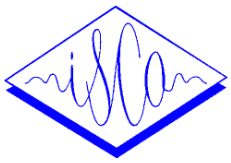

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

USING HIGH LEVEL KNOWLEDGE SOURCES AS A MEANS OF RECOVERING ILL-FORMED JAPANESE SENTENCES DISTORTED BY AMBIENT NOISE

K.H. Loken-Kim, Yasuhiro Nara, and Shinta Kimura

Fujitsu Laboratories Ltd.

ABSTRACT

In this paper, the authors present the performance of a large vocabulary Japanese spoken sentence recognition system tested under several noise levels (S/N (Signal to Noise Ratio) : 10dB, 20dB, 25dB, 30dB, ∞), and discuss the ability of the system to recover input sentences. The result of the performance evaluation was that the system recovered over 90% of the test sentences when the noise level was low (∞, 30dB), but recovered only 50% of the test sentences when the noise level reached to an S/N of 20dB.

1. INTRODUCTION

At Fujitsu laboratories, we are developing a large vocabulary (100,000 words and phrases) spoken sentence recognition system for Japanese language (speaker dependent isolated words). The overview and the performance of this system appears in [1]. The acoustic processing subsystem is described in [2], and the linguistic processing subsystem in [3]. In this paper, the authors present the performance of the acoustic processor tested under different ambient noise levels (mainly computer fan noise), and discuss the ability of the linguistic processor to recover sentences from the output of the acoustic processor.

2. LARGE VOCABULARY JAPANESE SPOKEN SENTENCE RECOGNITION SYSTEM

The large vocabulary Japanese spoken sentence recognition system (called recognition system hereafter) consists of an acoustic processing subsystem (called acoustic processor hereafter) and a linguistic processing subsystem (called linguistic processor hereafter) (Figure 1). The acoustic processor receives spoken input and generates a series of candidates collectively called a bunsetsu lattice. Figure 2 is an example of a bunsetsu lattice generated after speaking "watashi-wa hon-o yomimasu" (I am reading a book) .

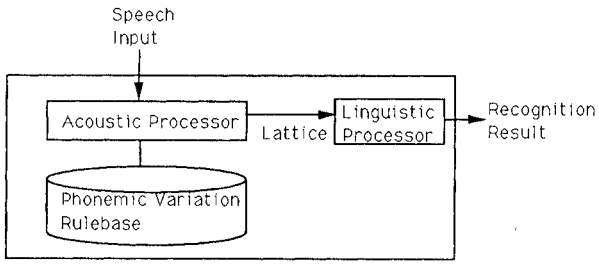

Fig. 1- Large Vocabulary Japanese Spoken Sentence Recognition System

The recognition system accepts sentences in a bunsetsu unit. Despite the absence of a clear definition of bunsetsu in Japanese, one can say that a bunsetsu generally consists of a function word and a concept word (Figure 3), and it is spoken in a single breath [4]. Uttering sentences in bunsetsu units minimizes the need to utter short function words separately, thus allowing a user to speak smoothly.

The linguistic processor receives a bunsetsu lattice and applies syntactic knowledge to recover what the speaker said. The linguistic processor, first, divides the bunsetsu into several groups based on the magnitude of the distance score differences. This grouping operation works because a great match score difference between the first and second candidates, or the second and third candidates (e.g. greater than 250) implies the likelihood of either the first or second candidate being the bunsetsu actually uttered. Separating likely candidates from the entire candidate list, and searching the likely candidate list first, we should find the sentence with a higher probability of correctness quicker than without grouping.

watashi-wa	hon-o	yomimasu
watashi-wa (326)	hon-o (318)	sumimasu(282)
watashi-o (366)	hon-mo (320)	yomimasu (303)
watashi-mo (405)	hon-no (351)	yobimasu (317)
watashi-ni (406)	hon-to (398)	yomemasu (362)
akai (414)	mono-mo(404)	yomu(370)

(numbers are distance scores)

