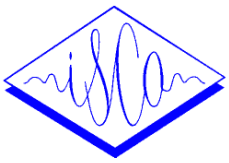

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

A REAL-TIME SPEECH DIALOGUE SYSTEM USING SPONTANEOUS SPEECH UNDERSTANDING

Yoichi Takebayashi[†], *Hiroyuki Tsuboi*[†], *Yoichi Sadamoto*[†],
Hideki Hashimoto[††] *and Hideaki Shinchi*[††]

[†]Toshiba Corporation, Research & Development Center
Saiwai-ku, Kawasaki, 210 Japan
[††]Toshiba Software Engineering Co., Ltd
Ome-shi, Tokyo, 198 Japan

ABSTRACT

We have developed a task-oriented speech dialogue system based on spontaneous speech understanding and response generation (TOSBURG) for unspecified users. The system consists of a noise-robust keyword-spotter, a semantic keyword lattice parser, a user-initiative dialogue manager and a multimodal response generator. After noise immunity keyword-spotting has been performed, the spotted keyword candidates are analyzed by a new keyword lattice parser to extract the semantic content of input speech. Using the dialogue history and situation, the dialogue manager understands input speech based on the semantic contents, and generates a confirmation message to the user about ambiguous points to help overcome difficulties due to the imperfection of speech understanding. The real-time dialogue system has been constructed for a fast food ordering task using two general purpose workstations and four DSP accelerators.

I. INTRODUCTION

The speech dialogue system is different from the dictation system; the aim is not to produce accurate transcription but to enable natural speech communication with a computer. In order to achieve such a speech dialogue system, a robust and real-time spontaneous understanding system should be developed in addition to an intelligent speech response system[1-3].

Several high performance HMM-based speech recognition systems have been developed to deal with speaker-independent large-vocabulary continuous speech[4]. However, their performance significantly decreases with the addition of background noise and spontaneous speech. Most systems deal only with grammatically correct spoken words; therefore they do not work for spontaneous speech in real-world applications. Due to the many possible variations in spontaneous speech, such as unintentional utterances and ellipses, spontaneous speech is difficult to represent by limited lexicon and grammatical rules.

Keyword-based speech recognition systems have been developed for specific applications. A telephone speech recognition and answering system[5] and an information retrieval system[6] were developed based on keyword-spotting. A keyword-based multilingual translation system[7] was also proposed based on isolated word recognition. In contrast with the above systems, we developed a keyword-based continuous speech understanding system.

In order to apply current speech recognition technologies to real-world speech dialogue systems, real-time performance and robustness should be improved. Furthermore, the inevitable ambiguity in the speech understanding process should be handled and dialogue control should be performed to achieve the dialogue goal of the user.

In this paper, our approach to real-world speech dialogue systems is first presented. Next, noise-robust keyword-spotting with Noise-Immunity learning is described. Keyword-based spontaneous speech understanding method is then described[8]. User-initiative dialogue management[9] and multimodal response generation[10] are also given. Finally, the real-time speech dialogue system based on spontaneous speech understanding is given.

II. SPONTANEOUS SPEECH UNDERSTANDING BASED ON KEYWORDS

2.1 Keyword-Based Approach to Speech Understanding

In contrast with the conventional HMM-based continuous speech recognition, we have employed a keyword-based approach to understand spontaneous speech which includes unintentional and unpredictable utterances. Since the keyword-based approach ignores details of utterances, it cannot, for example, describe the complete sentence. However, task-oriented spontaneous speech understanding is possible since the meaning of an utterance can be extracted by combining keyword-spotting with a keyword lattice parsing.

We have proposed a real-time keyword lattice LR parser for realizing a real-time task-oriented spontaneous speech understanding system. While conventional word lattice parsers continuously analyze the whole sentence from the beginning to the end[11], this new parser analyzes only keywords and ignores the rest of the sentence. Thus, the grammar of the new parser is relatively simple. Furthermore, this parser analyzes the keyword sequence semantically and syntactically at the same time.

2.2 Noise Immunity Keyword-Spotter

Performance of the keyword-based speech understanding method highly depends upon keyword-spotting accuracy. We previously developed the Noise Immunity word-spotter for speaker-independent isolated word recognition[12], and extended it to detect keywords in noisy continuous speech[13]. The powerful features of the system are a word-spotter based on the Multiple Similarity(MS) method for reliable keyword detection and the use of Noise Immunity Learning for improving robustness against spontaneous speech and background noise.

