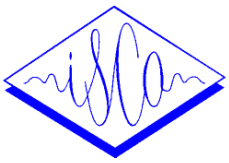

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

INVESTIGATION ON UNKNOWN WORD PROCESSING AND STRATEGIES FOR SPONTANEOUS SPEECH UNDERSTANDING

Atsuhiko KAI and Seiichi NAKAGAWA

e-mail:{kai,nakagawa}@slp.tutics.tut.ac.jp
Toyohashi University of Technology
Toyohashi, 441, JAPAN

ABSTRACT

In this paper, an unknown word processing method is investigated as an approach to achieve the robust spontaneous speech understanding. The unknown words are detected by a subword-unit based decoder and the process is incorporated into the One-Pass search algorithm. The preliminary experiments showed that the method is also effective to the utterance including interjections (or filled pauses). In addition, the phrase-spotting based approaches were compared with the One-Pass search method in which the unknown word processing was incorporated. The experiments showed that the One-Pass method attained the best performance on spontaneous speech.

1. INTRODUCTION

Recently, studies on spontaneous speech recognition and understanding have been done extensively. While the main approaches for realizing the analytical and verification methods of spoken language are also based on conventional methods used for read speech, these approaches have not been sufficiently evaluated for spontaneous speech.

In general, we should consider two aspects of speech recognition approaches for spontaneous speech: the verification method, which can process out-of-vocabulary words as well as restarts and filled pauses; and the parsing method for spoken language, which can accept the syntactically ill-formed sentences. Word spotting is one of promising approaches to the verification problem. However, the lack of higher-level knowledge in the verification process may make the recognition accuracy degrade. The other approaches to resolve the verification problem have been studied: the garbage modeling[1, 2] and the subword unit based modeling[2, 3] of the unknown words. Although the latter seems to be more promising in continuous speech recognition, it consumes much processing times in that all the probable unknown words in many grammatical states have to be verified independently. In this study, we first performed preliminary experiments to investigate the performance on utterances with unknown words or interjections using a fast approximate search method which is incorporated into the One-Pass search recognition algorithm[5].

The word- or phrase-spotting based approach seems to be effective in spontaneous speech since only the probable word candidates contribute to the sentence hypothesis. However, the parsing strategy has not sufficiently been studied when a heuristic search technique is used with weak syntactic constraints. In this paper, we first compare two strategies using a semantic grammar which has a weak syntactic constraint to allow ellipses of postpositional particles and inversion. Then, the phrase spotting based approach is compared with the one-pass search method which incorporates the unknown word processing method.

2. VERIFICATION OF OUT-OF-VOCABULARY WORDS

2.1. Incorporating into One-Pass Search

If we construct the acoustic models based on subword units, the unknown word can be modeled as an arbitrary sequence of subword units. To prevent from increasing a search space in continuous speech recognition, an approximate method of unknown word processing is employed.

An approximative method to reduce the computation for the syntax-directed connected word recognition has been proposed previously[4], which bundles the verification process of the word which corresponds to different grammatical states. Similarly, we can apply the idea to the unknown word verification process. Our preliminary experiment using syllable HMM acoustic models has showed an insignificant difference of the performance in compared with the exact verification method. Here we describe the approximative method briefly assuming that the subword unit is syllable.

In the dynamic time warping or the Viterbi scoring method for continuous speech recognition, e.g. the One-Pass algorithm[5], the accumulated likelihood score of the hypothesis which arrives at a grammatical state q is represented as

$$L_q(i) = \max_{p,m,n}\{L_p(m) + L^n(m+1:i)\}, \qquad (1)$$

where $L_q(i)$ is the maximum verification score among the Viterbi paths which arrive at state q in i-th frame, $L^n(m+1:i)$ is the accumulated verification score of the word n summed up between $m+1 \sim i$ frame and $\delta(p,n) = q$(the word n is outputted at the state transition from p to q) should be satisfied if a syntactic constraint is imposed. The accumulated likelihood score of the optimal syllable sequence without lexical constraints can be represented as

$$L(i) = \max_{m,s}\{L(m) + L^s(m+1:i)\}, \qquad (2)$$

where s is a syllable. This recursive formula is efficiently calculated by $O(n)$. We approximate the second term in equation (1) by equation (2) when n is an *unknown word*. That is, only the optimal boundaries obtained by equation (2) are considered as the candidate of m in equation (1) and $L^n(m+1:i)$ is calculated as $L(i) - L(m)$. In practice, we should impose a penalty on the likelihood score of

10.21437/Eurospeech.1995-501

unknown words because the score always equals or higher than the likelihood of any of the vocabulary words.

