STUDIA PHONOLOGICA XV (1981)

A Speech Understanding System with Dialogue Capability

Teruhiko UKITA, Sei-ichi NAKAGAWA and Toshiyuki SAKAI

SUMMARY

This paper describes a speech understanding system which has dialogue capability. Dialogue capability with a user is introduced to confirm user's messages as well as to respond to user's requests.

The system consists of the speech recognizer, the dialogue component, and the response-speech synthesizer. The system works for the task for "information service in a department store." At present the recognizer is set for a single male speaker and has ability whose recognition rate for sentences is 65%. The system can work in on-line mode and respond to a user in a few ten times of real time. A mechanism to deal with simple ellipses is introduced into the system. The dialogue component is capable of dealing with incomplete ability of the recognizer and returning an appropriate speech-response to a user's request.

1. INTRODUCTION

Speech is the most natural and convenient medium for us to communicate with each other, and hence it might be an attractive method for man-machine communication. To use speech as a channel between man and machine, a machine must recognize speech automatically. However, the art of automatic recognition of conversational speech is not enough today, because of the problems such as co-articulation, differences of speakers and so on. This is one of the reasons why the idea of 'speech understanding' was introduced, in which case a task of the system is restricted explicitly and a higher linguistic information can be used.[1] However, the ability of understanding is still imperfect and a score varies widely according to the system composition[2]. In this paper, we extend the line of 'speech understanding' and introduce dialogue capability with a user into a speech understanding system.

Motivations of introducing the dialogue capability are as follows:

1) It is necessary to compensate for the imperfectness of speech recognition.
2) For conversational speech, a speaker does not speak each word so clearly as isolated words, and some words may not be recognized completely.
3) In most applications, transcription of speech into written text is not a final

Teruhiko UKITA (浮田輝彦) Research Student, Department of Information Science, Kyoto University
Sei-ichi NAKAGAWA (中川聖一) Assistant Professor, School of Information Engineering, Toyohashi University of Technology
Toshiyuki SAKAI (坂井利之) Professor, Department of Information Science, Kyoto University

goal but to provide a user with information which he requires is rather the final.

Concerning man-machine communications by speech, few researches have been done. One was performed by Kohda et al.[3], and another by Levinson et al.[4] Though being suggestive, these studies are based on isolated word or phrase recognition and not on a full sentence recognition. Since the recognition of isolated word is an easier task than that of sentence, it may be said that their approach is practical but looses fundamental feature that speech originally has, i.e. convenience and naturality of sentence speech for a speaker.

In this paper, we propose a speech understanding system with a dialogue capability which is able to accept ordinary-spoken sentences and responds to a user by speech. As a first step, we designed the speech recognition part of the system to accept an utterance by a single male speaker, since acceptability of unlimited speakers is one of open and difficult problems for speech recognition research.

In the next section presented a configuration of the system. In Section 3, the system performance will be shown and discussed.

2. System Configuration

2.1 Task Delineation

The task of the system is "information service in a department store", which includes guidance for places of counters and goods and route guidance to the destination counter. The words managed in the speech recognition part are listed in Table 1. We call 'words' not as a meaning of linguistic one but as units to be recognized. The total number of words is 148 and they can be roughly divided into 3 classes; 1) counters of a department store (50 words), 2) goods (60 words), 3) predicates and others (38

Table 1. Words managed in the recognition part of the system.

goods	katsura (wig), kushi (comb), paipu (pipe), ...	[60]
counter names	biyoo-shitsu (beauty shop), keshoo-hin <uriba> (cosmetics) pureigaido (theater ticket agency), ...	[50]
predicates	desu (It is...), imasu (I am at...), nangaidesuka (On what floor...), dokodesuka (Where), chikakudesuka (Is XX near...), dokode-utteimasuka (Where is XX sold?), dokoni-oiteimasuka (Where XX puts?), mitai (I want to see...), kaitai (I want to buy...), douikeba-yoidesuka (How to go), ikitai (I want to go...)	[11]
objects	kita-kaidan (north stairs), kita-erebeetaa (north elevator), chuuoo-erebeetaa (center elevator), chuuoo-esukareetaa (center escalator), minami-esukareetaa (south escalator), minami-kaidan (south stairs)	[6]
floor	chika-nikai (2nd basement), chika-ikkai (1st basement), ikkai (1st floor), ..., hacikai (8th floor)	[10]
post-word	wa, ni, e, o, no	[5]
others	hai (yes), iie (no), chigaimasu (no), wakarimasu (I see.), wakarimasen (I don't know.), uriba (counter)	[6]

[ ] means a number of words for each category.

words).

Sentences which can be uttered by a user are limited by the syntactic rules which are shown in Table 2. Examples of inquiries for the system are, 'Tabako uriba wa dokodesuka. (Where is a tobacco shop?)', 'Omocha uriba e douikeba-yoidesuka. (How to go to the toy counter?)' and so on.

2.2 System Overview

Figure 1 shows an outline of the system. The system consists of three major components; the speech recognizer, the dialogue component, and the response speech synthesizer.

