An improved method of generating speech from concept and its application to a dialogue system of road guidance

Yuji Yagi[†], *Seiya Takada*[‡], *Keikichi Hirose*[††] *and Nobuaki Minematsu*[‡]

[†]Graduate School of Engineering, [‡]Graduate School of Frontier Sciences
[††]Graduate School of Information Science and Technology
University of Tokyo, Japan
{yagi, stakada, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A method of concept-to-speech conversion was developed and applied for the reply speech generation in a spoken dialogue system on road guidance. The method is an improved version of our formerly developed one, where a sentence with necessary information for speech synthesis is generated from the concept of reply and converted to a reply speech. By handling concept in phrase unit instead of sentence unit, an increased variety in generated sentences was realized. In order to realize the concept-to-speech conversion, linguistic information was handled keeping the syntactic structure throughout the process. Several improvements are also added to prosodic control in speech synthesis. The method was applied to a spoken dialogue system, where a user was guided by the system to reach a place marked on a map through conversation. Reply speech was evaluated mainly from the viewpoint of prosody through a listening test. The result clearly indicated a better prosodic control for the newly developed method as compared to the original method.

1. Introduction

Speech is known to be the most basic and important method of communication for humans, and, therefore, there is an increasing request to a scheme enabling man and machine interaction through speech. Corresponding to this request, a number of spoken dialogue systems have been developed. However, research works on speech output generation are rather few, and in most systems, text-to-speech (TTS) conversion devices are used for generating speech replies. During the process of reply sentence generation, the system has higher-level linguistic information of the generated sentence, such as its syntactic structure, important words carrying key information of the reply content, and so on. This information should be reflected on (prosody of) reply speech. However, this is rather difficult when we utilize commercially available TTS devices: a unified scheme of generating reply speech from the content of reply is necessary. Although this scheme was named concept-to-speech (CTS) conversion with a guide to its realization by Young and Fallside more than 25 years ago[1], works on its realization were rather limited. As for Japanese, no systems with CTS were reported other than those by the authors[2, 3]. Recently, we have developed an agent system and, in the system, realized a CTS conversion, where the syntactic structure of generated sentences and the position of important words were utilized to control prosodic features[4, 5]. In the system, an agent (stuffed bear) walks around in a virtual room constructed on a computer (display) to perform a task given from a user. When the agent finds some difficulties, it asks the user for help. To realize the CTS, syntactic structure of the user's input is kept and is utilized in the CTS process. Result of the listening test on the system's reply speech indicated that the prosody was properly controlled in the speech synthesis.

However, in the system, concept of reply is handled in sentence unit, which limits the variety of generated sentences; only one sentence style can be generated from a concept. The dialogue should proceed differently depending on the user's personality, and a concept need to be represented by a compound sentence, a complex sentence, or a set of simple sentences according to the dialogue situation. To realize this, we have developed a new scheme of CTS conversion, where a concept is handled in phrase units and summed up to generate a sentence. Henceforth, the newly developed CTS scheme is denoted as the new method while the previous one is denoted as the original method.

Since the dialogue in the agent system is limited to a simple one, it is not appropriate to show the validity of the new method. Therefore, a spoken dialogue system was newly constructed. It is a road-guidance system, where a user is guided by the system through spoken dialogue to reach a place marked on a map. The user can only have a short view around his/her location, and, thus, mis-understanding between the user and the system may occur. This situation requires an extended dialogue with a wide variety of sentence styles.

In order to generate reply speech, which is easy to be understood by users, higher-level linguistic information of the generated sentences need to be well reflected on the prosodic features of speech. To realize this situation, we adopted the F_0 contour generation process model (F_0 model)[6] for the control of F_0 contours of reply speech. The phrase and accent commands of the model are known to have a good correspondence with the linguistic information, and symbols representing them are inserted in the sentences according to the result of F_0 contour analysis of dialogue speech[7].

The rest of the paper is constructed as follows: Section 2 describes the sentence generation in the original and the new methods. After explaining the prosodic control in section 3, the dialogue system on road-guidance is explained in section 4 with an example of dialogue with a user and the result of the listening experiment. Section 5 concludes the paper.

