SCALABLE CONCATENATIVE SPEECH SYNTHESIS BASED ON THE PLURAL UNIT SELECTION AND FUSION METHOD

Masatsune Tamura, Tatsuya Mizutani[*]*, and Takehiko Kagoshima*

Multimedia Laboratory, Corporate Research and Development Center, Toshiba Corporation,
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki-shi 212-8582, Japan.
[*]Semiconductor Company, Toshiba Corporation, Oume-shi 198-8710, Japan.

ABSTRACT

Recently, concatenative speech synthesizers with large databases have been widely developed for high-quality speech synthesis. However, some platforms require a speech synthesis system that can work under the limitation of memory footprint or computational cost. In this paper, we propose a scalable concatenative speech synthesizer based on the plural speech unit selection and fusion method. To realize scalability, we propose the offline unit fusion method in which pitch-cycle waveforms for voiced segments are fused in advance. The experimental results show that the synthetic speech of the offline unit fusion method with half-size waveform database is comparable to that of the online unit fusion method, while the computation cost is reduced to 1/10.

1. INTRODUCTION

Concatenative speech synthesizers with large databases have been widely developed for high-quality speech synthesis[1]-[5]. Although some applications can use rich hardware resources, several TTS platforms have limitations of computational cost or memory footprint. Embedded platforms, which are used for car navigation systems, video games, robots, etc., are typical examples. In this paper, we propose a scalable concatenative speech synthesizer that can work on various platforms.

We have developed a single diphone based speech synthesis using the closed-loop training method[6][7]. In the method, pitch-cycle waveforms of speech units are trained so as to minimize the distortion of synthesized speech caused by prosodic modifications. Although it can synthesize stable and natural speech, even with a single unit per diphone, the synthetic speech is not quite human-like. One of the reasons is that variations of speech caused by prosody or context cannot be represented by the small set of speech units. In contrast, unit selection based speech synthesis[1]-[5] uses a large database to increase the variations of speech for synthesizing human-like natural speech. The unit selection based speech synthesis selects speech units that minimize cost functions and concatenates the selected speech units with or without prosodic modification. Potential problems for synthesizers of this type are discontinuity between two consecutive units and degradation of speech quality caused by prosodic modification or prosodic mismatch. To overcome these problems, we have been proposed a plural unit selection and fusion method[8] that combines the training based and unit selection based speech synthesis. In this method, first, plural speech units for each segment are selected based on cost functions. Then, the fused waveform that represents the selected plural speech units is generated (we call this "unit fusion"). Finally, the generated waveforms are concatenated to synthesize speech. Using this method, generated speech is both stable and human-like, and we have shown that this method outperforms the unit selection based and training based speech synthesis.

However, the system requires a large database, and high computational cost for the unit fusion process. Therefore, it is unsuitable for some platforms that have limitations, such as embedded platforms. In order to reduce the computational cost, this paper presents an offline unit fusion method that avoids the unit fusion process in the synthesis time. In the proposed method, fused pitch-cycle waveforms are generated in advance. The combinations of units to be fused are selected based on the frequencies that are calculated by synthesizing speech from a large number of test sentences. Additionally, the method can easily manage the number of units for a speech database so that it can synthesize speech using an appropriate database size for each platform.

2. THE PLURAL UNIT SELECTION AND FUSION METHOD

Figure 1 shows a block diagram of the speech synthesis process based on the plural unit selection and fusion method. First, a speech unit database, a phoneme sequence, and prosody (duration, F_0 contour) information are input to the system. Then, in the unit selection process, plural units for each segment are selected according to a cost function. The cost function consists of target cost and

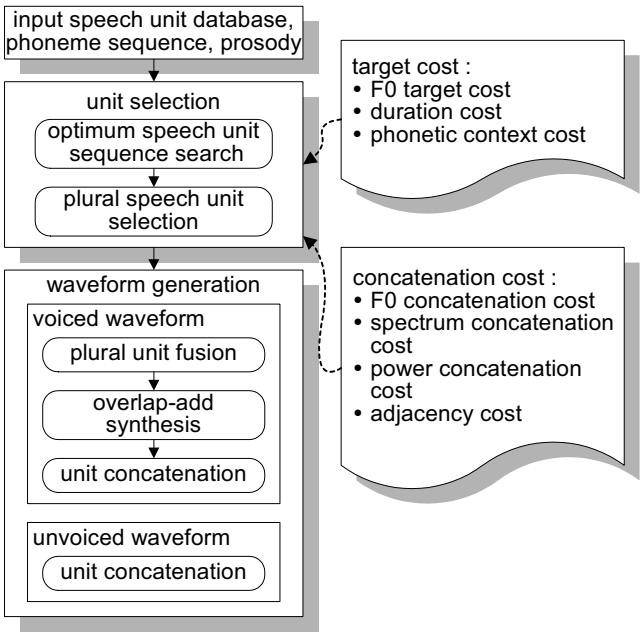

Figure 1. Speech synthesis process based on the plural unit selection and fusion method

concatenation cost. Target cost is defined by the weighted sum of F_0 target cost, duration target cost, and phonetic context cost. Concatenation cost is also defined by the weighted sum of F_0 concatenation cost, spectrum concatenation cost, power concatenation cost, and adjacency cost (set to 0 when two consecutive units are adjoining in the speech unit database, otherwise 1).

