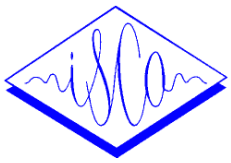

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

On the Unit Search Criteria and Algorithms for Speech Synthesis Using Non-uniform Units

Kazya TAKEDA
KDD R&D Laboratories
2-1-15 Ohara, Kamifukuoka
Saitama 356 Japan

Katsuo ABE
TOYO Communication Equipment
1-1 Koyato, Samukawa-mach
Kanagawa 253-01 Japan

Yoshinori SAGISAKA
ATR
Seika-cho Souraku-gun
Kyoto 619-02 Japan

Abstract

A selective use of non-uniform synthesis units for speech synthesis-by-rule is discussed focusing on an optimal unit selection method. In this paper, we propose two algorithms for unit selection. The first one uses one total measure reflecting contextual similarities and adequacy of unit concatenation. The second one combines top down control for concatenation points and bottom up search for the appropriate speech template. The high quality of both selection methods, compared to the conventional method using fixed units, is confirmed by both subjective and objective tests. Furthermore, the results of intelligibility tests are analyzed aiming at designing a quantitative measure to evaluate unit suitability.

1 Introduction

Although designing the unit set is one of the most important factors in concatenation type synthesis systems, no systematic method has been found to control unit attributes such as unit length (what the basic unit should be), unit extraction (what context to extract the unit from) and unit usage (modification or selective use of units). As for length of basic synthesis units, for Japanese, the CV (Consonant-Vowel) syllable has been commonly used because of its manageable size (about 100) . Longer phonemic sequences such as VCV or CVC [1,2] were reported to work better, but, the size of the unit set increases to several hundred and no compromise among CV and other longer units has ever been found. The second problem, what context to extract the unit from, has also been recognized as an important aspect of unit design. Preliminary experiments [3] suggested that the speech quality could be improved by taking extracting context into account, however, no general method has been proposed and unit extraction is still treated as *knowhow* of speech unit designing. As for the unit usage problem, selective use of speech unit templates is a recently introduced idea for Japanese speech synthesis. Most of these attempts however use phonemes as basic units, and not enough studies have been reported on the selection criteria [4,5].

The new synthesis-by-rule scheme that we are proposing [6] has an advantage over existing systems since it takes the above mentioned attributes of the units into account in a synthesis system. The basic idea of the system is the selective use of arbitrary phonemic sequences for synthesis units based on the context of the input text. Various kinds of acoustic realizations for speech templates can thus be used if we have an appropriate unit selection algorithm. In this paper, after describing the outline of the synthesis scheme, we propose and evaluate two algorithms for unit selection. Furthermore, the relationship between speech quality and some aspects of the unit usage is discussed based on the results of the intelligibility tests.

2 Selective Use of Non-uniform Synthesis Units

Basically, to utilize non-uniform speech chunks as units selectively, we need to make two decisions. The first is to find an optimal decomposition of the input phonemic sequence and the second is to find the most appropriate speech template for each sequence. The basic scheme of the system can therefore be outlined as shown in Figure 1. First, we obtain a lattice of phonemic sequences by decomposing the input sequence. Then, we check if each lattice element is in the speech database (the shadowed blocks in the figure indicate the available sequences). Finally, we select an optimal unit sequence and extract the corresponding speech chunks from the database. Although this is a simple extension of conventional fixed-length units, such as phonemes, dyads, diphones or Japanese CV syllables, we can use a greater variety of acoustic realizations from the speech database. Furthermore, we can increase the variety of acoustic realizations without explosion of the number of units by applying the phonotactic constraints of the languages to the speech database. Currently we are using an isolated word speech database consisting of 5,000 commonly used Japanese words.

