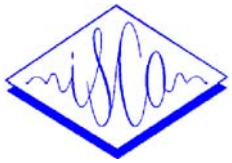

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A V-CV Waveform based Speech Synthesis Using Global Minimization of Pitch Conversion and Concatenation Distortion in V-CV Unit Sequence

Takao Koyama and *Jun-ichi Takahashi*
NTT DATA CORPORATION
Laboratory for Information Technology
Kayabacho-tower Bldg. 1-21-2 Shinkawa Chuo-ku Tokyo 104-0033 Japan
kym@rd.nttdata.co.jp

ABSTRACT

This paper proposes a new speech synthesis method for high-quality Japanese TTS(Text-to-speech) based on the waveform synthesis. The method uses V-CV as a basic synthesis unit to preserve the intelligibility of consonant. An efficient unit reconstruction method is newly adopted both to minimize pitch conversion and concatenation distortion when selecting waveforms. The minimization can provide fluency for synthesized speech. Furthermore, the proposed method enables to make a compact waveform dictionary keeping with high quality of synthesized speech. Using the waveform generation function of the method, the size of waveform dictionary can be drastically reduced by 1/40. Experimental evaluation using 32 ordinary peoples showed that high intelligibility of 97% was attained by the proposed V-CV speech synthesis method.

10.21437/Eurospeech.1999-504

1. INTRODUCTION

In speech synthesis, synthesis unit selection is very important to realize high quality synthesized speech. There mainly exist two types of synthesis units: phoneme and syllable (CV). The CV-type units are used for most Japanese TTSs[1][2][3], but there are several serious problems in the following: a large number of prosody patterns for various phonetic environments, noise caused by unit concatenation at spectral transitional point, and phonetic confusion. To solve these problems, we focused on the V-CV synthesis unit. Because this type of unit has the following advantages especially in synthesized speech quality: smooth concatenation of units at vowel steady position, a small number of vowel-based synthesis units for concatenation, and easy control of PSOLA-based pitch continuity due to V-CV's inherent feature of precise pitch contour. In the proposed method, several typical waveforms with different pitch patterns are selected for each kind of V-CV unit through clustering pitch variations. Because several waveform candidates for each V-CV unit enable to concatenate units smoothly with less pitch conversion without reducing naturalness of original waveforms. When concatenating units, we devised an effective waveform selection method in which spectral distortion is globally minimized for V-CV unit sequence through dynamic programming based on the Euclidian distance measure. This contributes to not only reduce concatenation noise but also increase naturalness of concatenated speech.

In the following section, principle of the proposed method is described. Section 3, 4, and 5 describes important technical issues of the method. Result of quality evaluation is also shown in section 6.

2. V-CV BASED SPEECH SYNTHESIS PROCESSING

The V-CV based speech synthesis processing flow is shown in Fig.1. Text analysis is carried out to get phonetic symbol sequences and word accent types from a Kanji string. Target prosody patterns, pitch patterns, and phoneme duration, are mainly set up through table look-up method at prosody setting process. These are registered in the prosody parameter table. The number of word mora and its accent type are used for determination of word pitch contour. The duration of each phoneme is determined using a duration table and is adjusted for desired utterance speed of speech. The pitch contour and duration of the selected waveforms are adjusted to the target prosody pattern by the PSOLA[4][5] method. After that, the waveforms are concatenated with smoothing. The processes of waveform selection and waveform generation for substitution are technically important for

