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

Ryo Mochizuki, Yasuhiko Arai* and Takashi Honda***
*AVC Research Lab., Matsushita Communication Industrial 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

ABSTRACT

In order to synthesize natural-sounding Japanese phonetic words, a novel VCV-concatenation synthesis with an advanced word database is proposed. The word database consists of VCV-balanced phonetic words which are uttered forcibly in type-0 and type-1 pitch accents. The advantage of using the advanced word database is that a variety of VCV-segments with the same phonetic chains and the different pitch patterns could be collected efficiently at the same time. The following pitch modification techniques are used to achieve the sound quality: (1) The optimal VCV-segment set which minimizes the pitch modification rate is selected. (2) Pitch waveforms are extracted by referring to excitation points. (3) Wavelengths of pitch waveforms are adjusted depending on the pitch modification rates. (4) Natural prosody in the VCV-segments in the database is effectively used. Superiority of the proposed database is ensured by means of the pitch pattern matching measurement and the subjective quality evaluation.

1. INTRODUCTION

In such speech output systems as used for car navigation or paging system, voice messages are composed by embedding some phonetic words in pre-recorded sentence patterns. In those systems, it is desired that the sound quality of phonetic words to be embedded should be as close as possible to that of pre-recorded sentences. Especially in the case that a very large or almost unlimited vocabulary is required, synthesis of very natural-sounding phonetic words would be desired since it is almost impossible for a single narrator to pre-record all words required for a whole system in unified quality. Therefore, new methods have been studied to synthesize the natural-sounding Japanese phonetic words based on the pitch waveform concatenation.

In the first approach of our study, the VCV-balanced phonetic word database, in which each word was uttered with a normal pitch accent, was used to synthesize new phonetic words by concatenation of VCV-segments[1]. It was made clear that some words synthesized in this way sounded natural, but others did not. A lack of VCV-variety in the word database was a major reason why some words did not sound natural.

In order to solve the above-mentioned problem, a novel VCV-concatenation synthesis method with an advanced database which includes pitch varieties of VCV-segments is proposed. The database consists of two sets of VCV-balanced phonetic words: one is uttered forcibly in type-0 accent and the other is uttered forcibly in type-1 accent. This database could have a variety of VCV-segments with the same phonetic chains and different pitch patterns, and therefore it could be used advantageously for synthesizing the very natural-sounding phonetic words.

2. DATABASE CONSTRUCTION

Generally in Japanese, an N-mora word could possibly take N+1 pitch accent types. For example, when an N-mora word has the type-n pitch accent ($1 \leqq n \leqq N$), its pitch frequency begins to decrease abruptly at the n-th mora. On the other hand, when it has the type-0 pitch accent, no abrupt pitch decreasing occurs anywhere. The pitch patterns of typical Japanese accent types are shown in Figure 1. 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. Therefore, it is considered that every pitch pattern could be produced by the combination of two typical accent types (type-0 and type-1) based on the pitch modification method.

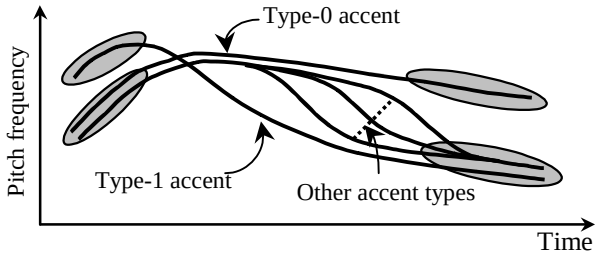

Figure 1: Pitch patterns of typical Japanese accent types.

Therefore, the word database which consists of two Japanese VCV-balanced word sets is proposed: one is uttered in the type-0 accent and the other is in the type-1. Both are uttered with different pitch accent from the normal. Each set consists of approximately 500 VCV-balanced words[2] which were uttered by a professional female narrator for this study. VCV-segments required to synthesize Japanese words are almost included in the above mentioned word database. The advantage of using the new word database is that VCV-segments which are the same in

phonetic chains, but different in pitch patterns could be collected at the same time in one database efficiently.

3. PITCH MODIFICATION

In the process of synthesizing a new word, it is required that the pitch frequency should be changed from original one to the target. Since the sound quality of synthesized words depends much on the pitch modification rates, it is important to use a pitch modification method by which synthesized words could retain the original natural sound, in addition to the effective construction of database.