3 Criteria and algorithms for unit selection

3.1 Unit selection criteria

The most significant factors for high quality speech synthesis in concatenation type synthesis-by-rule systems are continuity between units and the appropriate acoustic realization adapted to the context. We need to evaluate the unit suitability from these two standpoints in order to select the optimal unit sequence. In the following sections, we will propose two selection algorithms based on different strategies to search for the units which satisfy these two criteria. In order to implement the algorithms, costing functions are needed to evaluate these criteria of the unit attributes. Although, for each criterion, constructing such a costing function requires many more acoustical and perceptual studies, phonetic environments of the units can be reasonable variables for these functions. For example, certain phonetic environments such as a low power portion in consonants or a steady portion of vowels are regarded as a good concatenation bound-

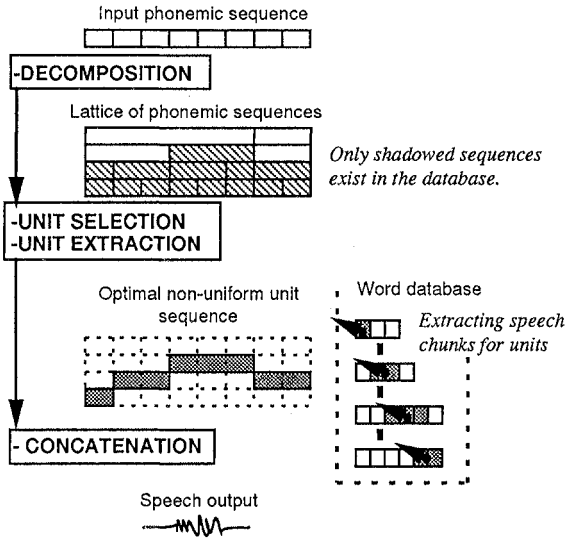

Figure 1: Basic scheme of synthesis-by-rule based on non-uniform units

ary. The degree of coarticulatory effects can also be estimated from the phonetic environment of the units. Thus, as a first approximation, we are currently controlling the criteria using a relative scoring based on the phonetic environment of the units, e.g. the closure portion of the stop consonants is a better concatenation boundary than the transient portion of vowels. Some analyses to design quantitative measures will be discussed later in this paper.

On the other hand, degradation due to the modulating fundamental frequency of the unit is also an important problem in concatenation-type synthesis-by-rule systems. It is desired to use a speech unit recorded in the fundamental frequency similar to the target frequency to be synthesized. As the third criterion, we introduce the difference of the original and the target fundamental frequency as a quantitative function.

In the next two sections two implementations of the selection algorithm are described based on these criteria.

3.2 SCF method–Selection based on a Single Costing Function

The first selection algorithm uses a single costing function combining two of the above criteria by simply adding given penalty costs for the concatenation environment and the context similarity. The penalty cost is calculated as follows for the example shown in Figure 2, where a phonemic sequence /kuyo/ is extracted from a word /mo*kuyo*/ (Thursday) and used in the context /toro*kuyo*shi/ (registration form).

$$\begin{aligned} CostValue &= \sum_{i=-3,-2,-1,1,2,3} D(e_i, u_i) W_i \\ &= D(\#, o) \times 10 + D(m, r) \times 100 + \cdots + D(\#, \#) \times 10 \\ &= 10190 \end{aligned}$$

where, $D(A, B)$ is the value defined for every combination of two phonemes taking the above mentioned two criteria, inter-unit continuity and coarticulatory effects, into account. Since

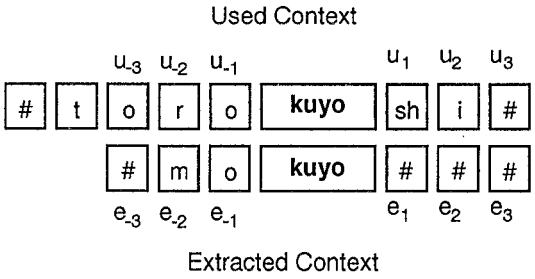

Figure 2: Calculation of cost value

voiceless fricatives are considered to be a better concatenation boundary than voiced fricatives, for example, $D(/p/, /b/) = 5$ is designed to be greater than $D(/p/, /s/) = 3$. The weighting factor $W_i = 10^{4-|i|}$ is introduced to emphasize the effects of the neighboring phonemes.