The keyword-spotting is performed based on time-frequency word patterns. The MS values are time-continuously

10.21437/ICSLP.1992-215

computed for keyword-spotting through pattern matching of the fixed-dimensional word pattern vectors and reference pattern vectors. During pattern matching between an input vector and reference vectors, which are generated by the noise immunity learning process, an end point candidate (tj) is assumed. A series of start point candidates corresponding to tj are determined based on the maximum and minimum duration for each word class. Fixed-dimensional word pattern vectors are then extracted through uniform sampling of the time series spectrum. Time-continuous calculation of MS values is performed to detect keyword candidates. Figure 1 shows an example of keyword lattice.

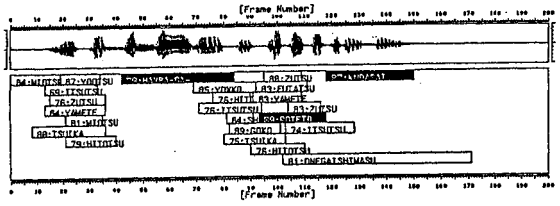

Fig.1 An example of keyword lattice
(Eh... one hamburger and... uhm... one french fries, please.)

2.3 Keyword Lattice Parser

Our new parser analyzes time-discrete keywords and performs keyword-sentence-spotting to obtain multiple sentence candidates of an input utterance. The parser is driven whenever a new keyword candidate is spotted to produce new keyword-sentence candidates. The keyword spotting and parsing processes can be performed in parallel which makes this approach suitable for real-time processing.

Each keyword candidate is assigned a likelihood and start- and end-point frame number. Our new parser deals with isolated keywords to understand spontaneous speech. A time-discrete keyword can be connected with other keywords within a predetermined time range. An example is shown in Fig.2; the connectable range (L) is specific to each keyword and is determined from training data. The sentence candidate table stores temporal sentence candidates during the parsing. The context-free grammar is a set of production rules and semantic processing functions.

The parser utilizes LR parsing tables obtained from the grammar. The parser comprises the following functional components: initial-state processing to check if a keyword event can be an initial keyword and to create a parsing stack, sentence connection processing to check if a current keyword can connect with a subsentence candidate, and accept processing to check if a subsentence candidate can be accepted as a sentence. Figure 3 shows an example of semantic representation of an utterance, which is obtained

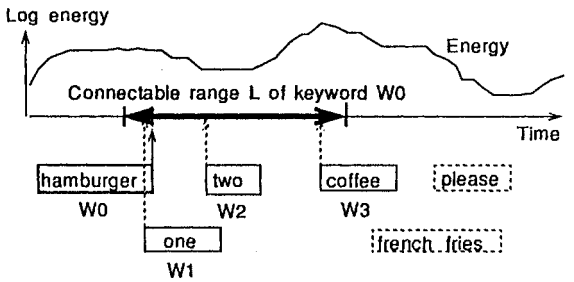

Fig.2 An example of keyword connection check

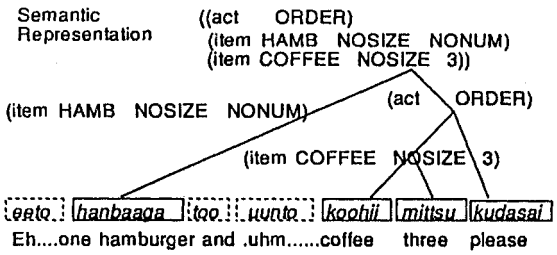

Fig.3 Example of tree structure and semantic representation

by the parsing. During the understanding process, the generated sentence candidates are sorted via their likelihood using a beam search. In addition, these sentence candidates are fed into a dialogue system to determine the content of the input speech using expectation and scoring.

III. USER-INITIATIVE DIALOGUE MANAGEMENT

3.1 Dialogue Model

A fast food ordering task was chosen to evaluate human factors of dialogue for many users, because most people know how to order hamburgers at a hamburger shop in everyday life. In order to achieve natural human-computer interaction, the system should deal with spontaneous speech and minimize restrictions concerning the user's speaking manner[14]. Furthermore, the system should generate adequate responses to the user depending upon the dialogue situation. Many human-computer dialogue systems employ a computer-initiative dialogue model. The user should responds to request messages from the computer and cannot speak spontaneously. By contrast with the conventional computer-initiative human-computer dialogue model, we have represented the dialogue model using user states and system states to enable spontaneous interaction.

