

Robust Speech Understanding Based on Expected Discourse Plan

Shin-ya Takahashi, Tsuyoshi Morimoto, Sakashi Maeda, Naoyuki Tsuruta

Department of Electronics Engineering and Computer Science
Fukuoka University, Fukuoka 814-180, Japan

{takahasi,morimoto,maeda,tsuruta}@tl.fukuoka-u.ac.jp

Abstract

This paper reports spoken dialogue experiments for elderly people in the home health care system we have developed. In spoken dialogue systems, it is important to decrease recognition errors. The recognition errors, however, cannot be completely avoided with current speech recognition techniques. In this paper, we propose a robust recognition understanding technique based on expected discourse plans in order to improve a recognition accuracy. First, we collect dialogue examples of elderly users through a Wizard-of-Oz (WOZ) experiment. Next, we conduct a recognition experiment for collected elderly speech using the proposed technique. The experimental result demonstrates that this technique improved a sentence recognition rate from 69.1% to 74.3%, a word recognition rate from 80.3% to 81.7% , and a plan matching rate from 88.3% to 92.0%.

1. Introduction

The aging of the population in Japan is proceeding more rapidly than in other countries. According to the Ministry of Health and Welfare, senior citizens, who are more than 65 years old, are expected to reach 20% of Japan's population by the year 2010. Especially the number of the elderly people in need of health care is growing. Under the circumstances described above, we have been developing a spoken dialogue system aiming at support for the home health care services[1]. The targets of this system are aged people who live alone with comparatively good health condition. In this system, the system watches user's daily physical conditions through medical examinations (e.g. checking the user's temperature and blood pressure) and some conversations (e.g. asking some questions about the user's condition).

Several spoken dialogue systems targeting elderly people have been developed, which are, for example, the Nursebot (Pearl Project) [2], the Care-O-bot[3] and the MedAdvisor [4]. In the Pearl project, they have conducted a preliminary experiment targeting elderly people in an assisted living facility. This system controls its behaviors using a hierarchical variant to a partially observable Markov decision process (hierarchical POMDP), but this is the part of a nursing robot and doesn't deal with conversations for health care service yet. The Care-O-bot project is also targeting elderly people, but they are mainly developing mobile robotics architecture and are utilizing a simple man-machine interface with speech and touch-screen. In the MedAdvisor project, they are developing a system which provides healthcare information about user's medications and they have conducted a Wizard of Oz (WOZ) experiment for elderly people, but their system is still an early prototype.

On the other hand, several experiments of speech recognition for elderly people have been conducted and corpora of elderly speech have been collected [5],[6]. But their purpose

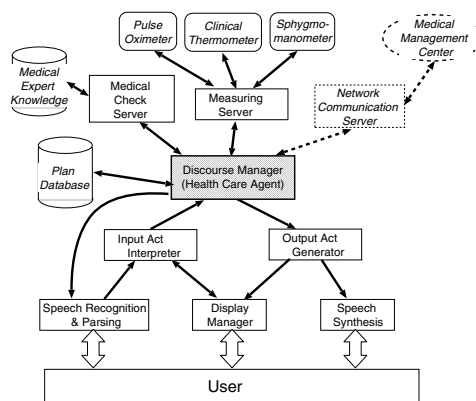


Figure 1: System organization.

is to construct acoustic models and they don't take account for interaction processes.

In order to develop a useful dialogue system, it is necessary to decrease the system's mis-understandings caused by the decoder's mis-recognitions. The conventional dialogue systems perform to detect these mis-recognitions on the basis of confidence score thresholds modeled with some verbal data after recognition process[7],[8]. However, since these techniques do not aim at dialogue-oriented application, the recovery from erroneous recognition results does not involve an interactive process with the user but a single utterance verification process inside the system. Besides, there are some dialogue systems which perform the recovery involving an interactive process with the user, but the recognition process and the recovery process are executed independently and not-cooperatively [9],[10],[11].