An optimal unit sequence obtained by this implementation minimizes the sum of these penalty cost values of used units. This is a simple compromise between a longer unit preference and the contextual similarity of units. As shown in Figure 3 (a), this algorithm calculates the sum of the costs among all possible decompositions of the input string. In this algorithm, accumulated costs can be effectively calculated by dynamic programming, and we did not use the difference of the fundamental frequency as a selection criterion.

3.3 TDH method–Selection based on Top Down Hypothesis

The second algorithm treats each factor independently. In this method, each criterion is used for the restriction of unit candidates. As shown in Figure 3 (b), at the first state, focusing only on the continuity between units, the concatenation points are hypothesized and some phonemic sequences are obtained as unit candidates. If no speech chunk corresponding to the sequence is found in the database, the sequence is re-divided into smaller sequences until a sequence is found. Then, we can obtain optimal unit phonemic sequences from the standpoint of unit continuity. Next, we focus on the similarity of the unit phonetic environment between the extracted and used contexts to reduce the number of word candidates from which the speech unit chunk should be extracted. At this stage, similar to the SCF method, the units' suitability is evaluated based on penalty scores calculated from phonetic similarities. Finally, the difference of the target and original fundamental frequency is calculated to decide the optimal unit speech chunk.

4 Selection result

To clarify the difference between two unit sequences selected by the above two selection algorithms, an experimental unit selection was performed using nine sentences of simulated telephone conversation as input texts. Figure 4 illustrates the distribution of the unit length selected by the two algorithms. The figure shows that not only CV syllables but also various phonemic sequences are utilized for synthesis units, and that the scheme is an extension of the conventional method of fixed length units. The mean length of the selected units for synthesizing the texts was

