An Automatic Broadcast System for a Weather Report Radio Program

Hiroyuki Segi, *Senior Member, IEEE,* Reiko Takou, Nobumasa Seiyama, Tohru Takagi, Yuko Uematsu, *Member, IEEE,* Hideo Saito, *Senior Member, IEEE,* and Shinji Ozawa

***Abstract*—Here we describe a speech-synthesis method using templates that can generate recording-sentence sets for speech databases and produce natural sounding synthesized speech. Applying this method to the Japan Broadcasting Corporation (NHK) weather report radio program reduced the size of the recording-sentence set required to just a fraction of that needed by a comparable method. After integrating the recording voice of the generated recording-sentence set into the speech database, speech was produced by a voice synthesizer using templates. In a paired-comparison test, 66 % of the speech samples synthesized by our system using templates were preferred to those produced by a conventional voice synthesizer. In an evaluation test using a five-point mean opinion score (MOS) scale, the speech samples synthesized by our system scored 4.97, whereas the maximum score for commercially available voice synthesizers was 3.09. In addition, we developed an automatic broadcast system for the weather report program using the speech-synthesis method and speech-rate converter. The system was evaluated using real weather data for more than 1 year, and exhibited sufficient stability and synthesized speech quality for broadcast purposes.**

***Index Terms*—Recording-sentence set, speech-rate conversion, templates, voice synthesizer.**

I. Introduction

THE LONG-running and historic weather report radio program has been transmitted by the Japan Broadcasting Corporation (NHK) since November 5, 1928 [1]. 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 high-quality speech-rate converter [2] is thus desirable.

Several speech-synthesis systems have been reported [3]–[7]. 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.

We therefore propose a speech-synthesis method using templates to produce natural-sounding synthesized speech that is similar to the human voice. In conventional speech-synthesis methods, speech processing for pitch conversion or parameterization, which compensates for the lack of appropriate synthesis units in small speech databases, degrades the speech quality. However, in the proposed speech-synthesis method, all of the synthesis units needed by the voice synthesizer can be included in the speech database, so speech processing for pitch conversion or parameterization is not necessary. Moreover, synthesis units can be positioned appropriately by the voice synthesizer using unified templates generated when the speech database is created.

Using the proposed speech-synthesis method and speech-rate converter, we developed an automatic broadcast system for the weather report program. We conducted a trial of the system for more than 1 year using weather data available on the internet. The results confirmed that the system has sufficient stability and synthesized speech quality for broadcast purposes.

The current paper is organized as follows. Section II describes the naturalness of conventional speech-synthesis methods. Section III gives an overview of our proposed speech-synthesis method. Section IV describes the sentence-generation method for creating unified templates and the speech database, which are required for producing synthesized speech. Section V evaluates our sentence-generation method. Section VI describes the voice synthesizer using unified templates. Sections VII and VIII detail subjective evaluations

Manuscript received August 10, 2012; revised April 12, 2013; accepted May 10, 2013. Date of publication July 31, 2013; date of current version August 21, 2013.

H. Segi, R. Takou, and N. Seiyama are with the NHK Science and Technology Research Laboratories, Tokyo 157-8510, Japan (e-mail: segi.h-gs@nhk.or.jp; takou.r-go@nhk.or.jp; seiyama.n-ek@nhk.or.jp).

T. Takagi was with NHK Engineering Services, Inc., Tokyo 157-8540, Japan (e-mail: takagi.t-fo@nhk.or.jp).

Y. Uematsu and H. Saito is with the School of Science for Open and Environmental Systems, Graduate School of Science and Technology, Keio University, Kanagawa 223-8522, Japan (e-mail: yu-ko@hvrl.ics.keio.ac.jp; saito@hvrl.ics.keio.ac.jp).

S. Ozawa is with the School of Media and Information, Graduate School of Technology, Aichi University of Technology, Aichi 443-0047, Japan (e-mail: ozawa@aut.ac.jp).

Digital Object Identifier 10.1109/TBC.2013.2272406

of the voice synthesizer. Section IX describes the automatic broadcast system for the weather report program developed using our speech-synthesis method and voice converter. Finally, Section X summarizes our findings.

II. Naturalness of Conventional Speech-Synthesis Methods

Before describing our method, this section reviews the performance of conventional speech-synthesis methods in terms of naturalness.