The dialogue manager employs as information-condensed semantic representation of both the user's utterance and the system's response. The semantic representations enable concise and efficient dialogue management. The user's speech can be classified into a number of speech acts[15] including order, addition, cancellation, replacement, affirmation, and negation to the confirmation. Similarly, the system response can be classified into several speech acts including confirmation, question and request. Therefore, the input speech and output response are represented using semantic frame representations. The following two sections explain dialogue speech understanding and response generation in user and system states, respectively, using an example shown in Fig.4.

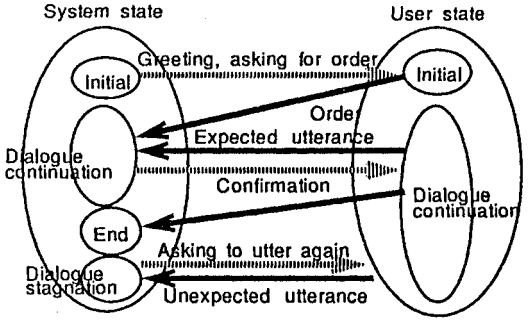

Fig.4 State-transition in dialogue

3.2 Dialogue Speech Understanding in User States

In a user state, the dialogue manager analyzes the semantic utterance representation candidates, obtained from the semantic keyword parser, in order to understand dialogue speech using the dialogue situation and dialogue history.

The dialogue manager first deals with ellipses by compensating values (object, size, number) for the ordered food items, which are not specified in the semantic utterance representation candidate, referring system response. If no suitable values are given in the semantic utterance representation, a default value is used.

The dialogue manager then evaluates semantic utterance representation candidates, obtained from semantic keyword parser, according to the semantic dialogue constraints. For example, when a conflict arises between the content of the system response and the content of a semantic utterance representation candidate, the dialogue manager decreases the score of the candidate. Thus, the dialogue manager rescores the semantic utterance representation candidates, and then selects the candidate which has the highest score as the result of understanding utterance.

Finally, the dialogue state arrives at the system's state according to the result of understanding utterance.

3.3 Semantic Response Generation in System States

In system states, the dialogue manager modifies the dialogue history based on the speech understanding result, and then produces a semantic response representation and transfers to a user state, as shown in Fig.5.

Since the semantic utterance representation might contain ambiguity or errors in keyword spotting, keyword lattice parsing and dialogue speech understanding, the dialogue manager deals with the ambiguity and produces an adequate response to confirm the system understanding result. The dialogue manager outputs the semantic response representation to the response generator.

Moreover, the dialogue manager has likelihood in understanding utterance and emotional information according to dialogue situations and output them to the response generator. The response generator controls the intonation of speech response and the facial expressions, altering them according to the certainty factors and the emotional information[16][17].

IV. MULTIMODAL RESPONSE GENERATION

In order to realize a friendly and efficient interface, both visual and audio media are utilized, as shown in Fig.6. Multimodal response, including synthesized speech, text, animated facial expressions and pictures of ordered food items, are generated from the semantic response representation and system internal state, obtained by the dialogue manager.

Synthesized speech and text are used as natural language output messages such as greeting and confirmation. The contents of speech understanding results such as ordered food items and number are visualized graphically. Animated facial expressions are changed according to dialogue states, and lip movement is synchronized with synthesized speech.

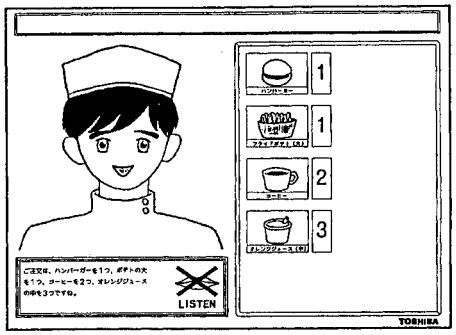

Fig.6 Examples of multimodal response

The emphasis of the synthesized speech is controlled using prosody according to the semantic utterance representation and the system internal state. For example, when the user's order does not include the number of food items and the system supplies a default number, synthesized speech is output with added prosody to emphasize the number.

