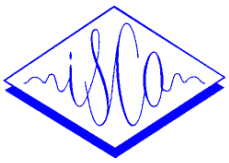


2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

INTERTALKER : An Experimental Automatic Interpretation System Using Conceptual Representation

Kaichiro HATAZAKI, Jun NOGUCHI, Akitoshi OKUMURA,
Kazunaga YOSHIDA, Takao WATANABE

C&C Information Technology Research Laboratories, NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki, 216 JAPAN

10.21437/ICSLP.1992-81

ABSTRACT

This paper presents an experimental automatic interpretation system named INTERTALKER, which recognizes speaker-independent naturally spoken Japanese and English, translates between the two languages, and converts the result into spoken output. In addition, the system also translates the input into French and Spanish at the same time. The system is composed of speech recognition, multi-lingual translation using the pivot method, and rule-based speech synthesis using the pitch controlled residual wave excitation method. Speaker-independent continuous speech recognition is achieved by using demi-syllable speech units. Speech recognition and language translation are tightly integrated using a conceptual representation, a language independent expression of a sentence. The system is robust in its ability to compensate for possible errors in speech recognition, as well as to be tolerant of the grammatical incorrectness and ambiguities in spoken language. Also, because necessary keywords in the network are directly mapped into a conceptual representation, variations in input sentences can easily be accommodated by adding them to the grammar network. The sentence recognition accuracy for the ticket reservation task of 500 word vocabulary was 83%, and its translation accuracy was 93%.

1. INTRODUCTION

Need for automatic translation between different languages has been increased, due to the globalization of society. Especially, a speech-to-speech interpretation system is desired, since speech is a natural and convenient communication means among people. Realization of the automatic interpretation system requires advanced technologies in regard to speech recognition and understanding, machine translation, and speech synthesis, as well as their integration method. Recently, there have been several research activities on these technologies, and some automatic interpretation systems have been built[1, 2, 3].

The authors also have researched methods of achieving speaker independent continuous speech recognition, multilingual machine translation, speech synthesis by rule, and their efficient integration. Based on these researches, a new experimental automatic interpretation system, INTERTALKER, was developed.

A language translation method should be considered first to build the interpretation system. Generally speaking, there are two kinds of translation method. One is the transfer method, which converts the source language structure to the target language structure directly. The other is the pivot method, or the interlingua method, which converts the source language into an intermediate language, an interlingua, and then generates the target language from the interlingua. The pivot method has an advantage in that the system can be efficiently extended to a multi-lingual system by adding the other languages independently, since the interlingua is independent from any languages. INTERTALKER uses the pivot method, because its ultimate goal is a multi-lingual interpretation system.

The next important problem to be solved is the integration method for speech recognition and linguistic processing. This is because of the difficulty in obtaining the correct meaning of an input utterance by parsing the text of the recognition result, which may contain some recognition errors. One of the integration methods studied recently is the N-best architecture[4, 3], where a speech recognizer produces a list of sentence hypotheses, each of which is parsed by the linguistic decoder in turn, in order to find the correct one. The other method is tight integration of speech recognition and linguistic processing, where both syntax and semantic knowledge are merged into a recognition grammar, by which the recognition is controlled in a top-down manner[5, 2]. The latter requires a lot of computation cost, but can achieve better speech recognition and understanding result.

The system presented in this paper achieves tight integration of speech recognition and language translation, using a conceptual representation, or interlingua. By incorporating semantic and task knowledge into a recognition grammar, the speech recognition part outputs the conceptual representation as the recognition result, which expresses a meaning for the input speech. Then, the translation part generates a sentence in the target language from the conceptual representation, directly using the pivot method described above.

This article first presents an overview of the INTERTALKER system. Next, it describes the speech recognition and understanding method used to output the conceptual representation, and then the translation method from the conceptual representation to the target language. Finally, experiment results obtained from automatic interpretation are shown.

2. SYSTEM OVERVIEW

This system recognizes speaker-independent naturally spoken Japanese and English, translates between the two languages, and converts the result into spoken output. In addition, the system also translates the input into French and Spanish at the same time. Figure 1 shows the system configuration. The system consists of work stations, speech recognition hardware and speech synthesis hardware, which are connected through a local area network.