In this chapter, new techniques, which play a significant role in reducing the sound deterioration caused by the pitch modification, are discussed. The following techniques are used to achieve the sound quality.

(1) The optimal VCV-segment set which minimizes the pitch modification rate is selected.

(2) The pitch waveforms are extracted by referring to excitation points.

(3) The wavelengths of pitch waveforms are adjusted depending on the pitch modification rates.

(4) The natural prosody in VCV-segments is used effectively.

The following sections deal with these techniques in further details.

3.1. Optimal VCV-segment selection

It is expected that the natural-sounding words could be synthesized by selecting the optimal VCV-segments so that the pitch modification rate would become as small as possible. Therefore, the pitch pattern matching distance between the target pitch frequency and original one is used as an criterion function to select the optimal VCV-segment set (Details are described in Chapter 4).

3.2. Pitch waveform extraction

Pitch waveforms are extracted from the selected VCV-waveforms by the method of excitation synchronous pitch waveform extraction[1], which is based on the Phase Equalized Residual Excited Linear Prediction (PE-RELP) model[3]. It takes advantages of being free from the extraction errors caused by the formant resonance and being fully automatic. The pitch waveform is extracted from two adjacent excitation intervals by using the asymmetrical Hanning window and rearranged along the target pitch pattern.

3.3. Wavelength adjustment of pitch waveforms

It is considered that the deterioration of sound quality is caused by the fact that spectral feature of the pitch-modified VCV-waveform does not match with the pitch frequency. The sound of the pitch-modified VCV-waveform could be improved by modifying wavelength of every pitch waveform based on the pitch modification rate. The relationship between the pitch modification rate and the wavelength modification rate is shown in Figure 2. The wavelength modification rate, 2^{-r}(r octave), is determined depending on the pitch modification rate, 2^{-b} (b octave), by the following equation.

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

where a is compensation coefficient, and it is determined by auditory tests.

A previous study showed that the quality improvement by the wavelength adjustment was seen especially in the downward pitch modification[4].

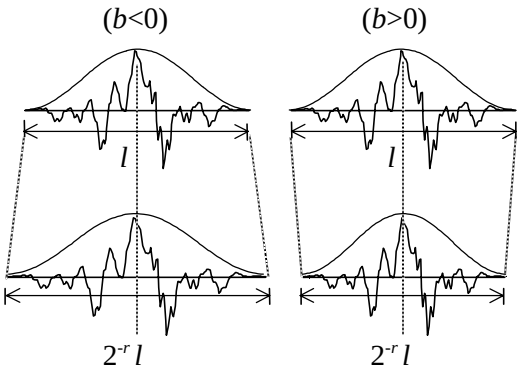

Figure 2: Wavelength adjustment depending on the pitch modification rate.

3.4. Application of VCV-micro prosody

In conventional methods for modifying the pitch frequency, every pitch waveform in selected VCV-waveforms is rearranged along the target pitch patterns. By these methods, however, VCV-micro prosody and the small perturbation of pitch period inherent in each VCV-waveform are not retained. In this study, global pitch contours of each VCV-waveform are modified along the target pitch patterns based on pitch modification rates only at VCV-concatenation points. Thus, the VCV-micro prosody and the small perturbation of the pitch period inherent in each VCV-waveform are held almost the same, as shown in Figure 3. For example, the modified pitch frequency $M_p(t)$ is calculated by the following equation as a function of time.

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

where $O_p(t)$ is the original VCV-pitch frequency in which VCV-micro prosody inheres and $D_p(t_2)$ and $D_p(t_1)$ are the difference in pitch frequency between the target pitch pattern and the original one at the VCV-concatenation points t_2 and t_1 respectively. The auditory test of five synthesized words in female voice resulted in an average preference score of 75% against the conventional method[4].

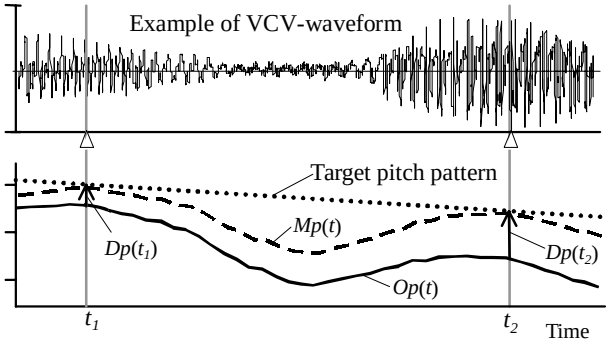

Figure 3: Pitch modification technique: use of micro prosody in a VCV-waveform.

