DEALING WITH OUT-OF-VOCABULARY WORDS AND SPEECH DISFLUENCIES IN AN N-GRAM BASED SPEECH UNDERSTANDING SYSTEM

Atsuhiko KAI, Yoshifumi HIROSE, and Seiichi NAKAGAWA

Faculty of Engineering, Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-chou, Toyohashi-shi, 441-8580 JAPAN

ABSTRACT

In this study, we investigate the effectiveness of an unknown word processing(UWP) algorithm, which is incorporated into an N-gram language model based speech recognition system for dealing with filled pauses and out-of-vocabulary(OOV) words. We have already been investigated the effect of the UWP algorithm, which utilizes a simple subword sequence decoder, in a spoken dialog system using a context free grammar(CFG) as a language model. The effect of the UWP algorithm was investigated using an N-based continuous speech recognition system on both a small dialog task and a large-vocabulary read speech dictation task. The experiment results showed that the UWP improves the recognition accuracy and an N-gram based system with the UWP can improve the understanding performance in compared with a CFG-based system.

1. INTRODUCTION

There exist many challenging issues for dealing with spontaneous speech in a spoken dialog system. In such situation, the performance of a system is greatly affected by the ability to deal with spontaneous speech phenomena. Also, we will encounter the out-of-vocabulary word and out-of-grammar sentence problems for the lack of task-specific knowledge and insufficient modeling of language model.

The approaches of garbage modeling and the use of phonetic typewriter have been reported to deal with unknown words and filled pauses[1, 2]. Although the detection performance of unknown words may be poor if the vocabulary size is large and when it is applied to a continuous speech task where the word boundary is ambiguous, an unknown-word processing method which is based on a subword sequence decoder or phonetic typewriter would have better performance when the acoustic model has high accuracy.

In this study, we investigate the effectiveness of an unknown word processing algorithm(hereafter, referred to as UWP algorithm), which is incorporated into an N-gram language model based speech recognition system for dealing with filled pauses and out-of-vocabulary(OOV) words. While a context-free grammar(CFG) is often used as a language model for read speech and written text, also ill-formed sentences and the linguistic phenomenon of an inversion should be treated for dealing with spontaneous speech. As a result, the increase in the complexity of such a language model may become serious. While we have already been applied the UWP to a speech recognition system based on the CFG and shown the effectiveness of UWP algorithm[4], this study applies the UWP to an N-gram based speech recognition system and compares the performance with the case of CFG language model to investigate the effectiveness of UWP algorithm. The evaluated utterances are spontaneous speech for a small task and read speech for a large vocabulary dictation task.

2. UNKNOWN WORD PROCESSING IN N-GRAM BASED SYSTEM

2.1. Principle of Unknown-Word Processing

If we construct the acoustic models based on subword units, unknown words can be modeled as an arbitrary sequence of subword units. Some researchers have already presented the results of the UWP method based on this principle. Asadi et al. introduced some new word models which consist of sequences of phonemes and their models are separately created for each open class that accepts new words[2]. A problem of computational cost would arises when a number of open classes are described since the verification of new word models for each open class has to be done independently. We have proposed an UWP algorithm which utilizes the intermediate result obtained during the process of subword sequence decoding or a phonetic typewriter[4]. This method allows a computational cost almost independent of the vocabulary size and the complexity of language models. In the following subsections, the algorithm for incorporating UWP into an N-gram based one-pass beam search algorithm is described.

2.2. Dealing with Out-of-Vocabulary Words

Our continuous speech recognition system is implemented based on a standard one-pass beam search algorithm. The search algorithm employs a tree-organized lexicon in which each node corresponds to a subword unit and generates copies of nodes to hold the cumulative likelihood of different word context which is remained within a beam width in the search process[6]. Recently, the search algorithm was improved by changing the search algorithm such that the tree nodes of different word context are shared. This approximation has already been applied in some continuous speech recognition systems[5] and shown that the computational cost can be reduced with no significant decrease in the recognition accuracy.

