Control of Prosodic Focuses for Reply Speech Generation in a Spoken Dialogue System of Information Retrieval on Academic Documents

Shinya Kiriyama† *Keikichi Hirose*‡ *Nobuaki Minematsu*⋆

†Graduate School of Engineering, ‡Graduate School of Frontier Sciences,
⋆Graduate School of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN
{kiriyama; hirose; mine}@gavo.t.u-tokyo.ac.jp

Abstract

We have been developing a spoken dialogue system of information retrieval on academic documents with a special focus on reply speech generation. In order to realize speech reply with its prosodic features properly controlled to express the dialogue focus, we had developed a concept-to-speech conversion scheme where the reply concept was directly converted to a sequence of phone and prosodic symbols. In our original system, however, a priority was given to the automatic processing, and the method for prosodic focus control was rather simplified. Aiming at improving the reply speech quality, new rules were constructed for prosodic focus control. Through the listening experiment, the new rules were evaluated to be revised further. The validity of the revised rules was verified through an evaluation experiment of the system.

1. Introduction

In order to realize smooth communication between men and machines, speech replies from spoken dialogue systems should be easily recognizable and understandable to users. Although, in most dialogue systems, a Text-To-Speech(TTS) conversion scheme is adopted to generate speech replies, it includes a serious problem from the above viewpoint. Unlike the case of text reading, during reply sentence generation process, the system may have rich information on the sentence to be synthesized, such as syntactic structures, discourse structures, and so on. This kind of information is tightly related to prosodic features of speech, and, therefore, prosodic control comes important in reply speech generation. However, since commercially available TTS devices usually have a very limited ability in linguistic analysis, they cannot deal with high-level linguistic information as above.

From this viewpoint, we already have realized a Concept-To-Speech (CTS) conversion scheme and succeeded in generating reply speech with its prosodic features properly controlled to express the dialogue focuses in our spoken dialogue system on document retrieval [1]. However, the rules of focusing were very simple and several problems were pointed out from users during the trial use of the system. Based on the comments from the users and experimental results, prosodic focus control rules were improved to realize a better understandability in the reply speech.

The following sections are organized as follows; first, methods of dialogue management and reply speech generation are briefly explained in section 2, then the prosodic focus control is discussed in section 3 followed by the results of evaluation experiments in section 4. Section 5 concludes the paper.

2. Speech reply generation

2.1. Dialogue management

In our system, dialogue management was conducted based on a state transition table. First, retrieval words and commands are extracted from the recognition results. Then the system operation is decided according to the table.

The system can answer not only simple questions on such as author names, years of issue and so on, but also more sophisticated ones, requiring higher semantic processing, such as questions on the number(s) of the latest (year of issue) document, on the journal name which appeared most frequently in the list, and so on. When answering these questions, their elliptic expressions should be handled. When information required to make a database access is not included in the user input, the missing information is searched in the dialogue record. If it is not found, the system asks back to the user. In order to decide the level of ellipsis of reply sentences, three levels were examined: without information element already known to the user (only with information asked by the user), with information elements not included in the last user's question, and with all the information elements including those known to the user. From the results of trial use of the system, the second level was selected.

Additionally, an efficient search function based on topic estimation was integrated to our system [2]. Topic (category) of documents which users are searching for is estimated from accumulated relevance scores between the topic and retrieval words included in user's input utterances. The system leads users properly to their goals using the estimated topic information. In the current system, the four topics are considered: 'speech (processing),' 'image (processing),' 'communication,' and 'others (not included in the above three topics).'

2.2. Reply speech generation

As mentioned already, our scheme to generate speech replies is based on a CTS conversion scheme, not on a TTS one. The contents of system reply represented by concept expressions are converted into a sequence of phone and prosodic labels via several levels of representation. Based on the current state and user's input, the abstract concept of reply sentence is first selected out of predefined seven concepts shown in Table 1. The abstract concept is converted to the sentence concept by adding answering information to the user's question.

An access to the sentence concept dictionary is conducted according to the code attached to the sentence concept to generate a prosodic phrase code sequence of the reply sentence,

Table 1: List of abstract concepts used in the system.

	Abstract concept	Example
A	Fixed style sentence	"Thank you for using."
B	Request for retrieval words	"What kind of documents are you looking for?"
C	Notification of operation	"Now showing abstract."
D	Confirmation of operation	"Do you need printed one?"
E	Instruction or guidance	"Say it again."
F	Notification of number of selected documents	"4 documents are matched."
G	Answer to user question	"The year of issue is 1997."

Table 2: An example of code expressions. 'SS' stands for 'Surface Sentence'. 'SC', 'PP' and 'WC' mean 'Sentence Concept code,' 'Prosodic Phrase code (sequence)' and 'Word code (sequence),' respectively.

