AUTOMATIC GENERATION OF SPEECH SYNTHESIS UNITS BASED ON CLOSED LOOP TRAINING

Takehiko Kagoshima[1] *Masami Akamine*[1]

[1]Toshiba R&D Center, 1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki-shi, Japan
kagosima, akamine@eel.rdc.toshiba.co.jp

ABSTRACT

This paper proposes a new method for automatically generating speech synthesis units. A small set of synthesis units is selected from a large speech database by the proposed Closed-Loop Training method (CLT). Because CLT is based on the evaluation and minimization of the distortion caused by the synthesis process such as prosodic modification, the selected synthesis units are most suitable for synthesizers. In this paper, CLT is applied to a waveform concatenation based synthesizer, whose basic unit is CV/VC(diphone). It is shown that synthesis units can be efficiently generated by CLT from a labeled speech database with a small amount of computation. Moreover, the synthesized speech is clear and smooth even though the storage size of the waveform dictionary is small.

1. INTRODUCTION

In a concatenative speech synthesis system with a small set of synthesis units, synthetic speech quality depends on the employed units. A context-oriented clustering method (COC) was proposed to generate synthesis units automatically from speech databases [1], where phonemic clustering with context dependence is performed on the basis of inner-cluster variance. This method minimizes error variance between the representative unit and training units in the cluster. However, the distortion in the synthesized speech cannot be minimized by COC, because the distortion caused by prosodic modification is not considered in the clustering criterion.

An alternative method was proposed to generate context-dependent CV(consonant and vowel) syllabic units[2]. This method improves the quality of the synthesized speech by using the CV units as the synthesis units. However, the above problem has not been addressed.

This paper proposes a new method for automatically generating synthesis units. The process of constructing synthesis units is formulated in two steps. Firstly, degradation caused by the prosodic modification of synthesis units is defined. The distortion is formulated as a difference between a natural speech segment prepared as training data cut off from the speech database and a synthesized speech segment with prosodic modification. In this formulation, the pitch period and duration of the synthesis unit are so modified to be equal to those of the training data. Secondly, the selection of the best synthesis unit is performed on the basis of a criterion to minimize the distortion in the cluster. The proposed method is called Closed-Loop Training method (CLT) because it is based on the evaluation and minimization of the distortion caused by the synthesis process. This method is so flexible that it can be used for various kinds of synthesizers. Thus the proposed method has been applied to both LSP-based and waveform-based synthesizers. This paper presents the generation of speech synthesis units for a Japanese synthesizer which is based on waveform concatenation. A basic unit of the synthesizer is CV/VC(diphone). It is shown that the proposed method produces synthesized speech having natural and smooth sounding quality.

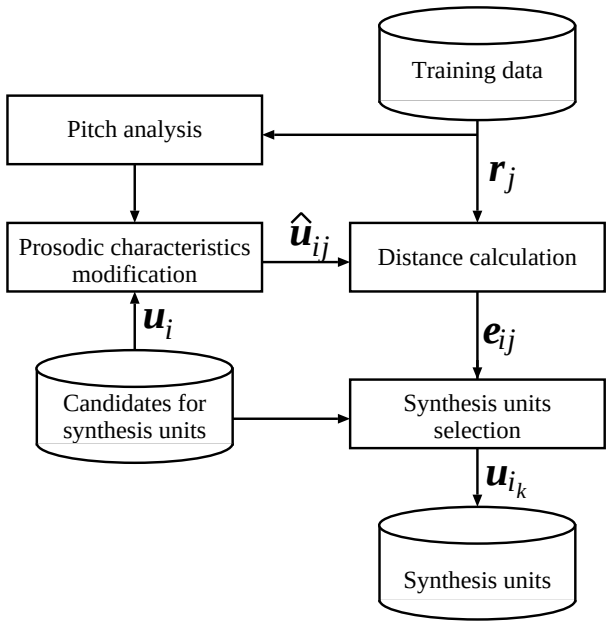


Figure 1. Procedure of synthesis unit selection

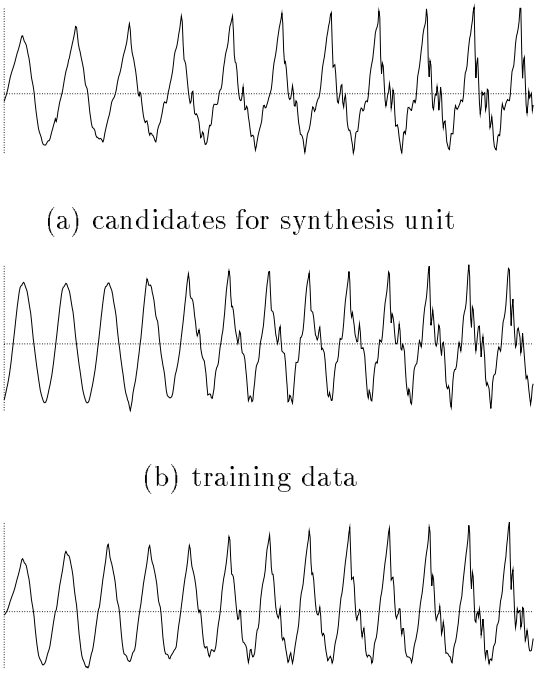
2. CLOSED LOOP TRAINING(CLT)

