# UNIT SELECTION ALGORITHM FOR JAPANESE SPEECH SYNTHESIS BASED ON BOTH PHONEME UNIT AND DIPHONE UNIT

*Tomoki Toda*[†‡], *Hisashi Kawai*[†], *Minoru Tsuzaki*[†], *and Kiyohiro Shikano*[‡]

[†]ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto, 619-0288 Japan
[‡]Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 Japan

## ABSTRACT

This paper proposes a novel unit selection algorithm for Japanese Text-To-Speech (TTS) systems. Since Japanese syllables consist of CV (C: Consonant, V: Vowel) or V, except when a vowel is devoiced, CV units are basic to concatenative TTS systems for Japanese. However, speech synthesized with CV units sometimes have discontinuities due to V-V concatenation. In order to alleviate such discontinuities, longer units (CV* or non-uniform units) have been proposed. However, the concatenation between V and V is still unavoidable. To address this problem, we propose a novel unit selection algorithm that incorporates not only phoneme units but also diphone units. The concatenation in the proposed algorithm is performed at the vowel center as well as at the phoneme boundary. Results of evaluation experiments clarify that the proposed algorithm outperforms the conventional algorithm.

## 1. INTRODUCTION

In Japanese, a speech corpus can be constructed efficiently by considering CV (C: Consonant, V: Vowel) syllables as synthesis units, since Japanese syllables consist of CV or V except when a vowel is devoiced. It is also well known that transitions from C to V, or from V to V are very important in auditory perception. Therefore, CV units are basic to concatenative TTS systems for Japanese. However, speech synthesized with these units sometimes have discontinuities due to V-V concatenation.

In order to alleviate these discontinuities, longer units have been proposed [1][2][3]. Kawai et al. extended the CV unit to the CV* unit to inhibit the concatenation at V-V boundaries [1]. Iwahashi et al. proposed non-uniform units[3]. In this algorithm, optimum units are selected from a speech corpus to minimize the total cost calculated as the sum of all target costs and concatenation costs. As the result of a dynamic programming search based on phoneme units, various sized sequences of phonemes are selected [4]. However, it is not realistic to construct a corpus that includes all possible vowel sequences, since countless vowel sequences exist in Japanese. If the coverage of prosody is also to be considered, the corpus becomes enormous beyond imagination. Therefore, the concatenation between V and V is unavoidable.

Formant transitions are more stationary at the vowel centers than at the vowel boundaries. Therefore, the discontinuities caused by concatenating vowels can be reduced if the vowels are concatenated at their centers. This view has been supported by our informal listening test. In this paper, we propose a novel unit selection algorithm incorporating not only phoneme units but also diphone units. The proposed algorithm permits the concatenation of synthesis units not only at the phoneme boundaries but also at the vowel centers. As a result of evaluation experiments, we state that the proposed algorithm outperforms the conventional algorithm.

The paper is organized as follows. In section 2, cost functions for unit selection are described. In section 3, the advantage of performing concatenation at the vowel centers is discussed. In section 4, the novel unit selection algorithm is described. In section 5, evaluation experiments are described. Finally, we summarize this paper in section 6.

## 2. COST FUNCTION FOR UNIT SELECTION

There are four kinds of sub-costs. These sub-costs are each classified in terms of the kind of measure (perceptual or acoustic) and whether it is the concatenation cost or the target cost. An overview of the sub-costs is showed in Fig. 1.

In unit selection, it is important to utilize a measure that corresponds to the perceptual characteristics [5]. Acoustic measures that effectively capture perceptual characteristics have been explored, but none has been found so far [6][7].

Therefore, we use not only acoustic measures but also perceptual measures as sub-costs in order to compensate for

0-7803-7402-9/02/$17.00 ©2002 IEEE

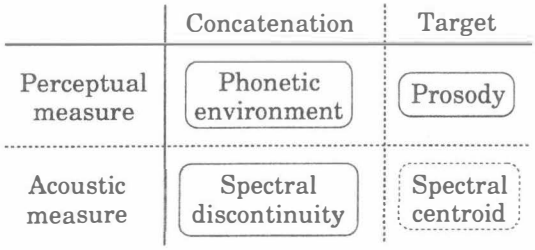


**Fig. 1**. Overview of sub-costs.

the shortcomings of the acoustic measures.

The total cost is calculated as the weighted sum of four sub-costs. The total cost $TC(u_i)$ at phoneme $u_i$ is given by,

$$TC(u_i) = w_{env} \cdot C_{env}(u_i, u_{i-1}) + w_{spec} \cdot C_{spec}(u_i, u_{i-1}) + w_{pro} \cdot C_{pro}(u_i) + w_{cent} \cdot C_{cent}(u_i), \quad (1)$$

where $C_{env}, C_{spec}, C_{pro}$, and $C_{cent}$ denote the sub-costs. $w_{env}, w_{spec}, w_{pro}$, and $w_{cent}$ denote the weights for individual sub-costs. In this paper, the sub-cost for the spectral centroid $C_{cent}$ is not used ($w_{cent} = 0$).