2. Sentence generation

2.1. Sentence generation in the original method[4, 5]

In order to realize CTS conversion, generated sentences should keep higher-level linguistic information such as its syntactic structure and role of its constituting words in the dialogue with the user. To keep the syntactic structure throughout the sentence generation process, all the concepts are represented in LISP forms with tags. Using this form, for instance, the concept of putting an item in a position can be written as:

(oku $PRED(o($ITEM))(ni($POS)))

Here, the tags $ITEM, $POS and $PRED depict an item, a position and a predicate, respectively. The $PRED tag means that the word "oku" works as a predicate when placed in a sentence.

Given a template (concept) frame in tag LISP form, a sentence (with syntactic structure) is generated by pasting words at tag positions. Words conveying important information are decided by referring to the tags and the preceding user utterance. The conjugation form of each content word in the sentence can be decided by the succeeding particle's identity. Therefore, the conjugation form can be controlled in a simple way when concatenating words according to the syntactic structure. For instance, the phrase "oite" is generated from the structure "(te(oku))."

2.2. Sentence generation in the new method

Although CTS conversion was realized by the method above, one template frame form could only generate one sentence in a style designated in the frame. We should say that the method still remained in the framework of filling words in slots of sentence templates. To solve this situation, the method is modified to accept phrases (in LISP form). (In the original method, phrases were allowed, but they were limited to special cases only.) This modification makes it possible to handle a concept with units smaller than a sentence; first generate phrases by inserting words to tag positions of phrase template frames and then concatenated them as designated in a sentence template frame, which is also represented in the LISP form. With this procedure we can realize various styles in generated sentences, not limited to simple sentences, but also to complex/compound sentences. For instance, a sentence "migini magatte ekimade ittekudasai (Turn right and go to the station.)" is generated through the following process:

1. Generate the noun phrase "(ni(migi))" from the frame "(ni($DIRECTION))".
2. Generate the noun phrase "(made(eki))" from the frame "(made($LANDMARK))".
3. Generate the verb phrase "(te(magaru(ni(migi))))" and "(te(iku(made(eki))))" from the frame "(te($VERB($NOUN_PHRASE)))".
4. Concatenate these two verb phrases to generate a (long) phrase "((te(magaru(ni(migi))))(te(iku(made(eki)))))."
5. Insert the phrase at "$VERB_PHRASE" position of the frame "(kudasai($VERB_PHRASE))" to generate "(kudasai((te(magaru(ni(migi))))(te(iku(made(eki))))))."

The generated sentence by the above process is a compound sentence, but it can be a set of two short sentences "migini magatte kudasai. soshite ekimade ittekudasai (Turn right. Then go to the station.)," if we slightly modify the process. The first sentence "migini magatte kudasai." is generated with a step similar to step 5: insert "(te(magaru(ni(migi))))" at "$VERB_PHRASE" position of "(kudasai($VERB_PHRASE))." The second sentence "sorekara ekimade ittekudasai" is also generated similarly, but a conjunction "soshite" is added to show the relation of the sentences.

3. Control of prosodic features

3.1. phrase/accent command symbols

The generated sentence should include information necessary for speech synthesis. For this purpose, the final sentence should be not only in the orthographic text form, but also in a form of a sequence of phone and prosodic symbols. The prosodic symbols are those indicating magnitudes/amplitudes of phrase/accent commands of the F_0 model. When all the command values are assigned, the model calculates the sentence F_0 contour. The symbols and the rules to assign them in a sentence were those formerly developed through the analysis of F_0 contours of dialogue speech by the multiple linear regression method[7]. Given "importance of word" and syntactic structure, the prosodic symbols are selected and inserted into the phone symbol string. For instance, the symbol sequence for the sentence "hidarie magatte jiNjamade ittekudasai (Turn left and go to the shrine.)" is given as follows:

P111212 hi F311 da ri e ma ga sx te A0 P11 D311 zi
A0 n zja ma de P21 i F413 sx te ku da sa A0 i P0 S1

Here, the symbols starting with P show the phrase command (onset) locations and magnitudes. Accent command (onset) locations and amplitudes are shown by the symbols starting with D and F: D for accent type with accent nucleus and F for one without. The digits included in these symbols indicate to which class each item of multiple linear regression analysis belongs.