An utterance of the user to the system first processed by the speech recognizer and transcribed into sentences with reliability scores. The speech recognition component has a word dictionary and syntactic rules as sources of knowledge. Sentences are recognized by accumulating syntactically correct word sequences using a parallel search algorithm. Beginning the process, the recognizer receives hypotheses of sentence types, from the dialogue component, which kind of utterances may come next.

The dialogue component investigates scores of the sentences which come from the

Table 2. Syntactic rules.

[UT0]	::=[Bla] | [B2] | [B3] | [B4a] | [B5a] | [B6] | [B8a]	
[UT1]	::=[B1] | [B2] | [B3] | [B4] | [B5] | [B6] | [B7] | [B8]	
[UT2]	::= hai | iie | chigaimasu	
[UT3]	::=[P] desu | [B2]	
[UT4]	::= hai | iie | wakarimasu | wakarimasen	
[B1]	::=[B1a] | [B1b]	
[B1a]	::=[P] wa dokodesuka | [P] wa nangaidesuka	
[B1b]	::= dokodesuka | nangaidesuka	
[B2]	::=[P] ni imasu	
[B3]	::=[P] e ikitai	
[B4]	::=[B4a] | [B4b]	
[B4a]	::=[P] wa [O] no chikakudesuka	
[B4b]	::=[O] no chikakudesuka	
[B5]	::=[B5a] | [B5b]	
[B5a]	::=[G] wa dokode-utteimasuka | [G] wa dokoni-oiteimasuka	
[B5b]	::=dokode-utteimasuka | dokoni-oiteimasuka	
[B6]	::=[G] o mitai | [G] o kaitai	
[B7]	::=[F] no dokodesuka	
[B8]	::=[B8a] | [B8b]	
[B8a]	::=[P] e douikeba-yoidesuka	
[B8b]	::= douikeba-yoidesuka	
[P]	::= biyoo-shitsu | keshoo-hin uriba | tabako uriba ...	(counters)
[G]	::= katsura | kushi | paipu ...	(goods)
[F]	::= chikanikai | chikaikkai | ikkai ...	(floor)
[O]	::= kita-kaidan | kita-erebeetaa ...	(objects)

UT0–UT4 are start symbols. UT0 is for the first time of the dialogue.
UT1–UT4 are for special cases.

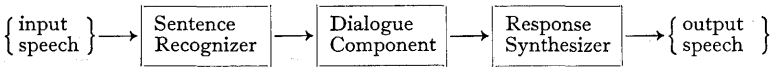

Fig. 1. Outline of the dialogue system.

speech recognition component. If the score of the top result of recognition might be considered as a reliable input, the dialogue component invokes the semantic analyzer which then analyzes and interprets the recognition result and produces an appropriate response. If the recognition results cannot be considered as reliable, the dialogue component tries to confirm the user's utterance through yes-no question.

The response synthesizer, receiving the type of a sentence and words to be generated, from the dialogue component, generates a response by speech. The response message is made by concatenating appropriate words stored in a disk file.

2.3 Recognition Component[5)]

There being many design choices to construct a speech recognizer, we pay attention to a level of matching between a physical and a symbolic world. The phoneme level is a possiblity of this stage, we don't take this course because co-articulation problem prevents us from recognizing phonemes satisfactorilly. On the other hand, matching a word against a simple sequence of acoustic parameters obtained at every sample period seems too redundant and consumes much of processing time. Considering these, we adopt a word level matching based on a sequence of acoustic segments that is a compressed representation of a uniformly sampled time sequence of feature parameters.

The configuration of the recognizer is illustrated in Figure 2. The recognizer consists of 5 modules; each of which will be described in the following.

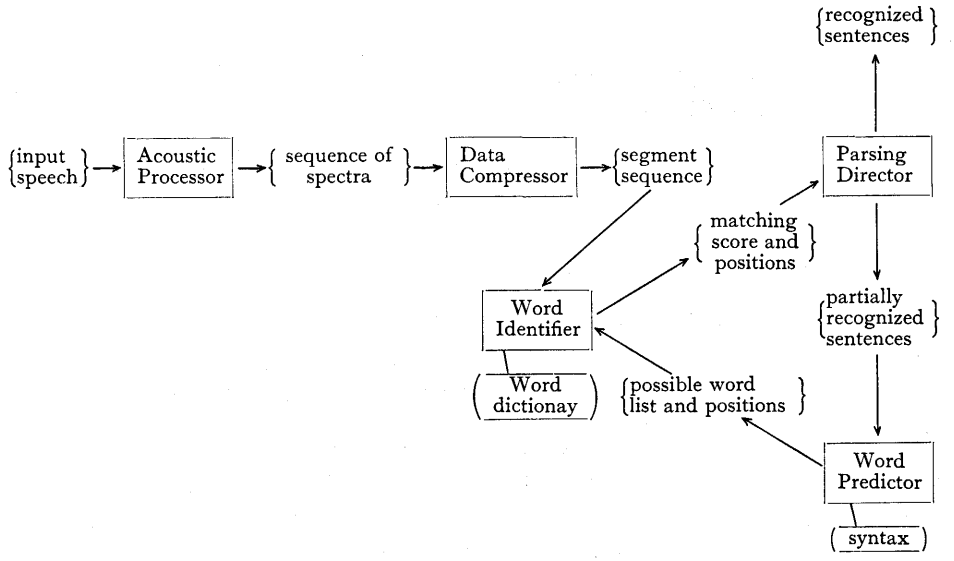

Fig. 2. Configuration of the recognizer.