Speech-synthesis method by compilation of recorded speech sound has been used for broadcasting for more than 20 years. This method is also used for airport and train announcement systems [8]. Although speech synthesized by this method has not been evaluated, it has been considered to achieve human voice quality because it has been used in broadcast systems. However, the contents of speech synthesized by this method are limited to combinations of the recorded phrases connected at silent sections. Thus, this method cannot be utilized for the weather report program, because its content is too wide-ranging. Moreover, this method does not take coarticulation into account, suggesting that the naturalness of synthesized speech is degraded without enough silence sections [9]. Indeed, 91% of synthesized speech with coarticulation was evaluated as more natural than synthesized speech without coarticulation [9].

Speech synthesized by the Hidden Markov model (HMM) method was evaluated [6]. Speech was synthesized by 14 methods, including the HMM system, under similar conditions using the same speech database and evaluation sentences. A subjective evaluation test with a five-point mean opinion score (MOS) scale was used, and speech synthesized by the HMM system was rated at about 3.1 by speech experts. The best speech-synthesis system was rated at about 3.7 by the speech experts, although its identity was not revealed as all systems were tested anonymously. The MOS for natural speech was about 4.7, so none of the systems achieved similar naturalness to the human voice. In addition, speech synthesized by Japanese methods developed within the past two years was evaluated [10], [11]. In these tests, none of the systems achieved similar naturalness to the human voice.

Speech synthesized by the concatenative speech-synthesis method was also evaluated [5], [12]. In these papers, two subjective evaluation tests were performed. In the first, speech synthesized by 10 commercially available systems and XIMERA, which is a proposed concatenative speech-synthesis method, was evaluated. The results showed a statistically meaningful improvement in performance using XIMERA compared with the other systems. In the second, speech synthesized by XIMERA using different-sized speech databases was evaluated using the five-point MOS scale. Speech synthesized using the largest database was rated at about 3.4, whereas natural speech was rated at about 4.8. Thus, XIMERA did not produce synthesized speech with similar naturalness to the human voice.

We therefore concluded that conventional speech-synthesis methods could not achieve synthesized speech with similar naturalness to the human voice, and that the best conventional speech-synthesis method was the concatenative method using a huge speech database like XIMERA.

III. Overview of the Proposed Speech-Synthesis Method

We thus identified a need to develop a speech-synthesis method that can generate natural-sounding synthesized speech, similar to the human voice.

Quality degradation of synthesized speech can be caused by speech processing used for pitch conversion or parameterization, which is intended to compensate for a lack of appropriate synthesis units using small speech databases [12], [13]. The best results are therefore achieved by concatenative speech-synthesis methods using huge speech databases like XIMERA, because the synthesis units used by the voice synthesizer are often included in the database and speech processing for pitch conversion or parameterization is not required.

In order to produce synthesized speech similar to the human voice, all of the synthesis units needed by a voice synthesizer must be included in the speech database.

The speech database is created by recording the voices of announcers or actors/actresses reading out a set of specific sentences. Natural-sounding synthesized speech similar to the human voice might therefore be achieved by generating a recording-sentence set that includes all of the synthesis units needed by the voice synthesizer.

In general, a huge number of synthesis units is needed by the voice synthesizer because the synthesis units must be treated as variants due to possible differences in factors such as coarticulation, pitch, the position of the sentence, accent, intonation, and emotion. However, a smaller number of synthesis units is needed for the weather report program, because the input texts can be described by 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 used 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. Thus, the recording-sentence set that includes all of the synthesis units needed by the voice synthesizer for the weather report program is relatively small, and could be recorded in a realistic time period.

Our speech-synthesis method consists of a sentence generator and a voice synthesizer. Fig. 1 shows the proposed method. It was necessary to create unified templates and a speech database before producing synthesized speech. This framework ensures that all the synthesis units needed by the voice synthesizer are included in the speech database, and speech can be produced without any pitch conversion or parameterization.

In the sentence generator, the unified templates and recording-sentence set are generated from input templates, which are able to describe all of the input sentences of

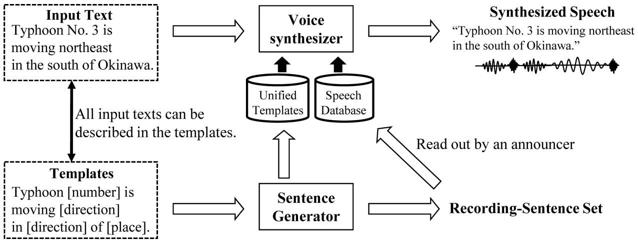
Input Text
Typhoon No. 3 is moving northeast in the south of Okinawa.
Voice synthesizer
Synthesized Speech
"Typhoon No. 3 is moving northeast in the south of Okinawa."
Unified Templates
Speech Database
All input texts can be described in the templates.
Read out by an announcer
Templates
Typhoon [number] is moving [direction] in [direction] of [place].
Sentence Generator
Recording-Sentence Set

Fig. 1. Overview of the proposed speech-synthesis method.

the voice synthesizer. Multiple input templates are thus permissible.