Since the phrase commands show different features depending on their locations in the sentence, the digits after P are differently assigned for the two cases; top of the (prosodic) sentence and middle of the sentence. The sentence initial phrase symbol has 6 digits, which indicate the yes-no answers to the following features of the sentence: opens FRD, contains an important word, changes the topic, follows to a conjunction, covers 7 morae or less, and ends with particle "ka." Here, FRD is the abbreviated form of "Fundamental Routine of Dialogue" and denotes a pair of user and system utterances, which are directly related to each other, such as a question and an answer. As for the in-sentence symbols, their first and second digits correspond to the second and the fifth digits of the sentence initial symbols. The in-sentence symbols are positioned in the phone string corresponding to the right branching syntactic boundaries, which are found easily by tracing the LISP form. The detail is given in section 3.2.

As for the accent commands, 3 digits included in the symbol indicate importance and novelty of the content word, position in the phrase, part of the speech of the content word, respectively. For each accent phrase, a symbol is selected according to its accent type and is inserted into the phone string at the position corresponding to the accent command onsets. The detail is given in section 3.3.

Symbols P0 and A0 are those indicating the ends of the phrase and accent commands started by the preceding symbols, respectively. Pauses are placed at the symbols starting with S. Symbol S1 corresponds to a long pause between two sentences.

3.2. Positioning of phrase command

In the original system, a sentence initial symbol is simply placed at the beginning of a sentence, while an in-sentence symbol is inserted at the right branching syntactic boundaries. This algorithm included a problem of "too" long phrase components, when left branching syntactic boundaries succeeded without right branching boundaries. In the new system, when a phrase component exceeds 12 morae, an additional phrase command is placed at the boundary where concatenation of preceding and succeeding words is loosest. The strength of concatenation is calculated as the following word bi-gram:

$$P = \frac{f(w_2, w_1)}{f(w_1)}$$

where $f(w)$ means frequency of w, and w_1, w_2 denote the preceding and succeeding words, respectively. The bi-gram is calculated for the Mainichi Newspaper corpus of the year 1997.

3.3. Positioning of accent command

Given the accent types of accent phrases, F or D symbols representing accent command onsets are inserted in the phone string: at the top of the accent phrase for type 1 accent and between the first and second morae for other types. (F symbols correspond to type 0 accent and always placed between the first and second morae.) Here, an accent phrase is defined as a sentence unit consisting of a content word and its following particle(s). In the tag LISP form, it corresponds to a unit with a tag, delimited by a set of parentheses. An accent type is assigned for each accent phrase by referring to the accent type dictionary. The dictionary has accent type and attribute information (for each word), and, using a system developed by the authors[8], the accent type can be automatically. Symbol A0 represents the accent command end and is placed immediately after the accent nucleus mora. For an accent phrase with type 0 accent, which has no accent nucleus, symbol A0 is placed at the end of accent phrase.

In Japanese, when no phrase command appears between two accent phrases, their two accent commands interact to each other. The second digit of accent command symbol is added to cope with this interaction. However, when the first accent phrase has type 0 accent, the two commands concatenate to produce a new command[9]. To cope with this phenomenon, the following concatenation rule is applied recursively from the top of the sentence before reaching a phrase command symbol:

1. For a sequence of F and D symbol, generate a new D symbol with its accent nucleus coinciding with the original D symbol.
2. For a sequence of two F's, generate a new F symbol.

4. Outline of the dialogue system

In order to show the validity of the new method, a spoken dialogue system was constructed[10]. This is a road-guidance system, where a user is guided by the system through a spoken dialogue to reach a place marked on a map. Fig. 1 shows an example of the map arranged for the system. The system has the full map and knows all the places shown as square symbols. Also it knows the distance between two places as a number attached to each path. Rectangular symbols are the signs of "road works (where passing is not allowed)." Such temporal information is not given to the system. On the other hand, the user can only know the start point and view a short distance around his/her current location, which is shown as a circle in Fig. 1. Because of limited information provided to the user, and lack of temporal information for the system, mis-understanding may occur between them. Also since the user's location is provided to the system only through the dialogue, the system may sometimes wrongly locate the user in the map. These situations require the system to generate reply speech in various contents and styles.

