PAPER

A study on the word synthesis method by using the VCV-balanced word database

Ryo Mochizuki,* Yasuhiko Arai,* and Takashi Honda**

* *AVC Research Lab., CED, Matsushita Communication Ind. Co., Ltd.,*

600 Saedo-cho, Tsuzuki-ku, Yokohama, 224-8539 Japan

** *School of Science and Technology, Meiji University,*

1-1-1, Higashi-mita, Tama-ku, Kawasaki, 214-8571 Japan

(Received 4 March 1999)

When a conventional word database is used to synthesize words, it was thought that a lack of pitch variety in synthesis segments was the major reason why the sound quality degradation occurred in the synthesized words. In order to reduce the sound quality degradation caused by pitch modification, we developed a word synthesis method taking into consideration the pitch variety in the database. The synthesis method uses a VCV-balanced word database consisting of words uttered forcibly in type-0 and type-1 accent patterns. This database contains the VCV-segments that are the same with regard to phoneme sequence but have different pitch patterns. We also improved the pitch modification process by using the techniques that the spectral features of each pitch waveform are adjusted according to the pitch modification, and the micro prosody inherent in the VCV-segment is retained. A pitch pattern matching measurement and a subjective evaluation confirmed the effectiveness of the new synthesis method as follows. (1) The total average of the pitch pattern matching distance obtained when the proposed database is used is approximately two thirds of that obtained when the conventional database is used. (2) The sound quality of synthesized words using the proposed database has marks more than 75% of preference score against that using the conventional database.

Keywords: Concatenation synthesis, Word database, Pitch accent type, Pitch pattern matching measurement, Pitch modification

PACS number: 43.72.Ja

1. INTRODUCTION

In speech output systems such as those used in car navigation systems and paging systems, voice messages are composed by inserting words into prerecorded sentence patterns. The sound quality of the inserted words should of course be as close as possible to that of the words in the prerecorded sentences. Especially when a very large vocabulary is required, a system that could synthesize natural-sounding words would be desired. This is because it would be almost impossible for a single narrator to maintain a consistent level of quality while all the required words were recorded. Pitch waveform

concatenation synthesis methods have therefore been widely studied in order to develop methods to synthesize natural-sounding Japanese words. The sound quality yielded by the synthesis methods available so far depends on the variety of the waveforms in a speech database. Therefore, how to construct the database is the problem.

One well known method for constructing a speech database is the context-oriented clustering method,¹⁾ which uses a segment quantization technique. This method generates new synthesis units consistent with spectral parameters, but does not take into account the variety of pitch patterns in these synthesis units. When a database consisting of such synthesis units is

used, the quality of the synthesized speech might be degraded due to the occurrence of the high pitch modification rate.

The concatenation synthesis method of the CHATR²⁾ which does not use pitch modification could synthesize natural-sounding speech if the speech database comprised speech segments that matched the synthesis environment. This method, however, is not suitable for use when the hardware size of the speech output system not extremely large (CHATR holds a huge database). Moreover, although this method takes the pitch frequencies in speech segments into account when selecting the optimal segment set, it does not necessarily consider the balance of pitch frequency in the database.

The purpose of our study is the synthesizing of any word with high-quality sound. Our first approach used a VCV-balanced word database made up of words that had been uttered with the usual pitch accent in the Tokyo dialect.³⁾ Some words synthesized using this way sounded natural, but others did not. And the lack of pitch variety in waveforms available in the word database was one of the main reasons some words did not sound natural.

In order to solve this problem, it will be necessary to construct a database that contains VCV-segments with various pitch patterns. We developed a VCV-concatenation word synthesis system with a word database that consists of two VCV-balanced word sets. The words in the sets were uttered in two pitch accent types typical of Japanese (the type-0 and the type-1). The system not only uses an im-

proved database, but also improves the pitch modification process by adjusting the spectral features of each pitch waveform according to the pitch modification rates and by using the micro prosody inherent in the VCV-segment.

This paper describes the system for word synthesis by VCV-concatenation in Section 2 and describes the new database construction method based on the pitch variety in Section 3. The pitch modification techniques reducing the degradation of sound quality are described in Section 4, and the results of pitch pattern matching measurement and subjective evaluation are reported in Sections 5 and 6.