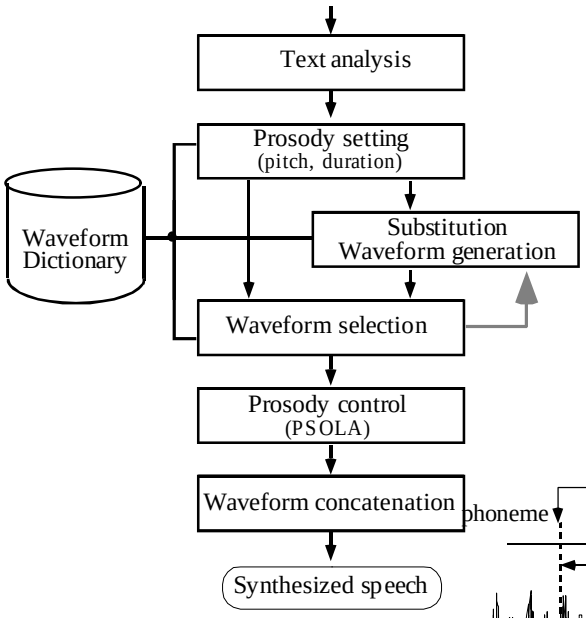

Fig. 1 Processing Flow of V-CV Waveform Speech Synthesis

high-quality of synthesized speech. For these processes, we devised effective methods in the following sections.

3. WAVEFORM SELECTION

There are two steps for selecting waveforms from waveform dictionary. At the first step, waveform candidates are selected suitable for target prosody pattern. At the second step, the optimum waveforms are determined for smooth spectrum continuity of waveform boundaries. The waveform selection is illustrated in Fig. 2. Waveform candidates are chosen using the parameters of pitch frequency of each waveform boundary and duration of each phoneme at the first step. The waveform candidates are evaluated using scores generated by the following evaluation function *Ef*.

The waveforms existed within the permitted region in pitch threshold are allowed to use as desired waveform candidates. The permitted pitch region is defined by the maximum pitch conversion ratio. The pitch conversion ratio is given as a constant. If there is no waveform within the region, the waveform with the minimum evaluated value is selected as a waveform candidate. At the second step, the Euclidean distance for waveform candidate sequences is calculated using cepstral coefficient parameters of each waveform candidates. Only a waveform sequence is determined as the best sequence

$$+ D_{fv}(d_{fvt} - d_{fvs})^2 + D_c(d_{cvt} - d_{cvs})^2 + D_{bv}(d_{bvt} - d_{bvs})^2$$

P_f P_b D_{fv} D_c D_{bv} : constant
p_{ft} p_{bt} : target pitch frequency
p_{fs} p_{bs} : waveform pitch frequency
d_{fvt} d_{cvt} d_{bvt} : target duration
d_{fvs} d_{cvs} d_{bvs} : waveform duration

with the minimum cepstral distortion in all combinations of candidates.