### 2.1. Cost on Substitution of Phonetic Environments

This sub-cost captures the naturalness degradation caused by a mismatch of phonetic environments between a candidate unit and the target. The sub-cost $C_{env}$ is given by,

$$C_{env}(u_i, u_{i-1}) = S_s(E_e(u_{i-1}), E_t(u_{i-1})) + S_p(E_e(u_i), E_t(u_i)) + B(u_i, u_{i-1}), \quad (2)$$

where $S_s$ denotes a mismatch of the succeeding environment and $S_p$ denotes a mismatch of the preceding environment. $E_e$ denotes the phonetic environment of the extracted phoneme (candidate), while $E_t$ denotes that of the target phoneme. $B$ denotes a bias to control the ease of concatenation between C and V, V and V, and at the vowel center. As a result of an informal listening test, the bias was tuned so that the concatenation at a C-V boundary, a V-V boundary, and at the vowel center would become harder in this order. When the units $u_{i-1}$ and $u_i$ are connected in the corpus, the sub-cost is set to 0.

The sub-cost was determined from results of perceptual experiments, in which listeners evaluated the naturalness degradation by listening to the speech stimuli synthesized by concatenating phonemes extracted from various phonetic environments.

### 2.2. Cost on Spectral Discontinuity

This sub-cost captures a spectral discontinuity at a unit boundary that may cause unnaturalness. The sub-cost $C_{spec}$ is given by,

$$C_{spec}(u_i, u_{i-1}) = D_{mc}(u_i, u_{i-1}) + D_{F_0}(u_i, u_{i-1}), \quad (3)$$

where $D_{mc}$ denotes the distance based on the mel cepstrum and its first order derivative. $D_{F_0}$ denotes the difference of log-scaled $F_0$.

### 2.3. Prosodic Cost

This sub-cost captures the naturalness degradation caused by the difference of prosody ($F_0$ contour and duration) between a candidate unit and the target.

In order to calculate the difference of the $F_0$ contour, a phoneme is divided into several parts, and the difference of an average of log-scaled $F_0$ in each part is calculated. In each phoneme, the prosodic cost is represented as an average of the costs calculated in each part. The sub-cost $C_{pro}$ is given by,

$$C_{pro}(u_i) = \frac{1}{N} \sum_{n=1}^{N} P(DifF0_n(u_i), DifDur(u_i)), \quad (4)$$

where $DifF0_n$ denotes the difference of the average of log-scaled $F_0$ in the n-th divided part. $DifDur$ denotes the difference of the duration, which is calculated for each phoneme and used in the calculation of the cost in each part. $N$ denotes the number of divisions. $P$ denotes the cost function.

The function $P$ was determined from results of perceptual experiments on the naturalness degradation caused by prosody modification, because the output speech is synthesized with the prosody modification. When the prosody modification is not performed, the function should be determined based on other experiments on the naturalness degradation caused by using a different prosody from the target.

### 2.4. Cost on Spectral Centroid

This sub-cost captures the distortion caused by utilizing outlying units. The sub-cost $C_{cent}$ is written by,

$$C_{cent}(u_i) = Dist(Cen(u_i), Cen(t_i)), \quad (5)$$

where $Cen$ denotes a spectral centroid, $Dist$ denotes the distance between centroids, and $t$ denotes the target. The target centroid $Cen(t)$ is calculated from all segments of each phoneme in the corpus in advance.

## 3. CONCATENATION AT VOWEL CENTER

Figure 2 compares spectrograms of vowel sequences concatenated at a vowel boundary and at a vowel center. In the former case, discontinuities can be observed at the concatenation points. This is because it is not easy to find a synthesis unit satisfying continuity requirements on both static

