TEMPLATE-BASED METHODS FOR SENTENCE GENERATION AND SPEECH SYNTHESIS

Hiroyuki Segi[†‡], *Reiko Takou*[†], *Nobumasa Seiyama*[†], *Tohru Takagi*[††], *Hideo Saito*[‡] *and Shinji Ozawa*[‡‡]

[†]NHK (Japan Broadcasting Corp.) Science and Technology Research Laboratory, Japan
[††]NHK Engineering Services, Inc., Japan
[‡]Graduate School of Science and Technology, Keio University, Japan
[‡‡]Graduate School of Technology, Aichi University of Technology, Japan
segi.h-gs@nhk.or.jp

ABSTRACT

Here we propose a sentence-generation method using templates that can be applied to create a speech database. This method requires the recording of a relatively small sentence set, and the resultant speech database can generate comparatively natural sounding synthesized speech. Applying this method to the Japan Broadcasting Corporation (NHK) weather report radio program reduced the size of the required sentence set to just a fraction of that required by comparable methods. We also propose a speech-synthesis method using templates. In an evaluation test, 66% of the speech samples synthesized by the proposed method using templates were preferred to those produced by the conventional concatenative speech-synthesis method.

Index Terms— speech synthesis, sentence generation, speech database, templates, recording sentences

1. INTRODUCTION

The long-running and historic weather report radio program has been transmitted by the Japan Broadcasting Corporation (NHK) since November 5, 1928. This 20-min program broadcasts temperature and wind velocity data for major cities in Japan and neighboring countries, as well as information about typhoons, low- and high- pressure systems, and so on.

Certain people, such as some mountain climbers and sailors, note down the weather conditions while listening to this radio program. Indeed, in some remote areas, AM radio is the only source of weather information. Some listeners in these areas create weather maps based on the broadcast data, and use them to forecast the weather for their location.

It is difficult for the announcers to regulate their speech rate during the weather report program, in order to allow sufficient time for the listeners to write down the information while also ensuring that all of the data are broadcast within the time available. An automatic broadcast system for the weather report program that can easily adjust the speech rate with a speech conversion technique [1] is thus desirable.

Several speech-synthesis methods have been reported [2-4]. These are not suitable for developing an automatic broadcast system, however, because the synthesized speech samples they produce sound unnatural, and listeners cannot tolerate them for long periods of time.

Creating more natural sounding synthesized speech requires a larger speech database. However, in view of the costs of recording and constructing speech databases, it is preferable to reduce the size of the sentence set. Hence, there is a need for a sentence-generation method that requires a relatively small sentence set to be recorded in order to create a speech database that can be used to produce natural sounding synthesized speech. Here we propose a sentence-generation method using templates that can be used to create a speech database. The proposed method reduces the size of the required sentence set to just a fraction of that required by comparable methods.

We also propose a speech-synthesis method using templates. In a subjective quality-evaluation test, 66% of the speech samples synthesized by the proposed method were preferred to those produced by the conventional concatenative speech-synthesis method.

2. SENTENCE GENERATION USING TEMPLATES

We previously developed a sentence-generation method for stock-price bulletins [5]. This method can be used with up to 1 billion recording sentence candidates. It is therefore unsuitable for the weather report program, which has around 10^{29} recording sentence candidates. Other sentence-generation methods have also been proposed [6-8]; however, the number of recording sentence candidates for these methods is no more than 1 million at the highest estimate, so they are also unsuitable for this purpose.

Although the number of recording sentence candidates is around 10^{29} for the weather report program, all of the input texts can be described as multiple templates.

For example, a representative input text for the weather report program can be described using the following template: “A low-pressure area will develop into a typhoon [number of hours] later on [date].” (note that English is used for explanatory purposes alone here as the real system can be performed only for Japanese). Here, [number of hours] and [date] are variables: the former is assigned values such as “1 hour”, “2 hours”, and so on; and the latter is assigned values such as “January 1”, “April 23”, and so on.

The following sections describe our proposed sentence-generation method using templates.

2.1. Template format

The templates described here include variables denoted as “[X1]”, branches denoted as “[X1] OR [X2] OR [X3]”, and abbreviations denoted as “<[X1]>”. An example template is as follows: “A low-pressure area will develop into a (typhoon OR hurricane) [number of hours] later <on [date]>.” This template allows both “…will develop into a typhoon…” and “…will develop into a hurricane…”, and the notation “<on [date]>” means that both “…[number of hours] later on [date].” and “…[number of hours] later.” are allowed.