In the unit selection process, the optimum unit sequence that minimizes the total cost function is obtained based on concatenation cost and target cost using the DP algorithm. Then, plural speech units are selected based on the cost function using consecutive units in the optimum unit sequence (described in 2.1.). As a result, we obtain plural speech units for each diphone.

Next, in the waveform generation process, speech waveform is synthesized using the selected plural speech units. This process is performed for each phoneme segment, and the process is divided into voiced segment generation, and unvoiced segment generation. For voiced segments, a fused pitch-cycle waveform sequence that represents the selected plural speech units is generated. Then, the voiced waveform is synthesized by overlap-adding the generated pitch-cycle waveforms at pitch marks, which are converted from input prosodic information. For unvoiced segments, we just concatenate the optimum units.

2.1. Plural unit selection algorithm

Figure 2 depicts the plural unit selection algorithm. Let $C^t(\boldsymbol{p}_i, \boldsymbol{u}_i)$ be a target cost function between target

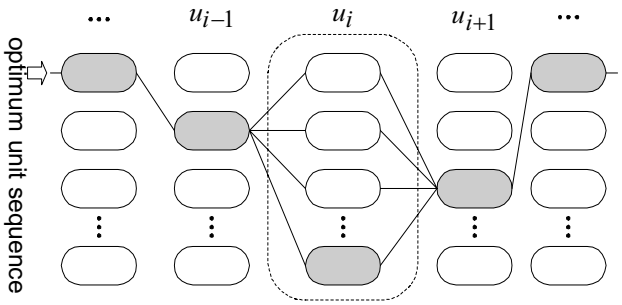

Figure 2. Plural unit selection

attribute $\boldsymbol{p}_i$ and unit $\boldsymbol{u}_i$, and let $C^c(\boldsymbol{u}_{i-1}, \boldsymbol{u}_i)$ be a concatenation cost function between $i-1$ th unit $\boldsymbol{u}_{i-1}$ and i th unit $\boldsymbol{u}_i$. The optimum unit sequence is obtained by minimizing the total cost function C,

$$C = \sum_i \{C^t(\boldsymbol{p}_i, \boldsymbol{u}_i) + C^c(\boldsymbol{u}_{i-1}, \boldsymbol{u}_i)\}. \quad (1)$$

In Figure 2, the gray marks represent the optimum unit sequence. Then, plural units are selected as i th segments based on the cost function C^u,

$$C^u = C^t(\boldsymbol{p}_i, \boldsymbol{u}_i) + C^c(\boldsymbol{u}^*_{i-1}, \boldsymbol{u}_i) + C^c(\boldsymbol{u}_i, \boldsymbol{u}^*_{i+1}), \quad (2)$$

where $\boldsymbol{u}^*_i$ represents the optimum i th unit. The candidate units for i th unit are sorted by C^u, and first N units are selected for plural N -best units.

2.2. Unit fusion algorithm

The unit fusion method we use is to average pitch-cycle waveform in time domain.

1. The pitch-cycle waveforms for each selected unit are extracted by multiplying Hanning window.
2. The number of pitch-cycle waveforms of each unit is adjusted to the number of target pitch marks by duplication or elimination.
3. The pitch-cycle waveforms for each target pitch mark are averaged in time domain to generate the fused pitch-cycle waveform.

In step 3, we average the band pass waveforms to reduce the attenuation of high frequency band. Step 3 is divided into decomposition of the pitch-cycle waveform by applying band-pass filters, alignment of the sub-band waveforms by searching the maximum correlation time lag, averaging the sub-band waveforms, and adding them into the whole band fused waveform. When we just use the pitch-cycle waveform of the optimum unit, the method becomes unit selection based speech synthesis using TD-PSOLA[9] based prosodic modification.

