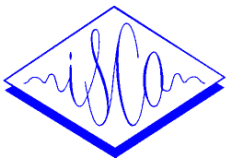

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

ENHANCEMENT OF ATR'S SPOKEN LANGUAGE TRANSLATION SYSTEM: SL-TRANS2

Tsuyoshi Morimoto, Toshiyuki Takezawa, Kazumi Ohkura, Masaaki Nagata, Fumihiro Yato, Shigeki Sagayama, Akira Kurematsu

ATR Interpreting Telephony Research Laboratories
Seika-cho, Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

This paper reports an overview of the recent enhancement of ATR's spoken language translation system that can translate Japanese speech to English. First, a speaker adaptation technique is introduced in the speech recognition module so that the system can be available for many users. Second, a phrase category predictor that uses inter-phrase context-free-grammar rules is adopted in the speech recognition module instead of a dependency analysis so that the system can deal with a large vocabulary size of nearly 1,000 words while keeping high recognition accuracy. Third, a new interface between the speech recognizer and the spoken language translator is proposed. Finally, the results of experiments are reported and current performance and future direction of our study are discussed.

10.21437/ICSLP.1992-82

1 Introduction

SL-TRANS is an experimental spoken language translation system that can translate Japanese speech to English. The original *SL-TRANS* system reported in [1] accepts spoken Japanese sentences uttered phrase-by-phrase by only one specific speaker who is a professional announcer. It then translates them into English utterances. Its vocabulary was quite small, less than 300 words. This paper reports an overview of the recent enhancement of *SL-TRANS*, now called *SL-TRANS2*.

First, we introduced a speaker adaptation technique in the speech recognition module so that the system is now available for many users. Second, we introduced a phrase category predictor that uses inter-phrase context-free-grammar (CFG) rules in the speech recognition module instead of a dependency analysis so that the system can deal with a large vocabulary of nearly 1,000 words while keeping high recognition accuracy. Third, we proposed a new interface between a speech recognizer and a spoken language translator. Finally, we reported the results of various experiments. This paper concludes with a discussion of current performance and future direction of our study.

2 System Overview

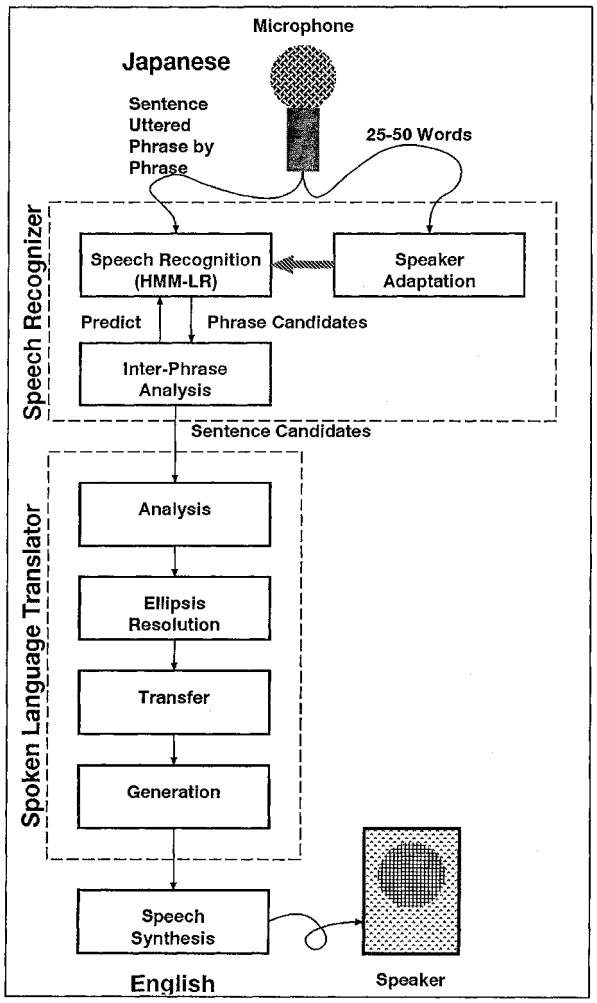

Figure 1: Construction of *SL-TRANS2*

Figure 1 shows the construction of the *SL-TRANS2* system. In the speech recognition module, we introduced a speaker adaptation technique and a phrase category predictor that uses inter-phrase CFG rules instead of dependency analysis. We also revised the spoken language translator so that it can deal with a large vocabulary size of nearly 1,000 words.