After the recording-sentence set is generated, it is read out by announcers. The recording voice is integrated into the speech database and the speech database includes more than one voice-waveform sample of all the synthesis units. The method used to create the unified templates and speech database is described in section IV.

In the voice synthesizer, speech is produced from an input sentence using the unified templates and speech database. Many combinations of the voice-waveform samples of synthesis units in the speech database can realize synthesized speech for the input sentence; however, only those combinations allowed by the unified templates are considered instead of using prosody information. The combination of voice-waveform samples with the largest cross-correlation is selected, the chosen samples are concatenated considering the phase shift, and the result is output as synthesized speech. The method used to produce synthesized speech is described in section VI.

The main technical contribution of this work is the development of a sentence-generation and speech-synthesis method in which the input texts of the speech-synthesis system can be described by multiple templates. Unlike conventional methods, our approach can generate a recording-sentence set from a huge number of recording-sentence candidates in a realistic time frame.

Moreover, our speech-synthesis method does not require estimations of pitch and phoneme duration with the use of unified templates; estimations of pitch and phoneme duration that is inherent in conventional speech-synthesis methods make the potential for errors. Therefore our method can produce synthesized speech with similar naturalness to the human voice.

IV. Sentence Generator Using Templates

We previously developed a sentence-generation method for stock-price bulletins [9]. 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 [14]–[18]; however, the number of recording-sentence candidates for these methods is only 1 million at the highest estimate, and so they are also unsuitable for this purpose.

As we mentioned in the previous section, although the number of recording-sentence candidates is around 10^{29} for the weather report program, all of the input texts can be described by multiple templates. Our sentence-generation method using templates can generate a recording-sentence set without selecting from 10^{29} recording sentence candidates. This section details our sentence-generation method.

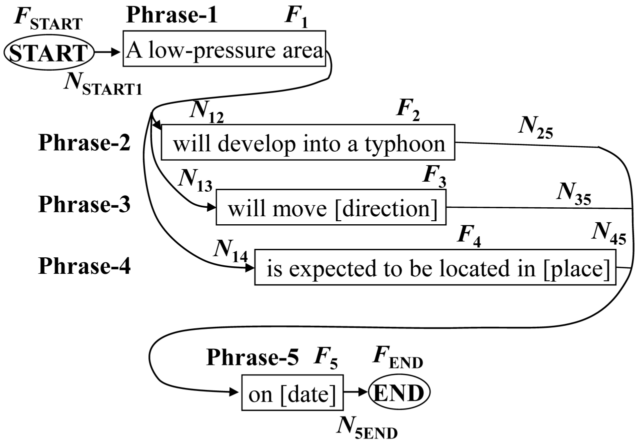

Fig. 2. An example of the unified templates.

A. Template Format

The templates described here include variables denoted as "[X1]", branches denoted as "[X1] OR [X2] OR [X3]", abbreviations denoted as "<[X1]>", and boundary marks denoted as "|". Here, 'abbreviations' indicates that there are optional units that can be either present or not present, and 'boundary marks' indicates that the coarticulation effect is reduced before and after this point. 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 "A low-pressure area | will develop into a typhoon … " and "A low-pressure area | 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. The notation "A low-pressure area | will develop..." means that coarticulation effect between "A low-pressure area" and "will develop..." need not be taken into account. The contents of the templates for the weather report program are decided by the Japan Meteorological Agency.

B. Comparison and Unification of Templates by Dynamic Programming (DP)

In order to reduce the size of the required recording-sentence set, templates that can describe all of the input texts are compared and unified using DP [19]. 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. 2 shows an example of the unified templates.

The comparison order is the input order. Initially, a comparison between the first input template and the second input template is performed. Next, a comparison between the third template and the unified templates, which is made from first and second input templates, is performed. Then, a comparison between the fourth template and the unified templates is performed, and so on. When a comparison is performed, the similarity P is calculated by following equations:

$$P = \frac{H - I}{H + S + D} \tag{1}$$

Here, H is the hit, S is the substitution error, D is the deletion error, and I is the insertion error in the DP result. If the similarity is more than a threshold that is decided beforehand, the input template is merged to the unified template with the highest similarity. If the similarity is less than the threshold, the input template is not merged and added to the unified templates. Thus, if the threshold is high (nearly 1.0), the input template tends to be unmerged and the number of unified templates increases. if the threshold is low (nearly 0.0), the input template tends to be merged and the number of unified templates decreases. In the performance evaluation described below, the threshold is 0.3, which means that if there is more than 30 % similarity between the unified template and the input template, the latter is merged with the former.