2.2. Comparison and unification of templates by dynamic programming (DP)

In order to reduce the size of the required sentence set, templates that can describe all of the input texts are compared and unified using DP. For example, in the case of the templates “A low-pressure area (will develop into a typhoon OR will move [direction]) on [date].” and “A low-pressure area is expected to be located in [place] on [date].”, DP gives a unified template described as “A low-pressure area (will develop into a typhoon OR will move [direction] OR is expected to be located in [place]) on [date].” Fig. 1 shows an example of unified templates.

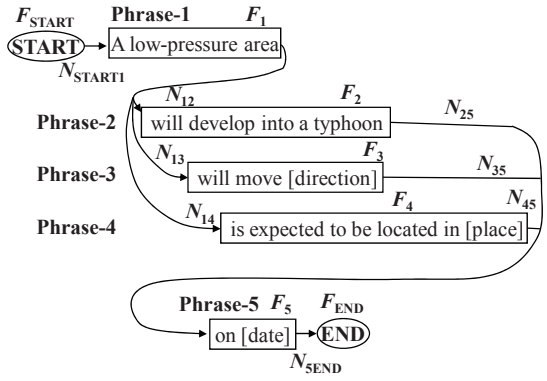

Fig. 1. An example of unified templates.

2.3. Optimization problem for generating a sentence set

To minimize the number of sentences that need to be recorded, the optimization problem for unified templates must be configured. The number of times that a phrase exists in a sentence set should be more than the number of the elements of the variables included in the phrase, because a sentence set should include all of the elements of all of the variables in all of the phrases over all of the templates. For example, in the case shown in Fig. 1, if the number of the elements of the variable [date] in phrase-5 “on [date]” is 366, then F_5, which denotes the number of times that a sentence set includes phrase-5, should be more than 366. Similarly, if the number of the elements of the variables in phrase-3 and phrase-4 is 16 and 100 respectively, then F_3 and F_4, which denote the number of times that a sentence set include phrase-3 and phrase-4, respectively, must be more than 16 and 100. Therefore, the inequalities are as follows:

$$F_1 \geq 1,\quad F_2 \geq 1,\quad F_3 \geq 16,\quad F_4 \geq 100,\quad F_5 \geq 366 \qquad (1)$$

In the speech-synthesis process, the synthesis unit at the start of the sentence should consist of the database unit at the start of the sentence, and the synthesis unit at the end of the sentence should consist of the database unit at the end of the sentence. This is because synthesized speech is more natural when the positions of the synthesis units and the database units are uniform. This means that only recording sentences that start from the “start node” and end at the “end node” are generated. Taking this condition into account, the number of times that a phrase is included is equal to the sum of all of the paths that lead up to that phrase and the sum of all of the paths that follow on from that phrase. The equations describing the process are as follows:

$$F_{START}=N_{START1},\quad N_{START1}=F_1,\quad F_1=N_{12}+N_{13}+N_{14},\quad N_{12}=F_2,\quad N_{13}=F_3,\quad N_{14}=F_4,\quad F_2=N_{25},\quad F_3=N_{35},\quad F_4=N_{45},\quad N_{25}+N_{35}+N_{45}=F_5,\quad F_5=N_{5END},\quad N_{5END}=F_{END}. \qquad (2)$$

Here, F_{START} and F_{END} are the number of times that the sentence set includes the “start node” and the “end node”, respectively. N_{ij} is the number of times that the sentence set includes the path from phrase-i to phrase-j.

Therefore, the proposed sentence-generation method involves the optimization problem of minimizing the F_{START} under the conditions of the inequalities and equations. This problem can be solved by using simplex methods to obtain the number of times that phrases and paths are included.

2.4. Sentence-generation process

Section 2.3 described how to determine the number of times that phrases and paths are included. To generate the recording sentences, in the beginning the “start node” connects to phrase-1, so the first element of the variable in phrase-1 is used. Then, phrase-1 connects to phrase-2, phrase-3 and phrase-4. Initially, phrase-2 is selected and the first element of the variable in phrase-2 is used. Next, phrase-2 connects to phrase-5, so the first element of the variable in phrase-5 is used. Finally, phrase-5 connects to the “end node”, so one sentence is generated to connect all of the elements selected from the “start node” to the “end node”. For example, if phrase-1 is “A low-pressure area”, phrase-2 is “will develop into a typhoon”, and phrase-5 is

"on January 1", then "A low-pressure area will develop into a typhoon on January 1." is generated.

