A Framework of Reply Speech Generation for Concept-to-Speech Conversion in Spoken Dialogue Systems

Seiya Takada[1], *Yuji Yagi*[2], *Keikichi Hirose*[1] *and Nobuaki Minematsu*[3]

[1]Graduate School of Information Science and Technology, [2]Graduate School of Engineering
[3]Graduate School of Frontier Sciences, University of Tokyo, Japan
{stakada, yagi, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

Due to recent advancements in speech technologies, a large number of spoken dialogue systems have been constructed. However, since most of them adopt existing text-to-speech synthesizers, it is rather difficult to reflect the linguistic information obtained during the reply sentence generation well in output speech. A framework is necessary for correctly reflecting higher-level linguistic information, such as syntactic structure and discourse information. We have constructed a spoken dialogue system on road guidance and realized concept-to-speech conversion, where output speech is generated in a unified process. Tag LISP forms keep the syntactic structures throughout the process in order to reflect the linguistic information in the prosody of output speech. Furthermore, by making it possible to insert not only words but also phrase templates in tags, various sentences were generated with a minor increase of templates. Validity of the methods is shown through experiments.
Index Terms: spoken dialogue system, concept-to-speech conversion

1. Introduction

Speech is known to be the most basic and important method of communication for humans, and therefore there is an increasing request for a scheme enabling man and machine interaction through speech. Insresponse 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. Such information should be reflected in the (prosody of) reply speech. However, this is rather difficult when we utilize commercially available TTS devices: a unified scheme of generating reply speech from concept of reply is necessary. Although this scheme was proposed more than 25 years ago and named as concept-to-speech (CTS) conversion by Young and Fallside[1], works on its realization were rather limited. As for Japanese, although a number of spoken dialogue systems have been developed, CTS conversion was not addressed except in the researches by the authors[2, 3]. In the agent dialogue system, where an agent (bear) in a virtual room is instructed to do a job, a scheme to keep syntactic structures and to assign important words throughout the sentence generation process was created[4]. Since the dialogue conducted in the agent system is limited to a simple one, we have newly constructed a spoken dialogue road guidance system, where a user is guided by the system through spoken dialogue to reach a place marked on a map[5]. In the system, a new method of sentence generation from concept was created: to handle a concept in phrase unit and to sum them up to a sentence. By doing so, style flexibility was added to generated sentences.

In order to generate reply speech which is easily understood by users, higher-level linguistic information needs to be well reflected in the prosody of speech as mentioned already. To accomplish this, 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[8].

The rest of the paper is constructed as follows: Section 2 describes the overview of the spoken dialogue road guidance system. After explaining the method of sentence generation[5] and the prosodic control in section 3, result of listening experiment is shown in section 4. Section 5 concludes the paper.

2. Outline of the dialogue system

The system has a full map, while the user can view a short distance around his/her current location. Figure 1 shows 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. The system knows all the places shown as square symbols. Also it knows the distance between two places as the number attached to each path. The circle of Fig. 2 shows the range viewable by the user. The user also knows the start point (coordinate: X=50, Y=470, for instance), and coordinate of the current location. The map legend (names of two letters in square symbols) is displayed at all times to the user. The rectangular symbol in the circle with "ROAD WORK" indicates that road work is going on and the user cannot pass through. Such temporal information is not given to the system. 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 incorrectly 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.2). It

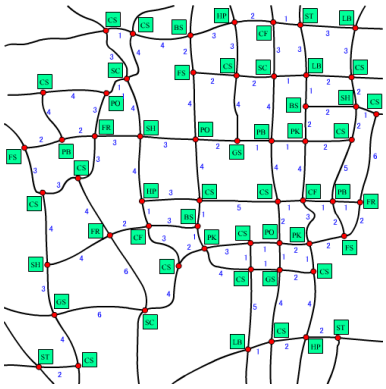

Figure 1: An example of full map, which the system has.

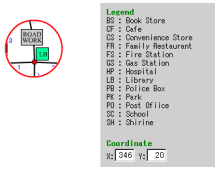

Figure 2: An example of range viewable for the user (left hand side), and map legend.

also has three dictionaries; a word dictionary, part-of-speech dictionary, and conjugation dictionary, which are necessary for dialogue management. The recognizer receives the speech input and converts it into a word string. The grammar-based version of speech recognition software, Julian[9], was used in this system. 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 rule 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 on how 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[10].

