A NEW METHOD OF GENERATING SPEECH SYNTHESIS UNITS BASED ON PHONOLOGICAL KNOWLEDGE AND CLUSTERING TECHNIQUE

Yuki YOSHIDA, Shin'ya NAKAJIMA, Kazuo HAKODA and Tomohisa HIROKAWA†

NTT Human Interface Laboratories
1-2356 Take, Yokosuka-shi, Kanagawa, 238-03 JAPAN
†NTT Multimedia Buisiness Department

ABSTRACT

This paper proposes a new method for generating synthesis units using context dependent phonemes to achieve high quality text-to-speech (TTS) synthesis. If all phoneme triplets (triphones) in Japanese are considered, the number of synthesis units is very large; therefore, we introduce two techniques to reduce the number of synthesis units. The first technique decreases approximately 15,000 triphones to about 6,000 triphones based on phonological knowledge. The second technique is based on a segment quantization, which reduces the number of units even more. Experimental tests show that the proposed method is effective in improving articulation and intelligibility scores, that the number of synthesis units can be decreased without significant loss in TTS quality, and that the preference score is proportional to the number of synthesis units.

1. INTRODUCTION

In text-to-speech (TTS) synthesis, the synthesis unit is one of the primary factors that greatly affects the quality of synthesized speech. Some kinds of synthesis units have been proposed such as CV, VCV, CVC, the Context-Oriented-Clustering (COC) method [1], and non-uniform phoneme sequence units [2]. In general, for these units, the longer they are, the more accurately they represent coarticulation effects. Thus, more fluent speech can be synthesized by using longer units which result in greater stability and naturality. However, even though this stability and naturality are guaranteed within a unit, these are not necessarily garanteed when these units are combined. In other words, there are some cases that because the discontinuity of the connection points between these units is harsh on the ear, it gives the impression that the synthesized speech is unstable. Moreover, although there are not many connection points because the unit is long, an enormous number of units is needed and it is not realistic to employ these units when the TTS system is applied to a limited capacity system such as a personal computer.

We propose a method for generating synthesis units using context dependent phonemes to achieve high quality TTS synthesis, and PSOLA-based [3] waveform speech synthesis [4] are used as a synthesis method in this paper.

To obtain natural and smooth synthesized speech, we need synthesis units that take into consideration the coarticulation between phonemes or between units. In addition, to obtain stable synthesized speech, it is desirable that all synthesis units are generated under the same conditions. We propose using context dependent phonemes (triphones) as synthesis units to improve the quality of synthesized speech. In Japanese, there are about 15,000 triphones and this number is so large that personal computers have great difficulty in achieving high quality TTS. Accordingly, the number of triphones must be reduced to a suitable number for a practical use, and using the proposed two techniques accomplishes this. The following describes the details of the two techniques.

Section 2. describes two ways of decreasing the number of units. Section 3. explains_how speech is synthesized using the proposed units. Section 4. shows the results of some listening tests for our synthesized speech.

2. GENERATING SYNTHESIS UNITS

If all triphones in Japanese are considered, the number of synthesis units is very large; therefore, we introduce two techniques to reduce the number of synthesis units. The first technique decreases approximately 15,000 triphones to about 6,000 triphones based on phonological knowledge. The second technique is based on a segment quantization, which reduces the number of units even more. The details are described in the pages that follow.

2.1. Synthesis Unit Reduction using Phonological Knowledge

We reduced the number of triphones by grouping the preceding and following phonemes of each triphone. The rules for reduction are as follows:

(1) Regard a long vowel as a normal vowel, if its position is first or third.

(2) Regard triphones whose last two phonemes are the same and whose middle phoneme is an unvoiced plosive as the same triphone.

(3) Regard triphones whose first two phonemes are the same and whose last phoneme is an unvoiced plosive as the same triphone.

Rule (2) is based on the phonological knowledge that, in Japanese, unvoiced plosives are affected mainly by the following vowels and that the preceding vowels' effects are negligible. Rule (3) comes from the consideration that the coarticulation effect on the vowels followed by unvoiced plosives is mainly explained by the factor "followed by unvoiced plosives" per se and it does not greatly depend on these unvoiced plosives.

