A JAPANESE TEXT-TO-SPEECH SYSTEM BASED ON MULTI-FORM UNITS WITH CONSIDERATION OF FREQUENCY DISTRIBUTION IN JAPANESE

Kimihito TANAKA, Hideyuki MIZUNO, Masanobu ABE and ShiN'ya NAKAJIMA
NTT Cyber Space Laboratories
Tel: +81-468-59-3938, Fax: +81-468-55-1054
E-mail: tanaka@nttspch.hil.ntt.co.jp

10.21437/Eurospeech.1999-217

ABSTRACT

This paper proposes our new text-to-speech (TTS) system that concatenates large numbers of speech segments to produce very natural and intelligible synthetic speech. One novel point of our system is its new synthesis unit, which is has three remarkable characteristics as follows;

(1) The synthesis units contain all Japanese syllables together with all possible vowel sequences, so very smooth synthetic speech is produced.

(2) Both previous and succeeding phoneme environments are considered when speech segments are concatenated, so natural sounding transients from a vowel to a consonant, which is the only concatenation point with the proposed unit, are present in the synthetic speech.

(3) Each unit has various fundamental frequency (F_0) contours. Therefore, F_0 modification rates are very small in any synthesis event, and the F_0 modification process causes only minor distortion.

To develop a unit database efficiently and effectively, we analyzed 4,850,000 Japanese phrases (breath-group) containing 87,810,000 phonemes and ranked them in order of appearance frequency. Listening tests confirm the high intelligibility and naturalness of speech produced by our new TTS system. It uses the 50,000 highest frequency units that cover over 77% of Japanese texts.

Keywords: text-to-speech system, multi-form unit, fundamental frequency

1. INTRODUCTION

In a Text-to-speech (TTS) system, the synthesis unit is one of the primary factors determining the quality of synthetic speech. Several kinds of synthesis units have been proposed such as CV, VCV, CVC, and tri-phones (context dependent phonemes)[1,2]. While the minimum number of these uniform units offers acceptable intelligibility, such restricted systems lack naturalness because concatenation is performed at transients or within phonemes. The recent increase in the performance of recent PCs has made it possible to increase the size of the speech segment inventory by utilizing non-uniform units. Some current systems yield more natural synthetic speech[3-6]. In such systems, high quality speech is synthesized when the context of the target is similar to the context of the speech corpus. To increase the coverage of such systems it is necessary to define an optimal function to select the most suitable unit and to develop a corpus that contains as many variations of units as possible. Moreover, the speech database of such a system would be too large. For example, it would contain many units rarely used in synthesis, or these were virtually the same as those of other units in the corpus. This paper proposes multi-form units, which can efficiently represent long phoneme sequences like non-uniform units. The multi-form unit approach has two main differences from the non-uniform idea. One is that only the necessary units are collected, and the other is that they were designed to solve two major problems. The first problem is acoustic discontinuity at the concatenation points, and the second is the quality degradation caused by large F_0 modification. Section 2 provides details on the multi-form unit design. Section 3 describes how to develop the unit database, and section 4 overviews our TTS system. In section 5, we show the evaluation results on intelligibility and naturalness.

2. MULTI-FORM UNIT

Japanese syllables consist of CV or V, and it is well known that transient parts from C (consonant) to V (vowel), or from V to V are very important for auditory sense. The quality of synthetic speech is degraded if speech segments are concatenated at C-to-V or V-to-V transients, because of the acoustic discontinuity so formed. Therefore, our new synthesis units were designed to include all Japanese syllables (CV) and all possible vowel sequences. The units consist of a consonant and a vowel chain placed after the consonant, such as /sa/, /bai/, /mjao/, and /kawaN/. Here, the vowel chain can be any sequence of vowels, semi-vowels, or syllabic nasals. This unit is named the multi-form unit, and is identified by symbol sequence C(V)k, where k denotes the number of vowels in the vowel chain. C(V)k units can produce natural sounding syllables and vowel sequences, because they cover all syllables and all possible vowel sequences.

In the C(V)k framework, the units are always concatenated at V-to-C transients. For smooth concatenation, we consider both the previous and succeeding phoneme environments of the C(V)k units. This consideration makes it possible to realize natural sounding V-to-C transients and to reduce distortion at the connections.

According to the speech production theory[7], the speech

spectrum characteristics are influenced by F_0, so the spectra of segments uttered with high, low, increasing, and decreasing F_0 contours differ quite a lot from each other. This means that the quality of synthesized speech is degraded when the F_0 of a synthesis unit is modified strongly[8]. In our approach, each phoneme sequence is uttered using the F_0 contours that appear frequently in Japanese speech, so the distortion caused by F_0 modification can be decreased by choosing the combination of unit and F_0 contour that is nearest to the target.

