approximation" uses the predecessor information, the number of word hypotheses increases as the number of predecessor words becomes large, even if many predecessor words have the same word ending portion (same phoneme sequence).
- **Drastic changes of language score near the root node of the lexicon tree**
  As the expected language score on each tree node is the "minimum language score" of the words sharing the initial phone sequence, the score near the root node is still large. Therefore, a rather large beam width is still required.

This paper proposes a spontaneous dialogue speech recognizer using cross-word context constrained word graphs. In this method, "cross-word context approximation" is introduced to solve the first problem, and "lenient language score smearing" is introduced to solve the second problem.

In section 2, we describe a system structure that uses word graphs in a multi pass search framework. Next, in section 3, we describe methods to reduce the number of word hypotheses in word graph generation. Then, we present experimental results obtained using a "travel planning corpus". Finally, we conclude with a discussion in Section 5.

2 recognizer overview

In our multi pass search approach, detailed acoustic models are employed in the first pass to avoid redundant acoustic matching in the second pass.

First pass Word hypotheses generation
A one-pass time-synchronous beam search generates a list of word hypotheses resulting in a word graph. The predecessor word information, the exact acoustic scores (including the cross-word context) using context-dependent HMMs, class bigram (variable-order N-gram [8]) language scores, and the start-time and end-time are kept.

Second pass Word hypotheses pruning
A more complex language model such as a trigram is applied and the score of each word hypothesis is re-evaluated time-asynchronously. Word hypotheses lower than the certain threshold are removed to reduce the overall graph size.

3 Approximation methods

3.1 Cross-word context approximation

When a detailed acoustic model is used, many predecessor words are expected to have the same word ending portion. However, as a "word pair approximation" uses the predecessor word when determining the word boundary, the number of word hypotheses becomes large, even if many predecessor words have the same word ending por-

tion. To solve this problem, we use the "cross-word context" as predecessor word information.

Figure 1 shows a simple example of "cross-word context approximation". If there exists many word hypotheses that have the same word-id and the same end-time, only the hypothesis having the largest score for each cross-word context will become *alive*. (Two word hypotheses which have the cross-word context "$x/a_1/a_2$" become "*dead*").

The advantage of this method is that the number of word hypotheses each with the same word-id does not exceed the number of phonemes.

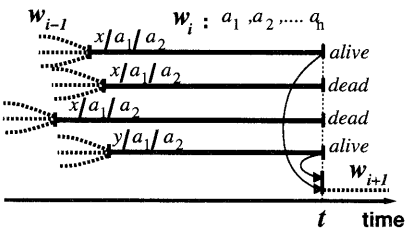


Figure 1: A simple example of cross-word context approximation

using 6,635 words lexicon is 215Mbyte.) However, as the distribution of unigram scores are quite different from that of bigram scores, unexpected pruning may occur. To solve this problem, we use class bigram (the classes are generated using variable-order N-gram procedure [8]) scores as the expected language score for each lexicon tree node. The storage effort of the class bigram scores are only 4% of the word bigram scores.

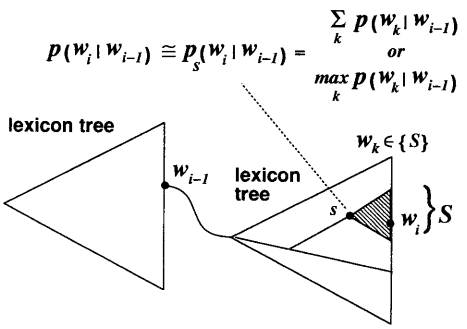


Figure 2: A simple example of lenient language score smearing (max,sigma)

3.2 Lenient language score smearing on lexicon tree

A tree organization that takes advantage of the fact that many words share the same initial phoneme sequence, and the expected language score for each lexicon tree node are applied before the word identity is known[5][7].

The proposed "lenient language score smearing" uses a smaller language score than the conventional method employing the "minimum language score" of words sharing the initial phone sequence. A simple example of phoneme level smearing is the use of the sum score instead of the minimum score (Figure 2). In this method, changes of the language score near the root node of the tree lexicon become smaller than those of the conventional method. Therefore, it is expected to use a narrower beam width. A further complex smearing, language score smoothing between the lexicon nodes is also considered to reduce the beam width. Figure 3 shows the example of language score smoothing. The language scores are re-estimated at each HMM state.