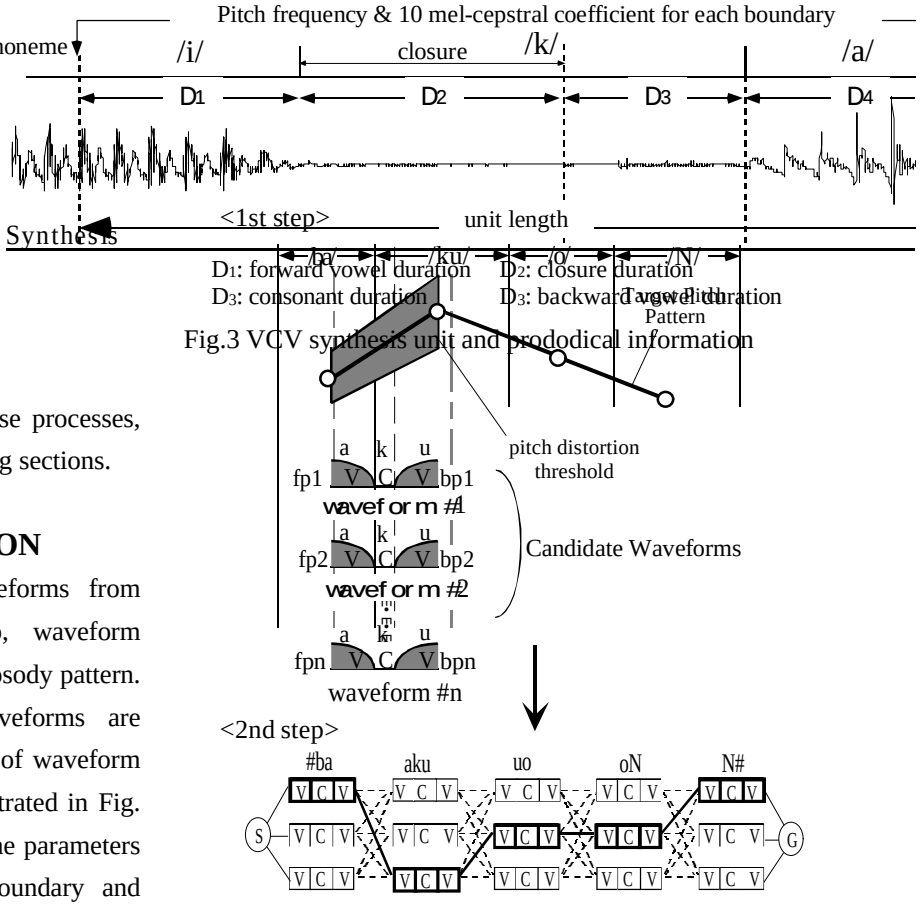

Fig.3 VCV synthesis unit and prododical information

Fig.2 Waveform selection process

4. DESIGN OF WAVEFORM DICTIONARY

The fundamental synthesis unit is a V-CV. The boundaries of V-CV unit are defined as the center of a vowel for left-side vowel V and the center of the next vowel for right-side vowel CV. Each boundary is determined manually at the zero-cross position according to the empirical method for each phoneme. Each unit has the following information: pitch frequency of each segment boundary, duration of each vowels and closure section and consonant, 10 mel-cepstral coefficients of each waveform boundary. An example of practical V-CV

unit is shown in Fig.3.

The specification of waveform dictionary is shown in Table 1. There are about 7100 typical V-CV waveforms. In addition, there are two types of specified synthesis units. One contains unvoiced syllable. Unvoiced syllable is one of the specified phenomena, but is so important to preserve the naturalness of Japanese. The other is the vowel- /i/-CV sequence unit. In this case, /i/'s duration is very short and the boundary between precede vowel and /i/ is not uttered clearly like diphthongs. They are also required to realize the fluent synthesized speech.

The speech data was digitized at a 12-kHz sampling rate. Acoustic segments with phoneme labels were obtained manually.

Table 1 Specification of Waveform dictionary
(number of waveform segments)

First mora segments	normal(voiced)	285	319
	unvoiced	34	
Medial mora segments	normal(voiced)	6809	8331
	vowel + /i/	1056	
	unvoiced	466	
End mora segments			38
Total			8688

dictionary size : 40 MB (12kHz-16bit sampling)

5. WAVEFORM GENERATION FOR SUBSTITUTION

When no V-CV units suitable for waveform candidates are found in the waveform dictionary, substitutive waveform is generated. Furthermore, if there are no waveform within the permitted pitch selection region or if the concatenation distortion is larger than the given threshold, the waveform generation for substitution is used.

As shown in Fig. 4, this process is carried out by dividing the V-CV unit into the precede vowel V and the followed segment CV and then reconstructing the required V-CV waveform. This process is performed only when the consonant C is unvoiced phoneme. When the consonant C is voiced phoneme, it is difficult to reconstruct the divided parts V and CV without noise superposition.

According to this waveform generation process for substitution, the number of waveform combination increases by 40 times. Therefore, about 320,000 waveforms can be used with decreasing concatenation distortions globally.