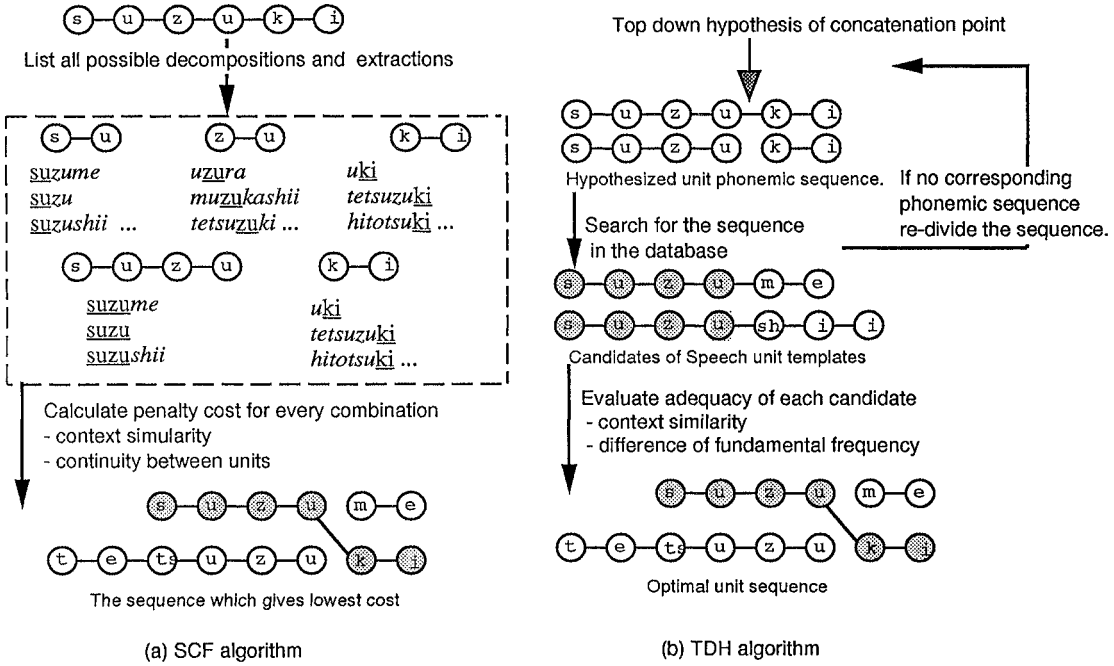

Figure 3: Two unit selection methods for the optimal unit sequence

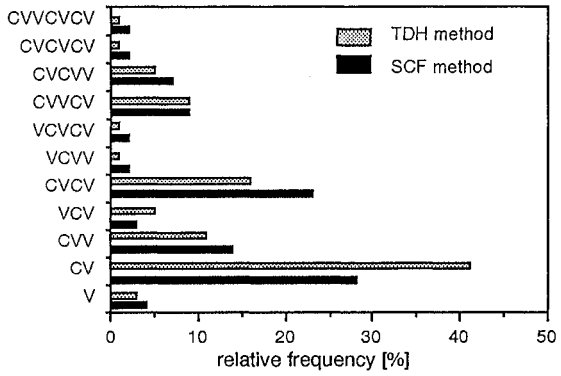

Figure 4: Distributions of unit length selected by the two algorithms

3.44 phonemes by the SCF method and 3.12 phonemes by the TDH method, respectively. On the other hand, the mean value of the context penalty costs and discontinuity between units, i.e. the Euclid distance of the thirty Cepstrum coefficients at the concatenation boundary, are lower in the TDH method than in the SCF method, as listed in Table 1.

Table 1: Difference of selected units by the two selection algorithms

	SCF method	TDH method
Context Penalty Cost	47.8	35.7
Inter-unit Discontinuity (Cepstral Difference)	2.05	1.80

5 Evaluation of algorithms

5.1 Intelligibility test

For the intelligibility test, we used four types of speech of 45 phrases consisting of technical terms. These four types are, i) natural, ii) synthesized from non-uniform units selected by the SCF method, iii) selected by the TDH method, and iv) synthesized from conventional fixed length units (CV syllables) extracted from a fixed context (a nonsense word "ateCVberi"). Twelve naive subjects who were unfamiliar with the used words listened to each speech sample three times and wrote down what they heard. Syllable intelligibility of the four types of speech was calculated from the results obtained (Figure 5 (a)). The figure shows the superiority of non-uniform unit synthesis over the conventional fixed length unit system. Furthermore, it can be seen that the units selected by the TDH method have better intelligibility than the SCF method.

5.2 Subjective preference test

Subjective evaluation tests were performed after the intelligibility tests using the same sample text and eleven of the twelve original subjects. Each subject was asked to indicate which type of speech they preferred. Figure 5 (b) shows the preference tendency obtained between the three types of synthetic speech. For both algorithms, the speech using non-uniform units was preferred in more than 80 % of the comparisons over the conventional fixed unit synthesis.

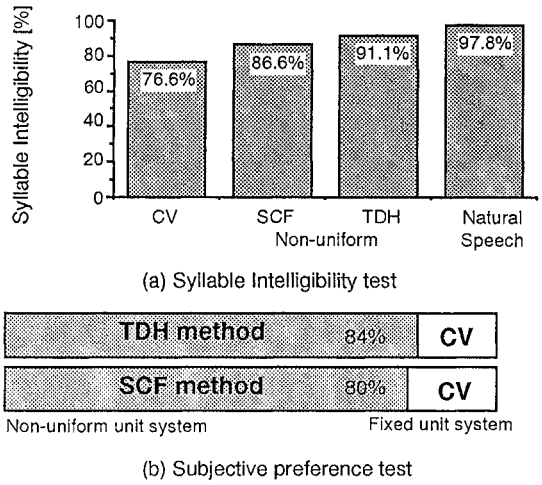

Figure 5: Evaluation tests on the two algorithms

6 Unit attributes and intelligibility

As mentioned in Section 3, it is desired to evaluate the suitability of unit usages by quantitative measures which take acoustic and perceptual knowledge into account. To clarify the relationship between unit usage and the resulting speech quality, the results of the intelligibility tests were analyzed in relation to unit attributes.