The above method can be extended to impose a general constraint for unknown words. We considered two constraints:

- length of unknown word (number of syllables)
- single occurrence of unknown words in an utterance

Obviously, the former constraint is easily incorporated into the above method. The latter is also accomplished by using two kinds of accumulated likelihood tables, either of which corresponds to the hypothesis including an unknown word or not including unknown words, respectively. However, the computation for updating likelihood tables also becomes about double while the overall increase of computation will be relatively small in practice.

2.2. Preliminary Results

We have developed a speech understanding system SPOJUS-SYNO-X which integrates a top-down context-free parser[6]. The acoustic models consist of 113 syllable based HMMs, which have 5 states, 4 Gaussian densities and 4 discrete duration distributions. The speaker-independent HMMs were adapted to the test speaker using 20 utterances for the adaptation[7]. The grammar used in our speech recognizer is represented by a context-free grammar which describes the syntactic and semantic information. The speech recognition task is "Sightseeing guidance around Mt. Fuji" with a vocabulary size of 500 words. To investigate the performance for processing of the interjections or unknown words in spontaneous speech, we incorporated the unknown word processing into our system.

The experiments for interjection and unknown word processing are separately performed using the different corpus. In the evaluation experiment of unknown word detection and recognition, several proper nouns, such as the name of mountains or lakes, are removed from the grammar. Instead, the special terminal symbol which represents unknown words is added to the rewriting rules of a wordclass to which an unknown word might belong. This operation yielded 6 kinds of unknown words and 21 sentences in all the 104 test sentences. The test set consists of 624 sentences (104 × 6) uttered by 6 male speakers. The test set perplexity is about 29 when all the unknown words are registered, that is, for the original grammar.

Table 1 shows the recognition accuracy of the sentences that include unknown words(UWs). To examine the effect of the constraints for unknown words mentioned above, the experiments are carried out for several conditions: "none"(with no constraints), "L"(with length constraint), "N"(with occurrence constraint) and their combinations. The false alarm rate(FA) denotes the rate of the number of sentences in which some words were detected as an unknown word despite being in the lexicon. This result shows that more than 70% of unknown words are correctly detected if we consider only the utterances which have no recognition errors when the unknown words are registered (52.8% > 69.4% × 70%). The added constraints improved both the detection accuracy of unknown words and the false alarm rate.

For utterances that include interjections(filled pauses), the unknown word processing(UWP) was evaluated in

Table 1: Performance on unknown words(UWs)

Constraint for UWs	include UWs (18 × 6 sent.)	not include UWs (86 × 6 sent.)	
	Correct(%)	Correct(%)	FA(%)
none	44.4	75.6	9.1
L	49.1	76.7	5.9
N	50.9	75.8	7.6
L,N	52.8	76.7	5.4
UWs registered & UWP not used	69.4	78.9	0.0

Table 2: Performance on interjections(interj.)

Grammar/ Method	not include interj.	include interj.	
	%Correct	%Correct	%Understanding
Original	86.0	4.0	32.0
+ 10 interj.	88.0	74.0	76.0
+ 30 interj.	86.0	68.0	74.0
+ 78 interj.	84.0	68.0	76.0
UWP used	84.0	56.0	72.0
+ 10 interj. & UWP used	84.0	70.0	74.0

