DEVELOPMENT AND EVALUATION OF A SPOKEN DIALOGUE SYSTEM FOR ACADEMIC DOCUMENT RETRIEVAL WITH A FOCUS ON REPLY GENERATION

Shinya Kiriyama† Keikichi Hirose‡ Nobuaki Minematsu†
† Department of Information Engineering, School of Engineering
‡ Department of Frontier Informatics, School of Frontier Sciences
University of Tokyo, Bunkyo-ku, Tokyo, JAPAN
{kiriyama,hirose,mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT: A spoken dialogue system has been developed for academic document retrieval. It generates speech replies with their important words emphasized by controlling prosodic features, viz., prosodic focusing. The important words are determined by referring to the dialogue flow. Validity of the prosodic focusing was proved through a system evaluation.

1. INTRODUCTION

A number of systems have been developed for information retrieval, such as internet search engines. Most of them, however, only show retrieved results with very limited (or no) guidances/advices for the following steps to users. Although several spoken dialogue systems were developed to realize this type of function through conversation with the user, their reply speech was in text-reading style with rather poor control of prosodic features. With the aim of realizing speech reply "easy to be understood by users," we constructed a system paying a special effort on generating speech reply with prosodic features properly reflecting dialogue focuses. For this purpose, we developed a scheme of concept-to-speech conversion, though a number of other spoken dialogue systems simply incorporating Text-To-Speech(TTS) conversion algorithms available elsewhere (commercially). By realizing this scheme and several functions useful to support users, we could make the spoken dialogue system one with high usability.

2. SYSTEM OUTLINE

Academic document retrieval was selected as the task of the system taking the aim of the current study into consideration: to realize a good man-machine interaction through spoken language and not to realize advanced retrieval schemes. Academic documents are usually well itemized and their retrieval is rather easy.

2.1. System structure

Figure 1 shows the structure of the system, which consists of five modules, speech recognizer, dialogue manager, data retriever, screen drawer, and speech synthesizer.

As for the speech recognizer, we adopted "Continuous speech recognition parser Julian(Lee et al., 1999)," which uses a language model based on the

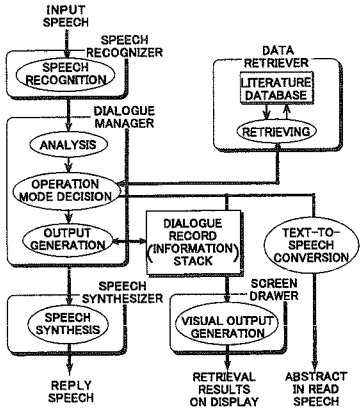

Figure 1: Configuration of the system.

context free grammar(CFG). We used a triphone model set (number of the states: 1,000, number of mixtures for each state: 4) included in "Japanese Dictation Toolkit -1997 version-(Kawahara et al., 1999)" as the parser's phone models, and provided a dictionary for the task. The CFG was constructed as task dependent. The vocabulary size was 535 words consisting of 375 retrieval words (nouns for data retrieval, defined in section 2.2) and 160 functional and other words.

The dialogue manager analyzes the recognition results and decides the system behavior. Dialogue management is conducted using a state transition table. In our previous work(Hirose & Kiriyama, 1999), the basic dialogue strategy of the system was discussed. Namely we adjusted the dialogue initiative in three levels, system-oriented, user-oriented and their mixture, and asked users to select the best one. As the result the mixture mode was selected. We also determined the elliptic expression level for the system reply when answer-

ing user's question. We compared three levels in a similar way (by asking to users). The result indicated that it was the best to add the information which did not appear in the adjacent user's question. Our present system was designed based on these results.

The data retriever selects articles in the database that match with retrieval words in user input. The database consists of 1,200 articles in Electronic Library Service by National Center for Science Information Systems.

The screen drawer shows six items on display: recognition result of input speech, retrieving formula(description on how to retrieve), number of articles found in the database, titles of the articles, details of an article selected by the user, and retrieval words related to the articles. The module selects the contents to display by referring to dialogue records.

The speech synthesizer is a formant synthesizer which we have formerly developed(Hirose & Fujisaki, 1993). Prosodic control was that for dialogue-like speech generation(Hirose et al., 1996).

2.2. System functions

The system has three functions, retrieval function, result output function, and answering-to-question function.