2. WORD SYNTHESIS SYSTEM

The outline of the system for word synthesis by VCV-concatenation is shown in Fig. 1. The system input is a letter sequence of the target word and a symbol designating accent type. The letter sequence is resolved into VCV-letter sequences that are used to find candidate VCV-segments (VCV-waveforms) in the word database. The word database consists of two VCV-balanced word sets (described in detail in Section 3). The average pitch patterns of each accent type, which are included in the pitch pattern database, are used for the target pitch patterns. The optimal set of VCV-waveforms is selected by measuring the pitch pattern matching between the original and target pitch frequencies. The pitch frequency of each selected VCV-waveform is modified according to the target pitch pattern. The modified VCV-waveforms are concatenated to yield the synthesized speech

Fig. 1 Outline of word synthesis system using a VCV-balanced word database.

waveform.

In the synthesis processes, because pitch modification is important to sound quality, the following techniques are used.

- (a) The VCV-segment set that minimizes the pitch modification rate is selected.⁴⁾
- (b) The pitch waveforms are extracted by referring to excitation points.³⁾
- (c) The wavelengths of pitch waveforms are adjusted in ways that depend on the pitch modification rates.⁸⁾
- (d) The micro prosody in VCV-segments is used.⁸⁾

Section 4 deals with these techniques in further detail.

3. WORD DATABASE CONSTRUCTION

3.1 Accent Types of Japanese Words

Generally, an N -mora Japanese word could take $N+1$ pitch accent types. For example, when an N -mora word has the type- n pitch accent ($1 \leq n \leq N$), its pitch frequency begins to decrease abruptly at the n -th mora. On the other hand, when it has the type-0 pitch accent, there is no abrupt decrease in pitch anywhere. The pitch patterns of typical Japanese accent types are shown in Fig. 2. The figure shows that every pitch pattern except type-1 is similar to that of type-0 at the beginning of the word and that every pitch pattern except type-0 is similar to that of type-1 at the end. It can be inferred from these similarities that pitch patterns other than type-0 and type-1 can be generated by combining the type-0 and type-1 patterns as follows. (1) Use the pitch pattern of the type-0 accent to represent the pitch pattern from the beginning of the word to the accent position (before the pitch frequency begins to decrease). (2) Use the pitch pattern of the type-1 to represent the pitch pattern from the position after pitch frequency has decreased to the end of the word.

Fig. 2 Pitch patterns of typical Japanese accent types.

(3) On the accent mora (where the pitch frequency is on the decrease) use linearly interpolated between the type-0 and the type-1 patterns. In the following subsections, an accent conversion experiment is performed to confirm that words with accent type other than type-0 and type-1 could be generated by combining the words with the same letter sequence but with different pitch patterns (the type-0 and type-1 patterns), and then a word database that consists of Japanese VCV-balanced word sets uttered forcibly in the two pitch accent types is proposed.

3.2 Experimental Test of Accent Conversion

The Japanese word “Kuyashimagire” with its usual accent (type-4 in this case) was generated by using the following methods. (a) Modify the pitch pattern of the word uttered forcibly in type-1 to that of the same word uttered in type-4. (b) Modify the pitch pattern of the word uttered forcibly in type-0 to that of the same word uttered in type-4. (c) Concatenate, by using a taper window on the accent mora (the fourth mora in this case), the beginning part of the word formed by modification (b) and the ending part of the word formed by modification (a). Each of the words generated was compared with the original naturally uttered word, and the perceptibility of quality degradation was evaluated on a 4-point scale ranging from the imperceptible to the prominent. The subjects were fifteen adults working in the field of speech processing. They heard the speech materials through binaural headphones.

The results of this experiment are shown in Fig. 3. Degradation of sound quality was recognized at the beginning of the word converted from the type-1 accent and at the end of the word converted from the type-0 accent. The sound quality of the word

Fig. 3 Results of pitch accent conversion experiment.

generated by the combination of the type-0 and the type-1 accents was very similar to that of the original word even though the concatenation process had been performed on the accent mora. The method (c) could reduce the perceptibility of the pitch modification. It was considered that the idea of using the words with the type-0 and the type-1 accents was adequate for the synthesis of natural-sounding words.

3.3 VCV-Balanced Word Database

Because the results of the accent conversion experiment showed the effectiveness of generating a word with a specified accent type by combining two words uttered in the type-0 and type-1 accent patterns, we constructed a word database consisting of two VCV-balanced word sets, one uttered in the type-0 accent and the other uttered in the type-1 accent. Each word in the two word sets was uttered with a pitch accent type different from its usual one. Each word set (WD-I)⁵⁾ consisted of approximately 500 VCV-balanced words that had been uttered for this study by a professional female narrator. The word database contained most of the VCV-segments required to synthesize Japanese words. The advantage of using this database was that it comprised VCV-segments with the same sequence of phonemes but with different pitch patterns.