comparison with the conventional approach where the frequent interjections(filled pauses) are registered in the system's lexicon. The context-free grammar was modified to accept sentences which could be intervened by interjections between phrases. The number of interjections registered in the system's lexicon was varied from 10(appeared in the test utterances) to 78(all of the interjections observed in database[9]). The testing data consisted of 50 utterances of one speaker, which was the subset of 104 utterances used before, and the different 50 utterances in which half of the sentences were preceded by an interjection and the rest of the sentences was intervened by an interjection. Table 2 shows the results of the recognition performance. We can see that the comparable performance for both of the approaches may be obtained. However, we should know that only the limited kinds of interjections are used while a great varieties of interjections would be used in practice. Our further experiments in the case that some interjections(appeared in test utterances) were not registered suggested that the combination of the unknown word processing and the registration of frequent interjections would be most effective (see the last row in the table).

3. COMBINATIONS OF VERYFICATION AND PARSING STRATEGIES

The word- or phrase-spotting based recognition seems to be effective in spontaneous speech since the implicit acoustic modeling of the extraneous speech such as restarts or interjections may not be required. However, in general, the parser should employ a heuristic search technique since the search space becomes extremely large, if a weak constrained grammar is used for dealing with spontaneous speech in a moderate task. In this section, we first describe the comparison of two parsing strategies using a semantic grammar which has weak syntactic constraints to allow ellipses of postpositional particles and

inversion. Moreover, the phrase spotting based approach is compared with the One-Pass search method which incorporates the unknown word processing method for verifying the occurrence of restarts or interjections.

3.1. Phrase-Spotting Based Approaches

A phrase spotter is used as the front-end of the parsers which are to be compared. The phrase in Japanese, which is called *bunsetsu*, is said to be constrained in its order much less than in English while the word order within a phrase is fairly constrained. Therefore, we represent the intra-phrase grammar by a finite state network and the inter-phrase grammar by a context-free grammar which allows ellipses of postpositional particles and inversion. The intra-phrase grammar also allows the arbitrary sequence of syllables followed by a bunsetu for dealing with extraneous speech. The initial and the final state of the network is connected by a null-arc to allow the spotter to recognize an optimal phrase and syllable sequence. The phrase lattice is produced by N-best backtracking of the partial phrase hypotheses like the lattice N-best method in the reference [8].

Island-Driven Search

While the island-driven search from a phrase lattice is promising for spontaneous speech, the algorithm becomes more complicated than in a usual left-to-right search. Therefore, we describe the rules of the syntactic and semantic knowledge by a case-frame like grammar. The rules consist of both *the sentence level description* and *the inter-phrase level description*. The former has at least one predicate(header) and some phrase non-terminal symbols for each rule. The phrase order isn't constrained in these rules. The latter represents the inter-phrase constraints by the correspondence between the phrase non-terminal and the legal phrase class sequence, and is the form of context-free rewriting rules. The phrase class is a representative of the set of phrases of which semantic function may be identical.

The search procedure first begins by selecting top N phrases which have predicate verbs to create new partial candidates. Then, each time a new partial candidate is created, top N phrases are selected to expand the candidates by checking if the phrase can connect to the candidate with the grammatical and temporal(i.e., phrase order and overlap length) constraints. While the N is used to adjust the beam width of the search, the number of candidates which will be expanded is also constrained according to the sum of the phrase scores for further pruning of the search space. The constraint is imposed on each set of candidates which includes the same number of phrases to avoid the normalization of the candidate scores. Finally, all the sentence candidates in the list should be ordered by an evaluation function. Since the candidate may include the skipped intervals or the overlaps by extraneous speech and the inaccuracies of the lattice, we introduce some score weights to the sentence score as shown below.

$$
\begin{aligned}
S_s \;=\; & \frac{\text{(sum of phrase scores)} + \text{(penalty for overlap)}}{\text{(frames in length covered by all the phrases)}} \\
& + \alpha \times \text{(frames in length covered by all the phrases)} \\
& - \beta \times \text{(number of phrases)}
\end{aligned}
$$

Left-to-Right Search

The left-to-right search strategy from a phrase lattice has been used in our conventional speech understanding system SPOJUS-SYNO I/II[10], which is based on an Earley-like top-down parser and processed in a frame-synchronous way. In this approach, the difference from the above approach is that we should provide the score for the extraneous speech during the search. We simply provide a constant score as the estimate of the verification score for the skipped intervals. The constant score is estimated by the normalized log-likelihood score of the connected syllable decoding over the input, which is obtained from the byproduct of the phrase spotting. Since the grammar should be the form of context-free grammar, the rules used in the previous method are converted to the context-free rewriting rules by enumerating all the possible phrase orders according to the previous sentence level description, resulting in the identical syntactic constraint.