Figure 1 shows the procedure of synthesis unit selection. The procedure of CLT is described hereafter. A set of synthesis units is constructed for every CV/VC. Firstly, speech segments $\boldsymbol{u}_i, (i = 1, \cdots, N)$ and speech segments $\boldsymbol{r}_j, (j = 1, \cdots, M)$ are prepared as candidates for synthesis units and training data, respectively, by cutting them off from a speech database. Then an error matrix $E = (e_{ij})$ is calculated. Here, e_{ij} is the distortion between training data $\boldsymbol{r}_j$ and output speech $\hat{\boldsymbol{u}}_{ij}$, which is synthesized from $\boldsymbol{u}_i$ by modifying its pitch period and duration so that they are identical with those of $\boldsymbol{r}_j$. The best set of synthesis units is selected from the candidate units $\boldsymbol{u}_i$ so as to minimize the average distortion of the synthetic speech for all the training data.

2.1. Prosodic Modification

In this paper, CLT is applied to a waveform-based synthesizer which is based on the pitch-synchronous overlap-add(PSOLA) method[3].

Figure 2 shows an example of the prosodic modification of a speech segment /ne/. The speech waveform $u(n)$ (Figure 2(a)) corresponding to a candidate for synthesis units is decomposed into a



(a) candidates for synthesis unit

(b) training data

(c) synthetic speech

Figure 2. An example of prosodic modification

sequence of short-term overlapping signals $u^m(n)$:

$$u^m(n) = w^m(n)u(n + t_m), \tag{1}$$

where $w^m(n)$ represents the Hanning window whose length is twice as long as the local pitch period. The successive instants t_m indicate pitch-marks. A sequence of short-term signals $\hat{u}^q(n)$ corresponding to pitch marks $\hat{t}_q$ of the training data $r(n)$ (Figure 2(b)) is generated by duplicating or eliminating the short-term signals $u^m(n)$ depending on the duration of the training data. The synthetic speech $\hat{u}(n)$ (Figure 2(c)) is obtained by overlap-adding the stream of short-term signals:

$$\hat{u}(n) = \sum_q \hat{u}^q(n - \hat{t}_q + \delta_q), \tag{2}$$

where δ_q is a term which plays the role of aligning the synthetic speech with training data. The term δ_q is so decided as to minimize the following cross-correlation function $\phi(\delta)$:

$$\phi(\delta) = \sum \hat{u}^q(n + \delta)r^q(n), \tag{3}$$

$$r^q(n) = \hat{w}^q(n)r(n + t_q), \tag{4}$$

where $\widehat{w}^q(n)$ represents the Hanning window whose length is twice as long as the local pitch period of the training data.

2.2. Distortion measure

The distance between the voiced portion of the training data and that of the synthetic speech is calculated to evaluate the candidate for synthesis units. Because the power of the synthetic speech is not controlled in the CLT procedure, it is required that the distance function is independent on the difference between the power of the synthetic speech and that of the training data. Therefore the authors define the following error function as the measure of the distortion between the training data $r(n)$ and the synthetic speech $\widehat{u}(n)$:

$$e = \sum_n \left(\frac{r(n)}{\bar{r}} - \frac{\widehat{u}(n)}{\bar{\widehat{u}}} \right)^2, \tag{5}$$

where

$$\bar{r} = \left(\sum_i r(i)^2 \right)^{\frac{1}{2}}, \tag{6}$$

$$\bar{\widehat{u}} = \left(\sum_i \widehat{u}(i)^2 \right)^{\frac{1}{2}}. \tag{7}$$

2.3. Synthesis Unit Selection

The best set of synthesis units is selected from the candidate units $\boldsymbol{u}_i$ so as to minimize the following cost function:

$$C(i_1, \cdots, i_n) = \frac{1}{M} \sum_{j=1}^{M} \min(e_{i_1 j}, \cdots, e_{i_n j}). \tag{8}$$

The clusters of the training data are represented by the following sets G_k corresponding to the synthesis units u_{i_k}:

$$G_k = \{\boldsymbol{r}_j | \min(e_{i_1 j}, \cdots, e_{i_n j}) = e_{i_k j}\}. \tag{9}$$

If the number of synthesis units per CV/VC is one, the cost function is written as follows:

$$C(i) = \frac{1}{M} \sum_{j=1}^{M} e_{ij}. \tag{10}$$

The selected synthesis units minimize the average distortion of the synthesized speech for arbitrary prosodic characteristics.