The effect of the unit length on intelligibility was found to be as shown in Figure 6 (a). This figure shows the proportional relation between unit length and better intelligibility. Especially, the most significant intelligibility degradation can be seen in mono-syllabic units. Based on this tendency, we can propose costing functions for evaluating unit length for the purpose of unit selection.

The analysis within the phonemic environment of unit boundaries and the resultant degradations, illustrated in Figure 6 (b), also gives us appropriate costing scores to evaluate inter-unit continuity.

7 Conclusion

In this paper, we described unit selection algorithms as a basic problem of non-uniform unit speech synthesis. We proposed two algorithms, the SCF method, which uses total measure reflecting contextual similarities and adequacy of unit concatenation, and the TDH method, which combines top down control for concatenation points and bottom up search for the appropriate speech template. Subjective and objective tests showed the superiority of the non-uniform synthesis units, and that the unit sequence selected by the TDH method is slightly better than the SCF method, using our tentative unit suitability evaluation functions. Finally, we discussed the relationship between unit attributes and resultant speech intelligibility based on the result of the intelligibility test. The results of the analyses are 1) The most important degradation related to unit length was found in mono-syllabic units, and 2) The phonemic contexts suited to unit concatenation are voiceless fricatives, steady portions of vowels

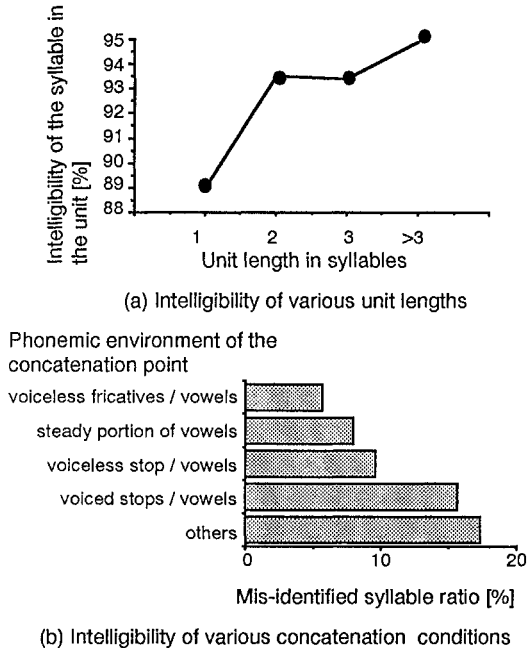

Figure 6: Unit attributes and resulting speech quality

and silent portions of stop consonants in that order.

For further development, improvement of algorithms and parameters used in the unit evaluation are needed on the basis of the results of the above analyses. Moreover, finding an optimal speech database for unit extraction including sentential speech, and decreasing its size, are also important issues to implement the system.

Acknowledgment

This work was done while the authors were in ATR Interpreting Telephony Research Laboratories. The authors are grateful to Dr. Kurematsu for his continuous support.

References

[1] **Saito, S., and Hashimoto, S.** ,"Speech synthesis system based on interphoneme transition unit" Proc 6th ICA, B-5-12, 1968

[2] **Sato, H.**, "Speech synthesis using CVC concatenation units and excitation waveform elements" Trans. Comm. Speech Res., ASJ, s83-69, 1984 (In Japanese)

[3] **Kishimoto, N. and To'kura, Y.**, "Speech Quality Improvement using phoneme environment dependent CV-files" Proc ASJ Fall Meeting, 1-6-20, 1980 (in Japanese)

[4] **Nakajima, S. and Hamada, H.**,"Automatic generation of synthesis units based on context oriented clustering" Proc. ICASSP'88, S14.2, 1988

[5] **Hirokawa, T.**, "Speech synthesis using a waveform dictionary" Proc Euro. Conf. Speech Comm., pp140-143, 1989

[6] **Sagisaka, Y.**,"Speech synthesis by rule using an optimal selection of non-uniform synthesis units" Proc. ICASSP'88, S14.8, 1988