In this paper, we proposed a robust speech understanding technique based on expected discourse plans in order to improve an accuracy of the speech recognition. In this technique, a dialogue context is used to weight a language score of the word candidate which can matched with expected dialogue plans. The speech recognition component communicates with the discourse manager and works cooperatively.

In order to evaluate our system, we collect dialogue examples of elderly users through a WOZ experiment and conduct a recognition experiment for collected elderly speech. The experimental result demonstrates that the proposed technique improved a sentence recognition rate from 69.1% to 74.3%, a word recognition rate from 80.3% to 81.7% , and a plan matching rate from 88.3% to 92.0%.

2. System Overview

2.1. System Organization

The basic organization of our system is shown in Fig.1. The system communicates between each component which works parallel and asynchronously. For example, the dialogue manager accomplishes the appropriate discourse plan according to the current topic, sends requests to or receives responses from the other components. Upon the requests from the Discourse Manager, the Medical Check Server prepares questionnaires and examinations to be performed, and sends them back to the Discourse Manager. The examination and questionnaire plans to be executed, that is, which data should be measured or what question should be asked, are determined by referring to the medical knowledge which is prepared based on the advice of human medical doctors in advance.

2.2. Efficient Discourse Management

Fig.2 shows a basic plan for a daily medical check. The system executes a dialogue using discourse plan given in forms of a tree structure, because the tree structure plan is suitable to deal with topic-shift in our domain. For example, when the user's physical condition is good, the system executes discourses according to the given basic plan¹. When the user complains his/her conditions to the system, an interruption from the user is sent to the Discourse Manager and the discourse plan to ask about the user's symptom is reconstructed.

In the parsing process of our system, a user's utterance is segmented by morphological analyzer, key phrases are derived from it. The key phrases are, for example, noun phrases, verb phrases, and negative expressions, and so on. Matching these key phrases with executable plans in the current state, the system instantiates the plan tree and executes the matched plan. If no plan is matched with the key phrases, the system asks the user what he/she said. Fig. 3 shows these processes matching plans and the user's utterance. In [1], we have shown the effectiveness of the above matching process between the key phrases and the discourse plan.

The system also displays menus to navigate a user's answer. The menu items are made from expected answers in the current plan. The user can answer the system by means of not only a speech but selecting the menu items. Selected answers from the menu items are also parsed with the same way as the speech inputs.

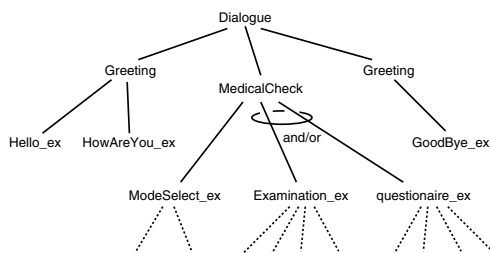


Figure 2: A plan tree for daily medical checks.

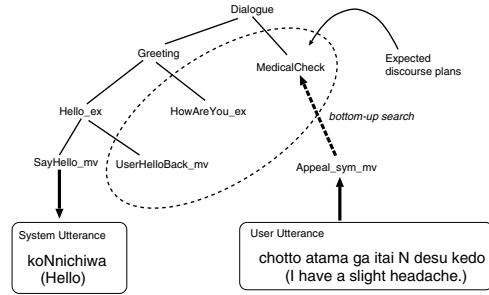


Figure 3: Matching process between expected plans and user's utterance.

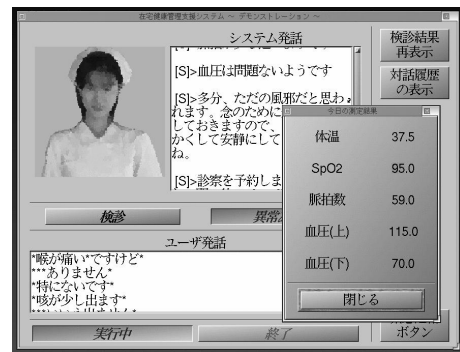


Figure 4: Prototype system.