(a) Acoustic processor

The acoustic processor, mainly hardware processor, consists of 1/4-octave 20-channel filter bank (center frequencies: 210–5660Hz). Speech signal is passed into a pre-emphasis circuit which has a slope of 6dB/oct below 1.6 KHz and then fed into the filter bank. The outputs of the filter bank are rectified, smoothed and sampled at every 10 msec. Hereafter we denote these data obtained at every 10 msec as a "frame." A frame consists of 20 dimensional spectrum vector (each element is denoted by $f_{i,j}$, $j=1-20$) and a root mean of sum of squares of the elements (denoted by AMP_i for the i-th frame).

(b) Data Compressor

The basic function of the data compressor is to cut the input sequence of frames into segments where acoutic parameters can be regarded as (quasi-)homogeneous. First, the processor calculates an auxiliary parameter: a spectrum change (DSP) defined by the sum of absolute differences between spectra of the present and the just previous frame (Equation 1).

$$DSP_i=\left\{\sum_{j=1}^{20}|f_{ij}-f_{i-1,j}|\right\}/AMP_i, \qquad i: \text{ frame number.} \tag{1}$$

Observing this parameter, the processor indicates the segment boundary as the logical OR of the followings: at the dip of a wave of AMP and at the peak of contour of DSP. The processor produces segmented speech chunks which have representative spectrum g_i (normalized by $g_i=f_{i,j}\times 100/AMP_i$, $j=1-20$), AMP_i and a length of the i–th segment. As a result, the compressor gives out many short segments for the transitional part of speech, and a few long segments for the stationary part.

(c) Word Identifier

The word identifier, receiving a list of possible words and range of matching from the word predictor and the parsing director, calculates matching scores for each word against the input sequence of segments. The identifier uses the dynamic programming method and has a word dictionary as a source of knowledge, in which a word is also described as a form of a segment sequence.

As depicted in Figure 3, the procedure of the identifier selects the route that has the best score for each 'grid' point recursively for steps of reference pattern (manually prepared segment sequence), using the following equation.

$$CS(i,j)=\max_{i-W\leq k\leq i}\{CS(k-1,j-1)+L(j)\cdot S(i:k,j)\} \tag{2}$$

where CS (i, j) is a cumulative score accompanied with the best route reached at (i, j) grid point, and W means the maximum number of test segments which can be merged into one reference segment (see permissible routes in Figure 3. W is set to 5 by preliminary experiments); I_B and I_E are positions of input segments controlled by the parsing director and Nd is a number of segments of reference word d. L(j) is a durational time of the j-th reference segment which will be normalized later. S(i: k, j) in Equation 2 is an averaged similarity score between segments and defined as follows:

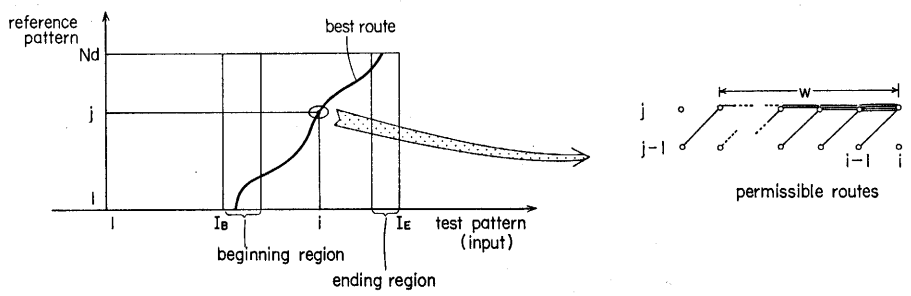

Fig. 3. Matching a test and a reference patterns.

$$S(i:\,k,\,j)=\left\{\sum_{n=k}^{i} l(n)\times S_0(n,\,j)\right\}/\sum_{m=k}^{i} l(m) \tag{3}$$

where $S_0(i,\,j)$ is a similarity score corresponding to the (i, j)-point and calculated as a negative Euclid distance between a test and a reference pattern of 20-dimensional spectrum vectors. And l(i) is a length (durational time) of the i-th test segment. The similarity S_0 (n, j) is calculated by Equation 4.

$$S_0(n,\,j)=100-0.4\times\left\{\sum_{p=1}^{20}(g_{np}-h_{jp})^2\right\}^{1/2} \tag{4}$$

where g_{np}, h_{jp} denote values of the n-th input and the j-th dictionary segment, respectively.

During the route calculation, durational condition is restricted. A segment of reference can basically match to a test segment of shorter duration, because reference segments are prepared to be longer than the test, except a case that a pause of the reference is forbidden to match to a segment whose durational time is less than a half of the reference. On the other hand, a reference segment cannot match to test segments whose sum of duration exceeds a twice of duration of the reference.