SS	saN baN wa,/ hirose-keekichi, minematsu-nobuaki/ desu. "(Author names of) number 3 is Keikichi Hirose and Nobuaki Minematsu."
SC	[0721]
PP	[13020 13019 21005]
W	[(100003)(11)(4)][(301468)(305343)][(21005)]

which is then converted to a word class code sequence using the prosodic phrase dictionary. Here, a word class indicates words belonging to one category. For example, word class 'author names' consists of author names. For each word class code in the sequence, a word entry is selected by referring to the information stack. An example of these code expressions of a reply sentence is shown in Table 2. The detail of the process is explained in [1].

Finally, a word code sequence thus obtained is converted into a phone and prosodic symbol sequence which serve as direct inputs to the speech synthesis engine. During this process, prosodic focus is placed on words with important information. This prosodic focus control is based on the F_0 model, where an F_0 contour is generated as combinations of phrase and accent components, which are generated as responses to phrase and accent commands, respectively [3]. In our speech synthesis engine, prosodic symbols indicating phrase and accent commands of the F_0 model include flags of focus. For phrase command, the flag indicates whether the phrase includes an important word (a portion with focus) or not, while, for accent command, it indicates whether the prosodic word is important or not. These flags shall be called 'importance flags' in the rest of the paper. The details of prosodic control using the flag information can be found in [4].

Dialogue-like prosody was realized in our system. Its control rules were those for F_0 contours and speech rate constructed through comparative study on human dialogue speech and read speech [3],[4].

3. Prosodic focus control

The following two points should be considered to properly realize dialogue focus in reply speech: 1) where to put dialogue focus in the reply sentence, and 2) how to control the prosodic features to realize a prosodic focus. Therefore, the developed rules for prosodic focus control can be divided into two groups: one group to decide focus position in reply sentences (henceforth, focus positioning rules), and the other group to control the prosodic features of the reply speech to put the focus on the position (henceforth, focus expression rules).

In this section, our original focus control rules are first described, and their problems were revealed through the trial use of the dialogue system. Then, the new rules are explained, which were constructed under the following guidelines: 1) to deal with reply forms newly added to the system, which is necessary to proceed dialogues based on the estimated dialogue topic knowledge (see section 2.1), and 2) to solve the problems of the original rules by taking the knowledge obtained through the trial use of the system into account[1].

3.1. Original rules

The original focus positioning rules are constructed related to the abstract sentence concepts in Table 1. They consist of the following three rules: 1) When the abstract sentence concept is 'notification of number of selected documents'(**F**), place a focus on the number of selected documents. When it is 'answer to user question'(**G**) place a focus on words conveying the answering information. 2) When the abstract sentence concept is 'notification of operation'(**C**) and the sentence concept includes document numbers, place a focus on them. 3) For other cases, place a focus on the verb of predicate.

The original focus expression rule was very simple. For all focuses placed, both of the importance flags in phrase and accent commands were set 'ON'.

3.2. Problems in the original rules

The following problems came clear for the original rules through the trial use:

1. Not a few reply sentences sounded unnatural when dialogue focuses were placed on verb of predicate (problem for rule 3). The followings are comments from the users.
   - Reply speech of 'notification of operation'(**C**) sounded unnatural.
   - Reply speech of 'confirmation of operation'(**D**) for which users might answer "Yes" or "No," sounded over-emphasized.
2. Reply speech of 'answer to user question'(**G**) occasionally sounded unnatural when F_0 gap between focused and unfocused parts was large. The gap came larger when the focused part (words with answering information) was rather long and/or located at the sentence beginning. The followings are comments from the users.
   - Journal names sounded over-emphasized when they were at the sentence beginning. This is because journal names compose a rather long phrase.
   - When two or more author names were enumerated and when they were focused, the reply speech came less comprehensive, because of large F_0 movements.

The above problems 1 and 2 can be solved by modifying focus positioning rules and focus expression rules as indicated in the next section, respectively.

3.3. Construction of new rules

Focus positioning rules: One or more focuses are placed for each reply sentence according to the original focus positioning rules. However, the new rules place focuses only on selected sentences. After dividing all the reply sentences into two