C. Sequential Greedy Algorithm for Consecutive Variables

In the nodes of the unified templates, if there are consecutive variables, the number of recording sentences increases. For example, the number of elements of the consecutive variables, “[thousand] [hundred] [tens] [ones] ” is 9 999.

There are many synthesis units in “[thousand] [hundred] [tens] [ones]”. However, if the recording-sentence set includes all of the synthesis units in “[thousand] [hundred] [tens] [ones]”, all of the elements in “[thousand] [hundred] [tens] [ones]” can be synthesized. Therefore, to reduce the size of the required recording-sentence set, a sequential greedy algorithm is employed using the following steps.

First, the initial value of the maximum number of synthesis units in one combination of consecutive variables is set to 0. Second, if the number of synthesis units in one combination of consecutive variables is more than the maximum, the combination of consecutive variables is added to the output, and the maximum is set to the number of synthesis units. In such cases, the system counts only those synthesis units not included in the output. If the number of synthesis units in one combination of consecutive variables is less than the maximum, no action is taken.

Third, all of the combinations of consecutive variables are tested.

Fourth, if the maximum number is still equal to 0, all of the synthesis units are included in the output; if not, steps one to four are repeated.

For example, the number of the recording sentence set for “[thousand] [hundred] [tens] [ones]” is 405 when a sequential greedy algorithm is applied for Japanese digits. Subsequently, consecutive variables are treated as a single entity including elements whose number equals the output generated by this method.

D. Optimization Problem for Generating Recording-Sentence Set

To minimize the number of recording sentences that need to be recorded, the optimization problem for unified templates must be configured. The coarticulation effect is reduced among the nodes of the unified templates. Thus, each element of the nodes is independent from the elements of the nodes that come before and after.

The number of times that a phrase exists in a recording-sentence set should be more than the number of the elements of the variables included in the phrase, because a recording-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. 2, if the number of elements of the variable [date] in phrase-5 “on [date]” is 366, then F_5, which denotes the number of times that a recording-sentence set includes phrase-5, should be more than 366. Similarly, if the numbers of elements of the variables in phrase-3 and phrase-4 are 16 and 100, respectively, then F_3 and F_4, which denote the number of times that a recording-sentence set includes phrase-3 and phrase-4, respectively, must be more than 16 and 100. Therefore, the inequalities are as follows:

$$F_1 \geq 1,\ F_2 \geq 1,\ F_3 \geq 16,\ F_4 \geq 100,\ F_5 \geq 366 \tag{2}$$

In producing synthesized speech, the voice-waveform samples at the start of the input text should comprise those at the start of the sentence in the speech database, and the voice-waveform samples at the end of the input text should consist of those at the end of the sentence in the speech database. This is because synthesized speech is more natural when the position of the selected voice-waveform samples in the speech database and the position where the voice-waveform samples are used in the input text 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:

$$\begin{aligned} &F_{\mathrm{START}} = N_{\mathrm{START1}},\ N_{\mathrm{START1}} = F_1,\ F_1 = N_{12} + N_{13} + N_{14}, \\ &\quad N_{12} = F_2,\ N_{13} = F_3,\ N_{14} = F_4,\ F_2 = N_{25},\ F_3 = N_{35}, \\ &\quad F_4 = N_{45},\ N_{25} + N_{35} + N_{45} = F_5,\ F_5 = N_{5\mathrm{END}}, \\ &\quad N_{5\mathrm{END}} = F_{\mathrm{END}}. \end{aligned} \tag{3}$$

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

Therefore, the sentence generator involves the optimization problem of minimizing the F_{START} under the conditions of the inequalities and equations. This problem can be solved by using the simplex method [20] to obtain the number of times that phrases and paths are included.

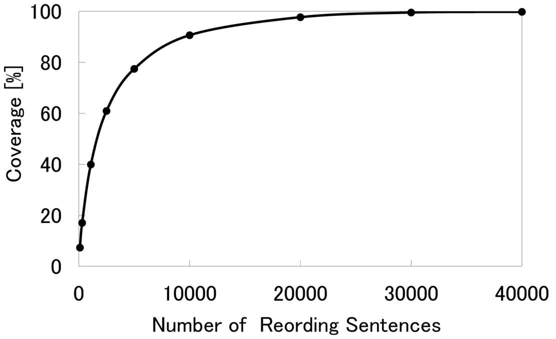

Fig. 3. Coverage of the elements in the variables by the random method of generating a recording-sentence set.

E. Sentence-Generation Process

To generate the recording-sentence set, in the beginning the "start node" connects to phrase-1, so the first element of the variable in phrase-1 is used. If there are no variables and only one fixed part in phrase-1, the system regards the fixed part as the variable including only one element. 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 [20], 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.