Fig. 2- Bunsetsu Lattice

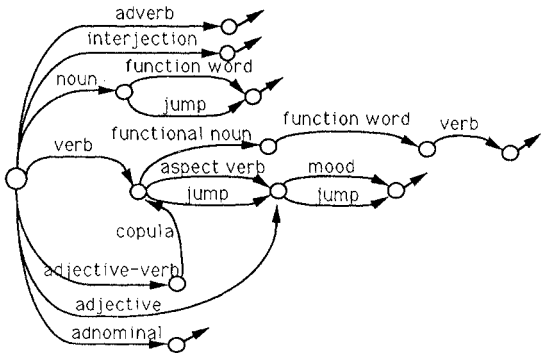

Fig. 3- Bunsetsu

After completing the grouping, the linguistic processor computes a priority index for each group from

$$PI_{hij} = (\,|\,MSG_i - MSG_j\,|\,/Average_h) \times e^{n+R}$$

where PI_{hij} is a priority index, MSG_i is the distance score for the last bunsetsu of group i, MSG_j is the distance score for the first bunsetsu of group i + 1, and R is a group number, n is an integer, h is a bunsetsu sequence number in a sentence, $Average_h$ is the average distance score difference for bunsetsu sequence number h. The linguistic processor generates sentences by searching a group with the smallest priority index first. The sentence search process is conducted in two phases: inter-group best-first-search, and intra-group depth-first-search (Figure 4). The generated sentences are tested for their grammatical correctness against 110 grammar rules (Figure 5) expressed in the Definite Clause Grammar formalism (DCG) [5]. We compiled these grammar rules for simple Japanese sentences using Junior high school textbooks.

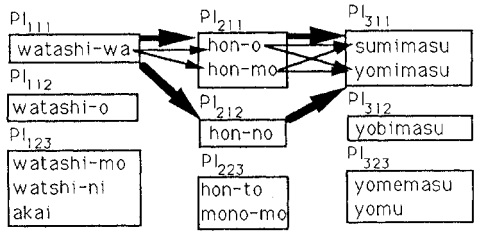

Fig. 4 - Sentence Search

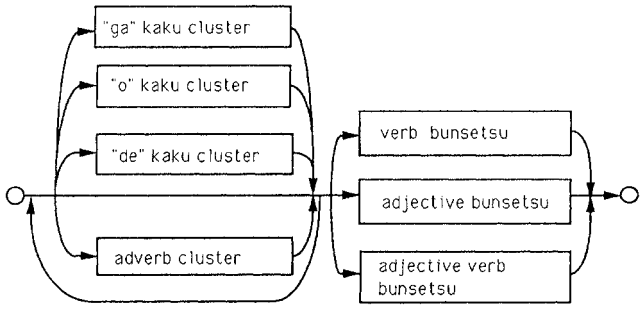

Fig. 5- Grammar

The parser parses a sentence in bottom-up fashion [6] while generating multiple phrase structures through backtracking. The resulting phrase structure is, then, used to inspect the case agreement between bunsetsu phrases. For example, in a sentence "watashi-wa hon-o yomimasu", the bunsetsu "yomimasu" contains a main verb "yomu" that subcategorizes bunsetsu phrases with particles "wa" and "o" in Japanese. The bunsetsu "yomimasu", therefore, can subcategorize the bunsetsu "watashi-wa" and "hon-o." The case analysis is conducted based on a unification grammar (Lexical Functional Grammar [7]) due to its transparent manipulation of syntactic information (see [1] for detail).

3. PERFORMANCE EVALUATION

We evaluated the performance of the linguistic processor by observing how fast it could recover a spoken sentence from a bunsetsu lattice. To evaluate the performance, we selected 4000 bunsetsu phrases from the same Junior high school textbooks used to compile the grammar, and composed 100 test sentences by using the selected bunsetsu phrases (average sentence length = 3.3 bunsetsu phrases).