Retrieval function: Retrieval words were selected manually in advance from noun words (including compound words) included in the document database. Depending in which part of database they were included, they could belong the following 6 categories: 1) article title, 2) keyword, 3) abstract, 4) author names, 5) year of issue, and 6) journal name. Category number to which a retrieval word belonged was attached to the word to facilitate the retrieval process: the data retriever searching the database partly indicated by the number. Additional category number "7" was also attached to words belonging categories 1 to 3. This is because users may request the document search by indicating retrieval word(s) only, such as "On speech." In this case, search should be done for the category 7. Retrieval process is possible for any of "and" and "or" combinations of retrieval words. When retrieval words are belonging to categories 1), 2), 3), 4) or 7), "and" is selected as the default, while they are belonging to categories 5) or 6), "or" is selected. An "and" search is possible for retrieval words each belonging to a different category. The system indicates the "current" retrieval words with "and/or" conditions, and users can change them easily by speech input. When dialogue proceeds, the retrieval words may be added or changed. These words are indicated on display by a color different from that of unchanged words.

Result-output function: The retrieved articles are listed on the display with their titles. By selecting one article form the list, users can see its details (such as author names, abstract and so on), which will be useful for users to proceed further the selection process. If many articles are matched through the retrieval process, it will confuse users to list up all of them. Currently, if the number of matched articles is above 10, the system indicates only the number by voice and asks users to add retrieval words by showing their candidates on the display. With this procedure, the article numbers can be efficiently reduced, and users can reach their goals easily. Users can ask the system to read abstracts of articles shown on the display. For this purpose, a TTS conversion software was incorporated in the system. Contents on the display are renewed dynamically as dialogue proceeds. Renewed parts are indicated by changing their color from that of other parts, in order to facilitate user's understanding.

Answering-to-question function: Users can make questions aurally on the retrieved results. Questions include simple ones such as on titles, author names, years of issue, journal names, and existent of abstracts of individual articles. They also include rather sophisticated ones, which require a high-level semantic processing, such as on the newest (year of issue) article number(s), on the most frequently appeared journal name in the list, and so on. When user's question includes elliptical expressions, the system automatically complements lacking information by referring to dialogue record. If the complementation fails, the system asks the user for the missed information. Prosodic features of system answers (speech replies) are controlled to emphasize important words, such as those conveying answering information to user's question. This prosodic control is planned also to facilitate user's understanding.

3. SPEECH REPLY GENERATION

Although a number of spoken dialogue systems simply incorporate TTS conversion modules to generate speech replies, it is difficult to properly control prosodic features for various speech replies. During reply sentence generation process, the system may have rich information, such as important words, syntactic structure of the sentence and so on, which should be reflected on reply speech prosody. To solve the problem, we developed a scheme to convert concept of reply step-

by-step to a sequence of phone and prosodic symbols. In order to make the reference process during sentence generation to dialogue record easier, three kinds of semantic representations of reply sentences are prepared. Details are explained later with an example shown in Fig. 2.

3.1. Data structure of semantic expression

Dialogue management of the system is conducted using the state transition table. Based on the current state and user input, abstract sentence concept of reply sentence is first selected out of 7: greetings, retrieval word request, notice of system operation, confirmation on system operation, notice of selected article number, guidance to user, and answer to user's question. An abstract sentence concept is converted to a sequence of phone and prosodic symbols using the following information:

1. Words (not included in user's question and need) to be complemented.
2. Sentence style: declarative or interrogative.
3. Dialogue focuses.
4. Sentence, clause and phrase boundaries.
5. Readings, mora counts, parts of speech, and accent types of words constituting reply sentences.

Content of the sentence is decided by the item 1. Other items are necessary to generate prosodic symbols. In order to employ the above information for the reply sentence generation appropriately with a simple scheme, three levels of inner representations were arranged and represented by codes: sentence concept codes, prosodic phrase codes, and word codes.

Item 2 corresponds to the sentence concept codes which represent the sentence patterns. The prosodic phrase codes are used to represent a reply sentence as a sequence of prosodic phrases. Items 3 and 4 are preserved and referred together with the prosodic codes when generating prosodic symbols for reply speech. Word codes are used to make an access to word dictionary and to find word information listed in item 5. Three dictionaries (sentence concept dictionary, prosodic phrase dictionary, word dictionary) were also arranged for code conversion.