3. THE OFFLINE UNIT FUSION METHOD

The computation time for this synthesizer is mainly attributable to the unit fusion process. Table 1 shows the preliminary results for percentage of the computation time

Table 1. Time percentage of processes

%Time	Process
74.8 %	Unit fusion
17.9 %	Unit selection
7.3 %	Overlap-add synthesis, other

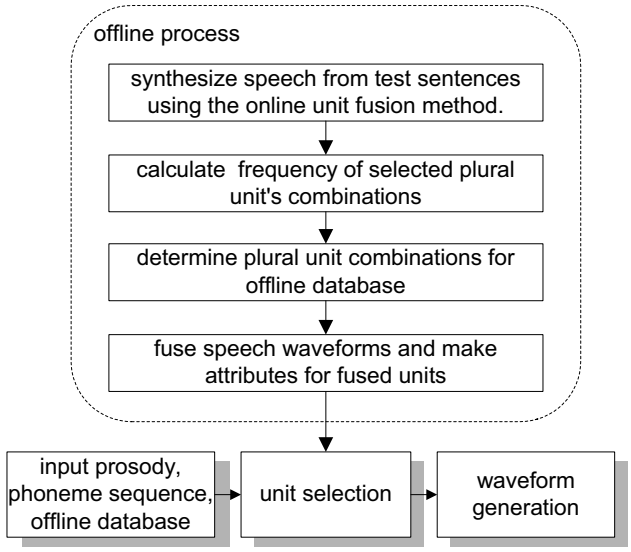

Figure 3. Offline unit fusion based speech synthesis

(female1 - Online in section 4.). From this table, it can be seen that the unit fusion process accounts for a large percentage. Therefore, avoiding the unit fusion process in synthesis time should decrease the computation time.

The offline unit fusion method performs the unit fusion process in offline mode. The speech unit database for the offline unit fusion method (offline database) consists of voiced speech units with fused pitch-cycle waveforms, unvoiced speech units and their attributes. Figure 3 shows the block diagram of offline unit fusion based speech synthesis. In the figure, we denote the synthesizer described in section 2 "online unit fusion method". The first step is to synthesize speech from a large number of test sentences using the online unit fusion method to get the information of selected plural unit combinations p^{ij}, $\boldsymbol{u}^{ij} = \{u_1^{ij}, u_2^{ij}, \cdots, u_N^{ij}\}$, where p^{ij}, u_n^{ij}, N represents the speech unit category (diphone, phoneme, halfphone, syllable, or so) of j th unit of i th sentence, the unit number in the category for n th best unit, and the number of plural units, respectively. Then we calculate the frequency of the plural unit combinations $F(p, \boldsymbol{u})$ by counting the occurrence of the combination of $\boldsymbol{u}$ in speech unit category p. Since the unit fusion method described in 2.2 does not depend on order of the plural units, we didn't consider the order of u_n. The third step is to determine the combinations for offline database. We set the maximum number of units for each category as L, sort the $F(p, \boldsymbol{u})$, and select the combinations $\overline{\boldsymbol{u}}_l (1 \le l \le L)$ from high frequency combinations in each category p. If the number of combinations for category p is smaller than L, all the combinations that appeared in the test sentences are used. Finally, the units in the each selected combination are fused into offline pitch-cycle waveforms to make an offline database. The attributes of F_0 contour, duration, and end-point spectrum parameter are also generated by averaging the selected plural unit's attributes, and those of phonetic context and adjacency information are created by just copying each unit's attributes.

In the synthesis time, after inputting offline database, target phoneme sequence, and prosody, unit selection is performed based on the attributes and optimum unit sequence is obtained. By overlap-adding the pitch-cycle waveforms of offline database, we generate the voiced waveforms. For unvoiced segments, we just concatenate the optimum units. In this method, the maximum number of units in the offline database can be easily specified as L. Thus, we can easily control the size of the offline database.

4. EXPERIMENTS

To evaluate the proposed method, we implemented a prototype system. We used diphones as speech units, and the diphone balanced speech database for two female speakers of Japanese (female1, female2) and a male speaker of Japanese (male1) is used. The databases consisted of 628, 937, and 876 sentences: 32, 54, and 44 minutes waveform without silence, respectively. The sampling frequency *fs* was 22.05kHz. We set the number of plural units N to 3. In the unit selection process, a maximum of 50 units for each segment were pre-selected using target cost, and optimum units searches using total cost function were performed for pre-selected units. The number of bands for sub-band averaging is four, and the boundaries are *fs*/16, *fs*/8 and *fs*/4. We used 10,000 sentences randomly selected from a newspaper database[1] and diphone balanced sentences for calculating the frequency of plural unit combinations.