The system consists of a speech recognizer, a syntax analyzer, a dialogue manager, and a speech synthesizer, together with a display controller showing the fragment of the map near the current location of the user (circled portion of Fig. 1). The recognizer receives the speech input and converts it into a word string. The grammar-based version of speech recognition software, Julian was used[11]. The syntax analyzer outputs the syntactic structure of the word string through morpheme and syntactic analyses. The morpheme analysis result is obtainable as the output of Julian. Syntactic analysis is conducted by a simple rules developed by the authors. The dialogue manager first extracts information on the user's current situation (such as current location) from the user's speech input, and, then, sends the user an instruction to reach the destination. It also generates reply content and converts it into a string of prosodic and phone symbols. The speech synthesizer generates output speech from the string. The speech synthesis is based on a waveform concatenation with TD-PSOLA prosody modification.

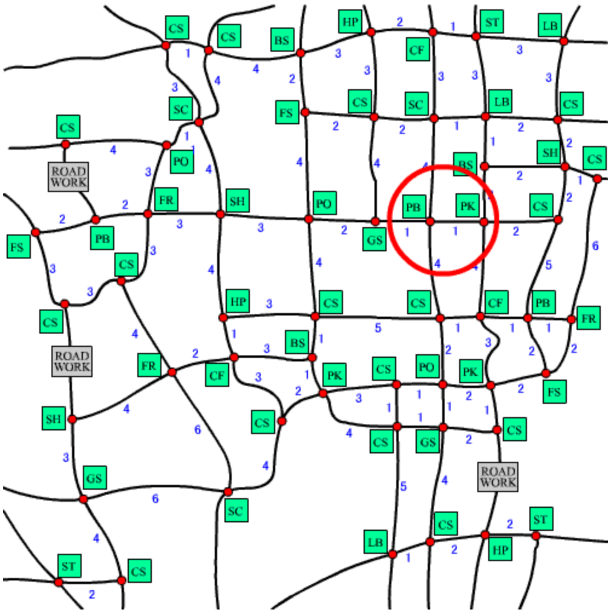

Figure 1: *An example of the map. Square symbols with two letters inside show places, which serve as landmarks in the dialogue between the user and the system. For instance, CS denotes convenience store, SH denotes shrine, and so on.*

After finding out the best (shortest) route to the destination, the system guides the user to follow it. When the system detects through the conversation with the user that the user deviates from the route, it re-calculates the best route from the user's current position (to the destination) and guides the user according to the new route. Depending on the current situation of the user estimated through the dialogue, the system selects one of the following processes to re-calculate the route to the destination:

1. Search around the previous position of the user.
2. Search around the user's current position the system estimates.
3. Ask the user the place(s) where he/she passed before reaching the current position. Then search the entire map where all the places (places related to the previous and current positions, places he/she passed) match.

In the following dialogue example, the words expressing directions and places (underlined) are assumed to be important.

Also, words in italic are novel (1st appearance in the dialogue). These words are focused using prosodic control method in section 3. In the following example "U" and "S" denote user's utterance and system's reply, respectively.

S: hidarie magatte *koNbini*made ittekudasai (Turn left and go to the convenient store.)

U: koNbinini tsukimashita ((I) reached the convenient store.)

S: hidarie magatte *jiNja*made ittekudasai (Turn left and go to the shrine.)

U: gasoriNsutaNdoni tsukimashita ((I) reached the gas station.)

S: jiNjawa mada sakidesune (The shrine seems to be still ahead.)

S: massugu jiNjamade ittekudasai (Go straight to the shrine.)

U: massugu susumemaseN ((I) cannot go straight.)

S: michiwa arimasuka (Does the road exist?)

U: koojichuudesu (It is under construction.)

S: migie magatte *kooeN*made ittekudasai (Turn right and go to the park.)

U: kooeNni tsukimashita ((I) reached the park.)

Using the new method in section 2, the above dialogue can proceed differently. For instance, the first FRD of the dialogue becomes as follows, if simple sentences are adopted as reply speech instead of compound sentences:

S: hidarie magattekudasai (Turn left.)