2.3. Prototype System

Fig. 4 shows the prototype system we have developed. This prototype system consists of the components drawn with the solid lines in Fig.1 (the parts with the dashed lines will be implemented soon).

For speech recognition, the system uses the HTK's Viterbi recognizer², and a Japanese acoustic model included Julius³. The language model is prepared as an FSA network grammar for our domain. The system used the health check plan such as shown in Fig.2. The number of the executable plans, which correspond to the leaves in the plan tree, is about 50. In the prototype system, the table of the pre-diagnosis rule consists of 36 disease names and 32 items of examination / questionnaire.

3. Speech Understanding Technique Based on Expected Discourse Plan

As mentioned in section 1, since the current speech recognition technology is far from perfect and cannot completely avoid the recognition errors, many researchers try to develop robust systems which can detect and recover from the system's misunderstanding. Almost these conventional systems are, however, do not aim at dialogue-oriented application.

In this paper, in addition to the previous method described in section 2.2, we investigate to utilize a current context in an

¹ While the basic plans are given in this case, the details of the plans are constructed dynamically as same as the other case.

² HTK (Hidden Markov Model Tool Kit) web page is "<http://htk.eng.cam.ac.uk>".

³ Julius is Japanese Dictation Tool Kit. Julius's web page is "<http://winnie.kuis.kyoto-u.ac.jp/pub/julius/>".

Table 1: an example of the values of $C_W^{(i)}$

$C_W^{(i)}$	word candidates for recognition
-1	koNnichiwa (hello), wakari-mashi-ta (I see.), . . .
0	hai (yes), iie (no), seki (cough), de-masu (I have), de-mase-N (I don't have)
1	zutsuu (headache), hukutuu (stomach-ache), hakike (nausea), . . .
2	otherwise

executing plan for recognition process. In order to achieve this idea, we try to increase a priority of the expected word which will be uttered in the current context and to decrease a priority of the un-expected word which has already been uttered in the dialogue. Specifically, for an observation sequence O and a word W , we compute a likelihood

$$p(O|W) \times P(W|context_i), \quad (1)$$

where $P(W|context_i)$ is a weight of the word W in the context when the i -th plan was executed and is assumed to have the following exponential distribution

$$P(W|context_i) = \begin{cases} \lambda e^{-\lambda C_W^{(i)}} & \text{if } C_W^{(i)} \geq 0, \\ 0 & \text{if } C_W^{(i)} < 0. \end{cases} \quad (2)$$

Here, $C_W^{(i)}$ is the category number of the expected plans. It seems appropriate to assume $P(W|context_i)$ as the exponential distribution, because the direct answer for the preceding question is the most expected response and the probability of the other answer is less than that of the direct one in our natural conversation.

$C_W^{(i)}$ is dynamically calculated according to the context. If the word W is the un-expected word, then $C_W^{(i)} = -1$ (i.e., $P(W|context_i) = 0$), if the word W is the expected word in the current plan which has already been instantiated, then $C_W^{(i)} = 0$ (i.e., $P(W|context_i) = -\lambda$), if the word W is the expected word in the future plan which has not been instantiated yet, then $C_W^{(i)} = 1$ (i.e., $P(W|context_i) = -\lambda e^{-\lambda}$), and if the word W is otherwise, then $C_W^{(i)} = 2$ (i.e., $P(W|context_i) = -\lambda e^{-2*\lambda}$). Here, the value of λ are determined after a preliminary experiment. For example, when the system asks the user, "seki wa de masu ka? (Do you have a cough?)", $C_W^{(i)}$ obtains the value shown in Table 1⁴. As shown in Table 1, the content words included in the system's question already obtain $C_W^{(i)} = 0$. As described later, the reason for this is that the user tends to repeat the content word.