A phonetic decoder is employed for generating the candidate of unknown words. The phonetic decoder is based on the one-pass DP algorithm and can decode a most likely phonetic sequence at each processing time. Thus, we investigate a method that both the main search process and the phonetic decoding process run in parallel and the latter hypothesizes unknown words which end at each processing time. At each frame time, a set of unknown-word candi-

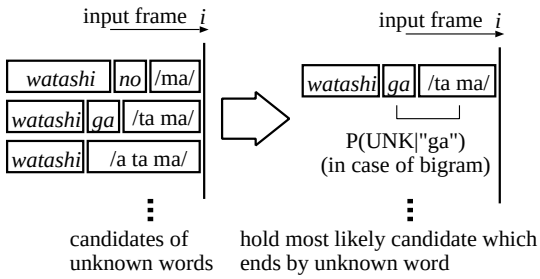

Figure 1: Concept of unknown-word processing

Notation:

N: vocabulary size

$start[x]$: beginning frame of the x-th candidate of unknown-word with the length of x subword units

$S_{\mathrm{UNK}}(i;x)$: acoustic score of the x-th unknown-word(UNK) candidate which begins from $start[x]$ and ends at the i-th frame

$L_n(i)$: cumulative acoustic and linguistic score of an optimal hypothesis which ends by word n, given an observation sequence of $1 \sim i$ frames

$L_{\mathrm{UNK}}(i;j,n)$: cumulative acoustic and linguistic score of an optimal hypothesis which ends by unknown-word(UNK), preceded by word n which ends at the j-th frame

$B_n(i)$: back-pointer of word n at the frame i

1. Execute steps 2,3,4 for all the unknown-word hypotheses($x = 1, 2, \cdots$) which end at i-th frame.
2. $j \leftarrow start[x] - 1$
3. Execute steps 4,5 for $n = 1, 2, \cdots, N$.
4. Update $L_n(i)$ with Viterbi algorithm
5. $L_{UNK}(i;j,n) = L_n(j) \cdot S_{UNK}(i;x) \cdot P(UNK|n)$
6. $\hat{j}, \hat{n} = \arg\max_{j \in \{start[x]:x=1,2,\ldots\},n} L_{UNK}(i;j,n)$
   $L_{UNK}(i) = \max L_{UNK}(i;\hat{j},\hat{n})$
   $B_{UNK}(i) = \hat{j}$

Figure 2: Unknown-word processing algorithm

dates is obtained by backtracking the optimal path that ends at that time and considering a part of subword unit sequences with a different start frame time as a different hypothesis. In this study, the length of unknown-word candidates is restricted to contain only 2∼10 syllables. Figure **1** illustrates the above UWP algorithm in which a most likely unknown-word candidate is selected at a processing frame.

Figure **2** shows the UWP algorithm at the i-th frame in the one-pass search algorithm. For simplicity, the algorithm assumes that a bigram language model is used and therefore the cumulative likelihood score and corresponding back-pointers should be updated and held for each word and frame time.

The N-gram probability $p(UNK|w)$, where UNK denotes unknown words, is estimated by replacing all the OOV words in a training text corpus with a special symbol. Since the UNK can represent a number of OOV words, the estimated probability $p(UNK|w)$ becomes higher than that of each different word which was replaced with

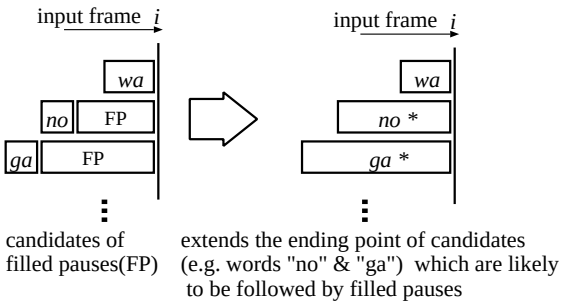

Figure 3: Concept of unknown-word processing for filled pauses

UNK. Therefore, a penalty score should be added to each unknown-word hypothesis when N-gram probability is applied. The penalty score may be empirically determined by considering the number of different words which belong to UNK in a training text.