4. PITCH MODIFICATION METHOD

When a new word is synthesized, the pitch pattern of the original word should be changed to the target pitch pattern. And because the sound quality of synthesized words depends much on the pitch modification rates necessary for their synthesis, it is important to use a pitch modification method in which the synthesized words retain their original natural sounds. In this section, the pitch modification techniques that could reduce the degradation of sound quality are described.

4.1 Selection of Optimal VCV-Segments

The spectral features of synthesis units are conventionally used for selection of the optimal unit, and the effectiveness of these conventional methods has been shown.^{6,7)} The pitch modification rates, however, should be used as a selection criterion since the database in this study is constructed based on the pitch variety. That is, pitch modification rates of the VCV-segments should be as small as possible.

Therefore, the pitch pattern matching distance between the target and original pitch frequencies was used as a criterion function when the optimal VCV-segment set was selected. (This pattern matching measurement is described in detail in Section 5).

4.2 Extraction of Pitch Waveforms

Pitch waveforms were extracted from the selected VCV-waveforms by the method of excitation synchronous pitch waveform extraction,³⁾ which is based on the Phase Equalized Residual Excited Linear Prediction (PE-RELP) model.⁹⁾ This extraction is both fully automatic and free of extraction errors caused by the formant resonance. The pitch waveform was extracted from two adjacent excitation intervals by using an asymmetrical Hanning window and was rearranged according to the target pitch pattern.

4.3 Wavelength Adjustment of Pitch Waveforms

It is considered that the degradation of sound quality is a result of that spectral features of the pitch-modified VCV-waveform not matching the pitch frequency. The sound of pitch-modified VCV-waveform could be improved by modifying the wavelength of each pitch waveform according to the pitch modification rate. The relationship between the pitch modification rate and the wavelength modification rate is shown in Fig. 4. The wavelength modification rate, 2^{-r} (r [octave]), is determined depending on the pitch modification rate, 2^{-b} (b [octave]), by the following equation:

$$2^{-r} = 1 + a(2^{-b} - 1), \quad (1)$$

where a is a compensation coefficient determined by auditory tests.

Fig. 4 Wavelength adjustment based on pitch modification rate.

A previous study⁸⁾ showed that the quality improvement due to the wavelength adjustment is especially prominent in the downward pitch modification.

4.4 Use of VCV-Micro Prosody

In the conventional method for modifying the pitch frequency, the pitch waveforms in the selected VCV-waveforms are rearranged according to the target pitch pattern. When this is done, however, neither VCV-micro prosody nor the small perturbations of pitch period inherent in each VCV-waveform are retained. In this study, global pitch contours of VCV-waveforms were modified by using the pitch modification rates at VCV-concatenation points. Thus the VCV-micro prosody and the small perturbations of the pitch period inherent in each VCV-waveform were almost unchanged (Fig. 5). For example, the modified pitch frequency $M_p(t)$ is calculated by the following equation as a function of time.

$$M_p(t) = O_p(t) + \{D_p(t_2) - D_p(t_1)\} \frac{(t - t_1)}{(t_2 - t_1) + D_p(t_1)}, \quad (2)$$

where $O_p(t)$ is the original VCV-pitch frequency, and $D_p(t_2)$ and $D_p(t_1)$ are respectively the pitch frequency differences between the target pitch pattern and the original one at the VCV-concatenation points t_2 and t_1 .

In order to confirm the proposed pitch modification technique, the auditory test (preference evaluation) has been performed. Five words synthesized using the proposed pitch modification method were compared with those synthesized using the conventional method. Consequently, the proposed method has marked a preference score of 75% against the conventional method.⁸⁾

Fig. 5 Use of micro prosody in a VCV-waveform

5. PITCH PATTERN MATCHING

If the synthesized word is to be natural sound, the pitch patterns of the selected VCV-segments should be as close as possible to the target pitch pattern. To do this we constructed a new word database consisting of words uttered in the type-0 and type-1 accent patterns. In order to confirm the fact that VCV-segments with various pitch patterns contained in the database, pitch pattern matching measurement is performed in this section.