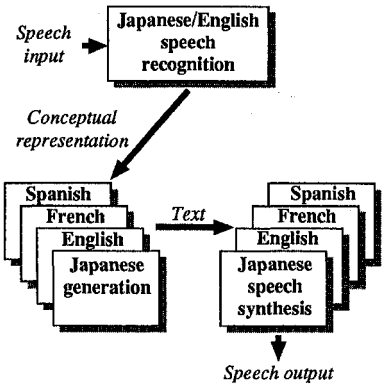


Figure 1: Automatic Interpretation System

Speech Recognition

The system operates in three stages. First, the input utterance is recognized, then a conceptual representation is obtained as the recognition result. This system accomplishes speaker-independent continuous speech recognition, controlled by a finite state network grammar. The Japanese speech recognition uses demi-syllable speech units[6]. The demi-syllable unit can efficiently treat contextual variations caused by the co-articulation effect, since it includes transitional part information of speech. There are 241 demi-syllable units, which are modeled by Gaussian mixture density HMMs. Speaker-independent HMMs are trained on task-independent training data: 100 talkers each speaking 250 phonetically balanced word utterances. The English recognition process is the same as the Japanese, except that 354 diphone units are used. A finite state automata grammar network and demi-syllable HMMs are compiled into a single network, where phonological variations and word juncture models are automatically expanded.

The finite state network is searched in order to locate a best path for the input utterance. Then, a conceptual representation is constructed as a recognition result, based on semantic and task knowledge incorporated into the network grammar. Bundle search [7, 8], a fast frame-synchronous search algorithm, is used for the network search to reduce the computational cost. Occurrences of the same word that appear in many different places in the network are "bundled" together, reducing the many calculations required for each occurrence by those required for just one occurrence. Recognition hardware has been built to achieve a real-time response, using parallel processing techniques[8].

Translation

In the next stage, a sentence is generated in each target language from the conceptual representation, directly using the multi-lingual translation method[11]. This is known as the pivot method. First, a conceptual representation is transformed to a syntactically dependent structure, using the target language dictionary. Then, a syntactic tree is generated, and a morphological processor generates the surface expression of the sentence.

Speech Synthesis

Finally, the text is converted to speech using a rule-based speech synthesizer. A Japanese text-to-speech conversion system with high intelligibility and natural sounding speech has been developed [9]. First, accentuation and pause insertion rules are applied to a sentence. Then, after the prosodic information has been created using prosody control rules, synthesized speech is generated using a pitch controlled residual waveform excitation technique. Real time response is realized by using application specific hardware in the form of a personal computer board. For English, French and Spanish speech synthesis, a commercial text-to-speech synthesizer is used.

System Features

System features are summarized as follows.

- Speaker-independent continuous speech recognition is accomplished, using demi-syllable speech units.
- Language translation is based on the pivot method, or an interlingua method.
- Speech recognition, understanding and language translation are tightly integrated, by using a conceptual representation, or interlingua.
- A new Japanese speech synthesis system is used, significantly improving the clarity and intelligibility of any required sentence.
- The real time response is achieved by means of the speech recognition hardware.

3. INTEGRATION USING CONCEPTUAL REPRESENTATION

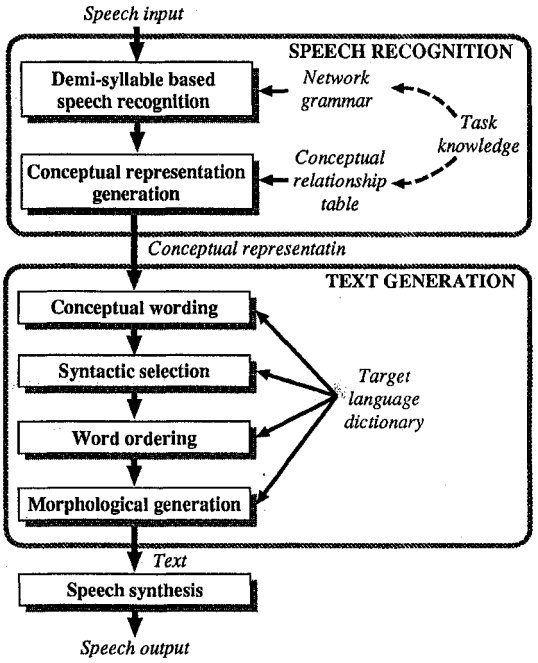
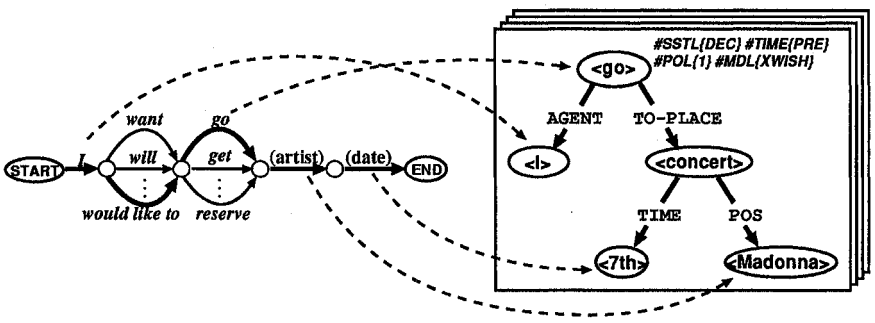


Figure 2: Integration using Conceptual representation

Tight integration of speech recognition part and language translation part is desirable, to achieve correct interpretation in spite of imperfections in speech recognition. It is hard to avoid recognition errors, particularly for short words like particles or for words at the end of a sentence, which are often uttered unclearly. Furthermore, application task knowledge is required to obtain a correct meaning for the input utterance, despite the ambiguity and grammatical incorrectness in spoken language.