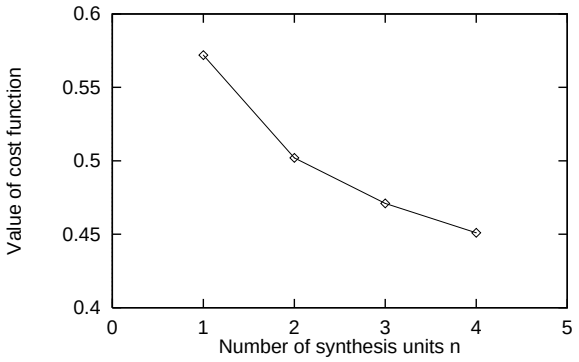


Figure 3. Relation between cost function $C(i_1, \cdots, i_n)$ and number of synthesis units.

3. EXPERIMENTAL RESULTS

CLT has been applied to generating a set of synthesis units of a Japanese text-to-speech system. A hand-labeled speech database was prepared for the training data and the candidates for synthesis units. Its sampling frequency was 11.025 kHz and its size was about 50 minute-long. Because a large number of training data and candidates for synthesis units with a variety of prosodic characteristics are required for CLT, all the speech segments extracted from the database were used for both (*i. e.* $\boldsymbol{u}_i = \boldsymbol{r}_i$). Figure 3 shows the relation between the average of the cost function $C(i_1, \cdots, i_n)$ and the number of synthesis units per CV/VC. It is seen that the distortion decreases as the number of synthesis units increases. When the number of synthesis units per CV/VC was one, the storage size of the waveform dictionary, which consists of 262 CV/VC units, was about 1.3 Mbytes and the training time was about 1.5 hours on a Sun Ultra2.

To confirm the validity of the proposed method, a comparative test was carried out. A set of synthesis units was generated by a training method which minimizes error variance between the representative unit and training units. This method can be regarded as Open-Loop Training(OLT) in contrast to CLT. The procedure to generate single units per CV/VC by OLT is carried out as follows:

step 1. Calculate the LSP parameters of all the candidates for synthesis units.

(a)male speech

(b)female speech

Figure 4. The result of comparative test between CLT and OLT.

Table 1. The condition of the comparative test

Utterances	10 sentence per speaker
Subjects	7 males
Number of units	1 per CV/VC
Prosody	Natural speech

step 2. Take the average of all the LSP parameters.

step 3. Select the synthesis unit whose LSP parameters are the closest to the average.

Figure 4 shows the result of the comparative test, whose condition is shown in Table 1.

4. DISCUSSION

The synthetic speech from the synthesis units by CLT was smoother than that by OLT. This is remarkable in female speech. Generally speaking, the distortion caused by prosodic modification is large in female speech. CLT is most effective for producing good quality female speech because CLT minimizes distortion caused by prosodic modification.

Synthetic speech quality can further be improved by increasing the number of synthesis units per CV/VC and introducing a criteria to select a suitable unit from the synthesis units according to prosody and/or context. The criteria can be extracted from the prosodic and/or contextual information of the clusters of training data.

The necessary training time is several hours if the number of synthesis units per CV/VC is less than 5. Computational requirements, however, grow exponentially with the number of synthesis units per CV/VC. Thus, it is required to improve the algorithm to minimize the cost function of Eqn. 8 in order to apply this training method to a larger set of synthesis units.

The proposed method can be improved by incorporating context dependent clustering such as the prosodic or neighboring phoneme dependent clustering and/or by introducing a distortion related to the transition of consecutive units. This topic will be discussed in a future study.

5. CONCLUSION

This paper has presented a new method for automatically generating synthesis units. The proposed Closed-Loop Training method (CLT) is based on the evaluation and minimization of the distortion caused by the synthesis process. This method is so flexible that it is applicable to various kinds of synthesizers. In this paper, CLT was applied to generate synthesis units for waveform concatenation based synthesizers, whose basic unit was CV/VC(diphone). It was shown that the synthesized speech was clear and smooth even though the storage size of the waveform dictionary was small.

REFERENCES

[1] M. Mohan Sondhi, *et al.*, "Advances in Speech Signal Processing", Marcel Dekker, Inc., 1992.

[2] T. Saito, Y. Hashimoto and M. Sakamoto, "High-quality speech synthesis using context-dependent syllabic units," Proc. of ICASSP, pp. 381–384, May 1996.

[3] C. Hamon, E. Moulines and F. Charpantier, "A Diphone Synthesis System Based on Time-Domain Prosodic Modification of Speech," Proc. of ICASSP, pp. 238–241, May 1989.