2.3. Dealing with Filled Pauses

In most spoken dialog systems, filled pauses are ignored in the process of speech understanding and dialog management. Although only limited kinds of interjections are frequently observed in a human-to-human dialog, the result also reveals that other about 10% of filled pauses can have many kinds of transcriptions and uncertainty in the pronunciation[7]. In our previous study, we have applied the UWP for dealing with filled pauses and attained the result which is comparable with an approach in which frequent 10 interjections are registered to a lexicon[4]. Thus, we also incorporate the UWP into our N-gram based search algorithm and investigate the effectiveness of this approach for dealing with filled pauses as well as unknown words.

In general, the N-gram statistics for filled pauses may not be estimated since it is not practical to collect a spontaneous speech and text corpus on a specific task. Thus, we ignore the N-gram statistics for filled pauses and the UWP algorithm for dealing with filled pauses was slightly changed from the one in the previous section. When a linguistic probability for a sentence hypothesis is calculated by N-gram model, all filled pauses are ignored and skipped. To prevent false alarms of filled pauses from increasing, we introduce a penalty score which controls the possibility of detecting filled pauses. Figure **3** illustrates the concept of a modified UWP algorithm for filled pauses.

At a processing frame time i, if a cumulative likelihood of the unknown-word hypothesis preceded by a word n is higher than that of the word n preceded by any words, we can assume that the word n is ending at the frame i with the former likelihood score while in fact we can detect that the word n is followed by filled pauses in this case. As in the case of out-of-vocabulary words, the length of unknown-word candidates for filled pauses is restricted to contain only 2∼3 syllables.

3. EXPERIMENTAL RESULTS

3.1. Systems for A Spoken Dialog Task

The experiments of evaluating language models and speech recognition systems are carried out using text and speech

Table 1: Language models and test conditions

Language model	CFG	bigram
Vocabulary size	241	
OOV rate(%)	0.9	
filled pauses(%)	4.4	
Acceptance rate(%)	87.0	93.0
Test set perplexity	78	24

corpora on a task domain of "Traveler's information guidance around Mt.Fuji." Language models of context-free grammar(CFG) are also used to compare with the word bigram language models. Furthermore, two types of text and speech corpora, which correspond to the read and spontaneous speech, respectively, are used.

The acoustic models used in this experiment consist of 113 syllable based HMMs(Hidden Markov Models). Each syllable-unit model has 5 states, 4 Gaussian densities with full-covariance matrix and 4 discrete duration distributions. Feature parameters consist of 10-th order LPC melcepstral coefficients and their regressive coefficients.

In the first experiment, a text corpus which contains 357 different words and 914 sentences is used for training and building language models, while a part of vocabulary words(116 words) are labeled as an OOV word in the training corpus and excluded them from a lexicon in the speech recognition system. Also a read speech corpus which includes various speech disfluencies and ill-formed sentences is used. The text corpus is collected by 53 subjects who are asked to think out and write down some query sentences in condition that available vocabulary words are presented in advance. A set of 115 sentences is uttered by 2 male speakers and they are used as the test set for evaluating language models and speech recognition systems. The test sentences include interjections(filled pauses) in 18 sentences and restarts in 17 sentences. The ill-formed sentences which are included in the test sentences and some other likely sentences are accepted in the CFG model. The summary of language models are shown in Table **1**.

Table **2** shows the experimental result of a bigram-based speech recognition system, with different penalty for unknown-word candidate. The penalty is added to a log-likelihood of the unknown-word candidate according to the number of subword units contained. To investigate the performance of the UWP, the following two measures are used:

$$Recall\ rate = NU_d/NU_c$$
$$Precision = NU_d/NU_h$$

where

NU_c: number of unknown words or filled pauses uttered
NU_d: number of unknown words or filled pauses correctly detected
NU_h: number of words output as unknown words or filled pauses

This result shows that the UWP algorithm used in a bigram-based system can also improve the recognition performance, attaining the word accuracy of 87.5% with the UWP, in compared with the word accuracy of 84.7% without the UWP. Although the unknown words often degrade the recognition accuracy in a bigram-based system, a recall rate of 65.8% and a precision of 60.7% were attained using the UWP for filled pauses.