Figure 1 shows some examples. For instance, using Rule (1), ${}^{A}M^{I}$ is a representation for ${}^{a}M^{I}$, ${}^{A}M^{i}$, ${}^{a}M^{i}$, and ${}^{A}M^{I}$. Here, ${}^{A}M^{I}$ denotes phoneme /M/ preceded by /A/ and is followed by /I/. Similarly, ${}^{U}K^{O}$ is a representation for ${}^{A}K^{O}$, ${}^{O}K^{O}$, ${}^{E}K^{O}$, and ${}^{I}K^{O}$ using Rule (2), and ${}^{P}E^{P}$ is a representation for ${}^{P}E^{T}$, ${}^{P}E^{K}$, ${}^{P}E^{P}$ using Rule (3).

In this way, about 6,000 triphones are generated consisting of 4,800 vowel units and 1,200 consonant units. These triphones are considered to be the basic synthesis units.

The following explains how the database of the basic synthesis units is created.

1. Make a meaningless word of each synthesis unit which contains a synthesis unit and vocal list.
2. Have a male and a female narrator record the vocal list. At this time, the meaningless words are recorded in a monotonous tone of voice.
3. Segment the recorded voice part which corresponds to a synthesis unit, and save it with the waveform and pitch marks.
4. Collect the segments to create the database.

2.2. Synthesis Unit Reduction using Segment Quantization

In [1], we proposed an automatic synthesis unit generation technique termed the COC method which has been applied to a Japanese TTS synthesis system, and it has been confirmed that COC synthesized speech is highly intelligible [5].

The COC method is developed to find an optimum set of phonetic contexts from a large speech database which includes various uncontrolled phonetic contexts. On the contrary, the basic synthesis units here covers most Japanese triphone contexts. Thus, to reduce the number of synthesis units, the segment quantization technique developed by [6, 7] can be easily applied.

Before starting synthesis unit generation, the upper limit of the total number of units (N_{max}), and the minimum number of sample segments per cluster (n_{min}) are specified. The synthesis units for each vowel is then obtained using the following procedure:

Step 1 : Let each initial cluster, W_i consist of the synthesis units with the same preceding phonemes. Compute the centroid segment of each cluster using the segment quantization technique, and select a *prototype segment*, γ_i, which is defined by the nearest segment to the centroid segment. Then, compute each *intra-cluster variance*, ν_i.

Step 2 : Of all the clusters, find cluster W_k such that it has a maximum ν_i and has more sample segments than n_{min}. If no such cluster exists, go to Step 5.

Step 3 : Using the LBG [8] algorithm, split cluster W_k into two clusters: W_{k1} and W_{k2}. Compute the intra-cluster variances of W_{k1} and W_{k2}.

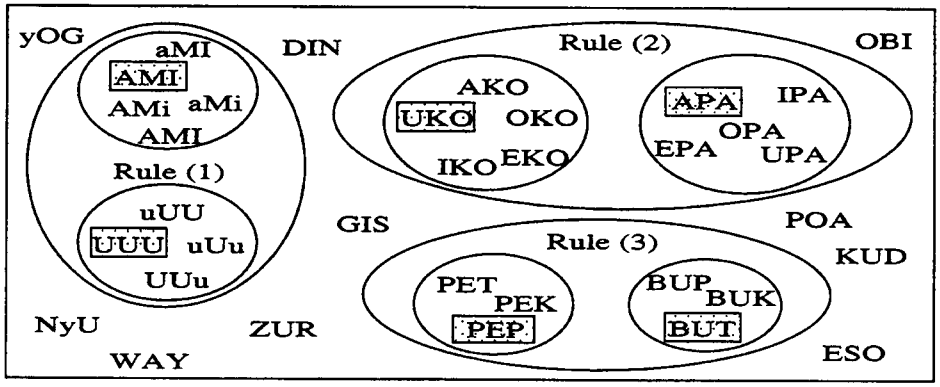
Step 4 : If the number of clusters is less than N_{max}, go to Step 2.

Step 5 : Each prototype segment, γ_i, is saved as a synthesis unit together with the phonetic contexts C_{ij} (j = 1, 2, ..., number of segments in W_i) of the other segments within the same cluster.

Since the basic Japanese syllable is a consonant-vowel concatenation, the initial cluster (in Step 1) is set to be a set of units having the same preceding phonemes.