The speech recognizer now runs on either a workstation such as *HP9000/750* or special hardware developed by ATR and Mitsubishi Electric Corporation [2]. The spoken language translator now runs on a workstation such as *SPARC2* or *HP9000/750*. We use a commercial English speech synthesizer *(DECtalk)* as the output device.

2.1 Speech Recognizer

We introduced a speaker adaptation technique. A mapping table for HMM phoneme models from a speaker to the standard speaker is made through an adaptation process [3, 4]. A fuzzy VQ mapping technique is used and satisfactory performance with continuous speech is obtained even with a small number of training words (25-50 words).

The speech recognizer accepts a sentence uttered phrase-by-phrase. The phrase level speech recognizer uses intra-phrase CFG rules for the prediction of the next possible phonemes, and verifies their existence by using HMM phoneme models [4]. Because each phrase was recognized independently, there were usually many ungrammatical sentence candidates. To eliminate such erroneous sentence candidates, a new phrase predictor has been introduced [5]. It predicts next possible phrase categories by using inter-phrase CFG rules [6] and drives the phrase level speech recognizer. Then, all recognized outputs are syntactically well-formed even at the sentence level.

2.2 Spoken Language Translator

We are also revising our spoken language translator [7] in order to deal with a large vocabulary size of about 1,500 words, which is enough to handle basic conversation in one specific domain.

Proper translation of an intention indicated in an utterance, besides propositional content, is very important for a high-quality dialogue translation system. In our system, an input utterance is analyzed by a unification-based parser. The outputs from the speech recognizer are sentence candidates, which are syntactically well-formed but not always semantically or pragmatically well-formed. The analysis module checks the validity of each sentence candidate, selects the most plausible one according to the speech recognition score, and generates a feature structure.

After analysis, ellipsis resolution is done using the pragmatics in Japanese honorific expressions. Then, a semantic representation is generated in the feature structure form. This feature structure is generally composed of two parts: intentional content and propositional content. The former indicates the speaker's intention and is expressed in terms of language-independent concepts. The latter is expressed in terms of language-dependent concepts. In the next transfer stage, only the propositional content is transferred to a target language concept. At the generation stage, it is combined with the intentional content and a final expression in a target language is generated. Finally, synthesized English speech is output from a speech synthesizer.

2.3 Interface between Speech Recognizer and Spoken Language Translator

In general, relatively long morphemes tend to be employed in a language analyzer because its primary objective is to extract the meaning of an input utterance. However, in speech recognition, the most significant concern is increasing recognition accuracy. For that purpose, syntactic rules applied to speech recognition tend to consist of relatively short morphemes.

For example, the sentence-final expression *"nakutewanarimasen"*, whose meaning is "have to do", is regarded as one auxiliary verb in the grammar for parsing. However, in the grammar for speech recognition, it is regarded as a sequence of the six morphemes *"naku te wa nari mase n"*. Here is another example:

Jūsho-wa Ōsaka-shi Higashi-ku Chayamachi 23desu.
My address is 23 Chayamachi, Higashi-ku, Osaka.

In the grammar for speech recognition, this sentence is assumed to be composed of five phrases because this sentence is usually uttered in separate parts by Japanese native speakers. However, in the grammar for parsing, the string *"Ōsaka-shi Higashi-ku Chayamachi 23"* is regarded as just one compound noun.

Therefore, there is always mismatching in the lexical entries of speech recognition and language processing.

The morpheme adjuster [8] proposed in our system, which is located between the speech recognizer and the language analyzer, adjusts word boundaries and attaches an appropriate word category name to each one. In other words, the speech recognizer, together with the morpheme adjuster, works as a front end morphological analyzer for a language analyzer. With this mechanism, not only can duplicated morphological analysis at the language analyzer be eliminated, but also unknown words, such as proper nouns in an input utterance, can be found and treated effectively by the back end language processor.