Table 2: Results of bigram-based system
UNK: unknown words **FP:** filled pauses

UWP & Penalty	Word accuracy (%)	Recall rate(%)		Precision (%)
		UNK	FP	
no UWP	84.7	–	–	–
−40	86.7	43.8	61.4	50.3
−45	87.3	31.3	65.7	56.3
−50	87.5	37.5	65.8	60.7

Table 3: Comparison of sentence-level performance

Language model	Sent.cor.(%)	Sem.acc.(%)
CFG	57.2	82.5
Bigram	55.9	86.9

Table 4: Language models for spontaneous speech

Language model	CFG	bigram
Vocabulary size	359	
OOV rate(%)	1.8	
filled pauses(%)	0.2	
Acceptance rate(%)	80.1	88.3
Test set perplexity	130	9

Table **3** shows the sentence-level recognition accuracy which is important for spoken dialog system. The semantic accuracy(Sem.acc.) is the same as sentence correct(Sent.cor.) except that the errors on postpositional particles are ignored, because these errors are corrected when a dialog system understands their intention for a simple spoken dialog task[8]. Although the CFG-based system is slightly better than the bigram-based system as for the sentence correct, the latter attained a better semantic accuracy. This result shows that the UWP can also be successfully applied for a bigram-based system.

In the second experiment, spontaneous speech and transcribed text data are used. Their data have been collected through the experiments in real condition that 24 naive users are asked to plan a travel using our spoken dialog system. Each text and speech corpus are separated into training and test set. Test set consists of 437 sentences uttered by 4 speakers and remaining 2063 sentences are used for training and building language models (359 vocabulary words).

Table **4** shows the test conditions in terms of language models. The significant difference in the perplexity is due to the fact that the test sentences contain only 5.9 words in average and all possible sentences with an inversion, which mean those with no constraints on the phrase order, are taken into account in CFG case, while the utterances of this sort is very few in the training text corpus.

Table **5** shows the experimental result of both the bigram-based and CFG-based systems. The UWP penalty per subword unit length was set to -60. In this experiment, the use of the UWP didn't lead to a significant improvement on the word accuracy. This is mainly due to the fact that only a few kinds of OOV words are included in the test sentences and a frequent OOV word could not be detected in most cases since it was similar to a registered word in its pronunciation. However, we find that the UWP does not degrade the performance in compared with a baseline sys-

Table 5: Experimental result of spontaneous speech

Language model & UWP	Word acc. (%)	Sent.cor. (%)	Sem.acc. (%)
bigram, no UWP	91.5	72.8	84.7
CFG, with UWP	73.3	39.3	76.9
bigram, with UWP	91.8	69.1	88.6

tem. Furthermore, the superiority of N-gram based speech recognition system is also observed in this experiment using the spontaneous speech corpus.

3.2. A Large-Vocabulary Dictation Task

In this study, we also performed an experiment to investigate the effect of the UWP on a large-vocabulary task. We used a newspaper article text database of the Mainichi Newspaper for building N-gram language models. A training set consists of about 86 million morphemes from a pool of 4 years articles. We built a bigram and trigram language model with a vocabulary size of 5000 words. A test set of speech data is a part of the ASJ continuous speech corpus, Japanese Article Sentences(JNAS), and this experiment used the utterances from 3 male speakers and about 100 different sentences for each speaker. The OOV rate of the test set is 10.6%. The test set perplexity and the adjusted perplexity(APP)[10], which the number of different unknown words in the test set is taken into account, are 74.8($APP = 141$) for bigram and 50.4($APP = 94.9$) for trigram, respectively.

In this experiment, we used a segment-unit input HMM[9], which has the same topology with the HMM described in Section **3.1.**. The output probability has used 4 Gaussian mixtures per state and the feature parameters consist of a 20-th order segmental feature, and both the linear and quadratic regressive coefficients of 10-th order mel-cepstral coefficients(ΔMCEP,$\Delta\Delta$MCEP) and energy(ΔE,$\Delta\Delta$E)[9].