4. PITCH PATTERN MATCHING

It is desired that the pitch contours of selected VCV-segments should be as close as possible to the target pitch pattern in order to synthesize the natural-sounding word. It is considered that a variety of VCV-segments could be collected efficiently by using the proposed database. Therefore, pitch pattern matching measurement is performed to confirm the effectiveness of the proposed database.

4.1. Pitch pattern matching measure

The pitch pattern matching measure is determined by the pitch modification rates at the centers of each vowel in the VCV-waveform. That is:

$$Pf_i = |\log_2(Tf_i) - \log_2(Of_i)| \quad \text{[octave]} \quad (3)$$

$$Pb_i = |\log_2(Tb_i) - \log_2(Ob_i)| \quad \text{[octave]} \quad (4)$$

where Pf_i and Pb_i are pitch pattern matching distances, Tf_i and Tb_i are target pitch frequencies, and Of_i and Ob_i are original pitch frequencies at front and back vowels in the i-th VCV-segments respectively (i is VCV-segment number). The average pitch pattern matching distance per VCV-segment (Ps_i) is calculated by the following equations.

$$Ps_i = Pb_i \qquad (i = 1) \quad \text{[octave]} \quad (5)$$

$$Ps_i = (Pf_i + Pb_i) / 2 \qquad (1 < i < n) \quad \text{[octave]} \quad (6)$$

$$Ps_i = Pf_i \qquad (i = n) \quad \text{[octave]} \quad (7)$$

where n is the number of VCV-segment used to synthesize the target word. In addition, the average pitch pattern matching distance per word (Pd) is calculated by the following equation.

$$Pd = \frac{1}{n}\sum_{i=1}^{n} Ps_i \quad \text{[octave]} \quad (8)$$

In this study, the pitch pattern matching distance (Pd) is used for evaluating the quality deterioration of synthesized words.

For example, the pitch pattern matching distance per VCV-segment (Ps_i) in the case of synthesizing "FURANSU" which means "France" in Japanese is shown in Figure 4. There are five segments used in this case: a CV-segment at the beginning of the word, three VCV-segments (including a VV-segment) in the middle and a V-segment at the end. The pitch pattern matching distance per VCV-segment (Ps_i) is calculated at each segment (for five segments in all) and the average pitch pattern matching distance per word (Pd) is calculated by averaging each Ps_i.

4.2. Experiment conditions

In this experiment, names of city and prefecture, which were synthesized by using the following databases, were evaluated.

(a) The word set with normal accent (conventional database).

(b) The word sets with type-0 and type-1 accents (proposed database).

(c) The combination of (a) and (b).

Each database was uttered by a single female narrator. The range of pitch frequency in her voice was between 150[Hz] and 430[Hz]. The range of pitch frequency in each word set is shown in Table 1.

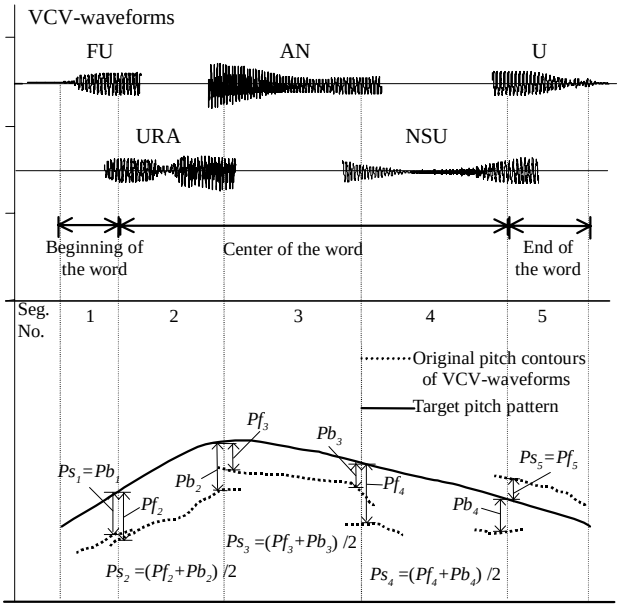

Figure 4: Pitch pattern matching distances in the case of synthesizing " FURANSU" meaning "France" in Japanese.