V. Performance Evaluation of the Sentence Generator

To examine the performance of the sentence generator described in section IV, we created a recording-sentence set from nine templates used in the weather report program [21]. Conventional methods [9], [14]–[18] could not be used in this case, because the number of recording-sentence candidates was around 10^{29}. We therefore compared the sentence-generation method with a method that randomly generated a recording-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".

The coverage of the elements in the variables in the recording-sentence set generated by the random method was calculated (Fig. 3). As the number of recording sentences increased, the coverage increased, and we found that 40 000 recording sentences achieved 99.7 % coverage. The size of the required recording-sentence set rapidly increased as the coverage increased. A similar tendency has been reported elsewhere [14], [18].

Our sentence-generation method was also performed with the same nine templates. Two unified templates were generated, and the number of required recording sentences was 1085. The coverage calculated from these 1085 sentences was 100 %. The number of recording sentences required by our method was just a few percent of that required by the random method, with coverage of more than 95 %.

To investigate the effects of DP, we compared "the sentence-generation method with DP" with "the sentence-generation method without DP." We created a recording-sentence set by the sentence-generation method without DP under the same conditions as with DP. The number of required recording sentences was 4 514 without DP compared with 1 085 with DP. Therefore, in this case, the use of DP reduced the number of required recording sentences to 24 % of that without DP.

VI. Voice Synthesizer Using Templates

In general, concatenative voice synthesizers search for the best combination of voice-waveform samples of synthesis units, which maximize the sum of the target score and the concatenation score [4], [5], [7], [13]. The target score is usually calculated as the similarity of the fundamental frequency and the phoneme duration between the voice-waveform candidates and the target values. The target values are estimated from the input text and it is difficult to avoid errors. Therefore, the selected best combination does not always yield natural sounding synthesized speech.

We therefore propose a voice synthesizer using unified templates created by the sentence generator described in section IV. In our voice synthesizer, the unified templates are used instead of the target score. The following section describes the voice synthesizer using the following sample input text: "Typhoon No. 3 is moving northeast in the south of Okinawa on August 1." (Note that English is used here for explanatory purposes again, and the real system can synthesize only Japanese).

Our voice synthesizer involves matching the input text and unified templates, which are usually multiple. Our example uses the following two unified templates: first, "Typhoon [number] (will move [direction1] OR is expected to be located in [place]) [number of hours] later."; and second, "Typhoon [number] is moving [direction1] in [direction2] of [place] on [date].". [number] is the variable assigned values such as "No. 1", "No. 2", and so on. [direction], [place], [number of hours],

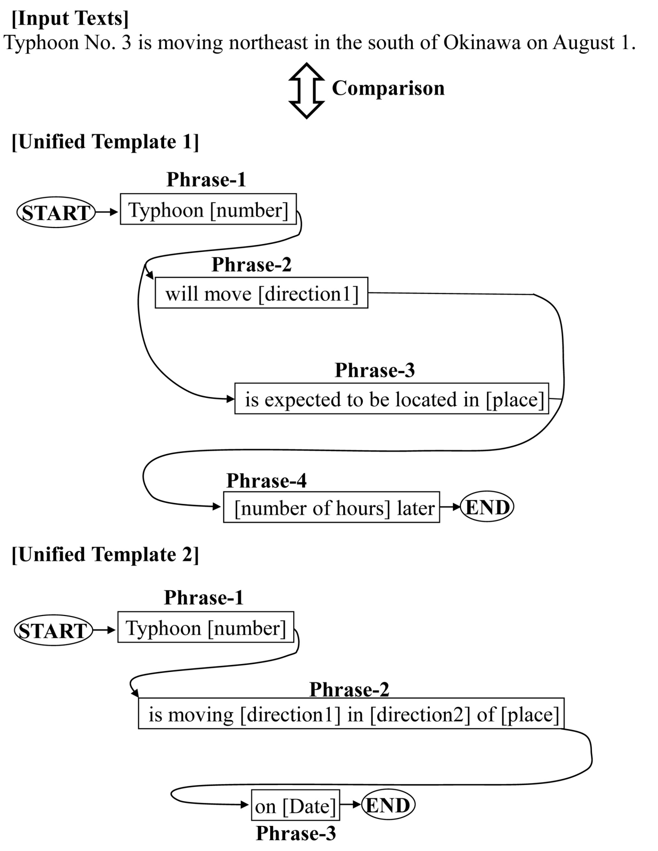

Fig. 4. Example of a comparison between the input text and unified templates.

and [date], are also variables. Fig. 4 shows the comparison process between the input text and unified templates.