3 DATABASE DESIGN

3.1 Motivations

The C(V)k unit can represent any Japanese phoneme sequence, but the number of units appears to be almost infinite if we are to synthesize any text; moreover, most units are rarely used. Our approach is to establish the unit database by considering the appearance frequency of C(V)k units in Japanese speech. Accordingly, it is desirable to analyze a large number of Japanese speech samples, but it is very difficult to collect enough speech samples whose phoneme boundaries and pitch marks are already labeled. Therefore, we analyze large amount of Japanese texts utilizing our text analyzer and a prosody generator, and rank the units in order of frequency of appearance. To synthesize phoneme sequences that are represented by C(V)k units stored in the database, we also prepared a simple diphone unit set that can synthesize any Japanese text.

3.2 Analysis of Japanese Text Data

About 4,850,000 Japanese phrases (breath-group) containing 87,810,000 phonemes were analyzed. The phrases were taken from a large text corpus on newscasts, novels, and magazines. We collected about 41,000,000 C(V)k units by text analysis and prosody generation, and they were classified into 100,000 different C(V)k units by quantizing the averages (5 grades) and slopes of the F_0 contours (10 grades). Table 1 shows the 20 most frequent C(V)k units. In table 1, the parenthesized phonemes mean phoneme environments, and symbol '#' indicates that the unit was derived from the top of a phrase when '#' was at the front of unit, or from the bottom of a phrase when '#' was at the end of unit. For example, the first unit "ta(#)" means that the unit, "ta", was extracted from the bottom of many phrases. The second unit "(a)s(t)" means that the unit, 's', was often placed after the previous vowel 'a' and in front of the succeeding consonant 't'; this is a special case in that there are no vowels in the unit.

Figure 1 shows the distribution of unit length, calculated from the 100,000 C(V)k units. The horizontal axis plots the length of units, and vertical axis indicates the distribution ratios for each length. Over 40% of all units consist of two phonemes, in particular, about 70% of them are the Japanese syllable CV. The remainders, about 60%, are long units that contain over two vowels. These yield good performance when continuous vowel sequences are synthesized.

Figure 2 shows the relation between of the number of units and the cover ratio of the units in Japanese text. The graph indicates that it is necessary to collect about 50000 units, to cover over 75% of Japanese text.

Table 1 The 20 most frequent C(V)k units

Rank	C(V)k	Rank	C(V)k
1	ta (#)	11	to (n)
2	(a) s (t)	12	ki (n)
3	te (#)	13	(a) s (#)
4	ko (t)	14	to (m)
5	ku (n)	15	to (s)
6	(o) s (t)	16	ta (m)
7	(i) ma (s)	17	ta (m)
8	ka (r)	18	tei (m)
9	ka (k)	19	(a) re (t)
10	(i) ma (s)	20	ka (n)