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

Table 1: Range of pitch frequency in each word set.

4.3. Measurement of pitch pattern matching distance

The pitch pattern matching distance (*Pd*) measured by the above mentioned method is shown in Table 2. The followings are made clear by the measurement.

(1) The pitch pattern matching distance in the case of synthesizing words with type-0 accent is smaller than in the case of synthesizing words with any other accent types.

(2) The total average *Pd* of the proposed database was reduced to approximately two thirds of the conventional one.

(3) The use of the database that consists of words with the three accent types (c) is not so efficient compared with the proposed database (b). It is thought that the reason why there is only a little difference between (b) and (c) is that the proposed database contains enough variety of VCV-segments to reduce the pitch pattern matching distance.

Consequently, it is shown that the proposed database is superior to the conventional one with regard to pitch pattern matching distance.

Synthesized words		*Pd* [octave]		
City and prefecture names	Accent type	a) Normal	b) Type-0 and type-1	c) Normal, type-0 and type-1
Yokohama	type-0	0.28	0.10	0.10
O-saka	type-0	0.19	0.12	0.10
To-kyo-	type-0	0.18	0.08	0.08
Ko-be	type-1	0.52	0.31	0.31
Nagano	type-1	0.32	0.20	0.19
SeNdai	type-1	0.39	0.29	0.27
Wakayama	type-2	0.48	0.29	0.27
Kanagawa	type-2	0.55	0.35	0.34
Okayama	type-2	0.44	0.33	0.30
Yamagata	type-2	0.27	0.37	0.27
Total averages		0.36	0.24	0.22

Table 2: Result of pitch pattern matching measurement for ten synthesized words.

5. PREFERENCE EVALUATION

In order to examine the sound quality of the words synthesized by using the proposed database, a subjective quality evaluation was performed. The words used for the evaluation were the same as those used for the pitch pattern matching measurement. The sound quality of the word synthesized by using the proposed database was compared with that of the word synthesized by using the conventional one. Ten evaluators were asked to select a word that sounded more natural in each pair of the synthesized words.

The result of this preference evaluation is shown in Figure 5. As indicated in the figure, the proposed database has marked more than 75% of preference score against the conventional one. There has been less roughness perceived in the words synthesized by using the proposed database. The subjective quality evaluation shows thus the effectiveness of the proposed database.

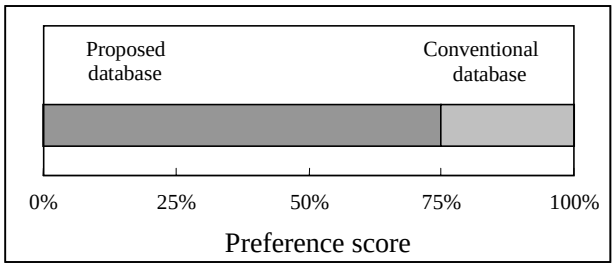

Figure 5: Result of preference evaluation of synthesized words.

6. CONCLUSION

The VCV-concatenation synthesis with the advanced word database which includes the pitch variety of VCV-segments is studied. The database consists of VCV-balanced phonetic words with the type-0 and type-1 accents, and is capable of having a variety of VCV-segments with the same phonetic chains and the different pitch patterns. In order to achieve the sound quality degraded by the pitch modification, new pitch modification techniques are also introduced. The pitch pattern matching measurement and the subjective quality evaluation are performed, and the superiority of the proposed database is ensured.

REFERENCES

[1] Y. Arai, R. Mochizuki, H. Nishimura and T. Honda, “An excitation synchronous pitch waveform extraction method and it’s application to the VCV-concatenation synthesis of Japanese spoken words” ICSLP 96, pp.1437-1440 (1996).

[2] S. Hayamizu, K. Tanaka, S. Yokoyama and K. Ohta, “Generation of VCV/CVC Balanced Word Sets for Speech Data Base” Bul. ETL, Vol.49, No.10 pp.803-834 (1985).

[3] M. Honda and T. Moriya,“Speech Encoding Based on Phase Equalization,” Trans. CSR, Acoustic.Soc. Japan: 33-40(S84-05), (1984).

[4] Y. Arai, R. Mochizuki and T. Honda, “A study on natural-sounding Japanese phonetic word synthesis based on the pitch waveform concatenation” Proc. ICA98, Vol.1 pp.267-268 (1998).