3.2. Focus of dialogue

We define the focus of dialogue as the information which a speaker wants most to be understood by a listener. In the current system, dialogue focuses are placed automatically based on the following three rules:

1. When the abstract sentence concept is "notice of selected article number," place a focus on the number of articles, and when it is "answer to user's question," place a focus on the words conveying answering information, respectively.
2. When the abstract sentence concept is "notice of system operation" and the concept sentence includes paper numbers, place a focus on these numbers.
3. For other cases, place a focus on verb of the object phrase.

3.3. Prosodic rules

The prosodic rules used for the control of prosodic features of reply speech are the dialogue speech version (Hirose et al., 1996) of those for read speech (Hirose & Fujisaki, 1993). The rules basically decide positions and magnitudes/amplitudes of phrase and accent commands of the generation process model of F0 contours. Phrase command magnitudes differ depending on its location in a sentence. These commands are represented as prosodic (phrase and accent) symbols in the rules; phrase symbols starting with "P," while accent symbols starting with "D, F or A." "D and F" are for the command onsets of two types of accent: with and without rapid downfall. "A" denotes command offsets. The command values are represented by linear regression equations, whose coefficients for various items are decided by the multiple regression analysis (Hirose et al., 1996). Information on these items being on or off for individual cases is attached as numerical suffices to "P, D, F." A flag for the word importance in the dialogue flow is included in these items.

3.4. Reply sentence generation process

Figure 2 shows an example of reply sentence generation process of the system. The system answers the question "Who is the author?" by complementing the number of document, as "Authors of No.3 are Keikichi Hirose and Goh Kawai."

A sentence concept was generated from its abstract sentence concept by appending information not included in the adjacent user's question. The information can be found in the dialogue record, stored in the information stack. The information stack stores contents (title, authors and so on) of the selected articles.

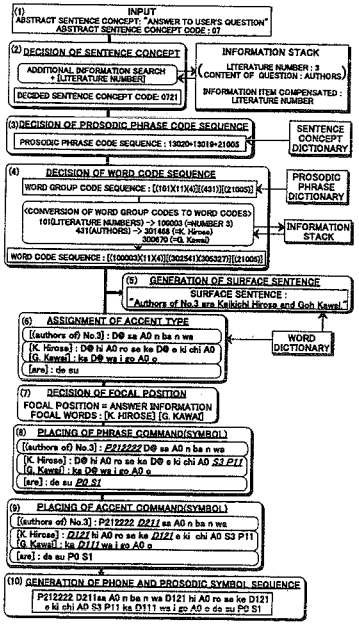

Figure 2: Process of reply sentence generation.

An access to the sentence concept dictionary is conducted following to the code attached to the sentence concept to generate the prosodic phrase code sequence of the reply sentence, which is then converted to the word group code sequence using the prosodic phrase dictionary. Here, a word group indicates words belonging to one category. For example, word group "author names (code: 431)" includes author names. For each word group code in the sequence, appropriate word is selected by referring to the information stack. When a word group has only 1 item, the same code is attached to the word group and the word. (Word selection process is not necessary.)

Finally, the word code sequence is converted to the phone and prosodic symbol sequence by placing phrase and accent commands (symbols) appropriately.

4. EVALUATION EXPERIMENTS

Quality of speech reply was evaluated from the viewpoint of prosodic focus control by comparing two versions of synthetic speech: one with the control (focused version) and the other without (unfocused version). Evaluation was done from the "acceptability" as dialogue system output, and from the "understandability (subjective intelligibility)" of the content of reply.

4.1. Method of experiments

Eight Japanese university students (subjects) were asked to evaluate the two versions of speech replies as a whole (experiment 1) and individually for each reply sentence (experiment 2). A 5-rank scoring scheme was adopted: point 2 when reply speech with prosodic focus control being clearly better, point -2 when it being clearly worse, and point 0 when no difference being perceived. Among the eight subjects, five subjects have experiences of using the system formerly and the others have no experiences. When using the system, the subjects were asked to perform a given task following to a given guideline. This procedure made dialogues by different subjects coming similar to each other. Figure 3 shows an example of dialogue between a user and the system. Prosodic focuses were placed on the underlined parts.

For experiment 1, the subjects were asked to use the system with speech replies in the two versions and to score them from the "acceptability." They were also asked to write down their impressions. (Experiment 1-a)

After about ten days, the same subjects did an evaluation experiment similar to the above, but with a reduced vocabulary size for the speech recognizer and simplified system functions. The scoring was on the "understandability." The reduced vocabulary size upgraded the recognition performance. This upgrading together with the simplification were planned to make the subject to concentrate on the speech replies. (Experiment 1-b)

When conducting the above experiments, four subjects were asked to use the system in the order of with and without focus control, while the other four were asked to use in the reversed order. This scheme was planned to reduce the influence of the order of the system use on the evaluation.