Subsequently, as described above, in the beginning the "start node" connects to phrase-1 and the second element of the variable in phrase-1 is used, because the first element has already been used. If all of the elements have already been used, the first element is re-used. Next, phrase-1 connects to phrase-2, phrase-3 and phrase-4. If the accumulated number of times of inclusion of the path from phrase-1 to phrase-2 is less than N_{12}, which is obtained by the simplex method, phrase-2 is re-selected and the next element of the variable in phrase-2 is used. If the accumulated number of times of inclusion of the path from phrase-1 to phrase-2 is more than N_{12}, phrase-3 is selected and the first element of the variable in phrase 3 is used.

These sentence-generation processes are repeated F_{START} times.

3. PERFORMANCE EVALUATION OF PROPOSED SENTENCE-GENERATION METHOD

To examine the performance of the proposed sentence-generation method, we generated a sentence set from nine templates used in the weather report program. Conventional methods [6-8] could not be used in this case, because the number of recording sentence candidates was around 10^{29}. We therefore compared the proposed sentence-generation method with a method that randomly generated a sentence set from templates according to the following procedure. First, a template was selected randomly. Second, a path of branches and abbreviations in the selected template was selected randomly. Third, an element in the variables in the phrase on the selected path was selected randomly. These operations were repeated until arrival at the "end node".

Nine templates for the weather report program were used for the evaluation. The coverage of the elements in the variables in the sentence set generated by the random method was calculated (Fig. 2). As the number of recording sentences increased, the coverage increased, and 40,000 recording sentences achieved 99.7 % coverage. We found that the size of the required sentence set rapidly increased as the coverage increased. A similar tendency was reported elsewhere [6, 7].

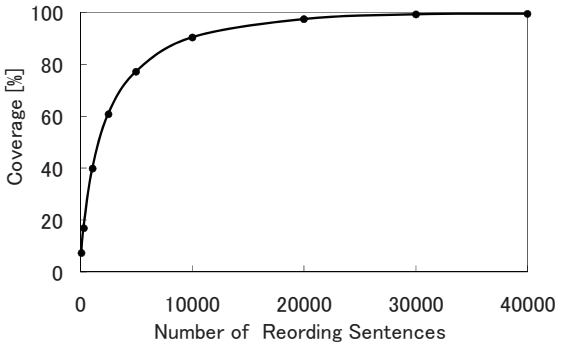

Fig. 2. Coverage of the elements in the variables by the random method of generating a sentence set.

Our proposed sentence-generation method was also performed with the same nine templates. Two unified templates were generated, and the number of required recording sentences was 1,085. The coverage calculated from these 1,085 sentences was 100%. The number of recording sentences required by our proposed method was just a few percent of that required by the random method, with coverage of more than 95 %.

4. SPEECH-SYNTHESIS METHOD USING TEMPLATES

Concatenative speech-synthesis methods search for the best combination of voice waveform samples which maximize the sum of the target score and the concatenation score. The target score is calculated as the similarity of the fundamental frequency and the phoneme duration between the candidates of voice waveform and the target values. If there are errors in the estimation of the target values, the selected best combination does not yield natural sounding synthesized speech, because the target score cannot be calculated correctly.

We therefore propose a speech-synthesis method using unified templates that are produced by the sentence-generation method described in section 2. In this speech-synthesis method, the unified templates are used instead of the target score. Synthesized speech is produced according to the following procedure, using an example input text described as "Typhoon No. 3 is moving northeast in the south of Okinawa." (note that English is used here for explanatory purposes again, and the real system can synthesize only Japanese). The proposed speech-synthesis method involves matching between the input text and unified templates. It is assumed that the matched template is described as "Typhoon [Number] is moving [direction] in [direction] [place].". The system searches for all the combinations in the speech database, and selects that which makes the acoustic features of the adjacent voice waveform samples as similar as possible. For example, for the "in [direction] [place]" part, the system searches for the combinations of the "(in)+the", "in-(the)+sou", "the-(south)+of", "th-(of)+Oki", and "of-(Okinawa)" voice waveform samples. Here, the notation "the–(south)+of" corresponds to the word "south" preceded by the word "the" and followed by the word "of".

Two points should be noted here. First, any word in all of the templates has more than one instance of voice waveform in the speech database, because the sentences that are read out by an announcer are designed to include any word in all of the unified templates. Second, the speech-synthesis method uses only the voice waveform that is included in the matched template. This means that although there are many voice waveform samples of "the-(south)+of"

in the speech database, only those in the matched template are used. This is because the voice waveform samples in the matched template might have a similar fundamental frequency or spectrum to the real value, whereas those in another template might differ.