The speech synthesizer consists of a response sentence generator, a phonological processor, a temporal acoustic parameter generator, and a speech waveform generator. A terminal analog speech synthesizer is employed to synthesize speech waveforms from the segmental features of CV syllable unit. The prosodic control is performed based on the Fijisaki-model[18].

Standard response sentence templates and words are pre-

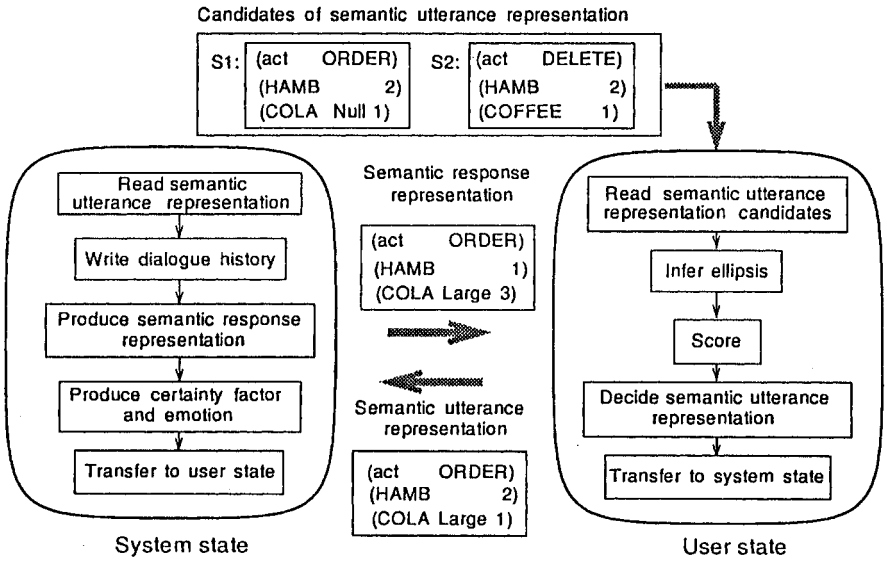

Fig.5 Process of the dialogue manager

pared as a database from which to generate response sentences. The response sentence is converted from a semantic representation to both syntactic boundaries with phonemic symbols and text. The latter is used for visual response text output. The phonological processor generates prosodic symbols and phonetic symbols from the syntactic boundaries with a phonetic symbol representation of the response sentence. The time series of the acoustic parameters are generated by concatenating the selected CV syllable templates. The speech waveform is synthesized from the time series of the acoustic parameters.

V. A REAL-TIME SPEECH DIALOGUE SYSTEM

We have built a task-oriented speech dialogue system based on spontaneous speech understanding and response generation (TOSBURG), based on the above approach as shown in Fig.7. The system consists of the keyword-spotter, the semantic keyword lattice parser, the dialogue manager and the multimodal response generator, as described in the previous sections. While each element of the system has been investigated separately in most cases, we have integrated the above elements and constructed a small task-oriented system, which can interact between human and computer, using 49 Japanese keywords. The system is user-initiated and friendly, since it utilizes a weight-sensitive floor mat to detect the user is presence, and a multimodal response including not only synthesized speech but also visual outputs of text, food items and facial expressions.

The real-time dialogue system has been implemented on two general-purpose workstations(AS4000) and four DSP accelerators(520MFLOPS) on a VME bus[19]. The A/D, D/A converters and the floor mat with pressure sensor are also connected to the VME bus in a UNIX environment.

VI. CONCLUSION

A real-time task-oriented dialogue system based on spontaneous speech understanding has been developed to facilitate natural human-computer interaction for unspecified users. The system deals with everyday spontaneous speech for a fast food ordering task. The system consists of not only spontaneous speech understanding but also user-initiative dialogue management and multimodal response generation. Experimental results obtained by the real-time system have shown the effectiveness and user-friendliness of the system. We are currently increasing reliability of speech understanding and improving robustness of the human-computer dialogue, by collecting real-world dialogue speech data.