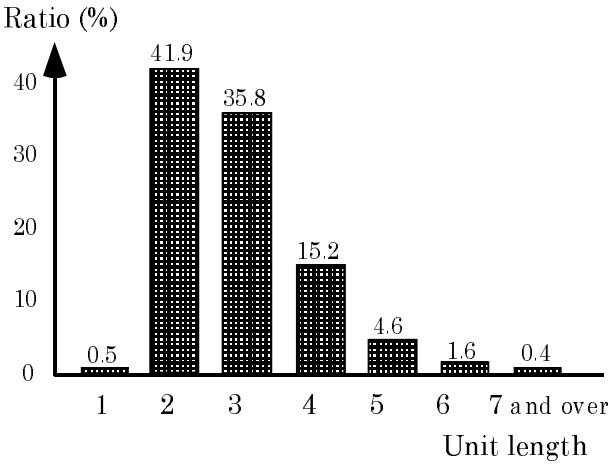

Fig.1 Distribution of unit length

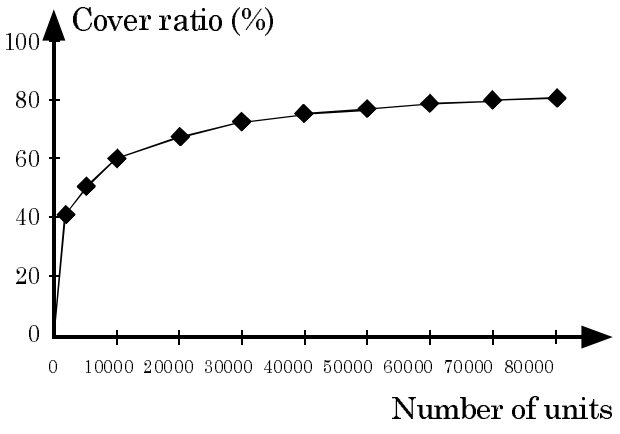

Fig.2 Cover ratio in Japanese text

4 SYSTEM OVERVIEW

Our TTS system has four modules: text analysis, prosody generation, unit selection, and speech synthesis. In the text analysis module, the input Japanese text, which consists of Kanji and Kana sequences, is transformed into phonetic symbols and accent patterns. The prosody generation module sets the F_0 contours, power patterns, and duration of each phoneme using prosody templates. In the unit selection module, the most suitable synthesis unit is determined by speech segment inventory. The speech synthesis module modifies the F_0 contours, power patterns, and duration of the speech segments to match those of the target; they are concatenated to synthesize the target speech. Sections 4.1 and 4.2 detail the unit selection module and the speech synthesis module, respectively.

4.1 Unit Selection Module

First we attempt to locate suitable C(V)k units and diphone units are used when this search fails. Figure 3 illustrates the block diagram of the unit selection procedure. Numbers in the following explanation correspond to the block numbers in Figure 3.

(1) Every C(V)k unit corresponding to the target phoneme sequence is searched in the speech segment inventory.

(2) The most suitable speech segment, whose F_0 contour most closely match that of the target, is selected from the candidates.

(3) If process (1) fails, diphone units are used for synthesis.

4.2 Speech Synthesis Module

The speech synthesis module consists of two parts, one is the F_0 modification part, and the other is the segment concatenation part. For F_0 contour modification and duration control, we use a TD-PSOLA-like algorithm[9]. The concatenation part uses one of the two below methods, depending on the segments being concatenated.

(a) Two C(V)k units are concatenated at the phoneme boundaries.

(b) A C(V)k unit and a diphone unit, or two diphone units are concatenated within the phonemes. The concatenation point in each phoneme is changed according to the features of the phoneme.

5 EVALUATION AND DISCUSSION

We have already developed a unit database that contains 50,000 C(V)k units and 10,000 simple diphone units. Its content, equivalent to 6.3 hours, occupies about 1GB. To evaluate intelligibility and naturalness, we conducted listening tests. Experimental conditions are shown in table 2; the sampling rate of speech segments was 22.05kHz, ten subjects participated in each test, and the numbers of C(V)k units and diphone units in the experimental system were 50,000 and 10,000, respectively.

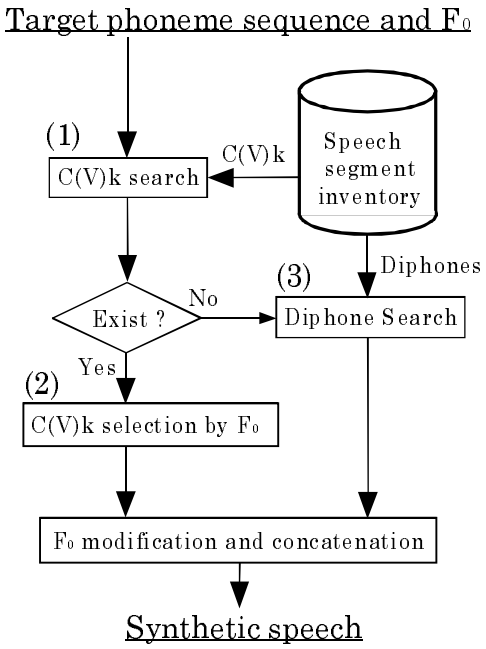

Fig. 3 Blockdiagram of unit selection

Table 2 Experimetal conditions

Sampling rate	22.05kHz
Number of C(V)k units	60,000
Number of diphone units	10,000
Number of subjects	10

5.1 Word Intelligibility

To evaluate the intelligibility of synthetic word speech, 100 Japanese family names were presented to the 10 subjects. Half of the names were very familiar, and the remainders are unfamiliar. Most names consisted of 3 or 4 syllables. The names were synthesized by the proposed TTS system using standard F_0 range and speed. In the tests, synthesized speech stimuli of familiar and unfamiliar names were presented at random; the subjects heard them only once using a closed headphone. Table 3 shows the intelligibility score so obtained. For comparison, the intelligibility score of our conventional TTS system[2], which has 6,000 triphone based units, are shown in the table[10]. All experimental conditions of the two dictation tests were the same. 95.1% of the speech samples synthesized by our new system were found to permit accurate hearing. In detail, 99.2% of familiar names and 91.0% of unfamiliar names were

heard correctly. Most mistakes were caused by the low quality of some kinds of voiced consonants. For example, they heard “Hamada” instead of “Yamada”, or “Nakashima” instead of “Nakajima”. These mistaken consonants were mainly synthesized by diphone units, and collecting more C(V)k units would be one solution. In conclusion, the proposed TTS system is slightly superior to the conventional system while offering high intelligibility.