4. Experiments

4.1. WOZ Experiment

For the purpose of the collection of dialogue examples and speech data, we conducted a WOZ experiment targeting elderly people. Subjects are seven 65-to-82-year-old females. Fig. 5 is a picture of the dialogue experiment for elderly.

Through the WOZ experiment, we collected 21 dialogues and got totally 137 user's utterances. Average of the dialogue time is 246 seconds. Fig.6 shows an example dialogue. The

⁴The score of key-phrases as shown in Table 1 is computed by calculating the weight of each word in the key-phrases.



Figure 5: Dialogue experiment for elderly

type of the user's utterance and the frequency of the occurrence of each type are shown in Table 2. This table also shows that the frequency of the user's answer with Yes/No and bare Yes/No answers is less than that of the answers without Yes/No. This result suggests that in spite of simple Yes-No question cases, the answers are not always simple. In other words, the user tends to response the content words corresponding to the system's question.

4.2. Recognition Experiment

Based on the dialogue example collected through the WOZ experiment, we prepared an FSA network grammar for our domain. The number of the words in the vocabulary is approximately 200 and the number of the node is approximately 2000. The perplexity of this grammar is approximately 4.5.

Using this network language model and collected elderly speech data⁵, we conducted a recognition experiment of the proposed method based on expected plan compared with the previous method. Table3 shows the result of the recognition experiment. As shown in Table 3, in spite of the perplexity of the FSA is low and the speech data are not so noisy, the recogni-

Table 2: List of type of the user's utterance

the name of type (example sentence)	frequency
greeting (ex. koNnichiwa. (Hello.))	24
acknowledge (ex. hai, wakari mashi ta. (All right, I see.))	16
response to Wh-question (ex. chotto atama ga itai N desu kedo. (I have a slight headache.))	29
bare Yes/No response (ex. hai. (Yes))	15
content word response with Yes/No (ex. hai, seki wa demasu. (Yes. I have a cough.))	19
content word response without Yes/No (ex. arimaseN. (I don't have))	32
Yes/No response with adverb (ex. hai, sukoshi. (Yes, a slight))	2

⁵Here, the average of the segmental SNR of speech data is 14.8 [dB] (the minimum is 5.6 [dB] and the maximum is 24.2 [dB]). This segmental SNR is approximated values by using an energy of sampling data where the utterance doesn't exist as a noise energy.

S1 koNnichiwa. (Hello)
U1 koNnichiwa. (Hello)
S2 taichou wa ikaga desu ka? (How are you feeling?)
U2 sukoshi kaze gimi desu. (I have a slight cold.)
S3 sou desuka. (I see.)
S4 hakike wa ari masu ka ? (Are you nauseated ?)
U4 hakike wa ari maseN. (No nausea.)
S5 mune no itami wa arimasu ka? (Do you have chest pain ?)
U5 mune no itami wa arimaseN. (I don't have chest pain.)
S6 iki wo hukaku suu to, mune no atari ga itamimasu ka ? (When you are breathing deeply, do you feel chest pain ?)
U6 eeto... (Let me see, ...)
S7 seki wo suru to, mune no atari ga itamimasu ka ? (When you are coughing, do you feel chest pain ?)
U7 iie. (No.)
S8 tan wa demasu ka ? (Do you bring up phlegm ?)
U8 sukoshi demasu. (a little.)
S9 tan no iro wa naniiro desu ka ? (What color of phlegm?)
U9 toumei desu. (It's no color.)
S10 tan ni ketueki ga majitte masu ka ? (Is it bloody phlegm?)
U10 iie. (No.)
S11 seki wa demasu ka ? (Have you been coughing ?)
U11 sukoshi demasu. (I have a slight cough.)
S12 kaze wo hiki kakete iru you desu. narubeku, atatakaku site ansei ni site ite kudasai ne. (You seems catch a cold. You had better let yourself warm and rest.