The prototype segment of each final cluster is the representative synthesis unit of the cluster. Thus, in the synthesis phase, the prototype segment, γ_i, is used for any phonetic context in C_{ij}. For instance, if a prototype segment, γ_i, having phonetic context ${}^{P}a^{d}$, and cluster W_i includes two more segments having phonetic contexts ${}^{P}a^{z}$ and ${}^{P}a^{s}$, respectively, then, γ_i is used for the phonetic context "preceded by /p/ and followed by /d/, /z/, or /s/."

The clustering technique was applied to the vowel units, and 2/3, 1/2, 1/3, and 1/4 the number of vowel units were obtained. (4,593, 3,693, 2,793, and 2,393 units respectively, including consonant units.)



A small letter represents a long vowel.

Figure 1. Grouping using phonological knowledge

3. SYNTHESIZED SPEECH

Waveform speech synthesis was used as the synthesis method and synthesis using the basic synthesis units was conducted in the following manner:

1. Extracted triphones from the input phonological series.
2. Determined the corresponding unit to each triphone. Here only one unit is selected for a triphone. In this case, however, the most suitable synthesis unit is selected from some candidates when using the COC method or the non-uniform phoneme sequence units.
3. Waveforms from different units are combined and then altered to fit the selected the pitch pattern [3]. Power and phoneme duration are set according to the phoneme enviroment.
4. The speech was synthesized.

4. ASSESSMENT OF QUALITY

The synthesized speech quality was evaluated using two kinds of listening tests. One evaluated the clarity of synthesized speech in order to assess the reduction effect from phonological knowledge. The other investigated personal preferences in order to assess the reduction effect from segment quantization. The details are described in the pages that follow.

4.1. Clarity Tests

We dictated 100 Japanese syllables to listeners five times. The experimental conditions and the articulation score are shown in Table 1.

In natural speech, it is said that an articulation score of 90% or more will be very satisfactorily understood but the same thing cannot be necessarily said in synthesized speech because there is distortion of the connection parts. Therefore, an inteligibility test of the speech synthesized using the proposal method was conducted.

In each two experiments, we dictated 100 Japanese family names to listeners: one experiment contained familiar Japanese names, the other had a mixture of familiar and unfamiliar names. The experimental conditions and the intelligibility score are shown in Table 2.

In the experiment using unfamiliar names, it was clear that the subjects misheard one syllable, for example they heard *Takada* instead of *Akada* or *Oosugi* instead of *Oosumi*. We believe the reason for this is that due to the subjects difficulty with the synthesized speech that the names were matched to familiar names. Incorrect recognition by only one syllable does not cause significant problems in actual applications.

Judging from the two results, the proposed method is effective in improving articulation and intelligibility scores and it is shown that this is an improvement on the current synthesized speech.

Table 1. Test Conditions and Articulation Score

Number of listers	3
Number of tests	5
Synthesis units	the basic synthesis units
Stimulus	100 Japanese syllable
Score	90.2% (male) 91.7% (female)

4.2. Preference Tests

We examined the relationship between the number of synthesis units and the quality of the synthesized speech. Pair comparison listening tests were carried out. One reference was the speech synthesized using the basic synthesis units and the other was the speech synthesized using the four reduced sets of speech units described in Section 2.2.. The number of basic synthesis units was 6,033, and the others were 4,593, 3,693, 2,793, and 2,393. Ten listeners selected one type of speech they preferred and 10 sentences were indicated. Figure 2 shows the relationship between the number of units and the preference score and that between the number of units and the cover ratio, i.e., the ratio of basic units not replaced due to reduction in the number of synthesis units to the total number of units.

As a result, that there is no significant difference in the quality of the synthesized speech, because the speech synthesized by half of the basic synthesis units is chosen at the rate of about 40% when compared to the speech synthesized by the basic synthesis units. Moreover, as the number of synthesis units decreases, the preference score decreases, namely it is clear that the preference score is proportional to the number of synthesis units. However, the preference score for the 2,393 synthesis units is higher than the one for the 2,793 synthesis units. This might be an error because there is only a slight difference and the two cover ratios are close to each other. Thus, when the TTS system is constructed, this result can be assumed to be one standard for setting synthesis units.