3. Reply generation

3.1. Linguistic information processing

In order to realize CTS conversion, generated sentences should keep higher-level linguistic information such as syntactic structure and role of its constituting words during the speech synthesis process. To realize this condition, all the concepts are represented in LISP forms. We adopted LISP forms, because they can easily keep the syntactic structure of a sentence by grouping words using parentheses. In order to represent a concept, the LISP form used here has tags as its elements instead of words/phrases. A tag corresponds to words/phrases with similar meanings or roles in sentences. Henceforth, this type of LISP form is called tag LISP form. Tags have some attributions such as "importance" and "novelty," which should be reflected in the prosody of output speech.

3.2. Sentence generation

Given a phrase template in tag LISP form, a phrase (with syntactic structure and information on important words) is generated by pasting words/phrases at tag positions of the template. Then the generated phrases are concatenated as designated in a sentence template, which is also represented in the tag LISP form. Words are assumed to be important when they are pasted at tags representing places and directions.

Through this procedure, we can create 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 processes:

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 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 sentence generated by the above process is a compound sentence, but it can be a set of two short sentences "migini magatte kudasai. sorekara 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 "sorekara" is added to show the relation of the sentences.

The flexibility in the style of generated sentences comes from the use of phrase templates as well as sentence templates.

3.3. Comparisons with existing methods

Although, our research using CTS, our research is exclusive to Japanese spoken dialog systems, there exist some related previous works such as template-based approach[7] in Dutch. In the research, each slot in the templates allows a limited number of words, and the variety of reply sentences depends directly on the number of templates. One of the distinguished differences between our method and their approach is the use of phrase templates. The introduction of phrase templates allows for a variety of resulting sentence templates and overcomes the inflexibility of rigid sentence templates.

3.4. Control of prosodic features

In order to create CTS synthesis, it also requires developing the speech synthesis system which reflects all the tag Lisp style information given in 3.2.

3.4.1. 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[8]. Given "importance of word" and syntactic structure, the prosodic symbols are selected and inserted into the appropriate positions of 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. The magnitude/amplitude for each phrase/accent command symbol is given by accessing the table arranged formerly by the authors[8].

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. Table 1 shows meaning of the 6 digits (items) included in the sentence initial phrase symbols. In Table 1, 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.

As for the accent commands, 3 digits have the meaning as indicated in Table 2. "Important" and "novel" are for the content word included in the accent phrase. 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. Details are given in 3.4.3.

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.

Table 1: Phrase command symbols.

Digit	Value	Meaning
1st digit	1	Opening FRD
	2	Closing FRD
2nd digit	1	Contains an important word
	2	Contains no important word
3rd digit	1	Changing the topic
	2	Keeping the topic
4th digit	1	Following to a conjunction
	2	Not following to a conjunction
5th digit	1	Covers 7 morae or less
	2	Covers 8 morae or more
6th digit	1	Ends with particle "ka"
	2	Ends without particle "ka"

Table 2: Accent command symbols.

Digit	Value	Meaning
1st digit	1	Important and novel
	2	Not important but novel
	3	Important but not novel
	4	Not important and not novel
2nd digit	1	At the phrase initial position
	2	Not at the phrase initial position
3rd digit	1	Noun
	2	Verb
	3	Adjective/Adverb
	4	Demonstrative/Interrogative pronoun
	5	Conjunction

3.4.2. Positioning of phrase command symbols

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, which are found easily by tracing the LISP form. This algorithm included a problem of too long phrase components, when left branching syntactic boundaries succeeded without right branching boundaries. In the current method, when a phrase component exceeds 12 morae an additional phrase command is placed at the boundary where concatenation of two consecutive words is weakest. The strength of concatenation is calculated as the word bi-gram. A sentence end symbol is simply placed at the end of the sentence, and, pauses are placed at the sentence boundaries.