Finally, the speech-synthesis system connects the selected voice waveform samples so as to realize the greatest similarity among all of the connection points, and outputs the results as synthesized speech.

5. PERFORMANCE EVALUATION OF PROPOSED SPEECH-SYNTHESIS METHOD

5.1. Listening Test

We conducted a paired comparison test to assess the naturalness of speech samples produced by the proposed speech-synthesis method using templates and those produced by the conventional concatenative speech-synthesis method as described elsewhere [9]. The speech database for both speech synthesis methods is created from 1,085 recording sentences that were generated by the proposed sentence-generation method, described in section 3.

The evaluation used 63 sentences that were not included in the speech database. In total, 126 speech samples were synthesized by the proposed method and the conventional method.

To conduct the test, a loud-speaker was set up in a sound-proof room. The subjects were five males and five females with no known hearing problems. They were asked to judge which of two speech samples with the same content they considered to sound more natural. They were not allowed to rate both samples in a pair as equally natural sounding. The order of presentation of the speech samples within each pair was randomized, as was the order of presentation of the sentence pairs.

5.2. Results

The experimental results (including the 95% confidence intervals) are shown in Fig. 3. In total, 66% of the synthesized speech samples produced by the proposed method were evaluated as sounding more natural than those produced by the conventional method.

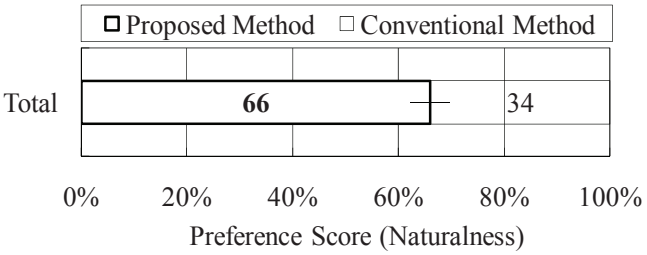

Fig. 3. Results of a paired comparison test between the proposed speech-synthesis method and the conventional method.

6. CONCLUSION

We proposed a sentence-generation method using templates that can be used to create a speech database. Applying this method to the NHK weather report radio program reduced the size of the required sentence set to a fraction of that required by comparable methods. We also proposed a speech-synthesis method using templates. In an evaluation test, 66% of the speech samples synthesized by the proposed method using templates were preferred over those produced by the conventional concatenative speech-synthesis method.

7. REFERENCES

[1] A. Imai, T. Takagi, and H. Takeishi, "Development of radio and television receiver with functions to assist hearing of elderly people," *IEEE Trans. Consumer Electron.*, vol. 51, no. 1, pp. 268–272, 2005.

[2] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: a new TTS from ATR based on corpus-based technologies," *Proc. 5th ISCA Speech Synthesis Workshop [SSW5]*, pp. 179-184, 2004.

[3] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," *IEICE Trans. Inf. Syst.*, vol. E91-D, no. 6, pp. 1764-1773, 2008.

[4] E. Moulines and F. Charpentier, "Pitch Synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones," *Speech Commun.*, vol. 9, pp. 453-467, 1990.

[5] H. Segi, R. Takou, N. Seiyama, and T. Takagi, "Development of a Prototype Data-Broadcast Receiver with a High-Quality Voice synthesizer," *IEEE Trans. Consumer Electron.*, vol. 56, no. 1, pp. 268–272, 2010.

[6] H. Kawai, S. Yamamoto, and T. Shimizu, "A Design Method of Speech Corpus for Text-To-Speech Synthesis Taking into Account Prosody," *Proc. ICSLP*, vol. 3, pp. 420-425, 2000.

[7] M. Isogai, H. Mizuno, and M. Kazunori, "Recording Script Design for Corpus-Based TTS System Based on Coverage of Various Phonetic Elements," *Proc. ICASSP*, vol. 1, pp. 301-304, 2005.

[8] J. V. Santen, "Diagnostic Perceptual Experiments for Text-To-Speech System Evaluation," *Proc. ICSLP*, vol. 1, pp. 555-558, 1992.

[9] H. Segi, T. Takagi, and T. Ito, "A Concatenative Speech Synthesis Method Using Context Dependent Phoneme Sequences with Variable Length as Search Units," *Proc. 5th ISCA Speech Synthesis Workshop [SSW5]*, pp. 115-120, 2004.