3.2. One-Pass Viterbi Search

The One-Pass search algorithm used in section 2.2 can often attain better recognition accuracies for read speech and the computational amount is fairly saved by integrating with beam search and pruning method[6]. The unknown word processing is employed here to verify the extraneous speech such as interjections and restarts. The grammar used in the left-to-right method is extended to allow unknown words(i.e., connected syllable sequence without language model) between all the adjacent phrases. For comparing with the spotting-based approaches, no interjections are registered in the system's lexicon.

4. EXPERIMENTAL RESULTS

Experiments were performed on the task "Accommodation guidance around Mt. Fuji" with the vocabulary size of 120 words, which is the subdomain of "Sightseeing guidance around Mt. Fuji" task. The testing data for evaluating the above methods consist of 70 utterances × 2 male speakers. The testing data includes interjections(filled pauses) in 20 utterances, ellipses of postpositional particles in 10 utterances, restarts in 10 sentences and inversion in 9 utterances (13 utterances are their combination). The rest 35 utterances have no such ill-formedness (referred to as legal utterances). The test set word perplexity is about 41 in the case that the interjections and restarts are ignored while the perplexity was about 29 for the 500-word vocabulary task mentioned in section 2.2 which accounted for only legal sentences. The acoustic model used in recognition systems consisted of 113 syllable-unit based HMMs which were also used in section 2.2. The parameters for above methods, which are penalty scores for unknown word processing, score weights for the island-driven method and beam search width, are determined by preliminary experiments on the test set.

Table 3 shows the quality of the phrase lattice used by spotting-based approaches. The rate in the column of the n-th rank shows the percent accuracy by which an uttered phrase is correctly identified as one among the best n spotted phrases in the neighborhood. The feature parameter MEL and RGC denotes the mel-scaled LPC cepstral coefficients and the regressive coefficients, respectively.

Table 3: Phrase spotting results

Feature parameter	Detection order and rate (%) Top	≤2	≤5	≤10	Missing phrases	Spotted phrases per sentence
MEL	58.4	79.0	95.1	98.7	3 (1.4%)	2451
MEL+RGC	62.9	85.3	96.8	98.2	4 (1.8%)	2429

Permissible detected boundary error = ±128 msec.

Table 4: Comparison of different approaches

(I-D: island-driven method, L-to-R: left-to-right method)

Feature parameter	Method	Phrase[1] %COR	%ACC	Sentence correct(%)	understanding(%)
MEL	I-D	55.0	52.9	22.9	65.7
	L-to-R	65.8	53.4	24.3	69.3
	One-Pass(1)	67.9	58.1	28.6	61.4
	One-Pass(2)	63.1	59.0	23.6	67.1
MEL+ RGC	I-D	63.8	57.5	30.7	70.7
	L-to-R	68.3	52.7	26.4	68.6
	One-Pass(1)	72.6	60.4	35.0	65.7
	One-Pass(2)	77.6	72.4	38.6	77.9

[1] %COR = (#correct phrases) / (#utterances) × 100
%ACC = (#correct phrases − #insertions) / (#utterances) × 100

The performance of different recognition methods are shown in Table 4. One-Pass(2) is the method used in section 2.2 and the One-Pass(1) is a variant of the One-Pass(2) method and it approximates the unknown word processing by using the constant penalty score which is proportional to the length as the estimate of the unknown word verification score. The *sentence correct* and *understanding* mean that the utterance was correctly recognized in the case that detected interjections and restarts were ignored and that the phrase class sequence and the meaning of noun phrases were correctly identified, respectively. The phrase accuracy shows that One-Pass(2) is superior to the other methods, in particular, if the dynamic feature parameter is used. Such results are also significant for the sentence understanding rate.