For each word, the process calculates the CS until the last reference segment finishes. Then it selects the greatest score, and divide the score by the whole length of the word (sum of L(j)) to obtain a normalized score. The process also searches for beginning and ending edge positions of the best route and gives them out as an output to the parsing director.

(d) Word Predictor

The word predictor receives a string of words (partial sentence maintained by the parsing director) and predicts a list of words which can come next to the string. Observing a word string and its production order in the grammar, which is kept in a 'partial sentence,' the predictor decides words which are permitted to come at right adjacent position of the word string. The syntax rules desrcibed in BNF form are listed in Table 2 and some sentences which can be generated by the rules are shown in Figure 5.

(e) Parsing Director

The parsing director controlls all components in the recognition part of the system and decides what partial sentences will be expanded next. The director first selects

three intermidiate results (='partial sentences') of the higher scores from some of them and passes them to the word predictor. Then the parser invokes the word identifier to calculate matching scores of words hypothesized by the predictor. Receiving the scores, the director then arranges intermidiate results so that they can obtain higher scores. These steps are repeated until three complete sentences are obtained which correspond to all the input segments.

The parsing director maintains the intermidiate results each of which consists of a partially rcoegnized word sequence, including an averaged score, an end point of matching of the last word, and rule numbers and positions in the syntactic rules.

Making the word identifier in motion, the parsing director decides a range where word matching is permitted. In order not to leave segments unmatched at a word boundary, the director controlls the word identifer to prepare a reference template supplementing the last segment of the previous word to a new word. The range of matching is decided by setting matching regions so that a starting region is from the position of the end point of the last word minus W (see section (c)) to the end point, and an ending region from the end point plus one time length of dictionary word to the point adding one and a half times of that.

Evaluating the reliability score of a whole sentence, it is not fair to evaluate functional words and content words equally, because functional words of Japanese consist of few phonemes and they carry less significant meanings for the semantic interpretation. Thus, we set an evaluation weight of content words be heavier than that of functional words; the corresponding rate to the sentence score of functional words is 0.8, and 1.0 for content words.

An example of parsing is shown in Figure 4 where the utterance is "Tabako uriba

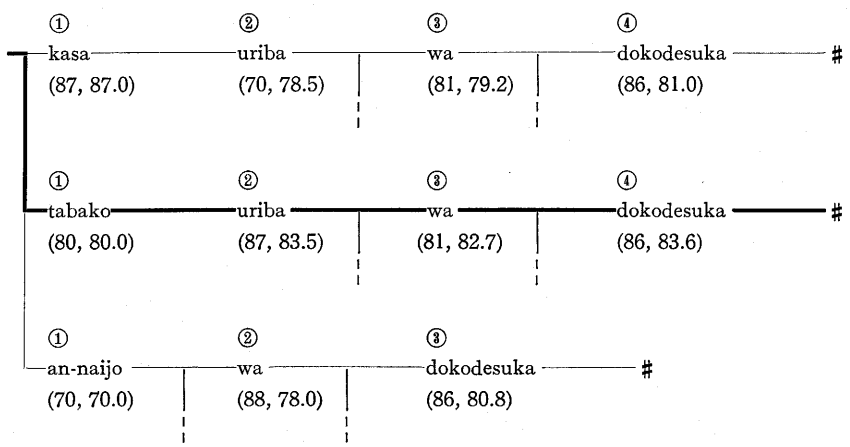

Fig. 4. An example of parsing, "Tabako uriba wa dokodesuka? (Where is a tobacco shop?)"
*Encircled number shows an order of expansion
**( ) means (score of the word match, average score of a partial sentence).

wa dokodesuka? (Where is a tobacco shop?)" and the sentence was correctly parsed.

In order to evaluate the ability of the recognizer, an experiment was performed in which sentences are tried to be recognized. The used sentences are those which can be uttered for the first time in the dialogue, shown in Figure 5. The total number of test sentences is 100 and they were uttered by the speaker and recorded in an ordinary room where mini-computers are set, using a high quality microphone. The words in the dictionary were prepared by segmenting the words manually which were uttered solely by the same speaker. The reference patterns were averaged for twice repetitions of the utterances. In the experiment, the number of intermidiate results is limited to 50.

The results are shown in Table 3. The number of correctly recognized sentences is 65 (65%) and cumulative rate including to the third rank of results is 78%. The rate of