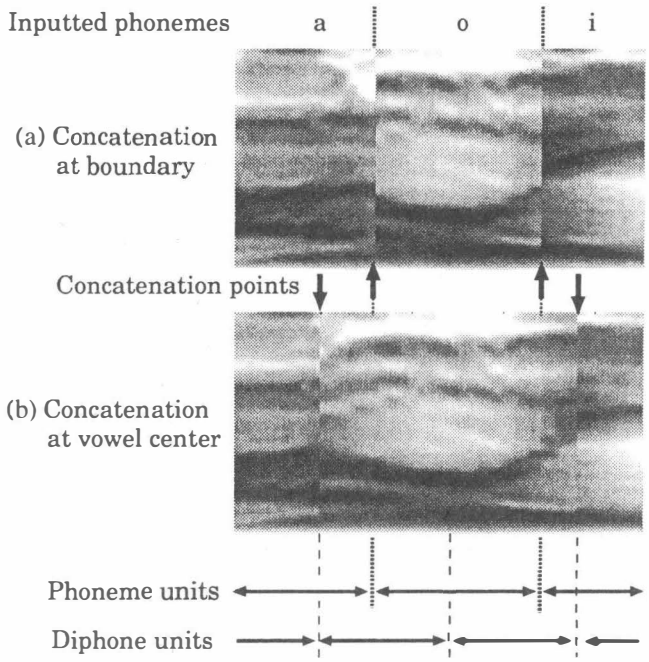


**Fig. 2**. Spectrograms of vowel sequences concatenated at (a) a vowel boundary and (b) a vowel center.

and dynamic characteristics of spectral features at once in a restricted sized speech corpus. In the latter case, in contrast, finding a synthesis unit concerns only static characteristics, because spectral characteristics are nearly stable. As a result, the formant trajectories are continuous at the concatenation points, and their transition characteristics are well preserved.

This example suggests that allowing concatenation not only at the vowel boundaries but also at the vowel centers can make the unit selection procedure more flexible, and consequently, reduce discontinuities at concatenation points. This flexibility can also lead to reductions in the corpus size.

## 4. UNIT SELECTION ALGORITHM WITH PHONEMES AND DIPHONES

Motivated by the considerations in the previous section, we developed a novel algorithm incorporating diphone units as well as phoneme units.

When concatenation at the vowel center is performed, the total cost on the vowel is calculated by replacing $u_{i-1}$ and $u_i$ with the first half-vowel and the last half-vowel, respectively in Eq. (1). The target sub-costs $C_{pro}$ and $C_{cent}$ are calculated as the averages of the sub-costs for the first half and the last half of a phoneme. The optimum sequence of units, which may be phonemes or diphones, is selected

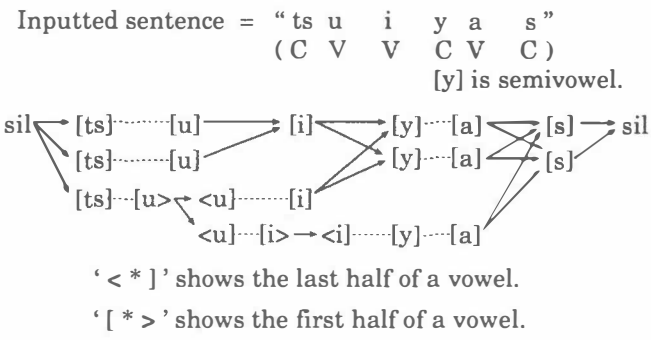


**Fig. 3**. An example of the proposed unit selection algorithm. The inputted sentence is "tsuiyas" ("spend" in English). By a proper setting of the bias component in Eq. (2), concatenation at the C-V boundaries and selection of isolated half vowels are inhibited.

from the speech corpus by minimizing the sum of the total cost with the dynamic programming method.

Diphone units that start from the middle of a vowel in front of consonants are used in not only transitions from V to V but also in transitions from V to a semivowel or a nasal. An example of the proposed unit selection algorithm is shown in Fig. 3. Diphone units such as /ts-u/, /u-i/, and /i-y/ as well as phoneme units are both considered in the unit selection.

## 5. EXPERIMENTAL EVALUATION

In order to evaluate the performance of the proposed algorithm, we compared the proposed algorithm with the conventional algorithm, which allows concatenation only at phoneme boundaries. Subjective and objective experiments were performed.

### 5.1. Subjective Experiment

We used a speech corpus comprising Japanese utterances of a male speaker, where segmentation was performed by experts and $F_0$ was revised by hand. The utterances are about 30 minutes (450 sentences). The sampling frequency is 16000 Hz.

A preference test was performed with synthesized speech of 10 Japanese sentences. The sentences were not a part of the speech corpus used in the unit selection. The speech was synthesized by the proposed unit selection algorithm or the conventional algorithm. In order to avoid increasing the number of concatenation points, the proposed algorithm was tuned so that half vowel units were not selected. In all of the synthesized speech, comprising 370 phonemes, 125 concatenations were performed between phonemes and 33

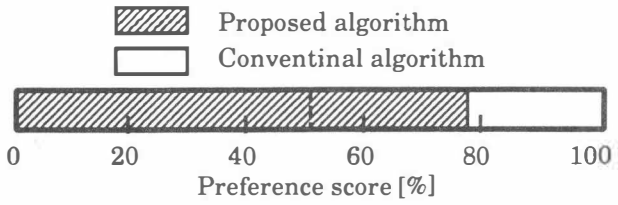


**Fig. 4**. Results of a preference test.

concatenations were performed at vowel centers by the proposed algorithm. On the other hand, 154 concatenations were performed between phonemes by the conventional algorithm.