Table 3 Intelligibility scores

Conventional	92.9 %
Proposed	95.1 %

5.2 Naturalness

Listening tests were carried out to evaluate the naturalness of the synthetic speech. Ten common Japanese sentences were synthesized by the proposed system and the conventional system[2] using the same prosody settings. To avoid the effects from the strange prosody sometimes created by the text-analyzer or the prosody generator, we used Sesign99[11], to reform the F_0 contours, power patterns and duration of synthetic speech. Ten subjects judged which was more natural, synthetic speech by the proposed system, or the one by the conventional system. Figure 4 shows the results of tests. 91% of subjects judged the synthetic speech by the proposed TTS system was more natural than that produced by the conventional system. This indicates that the new TTS system offers a high level of naturalness.

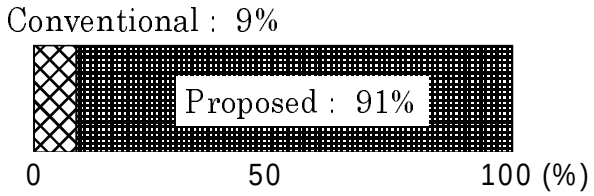

Fig. 4 Results of naturalness tests

5 CONCLUSION

A new text-to-speech system has been proposed based on the concept of the multi-form synthesis unit. Such units contain all Japanese syllables and all possible vowel sequences. They produce smooth concatenations because they consider previous and succeeding phoneme environments. Several F_0 contours were prepared for each phoneme sequence, so strong F_0 modification is not needed which reduces distortion. Listening tests using 60,000 synthesis units confirmed the new TTS system can produce very natural synthetic speech that is sufficiently intelligible. These results indicate that the multi-form units are very effective for high quality speech synthesis. In the future, we will extend our system by adding a new prosody generation algorithm[12] and a new F_0 modification algorithm[13]. We will also develop a male speech segment database, and apply the TTS system to various multimedia applications.

ACKNOWLEDGEMENT

We are grateful to the members of the Speech Synthesis Department for their helpful discussions. We also thank Mr. Yamamori, executive manager of the Media Processing Project, for his continuous support of this work.

REFERENCES

[1] K.Hakoda, S.Nakajima, T.Hirahara, and K.Kabeya, “Japanese text-to-speech synthesizer,” Journal of the American Voice I/O Society, Vol.6, pp.1-16 (1989).

[2] K.Hakoda, T.Hirokawa, H.Tsukada, Y.Yoshida, and H.Mizuno, “Japanese text-to-speech software based on waveform concatenation method,” AVIOS’95, pp.65-72 (1995).

[3] Y.Sagisaka, “Speech synthesis by rule using an optimal selection of non-uniform synthesis unit,” ICASSP’88, pp.679-682 (1988).

[4] K.Fujisawa and N.Campbell, “Prosody-based unit selection for Japanese speech synthesis,” 3rd ESCA/CCSDA International Workshop on Speech Synthesis, pp.181-184 (1998).

[5] M.Macon, A.Cronk, and J.Wouters, “Generalization and discrimination in tree-structured unit selection”, 3rd ESCA/CCSDA International Workshop on Speech Synthesis, pp.195-200 (1998).

[6] Y.Stylianou, “Concatenative speech synthesis using a Harmonic plus Noise Model,” 3rd ESCA/CCSDA International Workshop on Speech Synthesis, pp.261-266 (1998).

[7] G.Fant, “Acoustic theory of speech production,” Mouton, The Hague (1960).

[8] K.Tanaka and M.Abe, “A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F_0,” ICASSP’97, pp.951-954 (1997).

[9] H.Valbret, E.Moulines, and J.P.Tubach, “Voice transformation using PSOLA technique,” Speech Communication 11, pp.175-187 (1992).

[10] Y.Yoshida, S.Nakajima, K.Hakoda, and T.Hirokawa, “A new method of generating speech synthesis units based on phonological knowledge and clustering technique,” ICSLP’96, pp.1712-1715 (1996).

[11] H.Mizuno, M.Abe, and S.Nakajima, “Development speech design tool “Segin99” to enhance synthesized speech,” EUROSPEECH’99 (1999).

[12] M.Isogai and H.Mizuno, “A new F_0 contour control method based on vector representation of F_0 contour,” EUROSPEECH’99 (1999).

[13] S.Takano and M.Abe, “A new F_0 modification Algorithm by manipulating harmonics of magnitude spectrum,” EUROSPEECH’99 (1999).