To reduce the storage efforts the use of unigram scores instead of word bigram scores has already proposed (*unigram estimation*)[7]. (For example, the storage effort of the bigram scores

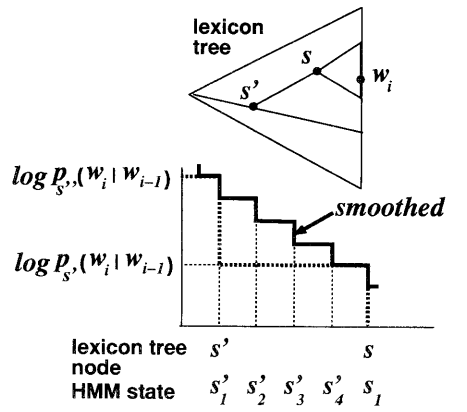


Figure 3: An example of further complex smearing (smoothed)

4 Experimental Results

Experiments evaluating the effect of "cross-word context approximation" and "lenient language score

smearing on lexicon tree" were carried out using spontaneous multi-lingual/monolingual dialogue speech of a "travel planning corpus"[9]. The state-shared context-dependent HMM (HMnet[10]) was used as an acoustic model. These acoustic models are adapted by the speaker specific utterances using Transfer Vector Smoothing (VFS)[11]. Three different size of lexicon are used for the experiment.

lex6600 vocabulary of whole task (utterances of customer, clerk and interpreter are included).

lex3000 vocabulary appeared in the customer's utterance.

lex1200 vocabulary appeared in the customer's utterance of the "hotel reservation task".

Other experimental conditions are summarized in Table 1.

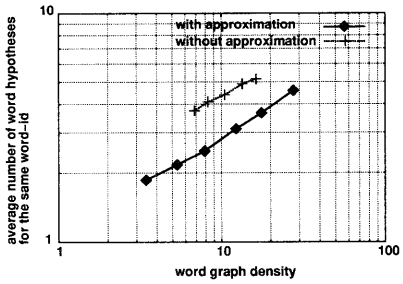
Table 1: Experimental conditions

Analysis conditions	
Sampling rate	12 kHz
Window	Hamming window (20 ms)
Frame period	10 ms
Analysis	log power + 16-order LPC-Cep + Δlog power + 16-order ΔLPC-Cep
HMnet	
State	401 states, 5 mixture
Training	2,620 words
Retraining	150 sentences (read speech)
Speaking-style Adaptation	128 utterances (non-read speech)
Speaker Adaptation	1 dialogue (non-read speech)
Language model (class bigram)	
Training	18,315 utterances (229,159 words)
Word preplexity	49.6
Lexicon	
lex1200 lex3000 lex6600	1,272 words 3,076 words 6,635 words
Recognition data	
Speaker	3 male, 4 female
Samples	100 utterances (983 words)

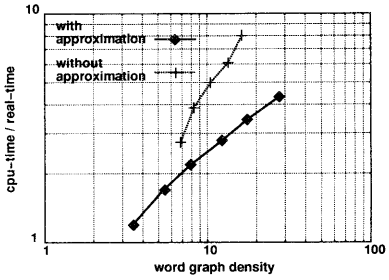
In the following subsection, we describe the recognition result obtained using first pass of the recognizer.

4.1 The effect of "cross-word context approximation"

- The average number of word hypotheses having the same word-id, but different start/end-time for the same word graph density is reduced 25-40%. The cpu-time requirement is also reduced 30-60% compared to the case of no approximation (Figure 4(a),(b)).
- The word graph density to achieve the same error rate decreases approximately 50% compared to the case without approximation(Figure 5).



(a) Word hypotheses number reduction



(b) CPU-time reduction

Figure 4: Effect of "cross-word context approximation" (lex3000)

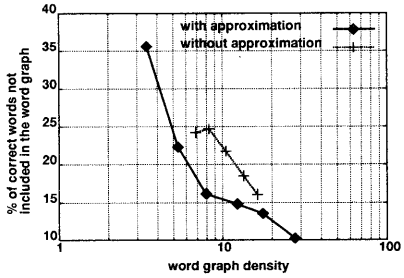


Figure 5: Effect of the reduction of word error rate by applying "cross-word context approximation" (lex3000)

4.2 The effect of "lenient language score smearing on lexicon tree"

- The beam width to achieve the same word error rate is reduced using proposed smoothed language model scores (*class bigram (smoothed)*). However, a simple lenient language model smearing that uses sum score instead of the minimum score (*class bigram (sigma)*) has no effect for beam width reduction (Figure 6).
- The use of class bigram scores achieved the word error reduction of 25-30% compared to the case of no approximation. Result of the conventional *unigram estimation* is worse than that of no approximation (Figure 7).

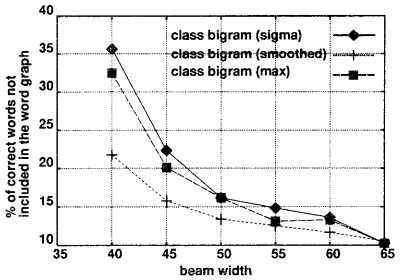


Figure 6: The reduction of beam width using smoothed language model scores (lex3000). Beam width is the fixed value of log likelihood. i.e. The hypotheses below the certain likelihood become dead.

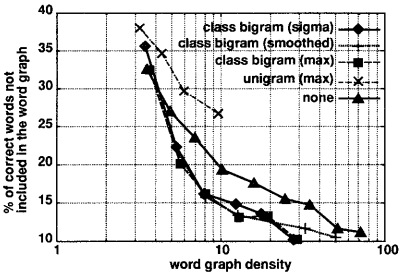


Figure 7: The word error rate reduction using class bigram language score as an expected language scores in the tree node (lex3000)

These result shows that the use of class bigram language scores as the expected language scores is more effective compared to the conventional *unigram estimation* from the point of view of recognition rate and storage effort.