1. XX wa dokodesuka. (Where is XX?)
   XX=biyoo-shitsu, tabako uriba, kasa uriba, kaban uriba, norimono uriba, tokei uriba, wakagu uriba, shashin-shitsu, bungoogu uriba, shokudoo.
2. XX wa nangaidesuka. (On what floor is XX?)
   XX=keshoo-hin uriba, shokuryoo-hin uriba, shinshi-gutsu uriba, bebii-yoohin uriba, shinshi-fuku uriba, kamera uriba, shokki uriba, rekoodo uriba, hon uriba, riyoo-shitsu.
3. XX ni imasu. (I am at XX.)
   XX=preegaido, yakuhin uriba, akusesarii uriba, annaijo, shinshi-yoohin uriba, nekutai uriba, gorufu-yoohin uriba, kimono uriba, gakki uriba, omocha uriba.
4. XX e ikitai. (I want to go to XX.)
   XX=handobaggu uriba, kodomo-fuku uriba, fujin-fuku uriba, megane uriba, hooseki uriba, koogee-hin uriba, hadagi uriba, hakimono uriba, choori-yoohin uriba, kissa-shitsu.
5. XX wa dokode-utteimasuka. (Where is XX sold at?)
   XX=kushi, paipu, raitaa, osake, okashi, sangurasu, sukaafu, beruto, sutekki, tebukuro.
6. XX wa dokoni-oiteimasuka. (Where XX puts?)
   XX=mafuraa, kegawa, gakusee-fuku, jiinzu, sanrinsha, jitensha, toreenaa, undoo-gutsu, mizugi, kontakutorenzu.
7. XX o mitai. (I want to see XX).
   XX=booenkyoo, kabin, chadoogu, okinomo, furoshiki, yukata, pajama, geta, tabi, beddo.
8. XX o kaitai. (I want to buy XX.)
   XX=honbako, shokkidana, tansu, mahoobin, hoochoo, toosutaa, kaapetto, kaaten, gakufu, kasetto-teepu.
9. XX e douikeba-yoidesuka. (How to go to XX?)
   XX=fujin-yoohin uriba, fujin-gutsu uriba, kodomo-yoohin uriba, undoo-yoohin uriba, yookagu uriba, interiya-yoohin uriba, onkyoo-seehin uriba, shingu uriba, jimuyoo-hin uriba, shinshi-yoohin uriba.
10. XX wa YYno chikakudesuka. (Is XX near YY?)
    (XX, YY)=(biyoo-shitsu, kita-kaidan), (keshoo-hin uriba, kita-erebeetaa), (preigaido, chuuoo-erebeetaa), (tabako uriba, chuuoo-erebeetaa), (shoku-ryoo-hin uriba, minami-esukareetaa), (yakuhin uriba, minami-kaidan), (akusesarii uriba, kita-kaidan), (shinshi-yoohin uriba, kita-erebeetaa), (handobaggu uriba, chuoo-erebeetaa), (nekutai uriba, minami-esukareetaa).

Fig. 5. Sentences used in the recognition experiment.

Table 3. Results of recognition experiment.

recognition		
rank of correct result	number of correct result	number of cumulative correct results
1st	65	65
2nd	12	77
3rd	1	78
understanding		
1st	68	68
2nd	11	79
3rd	2	81

Total number of sentences=100.

understanding, which means the resultant predicate is different from the correct one but has the same sentence type (see section 2.4), is 68% and the cumulative rate to the third is 81%.

The reason why the rate of correct recognition is low comparing with results obtained by the other system[6], is that we prepare a dictionary word averaging only twice and no facility is considered to deal with a phenomenon of variable ways of speaking words in sentence speech, eg. devocalization of vowels. The cumulative recognition rate shows a better score and this keeps a possibility of correcting the message by using dialogue capability.

Since the dialogue component must judge whether it rejects the input message or not, the scores of recognition results should be examined. Figure 6 shows the scores of the recognized sentences where correct results are shown by white bars,

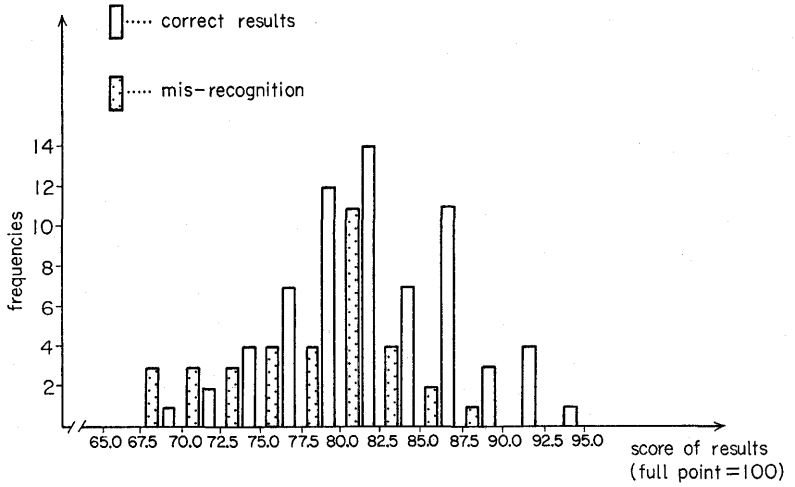

Fig. 6. Historam of recognition scores.

and the mis-recognition by dotted bars. The histogram shows that scores of correct recognitions are greater than those of the mis-recognized. This tendency may due to the fact that correct recognition succeeds when all words match well and then have higher scores. This tendency can also give a possibility that the dialogue component is able to exclude the correct messages from the results.

2.4 Dialogue Component

The dialogue component analyzes a recognized sentence and generates appropriate response codes which are passed to the response-speech synthesizer. Recognized sentences which the dialogue component receives consist of a sequence of words, score of the result, and sentence type which corresponds to the responding action of the system. The sentence types managed in the recognition component are listed in Table 4. Analyzing the recognition results and deciding what response is the most appropriate one, the dialogue component generates a sequence of words which will be decoded and transformed into speech form by the speech synthesizer. The dialogue component consists of two subsequent modules: the dialogue control module and the semantic analyzer.