A male speaker from the Tokyo area spoke the test sentences in a quiet studio. The spoken sentences were recorded and later added with different levels of ambient noise (S/N= 10dB, 20dB, 25dB, 30dB, ∞ , S/N=$10\log_{10}$(average power of speech / average power of noise)). We trained the acoustic processor with 200 VCV (Vowel Consonant Vowel) balanced bunsetsu phrases (without any noise) selected from the 4000 bunsetsu phrases. The test sentences with added noise were presented to the acoustic processor to generate bunsetsu lattices. These bunsetsu lattices were, then, supplied to the linguistic processor to evaluate its performance.

For a spoken bunsetsu, the acoustic processor generates as many candidates as the size of the vocabulary. We used only the top 10 recognition candidates to save the computation time since the average recognition accuracy for 2771 bunsetsu phrases showed that recognition accuracy does not improve much after the 10th candidates for all noise levels (Figure 6).

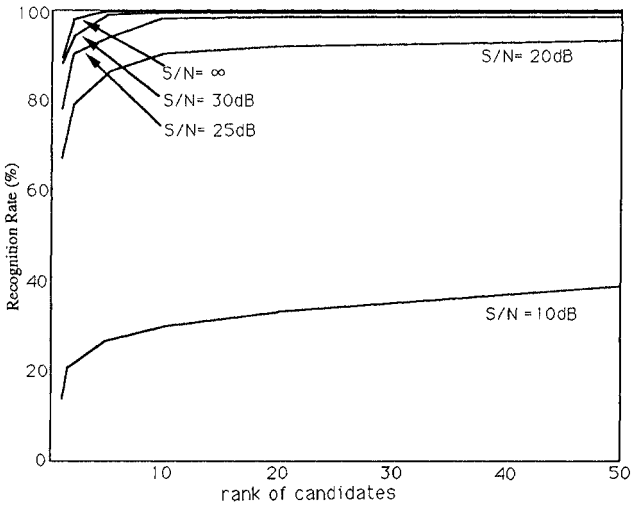

Fig. 6- Rank of Candidates and Recognition Accuracy

4. RESULTS

A) Performance of the acoustic processor

Table 1 illustrates the performance of the acoustic processor. When we did not add noise (S/N = ∞), the acoustic processor recognized 88% of the bunsetsu phrases as the first candidates, resulting in a 66% sentence recognition. In this case, 99% of the bunsetsu phrases were recognized as one of the top 10 candidates. When we added noise (S/N = 20dB), the acoustic processor recognized 62% of the bunsetsu phrases resulting in a 22% sentence recognition. In this case, over 90% of the bunsetsu phrases were still recognized as one of the top 10 candidates. However, when we increased noise (S/N = 10dB), the acoustic processor recognized only 15% of the

bunsetsu phrases as the first candidates, resulting in a 5% sentence recognition. In this case, only 40% of the bunsetsu phrases were recognized as one of the top 50 candidates (Figure 7).

B) Performance of the linguistic processor

Table 2 summarizes the performance of the linguistic processor. The evaluation was conducted by counting the first 20 candidate sentences, and checking the ranks of the spoken sentences within the candidate sentence list. For an S/N of ∞, the linguistic processor improved the sentence recognition rate by 30% over the acoustic processor reaching 96%. However, when we increased noise (S/N = 20dB), although the linguistic processor improved the sentence recognition rate by 32 %, the total sentence recognition rate did not exceed 50%. The results show that when noise level reached to an S/N of 10dB, the linguistic processor was ineffective in recovering input sentences (Figure 7).