Initially, the voice synthesizer compares the input text with the first unified template, and examines whether they are consistent. If they are not, the voice synthesizer finishes this comparison and then begins to compare the input text with the second unified template. In this case, the input text is found to be consistent with the second unified template. (Note that in English, words are separated by spaces, unlike Japanese words such as “IliveinTokyo”. Therefore, when applying these algorithms with Japanese, variants of word division must also be taken into account. For example, both “I saw it together.” and “Is a wit to get her?” should be considered.).

After the matched template is found, the voice synthesizer searches for all the combinations of adjacent voice-waveform samples from punctuation to punctuation that are included in the node of the matched template. Then, the voice synthesizer selects the combination of voice-waveform samples that makes the largest cross-correlation possible. For example, for the “in [direction] of [place]” part, the voice synthesizer 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, synthesis units in all of the templates have more than one voice-waveform sample in the speech database, because the recording sentences read out by an announcer are designed to include any synthesis units from all of the unified templates. Second, the voice synthesizer uses only the voice-waveform samples of the synthesis units that are included in the matched template. This means that although there are many voice-waveform samples of “the-(south) + of” synthesis unit in the speech database, only those in the matched template are used. This is because the voice-waveform samples of the synthesis units 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 adjusting the phase at the connection points, and outputs the results as synthesized speech.

VII. Performance Evaluation (1)

A. *Listening Test*

We conducted a paired comparison test to assess the naturalness of speech samples produced by the proposed voice synthesizer using templates and those produced by a conventional concatenative voice synthesizer as described previously [7]. The speech database for both voice synthesizers was created from 1 085 recording sentences that were generated by the sentence generation method described in section V. When the announcer read the recording sentences, we asked him to reduce the coarticulation effect at boundary marks.

The evaluation used 63 sentences that were not included in the speech database. In total, 126 test 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 without any known hearing problems. They were asked to judge which of two test speech samples with the same content they considered to sound more natural. They were not allowed to rate both test speech samples in a pair as equally natural sounding. Each speech sample of a pair was arranged in random order, and the order of the sentence pairs was also randomized. They were asked to listen to the test speech samples only once, because of their long duration. The subjects rested intermittently.

B. *Results*

The experimental results (including the 95 % confidence intervals) are shown in Fig. 5. 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.

The advantage of our approach compared with the conventional method is that only the voice-waveform samples included in the matched template are used. If the cost function in the voice-waveform search corresponding to the perceptual characteristics was known, the pitch and phoneme duration could be estimated without error, and if complete searching without pruning was realistic, the synthesized speech produced by the conventional method would be as natural as that

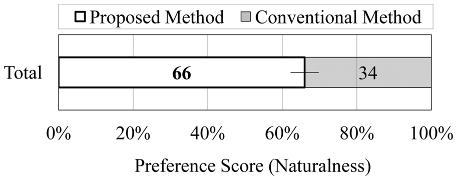

Fig. 5. Results of a paired comparison test between the proposed voice synthesizer and the conventional voice synthesizer. The 95% confidence interval is also shown.

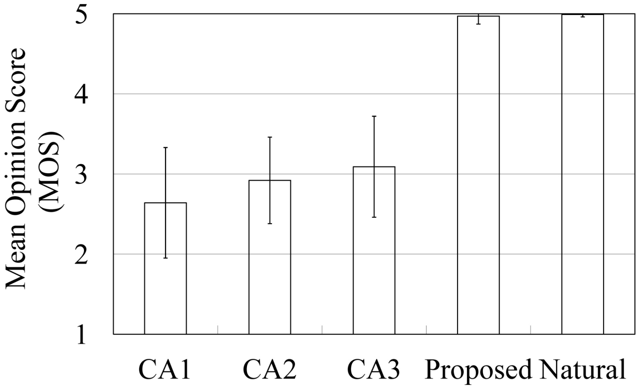

Fig. 6. MOS for voice synthesizers. CA1, CA2, and CA3 denote the commercially available voice synthesizers.

generated by the proposed method. This is because the search scope of the conventional method includes that of the proposed method.

VIII. Performance Evaluation (2)

A. Listening Test

We conducted a subjective quality-evaluation test using a five-point scale to assess the naturalness of the speech samples produced by the proposed voice synthesizer.

The evaluation used 63 sentences that were not included in the speech database. Thus, 252 test speech samples were synthesized using four methods: three commercially available voice synthesizers and the proposed voice synthesizer. Natural speech samples of 63 sentences were also included. Hence, a total of 315 test speech samples were prepared for the evaluation.