(a) Representation of World Knowledge

The information of the organization of the department store ("world knowledge") represented as three kinds of relations, which are relations between counters and their existing floor, "near-by" relations, and "visible" relations. The first relation is defined arbitrarily but simulates a certain department store and represented in a form of 'table' in the system. Near-by relation defines a relation between a counter and its nearest means to move up or down the floors (we call these means to move up and down the floors such as stairs, escalators, elevators as 'objects' hereafter). The relations of near-by are designed in order to provide a user with information about a route and means to move to his destination counter. Each entry of counters has a near-by relation to the nearest object.

Table 4. Sentence types for user messages.

sentence types	corresponding predicates and words
1. YES	hai
2. NO	iie
3. DESU	desu
4. DOKOKA	dokodesuka*, nangaidesuka
5. IRU	imasu
6. IKU	ikitai
7. CHIKAKU	chikakudesuka
8. URU	dokode-utteimasuka, dokoni-oiteimasuka
9. KAU	mitai, kaitai
10. DOKO2	dokodesuka*
11. DOU	douikeba-yoidesuka

* They are different in a sence that different kinds of words can precede.

Visible relation is designed so that the route guidance routine can provide a user with comprehensible information to guide him to his destination. Every counter has a list of all visible counters where a user can see from the point where he is standing. These three relations are represented as depicted in Figure 7. In addition to these relations, the relations between counters and goods are designed so that each kind of goods is sold at one counter.

(b) Dialogue Control Module

The dialogue control module controlls the other module and has a capability of dealing with incompleteness of the recognition component. The flow of the process of the dialogue control module is outlined in Figure 8. In dealing with the uncertainty of the recognition results, the control module makes use of two thresholds corresponding the scores of the results. One threshold (THaccept) is used in the module to examine whether the recognized sentence is reliable enough to accept as correct, and another one (THreject) is used to judge whether the module reject the input or not. The module first investigates the score of top recognized sentence and returns the reject message to the user and requests him to utter once more, if the score does not satisfy the THreject. If the best result satisfies the threshold THreject and does not satisfy THaccept, the module will show the recognized utterance to the user and asks him to confirm his message by forced binary choice 'YES' or 'NO'. If the response from the user is YES, the module passes the confirmed result to the semantic analyzer. If NO, the module then makes a confirmation question to the user for the second original result when it satisfies the threshold THreject.

When the dialogue control module asks the user to utter the original message once more, it gives instructions to the speech recognizer that sentences denied through yes-no questions would not be recognized for the next time. This mechanism is designed in order to guarantee a convergence of confirmation steps.

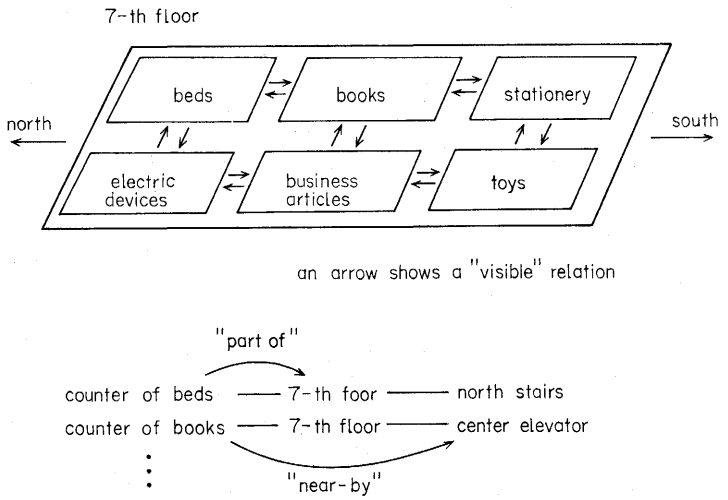

Fig. 7. An example of knowledge representation of the department store.