Figure. 4 shows the database size and process time for female1 database. The black and white bar represents the database size (MB) and the computation time (sec.), respectively. "Online" represents graphs of the online unit fusion method and the others are max number of units for each diphone (L). The computation time was measured on PC platform (Pentium 4, 3.06GHz). It was measured

[1] The Yomiuri Shimbun database

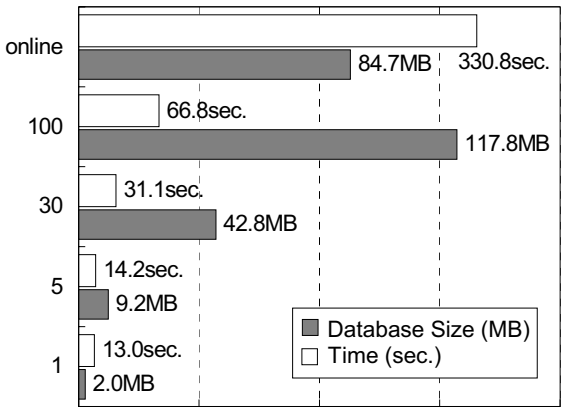

Figure 4. Database size and computation time

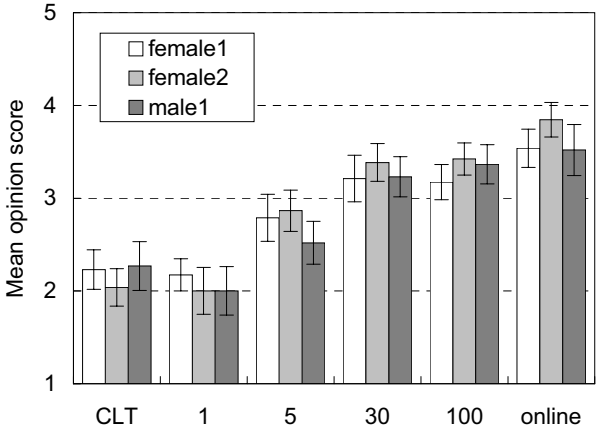

Figure 5. Subjective evaluation

by synthesizing speech from 1081 test sentences. It can be seen that the computation time is greatly reduced by using the offline unit fusion method. If the maximum number of units is set to 100, the computation time is reduced to approximately 1/5. It can also be seen that the computation time is getting smaller as decreasing the database size.

Figure 5 shows the results of subjective evaluations. The number of horizontal axis represents the maximum number of units for each diphone. "CLT" and "online" denote a single diphone synthesizer based on close-loop training[7] and the online method, respectively. The number of subjects was 13, and subjects listened to the synthetic speech and gave the 5-point MOS value. Four sentences were used for evaluation for each category. It can be seen that the MOS of L =100 was close to the MOS of the online method and higher than that of the conventional CLT method. The MOS of L =30 was also close to L =100. Consequently, the offline method with half-size dictionary reduces the computation time to 1/10, and the quality of synthetic speech is comparable. In the case of L =1, the MOS is lower than CLT. One of the reasons is that the power of synthetic speech for L =1 is unstable. In case of L =5, database size is about 10MB, computation time is 1/25, and the MOS value is higher than conventional CLT method. It also can be useful for some platforms.

5. CONCLUSIONS

In this paper, we proposed a scalable speech synthesis method based on the plural unit selection and fusion method. We proposed an offline unit fusion method in which pitch-cycle waveforms for voiced segments are fused in advance to avoid the unit fusion process in the synthesis time. The experimental results show that the offline unit fusion method with half-size waveform database is comparable to the online unit fusion method in terms of speech quality and the computation cost is reduced to 1/10.

One of our subjects for future work is multilingual extension. Since the proposed method is language independent, we can apply it to other languages.

6. REFERENCES

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," Proc. ICASSP-96, pp. 373–376, 1996.

[2] 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.

[3] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit," Proc. ICASSP 2002, pp. 465–468, 2002.

[4] E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith, M. Viswanathan "Recent Improvements to the IBM Trainable Speech Synthesis System," Proc ICASSP 2003, pp.701-708, 2003.

[5] A. K. Syrdal, C. W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K. Lee, M. J. Makashay, "Corpus-based techniques in the AT&T NEXTGEN synthesis system", Proc. ICSLP2000, pp. 410-415, 2000.

[6] T. Kagoshima and M. Akamine, "Automatic generation of speech synthesis units based on closed loop training," Proc. ICASSP-97, pp.963–966, 1997.

[7] M. Akamine and T. Kagoshima, "Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS Drive TTS)," Proc. ICSLP-98, pp.1927–1930, 1998.

[8] T. Mizutani and T. Kagoshima, "Speech synthesis based on selection and fusion of a multiple unit," 1-7-3, Proc. 2004 Spring Meeting of ASJ, pp.217–218, 2004 (in Japanese).

[9] C. Hamon, E. Moulines, and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modifications of speech," Proc. ICASSP-89, pp.238–241, 1989.