To conduct the test, a loud-speaker was set up in a sound-proof room. The subjects were seven males and five females without any known hearing problems. They were asked to listen to the test speech samples only once because of their long duration. The subjects were instructed to evaluate the presented test speech sample in terms of its perceived naturalness on a scale of 1 to 5, with 5 denoting “entirely natural”, 4 denoting “negligibly unnatural”, 3 denoting “slightly poor”, 2 denoting “poor”, and 1 denoting “very poor”. For each trial, the speech samples were presented in a random order. The subjects rested intermittently.

B. Results

The results are presented as the MOS and standard deviation (Fig. 6). CA1, CA2, and CA3 denote the three commercially available voice synthesizers. The MOS of natural speech was 4.99. The MOS of the speech samples synthesized by the proposed voice synthesizer was 4.97, whereas the highest MOS achieved by the commercially available voice synthesizers was 3.09. The subjective evaluation confirmed the superiority of the proposed voice synthesizer compared with the commercially available ones. This was considered to be because commercially available voice synthesizers transform the voice-waveform into target values of acoustic features, reducing the naturalness of synthesized speech. By contrast, the proposed voice synthesizer uses voice-waveforms that are not transformed into target values.

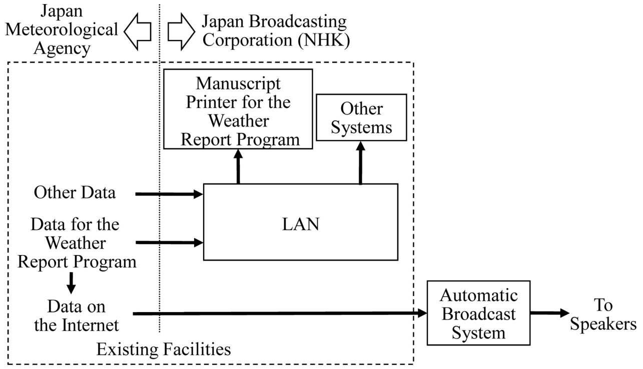
Japan Meteorological Agency
Japan Broadcasting Corporation (NHK)
Manuscript Printer for the Weather Report Program
Other Systems
Other Data
Data for the Weather Report Program
LAN
Data on the Internet
Automatic Broadcast System
To Speakers
Existing Facilities

Fig. 7. Data flow for automatic broadcast system. The system uses weather data that are publicly available on the internet instead of the exclusive data passed from the Japan Meteorological Agency to the NHK.

Our method does not use intonation, stress, and rhythm directly; rather, it takes account of them by using the matched templates. Only the voice-waveform samples that exist in the same place in the input sentence are used. This means that 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. As a result, the intonation, stress, and rhythm are re-created effectively.

IX. Automatic Broadcast System for Weather Report Program

We developed an automatic broadcast system for the weather report program using the proposed speech-synthesis method and speech-rate converter [2]. The system uses weather data that are publicly available on the internet, whereas exclusive data that are passed from the Japan Meteorological Agency to the NHK are used for the actual weather report radio program. This is to avoid the potential for errors in the prototype system to disrupt other broadcasting systems. As our main purpose was to verify the stability and quality of our system, the publicly available weather data were suitable despite the 1-day delay compared with the exclusive data. Fig. 7 shows the data flow for the developed automatic broadcast system.

The automatic broadcast system starts to analyze the weather data when it detects an update of the weather data. The system extracts the date information from the header and divides it into the following three parts: “weather in each place”, “weather from the ships”, and “fishery weather”.

For the first two, the recording compilation speech-synthesis method can be used, and the weather data must be split into words or phrases. For the third, the proposed speech-synthesis method can be used and the weather data can be split into sentences. It takes only 20 seconds to analyze the weather data and produce synthesized speech. The system is therefore ready to broadcast just 20 seconds after receiving the weather data.

When the play button is pushed or the broadcast time arrives, the system plays synthesized speech using a speech-rate converter. Even if sentences are skipped, a temporary stop occurs, or the broadcast time is changed, it is possible to fit the remaining synthesized speech into the remaining broadcast time frame, because the system controls the synthesized-speech rate in real time.

We have examined the system every weekday from June 6, 2011 to the present. Three sets of weather data (for the morning, evening, and night) are uploaded daily. In our trial, a single subject has listened to the weather report program produced by our proposed system approximately 750 times. The results have identified no significant problems with the stability of the system or the quality of the synthesized speech. We are therefore planning to use the system for the real broadcast in spring, 2014.