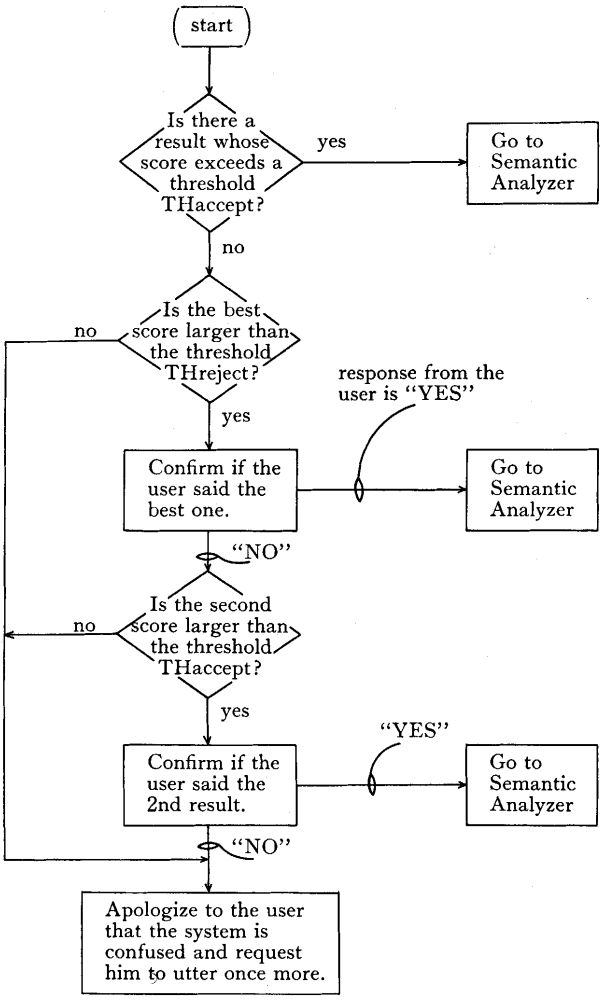

Fig. 8. Flow diagram of the dialogue control module.

(c) Semantic Analyzer

The semantic analyzer, receiving user's request from the dialogue control module, tries to analyze the sentence and take an action required by the user. The actions which the analyzer can take place correspond to sentence types, where each one of them has specific sequence of processes.

In the semantic analyzer, a stack called "F-stack" is designed as a memorandum to keep track of user's destination places and so on. F-stack is a tool to deal with a simple ellipsis which is admitted by the system. A user can say just "Douikeba-yoidesuka (How to go?)" instead of "Tabako uriba e douikeba-yoidesuka (How to go to the tobacco shop?)", if the tobacco shop appears in the preceding conversation.

Entry items of the stack are names of counters and goods which have attributes of FROM-LOC, TO-LOC, and GOODS. For example, if the input recognized sentence type is "DOKOKA (dokodesuka, where)" and if it accompanies the item of a counter, it

is stacked into the F-stack with attribute of TO-LOC. In a similar manner, if the type of the sentence is "KAU (kaitai, I want to buy..)", the name of goods will be registered. These items must be memorized in order that the system can deal with elliptic expressions.

The route guidance routine which is a constituent of the semantic analyzer is invoked when an input sentence type is "DOU (douikeba-yoidesuka, How to go?)". The guidance module searches for the user's original position and the destination in the F-stack. If these places have not specified yet, the module makes a question to the user, asking him where he is going or where he is now standing. The answer by the user is processed in an ordinary manner by the dialogue control module and passed to the route guidance module which then fills the empty items necessary to serve a path. When the original place and the destination are specified, the guidance routine tries to find a path which is basically composed by a sequence of counter names in that from the former counter the user can see the latter counter, examining near-by relations and visible relations.

2.5 Response-Speech Synthesizer

The output speech of the system is synthesized by concatenating words stored in a disk file. The words are prepared by recording solely uttered word samples which are of 8-bit linearly pulse-code-modulated representation (sampling frequency=10 KHz). The total number of words prepared for the response is 165. The number of sentence types for voice output is 17.

The quality of the synthesized speech might be awkward and unnatural, since no compensation is taken place that smoothes fundamental frequencies and durations over a sentence. But for comprehension, the least intelligiblity of speech is presented.

3. DEVELOPMENT AND PERFORMANCE

The system is installed on a complex of mini-computers. The hardware configuration is shown in Figure 9. The sentence recognizer and the response speech synthesizer are built on MELCOM70 and the dialogue component on FACOM U-200.

The system can respond to a user's utterance in few times or several ten times of

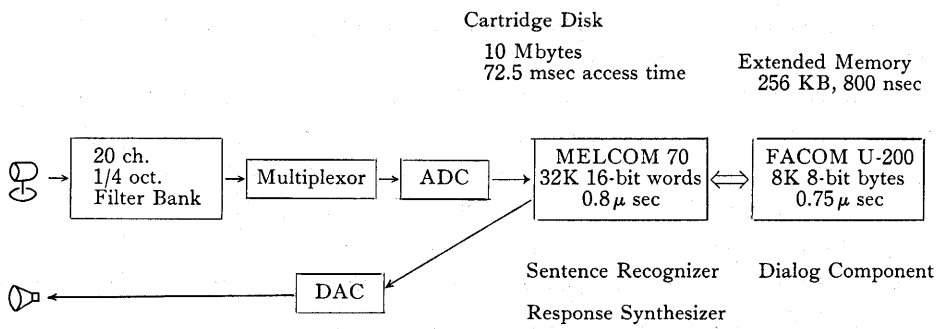

Fig. 9. Hardware configuration of the system.

User : Shinshi-fuku uriba wa dokodesuka? (Where is a counter of men's clothes?)

System : "Shinshi-gutsu uriba wa dokodesuka?" to osshaimashitaka? (Did you say, "Where is a counter of men's shoes?")

U : Iie. (No.)

S : "Shinshi-fuku uriba wa dokodesuka?" to osshaimashitaka? (Did you say, "Where is a counter of men's clothes?")

U : Hai (Yes.)

S : Shinshi-fuku uriba wa sangai desu. (The counter of men's clothes is in the third floor.)