(a) Network grammar example *(b) Conceptual representation example*

Figure 3: Conceptual representation generation

In this system, semantic and task knowledge for an objective application is incorporated into the network grammar for speech recognition. This makes it possible to determine the correct meaning for the input utterance, without expending much computation time. In addition, the system becomes robust in its ability to compensate for possible errors in speech recognition, as well as to be tolerant of the grammatical incorrectness and ambiguities in spoken language, because necessary keywords in the network grammar for recognition are directly mapped into a conceptual representation. Furthermore, the result is that variations in input sentences can easily be accommodated, by adding them to the grammar network. Figure 2 shows the process involved in speech recognition and language translation.

Network Grammar

The finite state network grammar explicitly describes various expressions for input sentences. The expressions include various verbs, auxiliary verbs, particles, dummy word utterances, and word order variations. An example of the network is shown in Figure 3(a). The network consists of arcs and nodes. Arcs express words, subnetwork, or just null transition between the nodes.

The network corresponds various kinds of conceptual representations. To construct a conceptual representation from the input utterance, the conceptual primitives for key words in the network, and in addition, their semantic dependency relationship, are described in a conceptual relationship table. These key words are those which are necessary to determine the meaning of the input utterance, namely to build an appropriate conceptual representation. In other words, the conceptual relationship table has direct mapping information between nodes in the network grammar and nodes in the conceptual representations. This table also includes pragmatic information regarding the sentences in the network.

Conceptual Representation

The conceptual representation is a language independent expression of a sentence, based on the PIVOT interlingua, which has been developed for text-to-text machine translation system[10]. PIVOT interlingua is a directional acyclic graph, in which all nodes are associated with conceptual primitives for such concepts as location, time, objects, things, aspect, intention, etc. Arcs indicate semantic dependency relations between the conceptual primitives. In addition, some pragmatic information, like topic, focus, theme and scope, is attached to the conceptual primitive of a predicate. Figure 3(b) shows its example.

Conceptual Representation Generation

First, the network grammar, into which demi-syllable HMMs are compiled, is searched in order to determine the best path for the input utterance by means of the Bundle search method. The best path obtained is an arc sequence, from the start node to the end node in the network. Next, both conceptual primitives for key words contained in the best path and their semantic dependency relationship are extracted from the conceptual relationship table. Then, the system tries to build conceptual representations, each of which is an acyclic graph, using these conceptual primitives and their relationship. The result is a conceptual representation which is constructed successfully. Finally, pragmatic information is attached to the nodes in the conceptual representation. In this stage, some conceptual primitives, which were not included in the input utterance, may be added to complete the conceptual representation.

Sentence Generation

Sentences in the target languages are generated from the conceptual representation using the PIVOT method[11], as the following step.

Conceptual wording: First, the target sentence structure is determined pragmatically and stylistically. In some cases, the language independent conceptual representation may be transformed into a language-dependent semantic structure, in order to generate simpler target sentences. In addition, a suitable clause or phrase, as well as the subject and predicate, may be selected in this stage.

Syntactic selection: The syntactic information is given to the nodes in the semantic structure. In addition, the morphological information for surface cases, modals, voice and so on are produced. As a result, a syntactic structure is created.

Word ordering: Word-order properties for the syntactic structure are determined.

Morphological generation: The nodes in the syntactic structure are arranged according to the Word-order properties. Then, surface morphemes for each node are generated. Finally, the morphemes are combined into words.

	Input utterance		*System output*
Clerk:	Hello. How can I help you?	→	こんにちわ。ご用件は何ですか?
Customer:	26日のマドンナのコンサートのチケットを予約したいんですが	→	I would like to reserve the Madonna's concert ticket on the twenty sixth.
Clerk:	Where would you like to sit?	→	どの席に座りたいのですか
Customer:	どんな席が余っていますか	→	Which seats are remaining?
Clerk:	We still have tckets for floor and balcony seats.	→	1階席と2階席の切符があります。
Customer:	2階席を2人分欲しいのですが	→	I would like two balcony tickets.
Clerk:	That'll be thirty dollars. Will that be cash?	→	30ドルです。現金ですか?
Customer:	はい、現金でお願いします	→	Yes, cah please.
Customer:	コンサートは何時に開演の予定ですか	→	When does the concert start?
Clerk:	It will start at seven p.m.	→	それは午後7時に始まります。

Figure 4: Interpretation example

4. EXPERIMENT

Two tasks were implemented on the system: a concert ticket reservation task and a tour guide task. The ticket reservation task has a 500 word vocabulary and a word perplexity of 5.5. Figure 4 shows an example regarding input sentences and the system output.