Experiment 2 was the listening test of each reply sentence speech with and without focus control. The sentences were those appeared in his/her dialogue of experiment 1. The subjects were asked to make evaluations on the "acceptability" and the "understandability." (Experiment 2)

4.2. Experiment results

Speech recognizer: Table 1 shows recognition rates of the speech recognizer for experiment 1. The left and right column indicate the results for experiment 1-a and experiment 1-b, respectively. "1st" and "2nd" mean the first and the second use of the system, while "focus" and "none" indicate the

S-1 What kind of papers are you looking for?
U-1 On speech.
S-2 Would you like to retrieve with this condition?
U-2 Yes.
S-3 Now retrieving.
S-4 As a result of retrieval, 144 papers are matched.
S-5 Do you want to add the condition?
U-3 Yes.
S-6 Enter keywords.
U-4 Speech synthesis.
S-7 Now, would you like to retrieve with this condition?
U-5 Yes.
S-8 Now retrieving.
S-9 As a result of retrieval, 8 papers are matched.
S-10 Do you need abstract for one of them?
U-6 Yes.
S-11 What number do you need?
U-7 Number 5.
S-12 Now showing abstract of number 5 article.
U-8 What is the newest paper in the list?
S-13 Number 2 and 8 issued in 1997 are the newest ones.
U-9 What journal is most frequently appeared in the list?
S-14 Journal of Acoustical Society of Japan appears twice and is the most.
U-10 Who is the authors of number 2 paper?
S-15 They are Keikichi Hirose and Goh Kawai.
U-11 What is the journal name?
S-16 (Name of) number 2 is Journal of Acoustical Society of Japan.
U-12 What is the year of issue?
S-17 Number 2 is (issued in) 1997.
U-13 How about number 5?
S-18 Number 5 is issued in 1995.
U-14 It's enough. Thanks.
S-19 Would you like to finish?
U-15 Yes.
S-20 Thank you for using.

Figure 3: An example of dialogue between user (U) and system (S). The underlined parts indicate the focus positions. The example is originally in Japanese, but is translated to English for readability.

Table 1: Speech recognition rates at Ex.1.

	Ex. 1a		Ex. 1b	
	WCR	SUR	WCR	SUR
1st	74.1	83.0	94.3	92.7
2nd	80.6	86.7	97.4	96.3
Focus	76.6	84.4	98.0	96.3
None	78.4	85.3	93.5	92.5
All	77.5	84.8	95.9	94.4

use of the system with focus control and the use without. Word correct rate(WCR) and sentence understood rate(SUR) are calculated by equation (1) and (2) are used as indices of recognition performance:

$$WCR = \frac{W_{all} - W_{sub} - W_{ins}}{W_{all}} \times 100(\%) \quad (1)$$

$$SUR = \frac{S_{und}}{S_{all}} \times 100(\%) \quad (2)$$

In the equations, W_{all}, W_{sub}, W_{ins}, S_{all}, S_{und} indicate total number of words in reply speech, number of substituted words, number of inserted words, total number of reply sentences, and number of sentences correctly understood by the system, respectively. Here "understood" means that all words necessary to extract semantic information are correctly recognized.

Evaluation of speech replies: Figure 4 shows the results of the experiments 1 and 2. The upper and lower panels indicate the results for the "acceptability" and the "understandability," respectively. The results indicated by UA and UB are those for experiments 1-a and 1-b, while S-1 and so on are those for experiment 2. "S-sp" indicates the results for sentence "Please say again," which does not appear in the example shown in Fig. 4. A-ave and B-ave are the averages of the scores for individual sentences. The results are shown as averages for experienced subjects, non-experienced subjects, and all subjects.

4.3. Considerations

Speech recognizer: In the experiment 1-a, recognition errors occurred mostly at retrieval words. When retrieval words in user's input were mis-recognized, they were repeated by the user, but still mis-recognized. This chain phenomenon largely degraded the recognition results. Higher WCR and SUR were obtained for the second use than the first use, indicating the users increased experience to the system from the first to the second uses. No clear differences were observed in the recognition rates depending on the focus control in reply speech generation. Increase in the "understandability" of reply speech may not affect speech recognition performance so much. In the experiment 1-b, all WCRs and SURs exceeded 90%, indicating our experimental set up worked as we planned; to make subjects to concentrate on the speech replies.