U: hai (Yes.)

S: soshite *koNbini*made ittekudasai (Then, go to the convenient store.)

U: koNbinini tsukimashita ((I) reached the convenient store.)

Although this style increased the number of FRD's, it will be preferred by users without previous experience of using the system.

A listening test was conducted on the reply speech to show the validity of the new method explained in section 3. Eight sentences were selected and their speech was synthesized by the original and new methods (totally 16 samples). Eighteen Japanese speakers were asked to select one from the two versions of synthetic speech for each sentence during the listening experiment. The selection was done changing the order of two versions: the version by the original method coming first and the version by the new method coming first. Score +1/-1 was assigned when the version by the new/original method was selected. When the selection was difficult, score 0 was assigned. The score for each sentence was averaged over 18 informants, and the result was, 1.00, 1.00, 0.89, 0.50, 0.56, 0.17, 0.78 and 0.78. The score 1.00 means that all the 18 informants selected the version by the new method. The low score for the 6th sentence may indicate that the effect of the new method did appear in the actual prosodic control. As a whole, we can conclude the result showing the advantage of the new method.

5. Conclusions

In order to increase the variety of generated sentences in concept-to-speech generation, a method of handling concepts in a unit of phrase instead of sentence was developed. Each phrase concept is represented as a frame in tag-LISP form. The generated phrases are concatenated to produce a sentence. Also, a better control of prosodic features using higher-level linguistic information was realized. A spoken dialogue system of road-guidance was constructed to show the validity of the new CTS conversion method. The listening test showed the better prosodic control by the new method.

For the future work, adaptability of the developed CTS conversion method to other sentence styles will be examined in detail. Also the improved system is planned after checking the usability of the current system. As for the prosodic control, we are planning to include "commonsense" on the dialogue topic of the user when assigning importance/novelty to each content word.

6. References

[1] Young, S. J. and Fallside, F., "Speech Synthesis from concept : A method for speech output from information systems," J. Acoust. Soc. Am., vol.66, no.3, pp.685-695, 1979.

[2] Asano, Y. and Hirose, K., "A dialogue processing system for speech response with high adaptability to dialogue topics," IEICE Trans, Information and Systems, Vol.E76-D, No.1, pp.95-105, 1993.

[3] Kiriyama, S. and Hirose, K., "Development and evaluation of a spoken dialogue system for academic document retrieval with a focus on reply generation," Systems and Computers in Japan, Vol.33, No.4, pp.25-39, 2002.

[4] Hirose, K., Tago, J. and Minematsu, N., "Speech generation from concept for realizing conversation with an agent in a virtual room," Proc. EUROSPEECH, Geneva, Vol.3, pp.1693-1696, 2003.

[5] Yagi, Y., Hirose, K. and Minematsu, N., "Improvement of response generation in a spoken dialogue system focused on prosody," Record of Spring Meeting, Acoustical Society of Japan, Vol.1, 3-8-7, pp.135-136, 2004. (in Japanese)

[6] Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Japan (E), vol.5, no.4, pp.233-242, 1984.

[7] Hirose, K., Sakata, M. and Kawanami, H., "Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features," Proc. International Conf. on Spoken Language Processing, Vol.1, pp.378-381, 1996.

[8] Minematsu, N., Kita, R. and Hirose, K., "Automatic estimation of accentual attribute values of words to realize accent sandhi in Japanese text-to-speech conversion," Proc. IEEE 2002 Workshop on Speech Synthiesis, Santa Monica, 2002.

[9] Hirose, K. and Fujisaki, H., "Accent and intonation in speech synthesis," J. of IEICE, vol.70, no.4, pp.378-385, 1987.

[10] Yagi, Y., Takada, S., Hirose, K. and Minematsu, N., "A method of response generation for a spoken dialogue system," Record of Spring Meeting, Acoustical Society of Japan, Vol.1, 3-5-14, pp.653-654, 2005. (in Japanese)

[11] Kawahara, T. et. al., "Product software of continuous speech recognition consortium -2001 version-," SIG Notes, Information Processing Society of Japan, 2002-SLP-43-3, pp.13-18, 2002. (in Japanese)