The recognition performance with respect to each set of the legal and illegal utterances is shown in Figure 4. It should be noted that the One-Pass(2) method is significantly better than the others with respect to the illegal utterances while the performance on the legal utterances is comparable. We can also see that, in particular, the left-to-right method and the One-Pass(1) method aren't robust to illegal utterances. In comparison of the required computations of the implemented systems, the One-Pass(2) method is required about only half of the processing time of the other methods.

5. SUMMARY

An unknown word processing method was investigated as a method for dealing with spontaneous speech. The preliminary results showed that the method could obtain a comparable performance for detecting interjections in continuous speech even if interjections were not registered to the system's lexicon. We also compared the phrase-spotting based approaches with the One-Pass search method in which the unknown word processing was incorporated. The experiments showed that the One-Pass method attained the best performance to spontaneous speech.

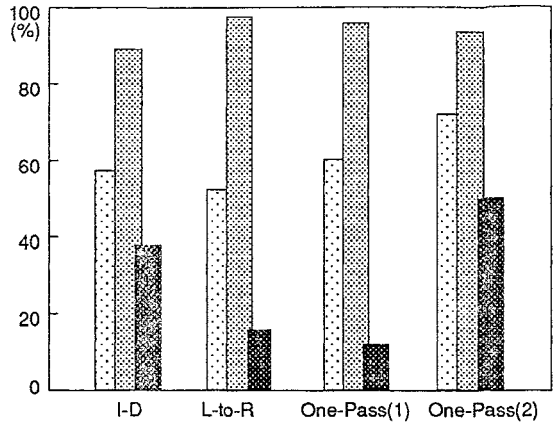

Figure 1: Comparison of recognition performance (MEL+RGC)

References

[1] J. G. Wilpon, L. R. Rabiner, C-H. Lee, E. R. Goldman: "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", *IEEE Trans.*, ASSP-38, Vol.11, pp.1870–1878 (1990).

[2] H. Bourlard, B. D'hoore and J-M. Boite: "Optimizing Recognition and Rejection Performance in Wordspotting Systems", *Proc. ICASSP*, I-373–I-376 (1994).

[3] A. Kai and S. Nakagawa: "Evaluation of Unknown Word Processing in a Spoken Word Recognition System", Proc. ICSLP, pp.2151–2154 (1994).

[4] S. Koga, R. Isotani, S. Tsukada, K. Yoshida, K. Hatazaki, T. Watanabe: "A Real-Time Speaker-Independent Continuous Speech Recognition System Based on Demi-Syllable Units", *Proc. ICSLP*, pp.1483–1486 (1992).

[5] I. S. Bridle, et al., "An Algorithm for Connected Word Recognition", *Proc. ICASSP*, pp.899–902 (1982).

[6] A. Kai and S. Nakagawa: "A Frame-Synchronous Continuous Speech Recognition Algorithm Using a Top-Down Parsing of Context-Free Grammar", *Proc. ICSLP*, pp.257–260 (1992).

[7] Y. Tsurumi and S. Nakagawa: "An Unsupervised Speaker Adaptation Method for Continuous Parameter HMM by Maximum A Posteriori Probability Estimation", *Proc. ICSLP*, pp.431–434 (1994).

[8] R. Schwartz and S. Austin, "A Comparison of Several Approximate Algorithms For Finding Multiple (N-BEST) Sentence Hypotheses", *Proc. ICASSP*, pp.701–704 (1991).

[9] S. Nakagawa and S. Kobayashi: "Phenomena and Acoustic Variation on Interjections, Pauses and Repairs in Spontaneous Speech", *Journal of ASJ*, Vol.51, No.3, pp.202–210 (1995) (in Japanese).

[10] S. Nakagawa, Y. Hirata, I. Murase and T. Tanoue: "The Syntax-Oriented Spoken Japanese Understanding System – SPOJUS-SYNO II", *Proc. EUROSPEECH*, pp.463–466 (1991).