B1 **What kind of** documents are you looking for?
Spectrum.
D1 Would you like to retrieve with the current condition?
Yes.
C1 Now retrieving.
F1 As a result of retrieval, **82 documents** are matched.
F2 In the field of **'speech'**, **24 documents** are matched.
D3 Would you like to limit the retrieval field on **'speech'**?
Yes.
C2 The retrieval field is set to **'speech'**.
D4 Do you need to reduce the number?
Yes.
E1 **Enter** the condition.
Noise.
D2 Now, would you like to retrieve with this condition?
Yes.
C1 Now retrieving.
F1 As a result of retrieval, **8 documents** are matched.
D5 Do you need abstract shown for one of them?
Yes.
D6 **Which number's** abstract do you need to be shown?
Number 3.
C3 Now showing abstract of **number 3**.
Which is the newest document in the list?
G4 **Number 1 and 4 issued in 1997** are the newest ones.
What (journal) is most frequently found in the list?
G5 **Journal of Acoustical Society of Japan appears twice** and is the most.
(a recognition error.)
E2 Please say again.
What are the author names of number 2?
G1a **Keikichi Hirose and Nobuaki Minematsu**.
Say it again.
G1c The author names of number 2 are **Keikichi Hirose and Nobuaki Minematsu**.
What is the journal name?
G2b Number 2 is in **Journal of Acoustical Society of Japan**.
What is the year of issue?
G3b Number 2 is **1995**.
How about number 5?
G3b The year of issue is **1995**.
It's enough. Thanks.
D7 Would you like to finish?
Yes.
A1 **Thank** you for using.

Figure 1: An example of dialogue used in the listening test. The underlined parts and the parts indicated by bold letters mean focus positions by the original rules and by the new rules, respectively. The sentences indicated by italic letters are user's utterances.

groups, 'notification' and 'guidance,' the sentences necessary with the focus control are selected. For each selected sentence, the focus position is determined. The rules can be summarized as follows:

1. When the reply sentence includes a word on the document topics, place a focus on it.
2. When the concept of reply sentence is 'notification' (concept **C** or **F** in Table 1), place a focus on the words representing the content of notification.
3. When the concept of reply sentence is 'confirmation' (concept **D**, except for the sentences for which users might answer "Yes" or "No") or 'instruction' (concept **E**), place a focus on the words representing the content of guidance.
4. No focus is placed for other cases.

Rule 1 is an additional rule related to the topic estimation function, which was included into the system after the former experiments. Therefore, the rule is not based on the users' comments. Other rules are those modified from the original rules to cope with the problems of reply speech indicated in section 3.2.

Focus expression rules: The focus expression rules were modified as follows:

I Set the importance flag 'OFF' for phrase command when the important word included in the phrase becomes longer than a threshold (10 morae in the current paper). These words are found in 'journal names' and 'year of issue.'

II Set the importance flag 'OFF' for phrase command when two or more author names are enumerated in the phrase.

The above rules are to solve the problem that F_0 contours of the focused words become higher when they are long. The high F_0 contours are considered to be one of the reasons of unnaturalness. Although, according to the users' comments, the threshold of rule I should be set smaller if the word locates at the beginning of the sentence, which was not tested here. This is because to avoid the rules come complicated.

4. Evaluation experiments

4.1. Listening test

The experiment was conducted by using 15 subjects. An example of typical dialogue shown in Figure 1 was offered to the subjects before the listening test so that they could know the acceptable utterances to the system. They were asked to compare the two versions of reply speech with identical content but with different prosodic focusing. The example of dialogue in Fig.1 includes almost all the sentence concepts of the system, except those of abstract concept **G**. Since there are a number of sentence concepts belonging to **G**, only a few examples are listed in the figure. All the reply sentences appeared in the figure and all the possible sentences of **G** were checked in the experiment.

Three versions of speech were synthesized for each test sentence (reply sentence): **P**) using the new rules (new version), **Q**) using the original rules (original version), **R**) without focus control (unfocused version). Out of these three versions, combinations of '**P** and **Q**' and '**P** and **R**' were selected and used for the comparison by the subjects. When both versions in a combination had the same prosodic features, such a combination was not included in the evaluation. The ratio of the excluded combinations to all was 15%. The comparison was done by the 5-rank scoring, from the viewpoint of 'understandability' (when the abstract concept of reply sentence was **F** or **G**) or from the viewpoint of 'acceptability' (otherwise).

In our system, two types of speech synthesizers are available: formant synthesizer [3] and waveform concatenation synthesizer [5]. The former experiment [1] was conducted using the formant synthesizer, however, the current experiment was

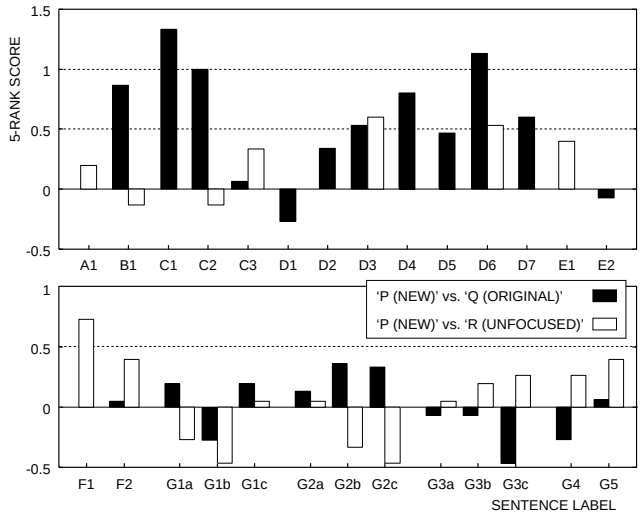

Figure 2: Results of 'acceptability' and 'understandability' of reply speech synthesized by the new rules.

done by the waveform concatenative synthesizer.