4.3 The result of large vocabulary spontaneous speech recognition

The experiment using large size lexicon (6,635 words) has been carried out.

The word error rate, cpu-time, the average number of word hypotheses for each word-id are shown for the various word graph densities in Figure 8,9,10 respectively.

- The cpu-time requirement could be estimated using word graph density and lexicon size (*lexsize*).
- The average number of word hypotheses for each word-id could be also estimated using word graph density and lexicon size (*lexsize*).
- The word error rate (% of correct words not included in the word graph) is also depend on the word graph density and lexicon size (*lexsize*).

These results show the possibility of simple estimation of recognition error rate for much larger or smaller size lexicon.

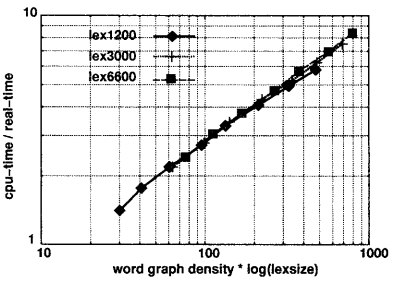


Figure 8: Cpu-time requirement dependency on word graph density and lexicon size

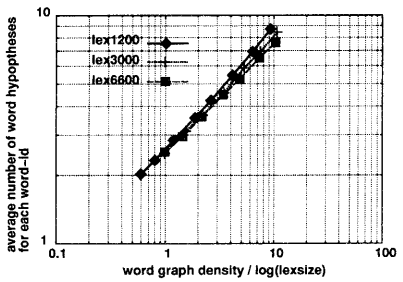


Figure 9: Complexity of time domain dependency on word graph density and lexicon size

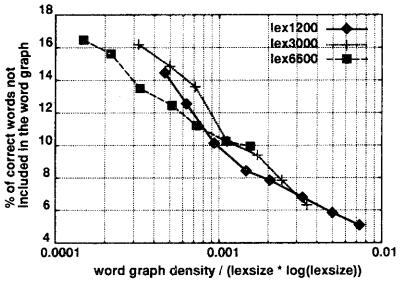


Figure 10: Word error rate dependency on word graph density and lexicon size

5 Conclusions

This paper proposes "cross-word context approximation" and "lenient language score smearing on lexicon tree" to reduce the number of word hypotheses (to reduce the computational cost) in word graph generation. The experimental results using spontaneous dialogue speech show an approximately 30% reduction of word hypotheses when applying "cross-word context approximation", compared to the case of no approximation. We also show the use of class bigram language model scores as the expected language scores in the lexicon tree achieve the lower error rate by the reasonable storage effort.

Acknowledgments

The authors would like to thank Hirokazu Masataki for supplying a class bigram language model on "travel planning corpus".

References

[1] T. Shimizu, S. Monzen, H. Singer, S. Matsunaga: "Time-Synchronous Continuous Speech Recognizer Driven by a Context-Free Grammar," Proc. of ICASSP'95, pp.584-587, (1995).

[2] H. Murveit, J. Butzberger, V. Digalakis, M. Weintraub: "Large Vocabulary Dictation using SRI's Decipher Speech Recognition System: Progressive Search Techniques," Proc. of ICASSP'93, pp.119-122, (1993).

[3] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker: "The LIMSI Continuous Speech Dictation Syatem: Evaluation on the ARPA Wall Street Journal Task," Proc. of ICASSP'94, pp.557-560, (1994).

[4] H.Ney, X. Aubert: "A Word Graph Algorithm for Large Vocabulary, Continuous Speech Recognition," Proc. of ICSLP'94, pp.1355-1358, (1994).

[5] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev, S.J. Young: "The 1994 HTK Large Vocabulary Speech Recognition System," Proc. of ICASSP'95, pp.73-76, (1995).

[6] R. Schwartz, S. Austin: "A Comparison of Several Approximate Algorithms for Finding Multiple (N-best) Sentence Hypotheses" Proc. ICASSP'91, pp. 701-704, (1991).

[7] V. Steinbiss, B.H. Tran, H. Ney: "Improvements in Beam Search," Proc. of ICSLP'94, pp.2143-2146, (1994).

[8] H. Masataki, S. Matsunaga, Y. Sagisaka: "Variable-Order Statistical Language Modeling for Continuous Speech Recognition," Technical Report of IEICE, SP95-73, pp.1-6, (1995).

[9] HMMsT. Morimoto et al.: "A Speech and Language Database for Speech Translation Research," Proc. of ICSLP'94, pp.1791-1794, (1994).

[10] J. Takami, S. Sagayama: "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. ICASSP'92, pp. 573-576, (1992).

[11] K. Ohkura, M. Sugiyama, S. Sagayama: "Speaker Adaptation based on Transfer Vector Field Smoothing with Continuous Mixture Density ," Proc. ICSLP'92, pp. 369-372, (1992).