Figure 2 shows the construction of the proposed interface between the speech recognizer and the unification-based parser. It consists of three parts: a *kana-kanji* converter, a morpheme adjuster and a feature structure generator. The *kana-kanji* converter accepts phrase or sentence candidates, which are the outputs from a speech recognizer, and outputs corresponding *kana-kanji* strings. This part is necessary for Japanese speech recognition because Japanese people are familiar with *kana-kanji* strings for reading and writing. The morpheme adjuster accepts a *kana-kanji* string with the morpheme information defined in the grammar for speech recognition. It then outputs a *kana-kanji* string with the morpheme

information defined in the grammar for parsing. The feature structure generator adds the corresponding feature structure, which is necessary for a unification-based parser, to each morpheme.

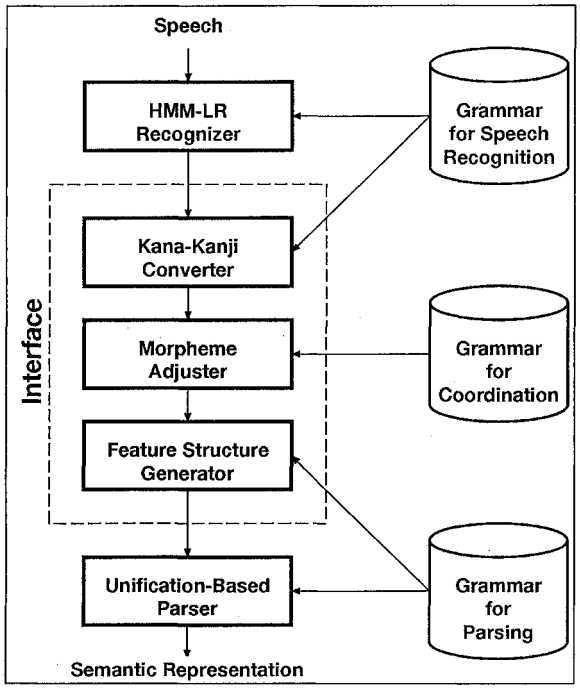

Figure 2: Interface between the Speech Recognizer and the Unification-Based Parser

3 System Performance and Experiment

Our system can accept conversational utterances in a domain concerning the secretarial service of an international conference. Table 1 shows examples of Japanese input and English output.

As for processing speed, it takes almost real-time to recognize one sentence on the special hardware. It takes about 10 or 20 seconds to translate one sentence on a *SPARC2* workstation. It is enough for a demonstration and experiments.

Table 1: Examples

Japanese Input	English Output
そちらは／会議事務局ですか。	Is this the conference office?
会議に／申し込みたいのですが。	I would like to apply for the conference.
登録用紙は／既に／お持ちでしょうか。	Do you already have a registration form?
それでは／登録用紙を／お送りいたします。	Then I will send you a registration form.

3.1 Analyzed Data

Table 2 shows the analyzed data. We used 137 Japanese sentences in a goal-directed dialogue concerning the secretarial service of an international conference. Utterances were made phrase-by-phrase, and thus recognition was performed for each phrase independently.

Table 2: Analyzed Data

Total number of sentences	137
Total number of phrases	353
Average number of phrases per sentence	2.6
Maximum number of phrases per sentence	8

The data is sampled at 12 kHz, pre-emphasized by $(1 - 0.97 \times z^{-1})$, and windowed using a 256-point Hamming window every 9 msec. Then, 12-order LPC analysis is carried out. A multiple VQ codebook for each feature was generated using 216 phonetically balanced words. Hard vector quantization without fuzzy VQ was performed for HMM training. Fuzzy vector quantization ($fuzziness = 1.6$) was used for test data.

3.2 Condition of Experiment and Result

Currently, about 700 Japanese words are defined and implemented in both the speech recognition and the spoken language translation subsystems. Table 3 shows speech recognition and translation rates. This experiment was done after speaker adaptation with 50 Japanese words uttered by each speaker. The maximum amount of whole beam width, the global beam width, is set at 250 and the maximum beam width of each branch, the local beam width, at 32 for this experiment.

Figure 3 shows the relationship between utterance speed and speech recognition. The highest recognition accuracy is obtained by a male speaker whose utterance speed is 6.3 mora/sec.