The results are shown in Figure 2. The 5-rank scores (the values are -2, -1, 0, 1 and 2) are averaged over the subjects. The score '0' means that there are no difference between the two versions. Positive values mean that **P** is better. Negative values mean that **Q** or **R** is better. The reply sentences from '**A1**' to '**E2**' are evaluated by 'acceptability,' and the others are done by 'understandability.' The sentence labels correspond to the example of dialogue in Fig.1. The characters '**A**' to '**G**' represent those in Table 1. The characters '**a**,' '**b**' and '**c**' attached to '**G1**', '**G2**' or '**G3**' mean that the number of compensated phrases in the reply sentence is 0, 1 and 2, respectively.

'Acceptability': As for almost all reply sentences, the new versions were judged to be better. Most subjects mentioned that the original versions sounded more unnatural. For **D1**, the original version with focusing was judged better, while, for **D2**, the new version without focusing (equal to the unfocused version) was preferred. One possible reason will be that **D1** appears in the dialogue earlier than **D2**. If a sentence with similar contents is repeated, it should not be focused. The reply **C2** comes after the user's confirmation to **D3** speech. Since the word 'speech' representing the topic already appeared in **D3**, focusing it again in **C2** was evaluated low.

'Understandability': As for 'notification of number of selected documents,' both of the new and original versions were supported. Here, considerations are necessary for the replies belonging to 'answer to user question.' For **G1**, when the sentence includes no or only one compensated phrase (**G1a**, **G1b**), the unfocused version was preferred the most. When the two compensated phrases "The author names" and "of number 2" were added on the top of the sentence (**G1c**), the new version was evaluated as best among the three versions. These results indicate that rule II in section 3.3 is operating as expected for the reply **G1c**. For **G2, G5** in which journal names are included, the new version was supported only when the journal name appeared at the sentence beginning. Other cases, the unfocused version was preferred. These results indicate that rule I in section 3.3 should be made active only for the first case. The results for **G3, G4** indicate that prosodic focus control related to the year of issue should be done by the original rules described in section 3.1.

4.2. Revised rules

Based on the results of listening test in section 4.1, the following rules were added to the new focus positioning rules: 1) Turn off the focus flag placed on the known information, 2) Place a focus on the words for confirmation in the sentence of 'confirmation of operation' for their first appearances, 3) Do not place a focus on proper nouns in sentences of 'answer to user question,' when they are journal names appearing at the middle of sentences, or when they are author names and the sentences has only one compensated phrase. Rules 1) and 2) are those related to dialogue flow control. Further rules were also added as modifications on rules 3 and 4 in section 3.3, whose details are not explained here.

As for the new focus expression rules in section 3.3, the experimental results indicated that the rule 1 should be limited to the journal names at the sentence beginning.

4.3. Evaluation using the system

8 subjects were asked to use the two versions of the system: one speech reply by the revised rules (revised version), and the other that by the original rules (original version). The two versions were totally compared after the use of the system. The 5-rank scoring scheme was adopted again on 'acceptability' and 'understandability.' The scores averaged over 8 subjects were +0.50 and +0.13 for acceptability and understandability, respectively. The results proved that our revisions of the rules for prosodic focus control were valid. No subject preferred the original version on 'acceptability,' however, a few subjects supported the original version on 'understandability' saying that the original version has clearer intonation. This point (difference in user's preference) was left for the future study.

5. Conclusions

In order to realize better speech reply in our spoken dialogue system, rules for focus positioning and prosodic control were improved according to the result of listening experiments. The system with the modified rules showed improvements in acceptability and understandability. A flexible focus control scheme enabling reply speech generation according to the user's preference will be studied.

6. References

[1] S. Kiriyama; K. Hirose, 2000. Development and evaluation of a spoken dialogue system for academic document retrieval with a focus on reply generation: *Proc. SST2000*, 32-37.

[2] S. Kiriyama et al., 2001. Use of Topic Knowledge in Spoken Dialogue Information Retrieval System for Academic Documents: *Proc. Eurospeech2001*, (2), 1315-1318.

[3] K. Hirose; H. Fujisaki, 1993. A System for the Synthesis of High-Quality Speech from Texts on General Weather Conditions: *IEICE trans. Fundamentals*, E76-A(11), 1971-1980.

[4] K. Hirose et al., 1996. Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features: *Proc. ICSLP96*, (1), 378-381.

[5] http://tcts.fpms.ac.be/synthesis/mbrola.html