Table **6** shows the performance of a speech recognition system with N-gram language model and the UWP. An N-best sentence candidate list was obtained from a bigram-based system and re-scored by a trigram language model. The unknown-word penalty per subword unit was fixed to -50 and the penalty for OOV bigram probability was changed from log 10 to log 290000 since there were about 290 thousands of different OOV words in a training corpus and the probability may be over-estimated. The values in parentheses are the performance on a part of utterances which do not include OOV words. In compared with the word accuracy on the *limited* utterances rejected or without OOV words, the result on *all* utterances shows relatively low performance for a small vocabulary size and a high OOV rate. We also evaluated a rejection performance with the measures of a recall and precision, which are calculated by a detection-per-sentence basis (i.e., one decides whether some unknown words are contained or not for each sentence) since the existence of unknown words will often affect the surrounding words in a continuous speech.

While the result shows that the improvement of the word accuracy is not significant, a high recall and precision rate is obtained. This is due to the fact that the UWP often detect an unknown word with an incorrect word boundary since it does not use strong constraint on the unknown-word hypothesis. However, this result also reveal that the UWP may be effectively used to deal with spontaneous speech since it can be used to detect the existence of OOV words.

Table 6: Result of a large-vocabulary dictation task

Language model, UWP & Penalty	ALL	Limited with rejection		
	Word acc.(%)	Word acc.(%)	Recall rate(%)	Precision (%)
bigram, no UWP	65.0	(85.9)	–	–
bigram, $-log10$	58.1	87.5	95.0	84.7
bigram, $-log5000$	64.5	74.5	60.3	91.1
bigram, $-log290000$	65.8	68.9	28.5	95.8
trigram, no UWP	66.7	(87.3)	–	–
trigram, $-log10$	62.6	86.8	95.8	82.4
trigram, $-log5000$	67.4	79.8	75.7	89.2
trigram, $-log290000$	68.1	72.4	43.1	93.6

4. CONCLUSIONS

We have investigated the effectiveness of an unknown-word processing method for dealing with spontaneous speech phenomena in N-gram based speech recognition system. The experimental result showed that the N-gram language model which includes OOV statistics and used in an unknown-word processing can be more effective for a spoken dialog system in compared with a grammar-based speech recognition system. We also showed that the unknown-word processing can improve the performance on a large-vocabulary dictation task, while the result also suggested that more detailed constraint for detecting unknown words should be introduced.

REFERENCES

1. J. G. Wilpon *et al.*, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *IEEE Trans. ASSP*, vol.38, no.11, pp.1870–1878 (1990).
2. A. Asadi *et al.*, "Automatic detection of new words in a large vocabulary continuous speech recognition system," *Proc. ICASSP'90*, pp.125–128 (1990).
3. N. Inoue *et al.*, "A method to deal with out-of-vocabulary words in spontaneous speech by using garbage HMM," *IEICE Trans.*, vol.J77-A, no.2, pp.215–222 (Feb. 1994)(in Japanese).
4. A. Kai, and S. Nakagawa: "Investigation on unknown word processing and strategies for spontaneous speech understanding," *Proc. EUROSPEECH'95*, pp.2095–2098 (1995).
5. S. Koga *et al.*,"A Real-Time Speaker-Independent Continuous Speech Recognition System Based on Demi-Syllable Units," *Proc. ICSLP*, pp.1483–1486 (1992).
6. G.Antoniol *et al.*, "Language model representations for beam-search decoding," *Proc. of ICASSP'95*, pp.588–591 (1995).
7. S. Nakagawa and S. Kobayashi: "Phenomena and acoustic variation on interjections, pauses and repairs in spontaneous speech," Jour. Acoustical Society of Japan, vol.51, no.3, 1995.(in Japanese)
8. T.Itoh *et al.*, "A robust dialogue system with spontaneous speech understanding and cooperative response," *Proc. of ACL Post-Conference Workshop "Interactive Spoken Dialog Systems: Bringing Speech and NLP Together in Real Applications"*, pp.57–60 (1997).
9. S.Nakagawa and K.Yamamoto, "Evaluation of segmental unit input HMM," *Proc. of ICASSP'96*, pp.439–442 (1996).
10. J.Ueberla, "Analysing a simple language model – some general conclusions for language models for speech recognition," *Computer Speech and Language*, Vol.8, No.2, pp.153–176 (1994).