The system was evaluated for the latter task by using 30 sentences by ten talkers (total 300 sentences). The average length of the test sentences is 7.1 words. The demi-syllable based speech recognition achieved good sentence recognition accuracy of 83% and word recognition accuracy of 95.5%.

However, in some cases, misrecognized sentences were interpreted correctly. This is the case when all the key words in a sentence were recognized, then a conceptual representation was obtained, in spite of the recognition errors appearing for unimportant words. This case includes both the situation when the system misrecognizes a grammatical input sentence and the situation when the user inputs a grammatically incorrect sentence, which results in the recognition errors. As a result, the translation accuracy, the percentage of the sentences which are translated correctly, is 93%.

5. CONCLUSION

A realtime experimental automatic interpretation system INTERTALKER, was developed, which integrates demi-syllable based speaker-independent continuous speech recognition, conceptual representation based language translation, and text-to-speech conversion. Tight integration of speech recognition and linguistic processing is desirable to cope with recognition errors. In addition, application task knowledge is required to cope with both the ambiguity and grammatical incorrectness in spoken language. The system proposed here tightly integrates speech recognition and language translation using a conceptual representation, a language independent expression of a sentence. This method makes the system robust in its ability to compensate for possible errors in speech recognition, as well as to compensate for grammatical incorrectness and ambiguities in spoken language. Also, because only certain key words in the input utterance are used to obtain the conceptual representation, variations in input sentences can easily be accommodated by adding them to the grammar network for speech recognition.

ACKNOWLEDGMENTS

The authors thank Mr. Yokio Mitome for his support of speech synthesis part of the system, Mr. Masaki Fujimoto for his contribution to implement of the interpretation system. The authors also wish to thank Mr. Masao Watari for his encouragement, and other members of the Media Technology Research Laboratory for their continuous support.

References

[1] T. Morimoto, K. Shikano, H. Iida, and A. Kurematsu. Integration of speech recognition and language processing in spoken language translation system (SL-TRANS). In *IC-SLP 90*, pages 21.7.1–21.7.4, 1990.

[2] D. B. Roe, F. Pereira, R. W. Sproat, and M. D. Riley. Toward a spoken language translator for restricted-domain context-free languages. In *Eurospeech 90*, pages 1063–1066, 1990.

[3] A. Waibel, A. N. Jain, A. E. McNair, H. Saito, A. G. Hauptmann, and J. Tebelskis. Janus: A speech-to-speech translation system using connectionist and symbolic processing strategies. In *ICASSP 92*, pages 793–796. IEEE, 1991.

[4] Y-L. Chow and R. Schwartz. The n-best algorithm: an efficient and exact procedure for finding the n most likely sentence hypotheses. In *ICASSP 90*, pages 81–84, 1990.

[5] S. E. Levinson and K. L. Shipley. A conversational-mode airline information and reservation system using speech input and output. *The Bell System Tech. J.*, 59(1):119–137, 1980.

[6] K. Yoshida, T. Watanabe, and S. Koga. Large vocabulary word recognition based on demi-syllable hidden Markov model using small amount of training data. In *ICASSP 89*, pages 1–4. IEEE, 1989.

[7] T. Watanabe, K. Hatazaki, and K. Yoshida. A fast continuous speech recognition algorithm based on a bundle search. In *ASJ spring meeting*, pages 125–126, 3 1991, (*in Japanese*).

[8] S. Koga, R. Isotani, S. Tsukada, K. Yoshida, K. Hatazaki, and T. Watanabe. A real-time speaker-independent continuous speech recognition system based on demi-syllable units. In *ICSLP 92*, 1992.

[9] K. Iwawa, Y. Mitome, J. Kametani, M. Akamatsu, S. Tomotake, K. Ozawa, and T. Watanabe. A rule-based speech synthesizer using pitch controlled residual wave excitation method. In *ICSLP 90*, pages 6.6.1–6.6.4. IEEE, 1990.

[10] K. Muraki. Conceptual dependency structure and English sentence generation. In *Proc. WGNLC of IEICE*, pages NLC44–3, 7 1984.

[11] A. Okumura, K. Muraki, and S. Akamine. Multi-lingual sentence generation from the pivot interlingua. In *MT SUMMIT*, pages 67–71, 1991.