U : Douikeba-yoidesuka? (How to go?)

S : Dokoni imasuka? (Where are you?)

U : Tokei uriba desu. (I am at the counter of watches.)

S : Mooichido onegaishimasu. (Once more, please.)

U : Tokei uriba desu. (I am at the counter of watches.)

S : Chuuoo-esukareetaa ga wakarimasuka? (Can you see the center escalator?)

U : Hai. (Yes.)

S : Chuuoo-eskareetaa de sangai e ittekudasai. Soko ga shinshi-fuku uriba desu. (Please go to the third floor by the center escalator. You will find the counter of watches there.)

Fig. 10. An example of guidance.

real time, according to the number of sentence types which can uttered and the magnitude of searching space in a recognition process. Most of processing time is consumed by the sentence recognizer.

An example of the system performance which is one of the experimental results is shown in Figure 10. In the example, the first message of the user was clarified by exchanging confirmation messages twice. The second request (How to go?) was accepted at once. Then, to provide the user with guidance, the system asked him where he stayed, by saying "Where are you?" The response of the user was once rejected and the system asked him to speak once more. Last three messages were exchanged for a route guidance. As a result, the user received information that he required.

To investigate a number of utterances required for message confirmation, an experiment was performed[8]. Inquiry messages used in the experiment are those which were used in the recognition experiment and 50 sentences out of 100 were chosen randomly (5 for each predicate). The experiment was performed in on-line mode and in the ordinary room where mini-computers are set (Thresholds were set as THaccept=90.0, THreject=70.0 by observing the results of Figure 6).

Of the fifty inquiries, most of them were confirmed correctly except only inquiry in which case the experiment breaks because the number of repetition for the original sentence exceeds five times. The required utterance for confirmation consist of 0.4 times (as an average) for the original message sentence and 1.8 times for YES-NO utterance, and hence 2.2 utterances as an average are required to confirm an inquiry message.

4. CONCLUSION

A speech understanding system with dialogue capability was described. The system can respond in a few ten times of real time and works for the task of information service at a department store. Through experiments using the system, following remarks must be noted.

1) A machine system can obtain almost complete understanding rate introducing dialogue capability with a user to confirm the message, using conversational speech.
2) A dialogue capability can compensate for an incomplete ability of the speech recognizer.
3) Required additional utterances to the user for message confirmation is not unbearable.
4) For the task of information service at a department store, the system can guide the user successfully, in which case some simple facilities of dealing with ellipses and of route guidance are implemented.

For future research, there are some problems to be considered. For speech recognition, acceptability of unlimited speakers and co-articulation problems must be solved. Acceptability of unlimited speakers means not only the requirement of speaker normalization at phoneme recognition level but also the requirement of mechanism to deal with wide variety of speaking the same contents.

As a machine dialogue system, more graceful ability of interacting with a user is required[7]. There are two major problems; problems of natural man-machine communication and of ability dealing with a fragmental input to the system. The former means ability of natural dialogue as a machine simulates a human listener[8]. The latter problem implies the difficult problem originated from conversational speech, which includes some meaningless sounds such as '..er..' and '..ah..', and also includes the problem of ellipses. One approach to deal with the latter problem is to spot a 'key word' from a stream of words or sentences and not to recognize a full sentence. And we consider this is an important item to develop a machine dialogue system by speech which can talk with a user more naturally.

ACKNOWLEDGEMENT

The authors wish to thank Mr. N. Ishikawa for his cooperation and development of the semantic analyzer. Thanks are also due to Miss K. Kimura who offered speech materials for the response synthesizer.

REFERENCES

1) A. Newell et al.: Speech understanding systems; Final report of a study group, North Holland (1973).
2) W. A. Lea, Ed.: Trends in speech recognition, Prentice-Hall (1980).
3) M. Kohda, R. Nakatsu, K. Shikano and K. Itoh: On-Line Question-Answering System by Con-

versational Speech, The Jounal of the Acoustical Society of Japan, Vol. 34, No. 3, pp. 194–203 (1978, in Japanese).

4) S. E. Levinson and K. L. Shipley: A Conversational Mode Airline Information and Reservation System Using Speech Input and Output, Conf. Record ICASSP80, 2-5.10, pp. 203–208 (1980).

5) T. Ukita, S. Nakagawa and T. Sakai: A Use of Pitch Contour in Recognizing Spoken Japanese Arithmetic Expressions, Trans. IECE Japan, Vol. J-63D, No. 11, pp. 954–961 (1980, in Japanese).

6) T. Sakai and S. Nakagawa: A Speech Understanding System of Simple Japanese Sentences in a Task Domain, Trans. IECE Japan, Vol. E-60, No. 1, pp. 13–20 (1977).

7) P. Hayes and D. R. Reddy: An Anatomy of Graceful Interaction in Spoken and Written Man-Machine Communication, CMU-CS-79-144 (1979).

8) T. Ukita, N. Ishikawa, S. Nakagawa and T. Sakai: Confirmation Methods of Utterances in a Dialogue System by Speech, Trans. Information Processing in Japan (1981, in Japanese, to appear).

(Aug. 31, 1981, received)