The natural prosody extracted from the original utterances was used to investigate the performance of the unit selection algorithms. The speech was synthesized with prosody ($F_0$ contour, duration, and power) modification by using STRAIGHT, which is a high-quality vocoder [8]. Ten listeners participated in the experiment. At each trial, a pair of utterances synthesized with the proposed algorithm and the conventional algorithm was presented in a random order, and the listeners were requested to choose the speech they felt was more unnatural.

Experimental results are shown in Fig. 4. The figure shows that the proposed algorithm can synthesize more natural speech than the conventional algorithm.

### 5.2. Objective Experiment

We compared these algorithms in terms of the cost for unit selection. A speech corpus of about 8 hours (10000 sentences) was used for the unit selection. The evaluation sentences were 53 sentences. The sentences were not a part of the corpus.

Experimental results are shown in Fig. 5. The figure shows that the proposed algorithm can reduce the average cost more than the conventional algorithm, because of the greater number of candidate units comprising not only phoneme units but also diphone units. Therefore, the proposed algorithm yields an equal performance to the conventional algorithm with a smaller corpus. The results also clarify that the effect is more remarkable as the corpus size increases.

## 6. CONCLUSION

In this paper, we proposed a novel unit selection algorithm with both phoneme units and diphone units in order to avoid the quality degradation caused by concatenation at perceptually important transitions between phonemes. We performed perceptual and objective evaluation experiments. The results showed that speech synthesized with the proposed

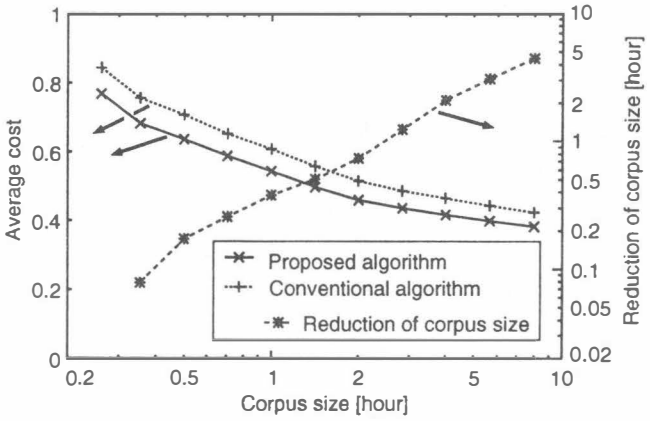


**Fig. 5**. Average costs as a function of the corpus size.

algorithm has better naturalness than that of the conventional algorithm. They also clarified that the proposed algorithm yields an equal performance to the conventional algorithm with a smaller corpus.

## 7. REFERENCES

[1] H. Kawai, N. Higuchi, T. Shimizu and S. Yamamoto, "Development of a text-to-speech system for Japanese based on waveform splicing," Proc. ICASSP, pp. 569–572, Adelaide, Australia, Apr. 1994.

[2] S. Takano, K. Tanaka, H. Mizuno, M. Abe and S. Nakajima, "A Japanese TTS system based on multiform units and a speech modification algorithm with harmonics reconstruction," IEEE Trans. Speech and Audio Processing, vol. 9, no. 1, pp. 3–10, 2001.

[3] N. Iwahashi, N. Kaiki and Y. Sagisaka, "Speech segment selection for concatenative synthesis based on spectral distortion minimization," IEICE Trans. Fundamentals, vol. E76-A, no. 11, pp. 1942–1948, 1993.

[4] A. Black and N. Campbell, "Optimising selection of units from speech database for concatenative synthesis," Proc. EUROSPEECH, pp. 581–584, Madrid, Spain, Sept. 1995.

[5] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," IEEE Trans. Speech and Audio Processing, vol. 9, no. 1, pp. 39–51, 2001.

[6] Y. Stylianou and A.K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," Proc. ICASSP, pp. 837–840, Salt Lake City, U.S.A., May. 2001.

[7] M. Tsuzaki, "Feature extraction by auditory modeling for unit selection in concatenative speech synthesis," Proc. EUROSPEECH, pp. 2223–2226, Aalborg, Denmark, Sep. 2001.

[8] H. Kawahara, I. Masuda-Katsuse and A.de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based $F_0$ extraction: possible role of a repetitive structure in sounds," Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.