Experiment 1: The result (UA) of "acceptability" was almost zero both for experienced and non-experienced subjects, indicating no difference between two versions of the system, though "acceptability" for individual sentences of the experiment 2 takes rather negative values. "Understandability" score (UB) took positive values. This may indicate clearer information transmission being possible with focus control on reply speech, showing the validity of the proposed method on focus control.

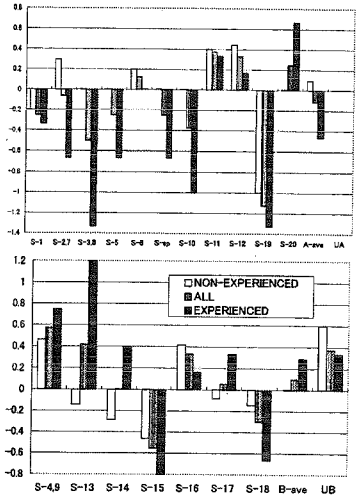

Figure 4: Evaluation results on speech replies. Upper panel shows the results by "acceptability," while lower panel does by "understandability," separately for each reply and also totally.

"Acceptability" in Experiment 2: For many sentences (S-1, S-2,7, S-3,8, S-5, S-10, and S-19), the unfocused version was judged as better in acceptability. One reason will be; most of these sentences (except S-1 and S-3,8) were those of confirmation to the user, and reply speech of focused version sounded as overstatement. As for S-6 and S-11, the focused version was evaluated higher. Several subjects make a comment; "The focused version is better, because it indicates clearly the user's next action." This results indicates that to place an emphasis on verb is effective for a smooth dialogue when the reply sentence is "guidance to user." S-12 is an example of prosodic focus being placed on the article number in the listing on display. This case, the focused version was clearly preferred.

"Understandability" in Experiment 2: For 4 types of sentences, the focused version of reply speech was judged as clearly better, while it was judged worse for 2 types. For S-4,9, most subjects made comments, such as, "Focused version is better, because we can hear the number of articles clearer." In the case of S-14, since the focus is placed on a rather long compound word, like a journal name, locating at the beginning of the sentence, the current prosodic control may over-emphasize the word. This problem can be solved rather easily, say through suppressing the accent command amplitude of the word. The focused version got a low score for S-18. This was because the complimented word "issued" sounded as emphasized. From this aspect, some modifications are necessary to the current prosodic control.

There are notable differences between results by experienced subjects and non-experienced subjects. The results by experienced subjects showed better consistency than those by non-experienced subjects. The experienced subjects tended to give extreme scores, while the non-experienced subjects tended to give zero (two versions are similar) score.

5. CONCLUSION

A spoken dialogue system on academic document retrieval was constructed with a special focus on reply speech generation. In order to realize a "good" prosodic control, we developed a scheme of converting concept of reply to speech sounds. Important words in reply sentences are emphasized by placing prosodic focuses on them. Evaluation experiments on reply speech quality showed the validity of the developed method of reply sentence generation. The experimental results including a number of comments from the subjects can be utilized to improve the system. Further improvements are also planned by introducing higher semantic processes, such as to extract abstract information from the documents and to show semantic differences between two documents.

6. REFERENCES

A. Lee et al. (1999) "Large Vocabulary Continuous Speech Recognition Parser Based on A* Search Using Grammar Category-pair Constraint," IPSJ Trans., Vol.40, No.4, pp.1374-1382 (in Japanese).

K. Hirose and H. Fujisaki, (1993) "A System for the Synthesis of High-Quality Speech from Texts on General Weather Conditions," IEICE trans. Fundamentals, Vol.1E76-A, No.11, pp.1971-1980.

K. Hirose, M. Sakata, and H. Kawanami, (1996) "Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features," Proc. ICSLP96, Vol.1, pp.378-381.

K. Hirose and S. Kiriyama, (1999) "Generation of speech reply in a spoken dialogue system for literature retrieval," Proc. ESCA Workshop on Interactive Dialogue in Multi-Modal Systems, pp.29-32.

T. Kawahara et al. (1999) "Japanese Dictation Toolkit – 1997 version –," J. Acoust. Soc. Jpn. (E), 20-3, pp.233-239.