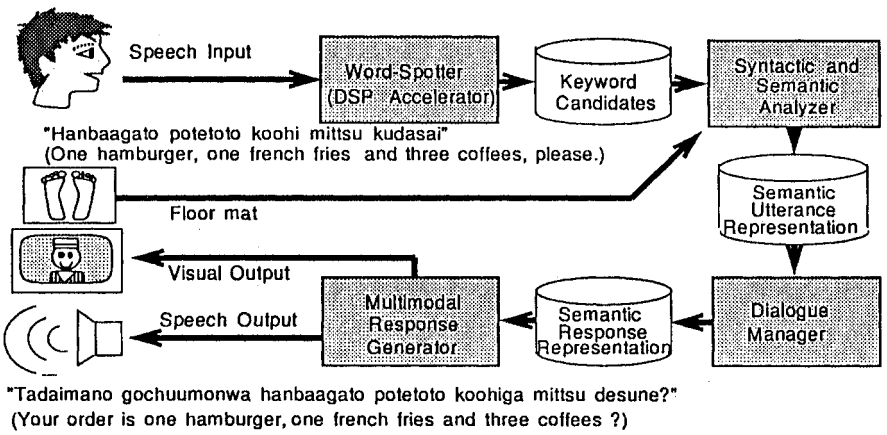

Fig. 7 Configuration of Speech Dialogue System

ACKNOWLEDGEMENTS

We would like to thank H.Kanazawa, Y.Nagata, Y. Yamashita, S.Seto and T.Nii for their contribution to the development of the dialogue system, and to thank Dr.D. Gleaves and Miss.L.Mayfield for their help during the preparation of this paper.

REFERENCES

[1] Ward W. : "Understanding Spontaneous Speech: The Phoenix System", ICASSP '91, pp.365-367 (1991).

[2] De Mori R., et al. : "A Probabilistic Approach to Person-Robot Dialogue", ICASSP '91, pp.797-800 (1991).

[3] Rudnicky A.I., et al.: "Spoken Language Recognition in an office management domain", ICASSP '91, pp.829-832 (1991).

[4] Lee K. : "Automatic Speech Recognition : The Development of the SPHINX System", Kluwer Academic Publishers, Boston, (1989).

[5] Wilpon J.G., et al. : "Improvements and Application for Key Word Recognition Using Hidden Markov Modeling Techniques", ICASSP '91, pp.309-312 (1991).

[6] Rose R.C., et al. : "Techniques for Information Retrieval from Voice Messages", ICASSP '91, pp.317-320 (1991).

[7] Stentiford F.W.M. et al. : "A Speech Driven Language Translation System", EUROSPEECH '87 , pp.418-421 (1987).

[8] Tsuboi H., Hashimoto H. and Takebayashi Y. : "Continuous Speech Understanding Based on Keyword-spotting", IEICE Technical Report, SP91-95 (1991).

[9] Takebayashi Y., Tsuboi H., Sadamoto Y., Hashimoto H. and Shinchi H. : "Speech Dialogue System for Unspecified Users", JSAI Technical Report, SIG-SLUD-9201, pp. 27-36 (1992).

[10] Yamashita Y., Seto S., Hashimoto H. and Takebayashi Y. : "Development of Real-Time Speech Dialogue System TOSBURG (4) Multimodal Response", IPSJ Spring Meet., 6N-8 (1992).

[11] Tomita M. : "An Efficient Word Lattice Parsing Algorithm for Continuous Speech Recognition",ICASSP'86, pp.1569-1572 (1986).

[12] Takebayashi Y., et al. : "A Robust Speech Recognition System using Word-Spotting with Noise Immunity Learning", ICASSP '91, pp.905-908 (1991).

[13] Takebayashi Y., Tsuboi H. and Kanazawa H. : "Keyword-Spotting in Noisy Continuous Speech Using Word Pattern Vector Subabstraction and Noise Immunity Learning", ICASSP '92, pp.II-85~II-88 (1992).

[14] Iida H., et al. : "Natural Language Understanding on a Four-Typed Plan Recognition Model", Trans. IPSJ, 31, 6, pp. 810-821 (1990).

[15] Searle J.R. : "Speech Acts", Cambridge University Press (1969).

[16] Grosz B., et al. : "Attention, intention, and the structure of discourse", Computational Linguistics 12, pp.175-204 (1986).

[17] Allen J.F. : "Recognizing Intention from Natural Language Utterances", Computational Model of Discourse, pp. 107-166 (1983).

[18] Fujisaki H., et al. : "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese", JASJ., 5, 4, pp.233-242 (1984).

[19] Tsuboi H., Kanazawa H. and Takebayashi Y. : "An Accelerator for High-Speed Spoken Word-Spotting and Noise Immunity Learning System", ICSLP '90, pp.273-276 (1990).