Figure 6: A dialogue example (U ans S mean a user and a system utterance, respectively)

Table 3: result of recognition experiment

	word recog. rate	sentence recog. rate	plan-matching rate
previous method in [1]	80.3%	69.1%	88.3%
proposed method	81.7%	74.3%	92.0%

tion rate is not high. The reason for this is that the qualities of speech data are not high so that the head of some speech data is missing and the end of some utterances is not clear. It is necessary to investigate whether such a low quality is due to the characteristics of elderly.

On the other hand, Table 3 shows that the word and sentence recognition rate were improved. The reason for this is that the proposed method effected to avoid the hypotheses which are not suitable for the state of the dialogue. It was also caused to improve the plan-matching rate. This result demonstrates that the proposed method can improve the recognition accuracy.

5. Conclusion and Future Work

In this paper, we described dialogue experiments for elderly people in the home health care system we have developed. In order to improve an accuracy of the speech recognition, we proposed a robust speech understanding technique which uses a dialogue context to weight a language score of the word can-

didate which can matched with expected dialogue plans. First, we collected dialogue examples of elderly users through a WOZ experiment. Next, using the proposed technique, we conducted a recognition experiment for the collected elderly speech. From the experimental result, we showed that the proposed technique could improve a sentence recognition rate from 69.1% to 74.3%, a word recognition rate from 80.3% to 81.7% , and a plan matching rate from 88.3% to 92.0%.

In the future, we intend to cope with the issues of speech recognition for the elderly[6]. We also intend to integrate image processing, such as the recognition of the user's image and the generation of graphical interfaces with a visualized agent, in order to achieve the more natural and user friendly conversation.

6. Acknowledgment

We would like to thank Dr. Yamada at School of Medicine, Fukuoka University, Dr. Ogawa at Hyaku-Nen-Bashi Clinic, and Dr. Eshita at Eshita Clinic for providing the medical knowledge.

7. References

- [1] S. Takahashi, T. Morimoto, S. Maeda, and N. Tsuruta, "Spoken dialogue system for home health care," in *Proc. of the ICSLP*, 2002, pp. 2709–2712.
- [2] M. Montemerlo, J. Pineau, N. Roy, S. Thrun, and V. Varma, "Experiences with a mobile robotic guide for the elderly," in *Proc. of the International Conference on Artificial Intelligence*, 2002, pp. 587–592.
- [3] M. Hans, B. Graf, and R.D.Schraft, "Robotic home assistant care-o-bot: Past-present-future," in *Proc. of the IEEE ROMAN*, 2002, pp. 380–385.
- [4] G. Ferguson *et al.*, "The medication advisor project: Preliminary report," *Technical Report 776, CS Dept., U. Rochester*, 2002.
- [5] S. Anderson *et al.*, "Recognition of elderly speech and voice-driven document retrieval," in *Proc. of the ICASSP*, 1999.
- [6] A. Baba *et al.*, "Elderly acoustic models for large vocabulary continuous speech recognition," in *Proc. of the EUROSPEECH*, 2001, pp. 1657–1660.
- [7] T.-H.Chiang and Y.-C.Lin, "Error recovery for robust language understanding in spoken dialogue systems," in *Proc. of the EUROSPPECH*, 1999, pp. 367–370.
- [8] E.K.Ringer and J.F.Allen, "A fertility channel model for post-correction of continuous speech recognition," in *Proc. of the ICSLP*, 1996, pp. 897–900.
- [9] K.Komatani and T.Kawahara, "Flexible mixed-initiative dialogue management using concept-level confidence measure of speech recognizer output," in *Proc. of the COLING*, 2000, pp. 467–473.
- [10] D.J.Litman, J.B.Hirschberg, and M. Swerts, "Predicting automatic speech recognition performance using prosodic cue," in *Proc. of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000, pp. 218–225.
- [11] J.Hirasawa, N.Miyazaki, M.Nakano, and K.Aikawa, "New feature parameter for detecting misunderstandings in a spoken dialogue system," in *Proc. of the ICSLP*, 2000, pp. 155–158.