X. Conclusion

We developed a speech-synthesis method using templates that can generate a recording-sentence set for a speech database, and produce natural sounding synthesized speech. Applying this method to the NHK weather report radio program reduced the size of the required recording-sentence set to just a fraction of that required by a comparable method. After integrating the recording voice of the sentence set into the speech database, speech was synthesized using templates. In a paired comparison test, 66 % of the speech samples produced by the proposed voice synthesizer using templates were preferred to those produced by a conventional voice synthesizer. In an evaluation test using a five-point MOS scale, the speech samples generated by the proposed voice synthesizer scored 4.97, whereas the highest score achieved by a commercially available voice synthesizer was 3.09. We also developed an automatic broadcast system for the weather report program using the proposed speech-synthesis method and speech-rate converter. The system has been evaluated using real weather data for more than 1 year, and has been shown to have sufficient stability and synthesized speech quality for broadcast use.

Acknowledgment

The authors would like to thank H. Tada at the Japan Meteorological Agency for advice on templates for the weather report program, and M. Seta for reading out the generated recording-sentence set.

References

[1] NHK, *50 Years of Japanese Broadcasting*. Tokyo, Japan: NHK, 1977, pp. 29–33.

[2] 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, Feb. 2005.

[3] E. Moulines and F. Charpentier, “Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech Commun.*, vol. 9, pp. 453–467, Dec. 1990.

[4] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 1. 1996, pp. 373–376.

[5] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, “XIMERA: A new TTS from ATR based on corpus-based technologies,” in *Proc. 5th ISCA Speech Synthesis Workshop*, no. 1057. 2004, pp. 179–184.

[6] 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, Jun. 2008.

[7] H. Segi, T. Takagi, and T. Ito, “A concatenative speech synthesis method using context dependent phoneme sequences with variable length as search units,” in *Proc. 5th ISCA Speech Synthesis Workshop*, no. 1038. 2004, pp. 115–120.

[8] J. Demeur, P. Nguyen, and M. Vanlieferinge, “Speech announcement in the SNCB’s major railway stations,” *ACEC (Ateliers de Constructions Electriques de Charleroi) rev.*, no. 2, pp. 14–17, Feb. 1987.

[9] 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, Feb. 2010.

[10] S. Takaki, K. Oura, Y. Nankaku, and K. Tokuda, “An optimization algorithm of independent mean and variance parameter tying structures for HMM-based speech synthesis,” in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 4700–4703.

[11] T. Nose and T. Kobayashi, “An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model,” *Speech Commun.*, vol. 55, pp. 347–357, Feb. 2013.

[12] H. Kawai, T. Toda, J. Yamagishi, T. Hirai, J. Ni, N. Nishizawa, M. Tsuzaki, and K. Tokuda, “XIMERA: A concatenative speech synthesis system with large scale corpora,” *IEICE Trans. Inf. Syst.*, vol. J89-D, no. 12, pp. 2688–2698, Dec. 2006.

[13] A. Conkie, “A robust unit selection system for speech synthesis,” in *Proc. 137th Meeting ASA/Forum Acusticum*, vol. 105, no. 2. 1999, p. 978.

[14] H. Kawai, S. Yamamoto, and T. Shimizu, “A design method of speech corpus for text-to-speech synthesis taking into account prosody,” in *Proc. Int. Conf. Spoken Language Process.*, vol. 3. 2000, pp. 420–425.

[15] J. V. Santen, “Diagnostic perceptual experiments for text-to-speech system evaluation,” in *Proc. Int. Conf. Spoken Language Process.*, vol. 1. 1992, pp. 555–558.

[16] T. Hirai, S. Tenpaku, and K. Shikano, “Manipulating speech pitch periods according to optimal insertion/deletion position in residual signal for intonation control in speech synthesis,” in *Proc. Int. Conf. Spoken Language Process.*, vol. 3. 2000, pp. 330–333.

[17] C. Kuo and J. Huang, “Efficient and scalable methods for text script generation in corpus-based TTS design,” in *Proc. Int. Conf. Spoken Language Process.*, vol. 1. 2002, pp. 121–124.

[18] M. Isogai, H. Mizuno, and M. Kazunori, “Recording script design for corpus-based TTS system based on coverage of various phonetic elements,” in *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 1. 2005, pp. 301–304.

[19] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes*, 3rd ed. New York, NY, USA: Cambridge University Press, 2007, pp. 555–558.

[20] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in Pascal*. New York, NY, USA: Cambridge University Press, 1989, pp. 351–364.

[21] H. Segi, R. Tako, N. Seiyama, T. Takagi, H. Saito, and S. Ozawa, “Sentence-generating system for speech synthesis using templates and application for the weather report radio program,” *J. Inst. Image Inform. Televis. Eng.*, vol. 65, no. 1, pp. 76–83, Jan. 2011.