3.4.3. Positioning of accent command symbols

Given the accent types, F or D symbols representing accent command onsets of accent phrases are inserted in the phone string: at the top of the accent phrase for the type 1 accent and between the first and second morae for other types. (F symbols correspond to the type 0 accent and are 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[11], the accent type can be decided 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 two accent phrases exist sequentially with no phrase command between, accent commands of these phrases may interact with each other. The second digit of an accent command symbol is added to represent the change in accent command amplitude due to this interaction[12].

4. Listening Experiment

4.1. Outline

A listening test was conducted on the reply speech to show the validity of our method: whether syntactic structure (kept through the sentence generation process by using tag LISP form) and discourse information (included in tag) are well reflected in the prosody of reply speech or not.

From the dialogue example of the system, eight sentences were selected and their speech was synthesized by the following three methods:

Proposed method: Uses syntactic and discourse ("importance" and "novelty" of words) information obtained through the sentence generation process. In the current system, content words conveying information on landmarks or directions are assumed as "important," while "novelty" means first appearance in the dialogue.

JUMAN+KNP analysis: Uses syntactic structures obtained by the sentence analysis by Japanese parsers (JUMAN[13] and KNP[14]) instead of those obtained through the sentence generation process.

No discourse information: Not using "importance" and "novelty" of words.

Twenty-four Japanese speakers were asked to evaluate the reply speech for each of 8 sentences generated by the three methods. The evaluation was conducted in 5 rank scoring: 1 when prosody of synthetic speech is poor, 3 when it is marginal and 5 when it is mostly natural. The three versions of speech were presented randomly.

4.2. Results

Table 3 shows scores averaged over the 24 speakers. The best scores were obtained by the proposed method for all the sentences. T-test (in 5% significance level) was conducted to check if the differences in scores are meaningful. Although, for sentences 2 and 4 there were no significant differences among the three versions, and for sentence 6 no significant difference between the versions by the proposed method and JUMAN+KNP method. Prosodic features are identical for the three versions in sentence 4 and for the two versions (proposed method and JUMAN+KNP method) in sentence 2.

As a whole, the proposed method is significant different from other two methods. Additionally, this advantage tends to expand according to the length of sentences.

Table 3: Average of the score for each synthesized speech

Sentence No.	1	2	3	4
Proposed method	3.58	3.42	3.46	2.42
JUMAN+KNP analysis	3.04	3.25	2.92	2.38
No discourse information	2.50	3.13	2.38	2.33
Sentence No.	5	6	7	8
Proposed method	2.79	3.83	3.88	4.04
JUMAN+KNP analysis	2.17	3.17	3.25	3.29
No discourse information	2.21	3.54	2.54	3.38

5. Conclusion

In the road guidance spoken dialogue system, we presented a framework of reply speech generation for concept-to-speech conversion. Tag Lisp form-based phrase template keeps syntactic structure and important word information to be used at the prosodic control of speech synthesis while it realizes style flexible sentence generation. The validity of the method was proven through a listening test of synthetic speech.

For the future work, better prosodic control is planned through analysis of the dialogue in the road-guidance situation. Correct assignment of the degree of importance and novelty to each word in generated sentence is also important for the better prosodic control. A scheme to change the style of the reply sentences according to user ' s preference and dialogue situation is also in the scope of the future work.

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. EUROSPEECH2003, Geneva, Vol.3, pp.1693-1696, 2003.

[5] Yagi, Y., Takada, S., Hirose, K. and Minematsu, N., "An improved method of generating speech from concept and its application to a dialogue system of road guidance," Proc. SPECOM2005, Vol.2, pp.703-706, 2005.

[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] Mcroy, S. W., Channarukul, S. and Ali, S. S., "An augmented template-based approach to text realization, Natural Language Engineering," Vol.9, No.4, pp.381.420, 2003.

[8] 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.

[9] 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)

[10] Moulines, E. and Charpentier, F., "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Communication, Vol.9, pp.453-467, 1990.

[11] 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 Synthesis, Santa Monica, 2002.

[12] Hirose, K. and Fujisaki, H., "Accent and intonation in speech synthesis," J. of IEICE, Vol.70, No.4, pp.378-385, 1987. (in Japanese)

[13] Japanese Morphological Analysis System JUMAN ver5.1:
http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

[14] Japanese Syntactic Analysis System KNP ver2.0:
http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html