5.1 Measurement of Pitch Pattern Matching Distance

The pitch pattern matching distance is measured by the pitch modification rates at the centers of each of the vowels in the VCV-waveform. That is, the pitch pattern matching distances Pf_i and Pb_i are given by

$$Pf_i = |\log_2(Tf_i) - \log_2(Of_i)| \quad (3)$$

and

$$Pb_i = |\log_2(Tb_i) - \log_2(Ob_i)|, \quad (4)$$

where Of_i and Ob_i are respectively the original pitch frequencies at the V1 and V2 in the i -th VCV-segments (V1 and V2 are the first and the second vowels of VCV-segment, and i is VCV-segment number), and Tf_i and Tb_i are respectively the corresponding target pitch frequencies. The average pitch pattern matching distance per VCV-segment (Ps_i) is calculated by using the following equations.

$$Ps_i = Pb_i \quad (i=1), \quad (5)$$

$$Ps_i = (Pf_i + Pb_i)/2 \quad (1 < i < n), \quad (6)$$

$$Ps_i = Pf_i \quad (i=n), \quad (7)$$

where n is the number of VCV-segments used to synthesize the target word. The average pitch pattern matching distance per word (Pd) is calculated by using the following equation.

$$Pd = \frac{1}{n} \sum_{i=1}^n Ps_i. \quad (8)$$

In the work reported here the pitch pattern matching distance Pd was used for evaluating the degradation of the sound quality of the synthesized words. For example, the pitch pattern matching distance per VCV-segment Ps_i obtained when "FURANSU" (which means "France" in Japanese) was synthe-

sized is shown in Fig. 6. Five segments were used in this case: one CV-segment at the beginning of the word, three VCV-segments (including a VV-segment) in the middle, and one V-segment at the end. The pitch pattern matching distance per VCV-segment Ps_i was calculated at each segment,

Fig. 6 Pitch pattern matching distances measured when “FURANSU” (meaning “France” in Japanese) was synthesized.

Table 1 Range of pitch frequency in each word set.

Pitch frequency	VCV-balanced word sets		
	Normal accent	Type-0 accent	Type-1 accent
Average (Hz)	270	305	255
Maximum (Hz)	430	420	430
Minimum (Hz)	150	180	150

Table 2 Result of pitch pattern matching measurement. One hundred words (twenty words for each of the five accent types) were evaluated.

Accent type of synthesized words	Pd (octave)	
	(a) Original accent (Conventional database)	(b) Type-0 and Type-1 accents (Proposed database)
Type-0	0.23	0.13
Type-1	0.34	0.22
Type-2	0.38	0.26
Type-3	0.36	0.25
Type-4	0.31	0.20
Total averages	0.32	0.21

and the average pitch pattern matching distance per word Pd was calculated by averaging them.

5.2 Experimental Conditions

In this experiment, the words synthesized using the following databases were evaluated.

- The word set with the usual pitch accent for each word (conventional database).
- The word sets with type-0 and type-1 accents (proposed database).

Each set, which included the same words, consisted of words uttered by the same female narrator. The range of pitch frequency in her voice was between 150 and 430 Hz, and the range of pitch frequency in each word set is listed in Table 1. The words (one hundred words in total: twenty words for each of the five accent types) synthesized in this evaluation were selected from the VCV/CVC-balanced word set (WD-II).⁵⁾

5.3 Result of Pitch Pattern Matching Measurement

The average Pd score for each pitch accent type is listed in Table 2, from which we can see that the total average Pd obtained with the proposed database was approximately two thirds of the total average Pd obtained with the conventional database. It is shown that the proposed database is superior to the conventional one with regard to pitch pattern matching distance. The Pd score for the synthesis of words with type-0 accent is smaller than that of words with any other accent types. It is considered that the proposed database contains many of the VCV-segments that match with the pitch pattern of the type-0 accent. On the other hand, in the case of synthesizing words with the accent types other than type-0, the Ps_i (pitch pattern matching distance per segment) tends to become larger near the accent mora compared to the Ps_i at

Fig. 7 Result of preference evaluation.

the other positions. It is also considered that there are not enough VCV-segments suitable to the pitch-decreasing pattern compared with those suitable to the type-0 pattern in the database.

6. PREFERENCE EVALUATION

In order to compare the sound quality of the words synthesized using the proposed database with the sound quality of the words synthesized using the conventional one, we asked ten subjects to select the more natural-sounding word in each of ten pairs of synthesized words. The result of this preference evaluation is shown in Fig. 7, which shows that the proposed database has marked more than 75% of preference score against the conventional one. Less roughness was perceived in the words synthesized using the proposed database. The subjective quality evaluation thus shows the effectiveness of the proposed database.