Table 1- Performance of the Acoustic Processor

(%)

S/N	10dB	20dB	25dB	30dB	∞
Bunsetsu	15	62	77	87	88
Sentence	5	22	49	64	66

Table 2- Performance of the Linguistic Processor

(unit: sentence)

Rank (R)	10dB	20dB	25dB	30dB	∞
R ≦ 5	0	26	25	26	28
5 < R ≦ 10	0	6	6	2	1
10 < R ≦ 20	0	0	2	1	1
20 < R	1	16	12	7	4
unrecoverable	94	30	6	0	0

5. DISCUSSION

The results of this study show that the linguistic processor performed well while the noise was relatively low (S/N ratio = ∞, 30dB). In this case, the linguistic processor recovered over 90% of the sentences as one of the first 5 candidate sentences. The linguistic processor, however, performed poorly when the noise level reached to an S/N of 20dB. One way to improve the performance is to extend the number of recognition candidates to 20. This, of course, will require longer computation time, and result in more grammatically well-formed candidate sentences. For example, when a sentence "watashiwa nemurimasen" (I am not sleeping) was spoken, the following sentences were recovered as grammatically correct candidate sentences.

1. "garasuwa nemurimasu."
2. "garasuwa nemurinasai."
3. "garasuwa nemurimasen."
4. "garasuga nemurimasu."
5. "watashiwa nemurimasu."
6. "garasuga nemurinasai."
7. "garasuga nemurimasen."
8. "watashiwa nemurimasen."

Although the linguistic processor accepted all the sentences as grammatically correct, only sentence 5 and 8 make any sense, since "garasu" (glass) cannot "nemuru" (sleep). In this case, employing surface level semantic information may help to reject syntactically well-formed but meaningless

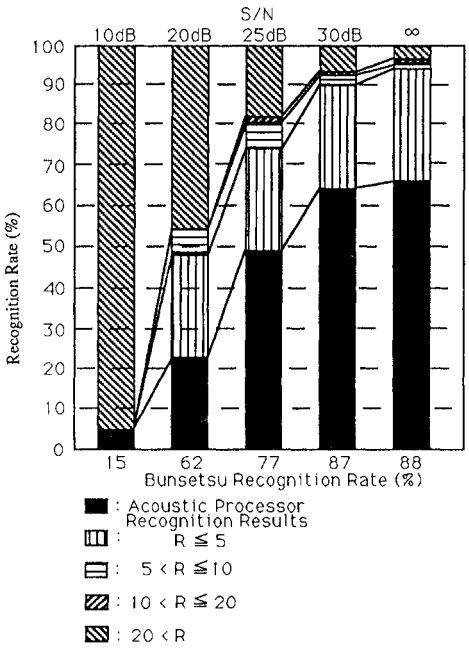

Fig. 7- Performance of the Japanese Speech Recognition System

candidate sentences, hence improving the overall performance.

6. CONCLUSION

In this study, the authors presented the performance of a large vocabulary spoken sentence recognition system. The system performed well in a quiet environment, but performed poorly when noise level reached to an S/N of 20dB. To improve performance, we are working on: 1) a noise adaptation procedure for the acoustic processor to improve recognition accuracy in a noisy environment, and 2) implementing semantic information for the linguistic processor to reject syntactically correct but meaningless sentences.

REFERENCES

[1] Loken-Kim, K.H., et al., A Large Vocabulary Japanese Speech Recognition System, Speech Technology, APRIL/MAY 1990

[2] Kimura, S., 100,000 Word Recognition using Acoustic Segment Networks, ICASSP'90, CH2847-2/90/0000-0061

[3] Loken-Kim, K.H., et al., A Postprocessor for a Large Vocabulary Japanese Speech Recognition System, EUROSPEECH'89, Vol. 2, 1-4

[4] Kido, K., Phoneme Recognition and Knowledge Based Japanese Continuous Speech Recognition (in Japanese), Research Report #59420031, Tohoku University, Japan, 1987

[5] Pereira, F., et al., Definite Clause Grammars for Language Analysis - A Survey of the Formalism and a Comparison with Augmented Transition Networks , Artificial Intelligence 13, pp. 231-278, North Holland Publishing Company, 1980

[6] Niimi, Y., et al., A Method for Word Prediction in a Speech Understanding System Based Upon Bottom-up Parsing, The Acous. Soc. of Japan, S84-84, 1985

[7] Bresnan, J., The Mental Representation of Grammatical Relations, MIT Press, 1983