5. CONCLUSIONS

This paper proposed a method for generating synthesis units using context dependent phonemes to achieve high quality TTS synthesis. Two techniques using phonological knowledge and segment quantization is shown to reduce the enormous number of synthesis units. Listening tests showed three aspects: the proposed method is effective in improving articulation and intelligibility scores; the number of synthesis units can be decreased with no significant loss in TTS quality; and the preference score is proportional to the number of synthesis units. There are two samples of the synthesized speech using the proposed method. One is male synthesized speech in [MSPEEH A295S1.WAV], and the other is female synthesized speech in [FSPEECH A295S2.WAV].

Acknowledgment

We are grateful to the members of the Speech Processing Department for their helpful advice. We also thank

Table 2. Test Conditions and Intelligibility Score

Number of listers	10
Number of tests	1
Synthesis units	the basic synthesis units
Stimulus	100 Japanese family names (mixed familiar and unfamiliar) 100 Japanese family names (familiar)
Score	89.9% (mixed : male) 92.9% (mixed : female) 99.1% (familiar : average)

Dr. Nobuhiko Kitawaki, director of the Speech and Acoustic Labs. for his continuous support of this work.

REFERENCES

[1] Shin'ya NAKAJIMA, and Hiroshi HAMADA, *"Automatic generation of synthesis units based on context oriented clustering,"* Proceedings of ICASSP'88, pp. 659–662,(1988-4).

[2] Yoshinori SAGISAKA, *"Speech synthesis by rule using an optimal selection of non-uniform synthesis units,"* Proceedings of ICASSP88 S14.8, pp. 679–682,(1988).

[3] Francis CHARPENTIER and Eric MOULINES, *"Pitch-Synchronous waveform processing technique for Text-to-Speech synthesis using diphones,"* Proceedings of Eurospeech89 volume 2, pp. 13–19, (1989).

[4] Tomohisa HIROKAWA, Kenzo ITO, and Hirokazu SATO, *"High quality speech synthesis based on wavelet compilation of phoneme segments,"* Proceedings of ICSLP92, pp. 567–570, (1992-10).

[5] Kazuo HAKODA, Shin'ya NAKAJIMA, Tomohisa HIROKAWA, and Hideyuki MIZUNO, *"A new Japanese text-to-speech synthesizer based on COC synthesis method,"* Proceedings of ICSLP90, pp. 809–812, 1990.

[6] S. ROUCOS, R.Schwartz, and J. Makhoul, *"Segment quantization for very low rate speech coding,"* Proceddings of ICASSP82, pp. 1565–1568, 1982.

[7] Yoshinao SHIRAKI, Masaaki HONDA, *"LPC speech coding based on variable-length segment quantization,"* IEEE Transaction on Acoustic Speech Signal Process., Volume 36, No. 9, pp. 1437–1444, 1988.

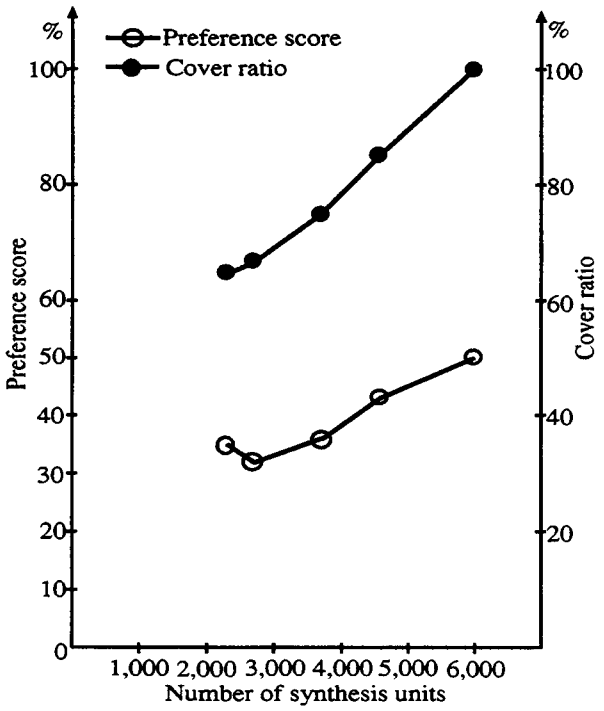


Figure 2. Relationship between the number of units and preference score and that between the number of units and the cover ratio

[8] Yoseph LINDE, Andres BUZO, and Robert M. GRAY, *"An algorithm for vector quantizer design,"* IEEE Transaction on Communications, COM-28,1,pp. 84–95 (Jan. 1980).