7. CONCLUSION

In this study, the VCV-concatenation word synthesis system which employs the word database and the pitch modification method is proposed. In order to reduce the sound degradation, the word database that consists of VCV-balanced words with the type-0 and type-1 accents is studied. It is capable of having a variety of VCV-segments with the same sequence of phonemes but the different pitch patterns. To achieve higher sound quality, pitch modification techniques are also introduced. Then the pitch pattern matching measurement and the subjective quality evaluation confirm the effectiveness of the proposed word synthesis method as follows.

- (1) The total average of pitch pattern matching distance Pd obtained with the proposed database is approximately two thirds of that obtained with the conventional database.
- (2) The sound quality of synthesized words using

the proposed database has marks more than 75% of preference score against that using the conventional database.

Thus, the effectiveness of the proposed database has been shown in this paper, but spectral parameters at waveform connecting points in each VCV-waveform are not considered in selecting the optimal VCV-waveform set. There is a possibility that the synthetic speech quality could be further improved by considering the spectral parameter. The problem to be solved in the future is to find a more effective speech segment selection method.

REFERENCES

- 1) S. Nakajima and H. Hamada, "Automatic generation of synthesis units based on context oriented clustering," Proc. ICASSP 88, S14.2, 659-662 (1988).
- 2) W. N. Campbell and A. W. Black, "Chatr: a multilingual speech re-sequencing synthesis system," Tech. Rep. IEICE SP96-7, 45-52 (1996) (in Japanese).
- 3) Y. Arai, R. Mochizuki, H. Nishimura, and T. Honda, "An excitation synchronous pitch waveform extraction method and its application to the VCV-concatenation synthesis of Japanese spoken words," Proc. ICSLP 96, Vol. 3, 1437-1440 (1996).
- 4) R. Mochizuki, Y. Arai, and T. Honda, "A study on the natural-sounding Japanese phonetic word synthesis by using the VCV-balanced word database that consists of the words uttered forcibly in two types of pitch accent," Proc. ICSLP 98, Vol. 5, 2011-2014 (1998).
- 5) S. Hayamizu, K. Tanaka, S. Yokoyama, and K. Ohta, "Generation of VCV/CVC balanced word sets for speech data base," Bul. ETL 49, 803-834 (1985) (in Japanese).
- 6) Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," Proc. ICASSP 88, S14.8, 679-682 (1988).
- 7) N. Iwahashi, N. Kaiki and Y. Sagisaka, "Speech segment selection for concatenative synthesis based on spectral distortion minimization," IEICE(A) E76-A, 1942-1948 (1993).
- 8) Y. Arai, R. Mochizuki and T. Honda, "A study on natural-sounding Japanese phonetic word synthesis based on the pitch waveform concatenation," Proc. ICA98, 267-268 (1998).
- 9) M. Honda and T. Moriya, "Speech encoding based on phase equalization," Trans. CSR, Acoust. Soc. Jpn. S84-05, 33-40 (1984) (in Japanese).

Ryo Mochizuki received his B. E. and M. E. degrees in electrical engineering from Meiji University in 1995 and 1997, respectively. In 1997, he joined the AVC Research Laboratory, Matsushita Communication Industrial Co., Ltd. His current research and development interests are the design of the speech synthesis systems. He is a member of ASJ and IEICE.

Yasuhiko Arai received his B. E. and Dr.Eng. degrees from Waseda University in 1963 and 1988, respectively. In 1963, he affiliated with Matsushita Electric Co., Corporate Eng. Div., Matsushita Communication Industrial Co. Engaged in development of automatic public address systems and speech synthesis systems. Currently he is a professor of the Kanazawa Institute of Technology since 1999. His current inter-

ests are the assessment of human interface by speech, the sound quality evaluation of synthesized speech, and signal processing for hearing aids. He is a member of ASJ, Japan Audiological Society, JSST, IEICE, ASA and IEEE.

Takashi Honda received the B. E. and the Dr. Eng. Degrees in Electrical Engineering from Meiji University, in 1961 and 1968, respectively. He was a Lecturer and an Associate Professor at the Department of Electrical Engineering, School of Engineering, Meiji University, from 1968 to 1973. He is currently a Professor at the Department of Electrical and Electronic Engineering, School of Science and Technology, Meiji University. His current interests are in the area of speech synthesis and recognition systems. He is a member of ASJ, IEICE and JSST.