Table 3: Speech Recognition and Translation Rates

Speaker	Phrase Accuracy (%)			Sentence Accuracy (%)			Translation Accuracy (%)		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
Speaker (FAK)	85.0	94.1	96.3	68.6	79.6	83.2	69.3	74.5	74.5
Speaker (FEO)	79.3	89.0	92.1	65.7	74.5	77.4	66.4	73.0	75.2
Speaker (MIK)	93.5	97.5	98.0	86.1	94.2	94.2	80.3	88.3	88.3
Total (Average)	85.9	93.5	95.5	73.5	82.8	84.9	72.0	78.6	79.3

3.3 Discussion

We can find large differences in Figure 3. The highest recognition accuracy is obtained by a male speaker whose utterance speed is 6.3 mora/sec. This is the nearest to that of the standard speaker's 216 phonetically balanced words. This result suggests the vulnerability of speech recognition to utterance speed.

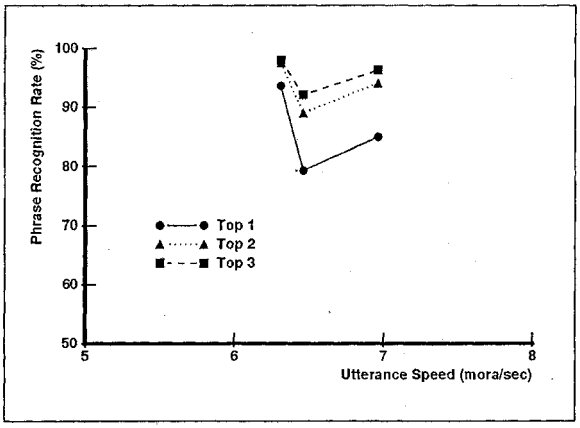

Figure 3: Relationship between Utterance Speed and Speech Recognition Rate

4 Conclusion

We have reported an overview of the recent enhancement of our spoken language translation system that can translate Japanese speech to English. First, we introduced a speaker adaptation technique in the speech recognition module. Second, we introduced a phrase category predictor that uses inter-phrase CFG rules in the speech recognition module instead of a dependency analysis. Third, we proposed a new interface between the speech recognizer and the spoken language translator. Finally, we also reported the results of experiments. These results suggest the vulnerability of speech recognition to utterance speed.

We are now trying to extend the vocabulary up to about 1,500 Japanese words in both speech recognition and spoken language translation subsystems. We are also studying automatic speech translation among Japanese, English and German.

Acknowledgments

The authors wish to thank Mr. J. Takami and Mr. H. Hattori for their help in transporting the speaker adaptation module. The authors are also grateful to all of the members of the laboratories for their constant help and encouragement.

References

[1] Morimoto, T., Shikano, K., Iida, H. and Kurematsu, A.: "Integration of Speech Recognition and Language Processing in Spoken Language Translation System *(SL-TRANS)*", *Proc. of ICSLP 90,* pp. 921-924 (1990).

[2] Nagai, A. et al.: "Hardware Implementation of Realtime 1000-Word HMM-LR Continuous Speech Recognition", *Proc. of ICSLP 92* (1992).

[3] Nakamura, S. and Shikano, K.: "Speaker Adaptation Applied to HMM and Neural Networks", *Proc. of ICASSP 89,* pp. 89-92 (1989).

[4] Hanazawa, T., Kita, K., Nakamura, S., Kawabara, T. and Shikano, K.: "ATR HMM-LR Continuous Speech Recognition System", *Proc. of ICASSP 90,* pp. 53-56 (1990).

[5] Kita, K., Takezawa, T., Hosaka, J., Ehara, T. and Morimoto, T.: "Continuous Speech Recognition Using Two-Level LR Parsing", *Proc. of ICSLP 90,* pp. 905-908 (1990).

[6] Hosaka, J. and Takezawa, T.: "Construction of Corpus-Based Syntactic Rules for Accurate Speech Recognition", *Proc. of COLING 92* (1992).

[7] Morimoto, T., Suzuki, M., Takezawa, T., Kikui, G., Nagata, M. and Tomokiyo, M.: "A Spoken Language Translation System: *SL-TRANS2*", *Proc. of COLING 92* (1992).

[8] Nagata, M., Takezawa, T. and Morimoto, T.: "A Loosely-Coupled Hierarchical Interface between Speech Recognition and Natural Language Processing", *Proc. of the 43rd Meeting of the Information Processing Society of Japan,* Vol. **2**, pp. 549-550 (1991) *(in Japanese).*