6. QUALITY EVALUATION

We compared the proposed method with the waveform-CV method[6] in synthesized speech quality. The evaluation was conducted with a headphone environment using transaction names generally encountered during actual banking service. Intelligibility

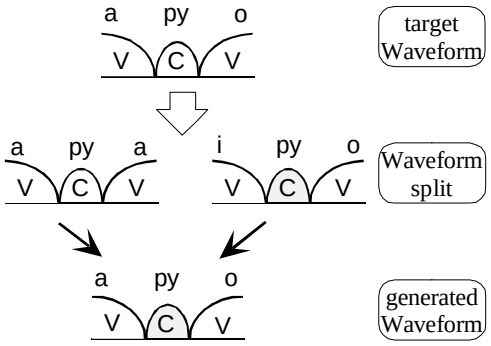

Fig.4 Substitution Waveform Generation

test was conducted by 32 ordinary people (equally spread in age) who had never heard any synthesized speech. The intelligibility was evaluated by a written test using 200 Japanese family names, consisting of 100 familiar names and 100 unfamiliar names. 3- and 4-mora Japanese names were used to exclude anamnesis as a factor. The familiar names were aiming to evaluate the outline of intelligibility in practical service, and the unfamiliar ones were used to evaluate the actual intelligibility of each synthesis method. Although the well-known Japanese TTS evaluation guidelines [7][8] recommend the use of at least two thousand words for an intelligibility test, in order to save a lot of time for evaluation we conducted the evaluation under the minimum conditions for obtaining sufficient results to compare the two methods. The results are shown in Table 2. This shows that high intelligibility of 97% was achieved compared with 94% for the waveform-CV method.

Table 2 - Results of Intelligibility Tests

	Name Category	Accuracy	Total Accuacy
Waveform CV method	familiar	97%	94%
	unfamiliar	91%	
Waveform V-CV method (proposed method)	familiar	98%	97%
	unfamiliar	96%	

7. CONCLUSION

A new speech synthesis was proposed, which can produce high-quality synthetic speech based on V-CV waveform synthesis units. Our method can eliminate noise superposition and preserve the articulation of consonants. The waveform generation for substitution plays an effective role of minimizing the pitch conversion and preserving the articulation. Furthermore, our waveform selection method enables to minimize the concatenation distortions globally. Therefore, we found that our method can provide a very high intelligible synthesized speech compared with the conventional waveform-CV method. In the intelligibility evaluation using 200 transaction names of practical banking service, high intelligibility of 97% was attained. For naturalness, we have already made sure that our method is more natural than the other methods.

REFERENCES

[1] M. Kitai, K. Hakoda, and S. Sagayama, "Trends of ASR and TTS Applications in Japan," *Proc. of IVTTA96*, pp. 21-24 (Sep. 1996).

[2] T. Hirokawa, K. Itoh, and H. Sato, "High Quality Speech Synthesis System Based on Waveform Concatenation of Phoneme Segment," *IEICE Trans. Fundamentals*, Vol. E76-A, No.11, pp. 1964-1970 (1993).

[3] T. Hakoda, T. Hirokawa, H. Tsukada, Y. Yoshida, and H. Mizuno, "Japanese Text-To-Speech Software based on Waveform Concatenation Method," *Proc. of AVIOS95* pp. 65-72 (1995).

[4] T. Hirokawa, and K. Hakoda, "Segment Selection and Pitch Modification for High Quality speech Synthesis using Waveform Segments," *Proc. of ICSLP90* pp. 337-340 (1990).

[5] F.J. Charpentier and M.G. Stella, "Diphone Synthesis using an Overlap-Add Technique for Speach Waveforms Concatenation," *Proc of ICASSP86* pp. 2015-2018 (1986).

[6] T. Koyama, T. Horie, T. Yoshioka, F. Yoshitani and J. Takahashi, "A Highly Intelligible Speech Synthesis for Banking Services in Financial Network System ANSER," *Proc. of IVTTA98*, pp. 87-90 (Sep. 1998).

[7] Hideki Kasuya, "Assessment of Speech Synthesis Technology," *The Journal of the Acoustical Society of Japan*, Vol. 48, No. 1, pp. 46-51, 1992 (in Japanese).

[8] JEIDA "JEIDA Guideline for Speech Synthesizer Evaluation," *Research Report on Office Automation Equipment Standardization* pp. 221-241 (Mar. 1995).