Analytic Generation of Synthesis Units by Closed Loop Training for Totally Speaker Driven Text to Speech System (TOS Drive TTS)

Masami Akamine and Takehiko Kagoshima

Toshiba Kansai Research Laboratories

8-6-26, Motoyama-minami-machi, Higashi-nada-ku, Kobe, 658-0015 Japan

ABSTRACT

This paper provides a new method for automatically generating speech synthesis units. The algorithm, called Closed-Loop Training (CLT), is based on evaluating and reducing the distortion in synthesized speech. It minimizes distortion caused by synthesis process such as prosodic modification in an analytic way. The distortion is measured by calculating the error between synthesized speech units and natural speech units in a large speech database (corpus). The CLT method effectively generates the synthesis units that are most resembling of natural speech after synthesis process. In this paper, CLT is applied to a waveform concatenation based synthesizer, whose basic unit is a diphone. By using CLT, the synthesizer generates clear and smooth synthetic speech even with a relatively small volume of synthesis units.

1. INTRODUCTION

A speaker-driven approach is a promising way to produce high quality and highly natural speech [1][2]. Figure 1 shows a totally speaker-driven text-to-speech system, which has been developed in Toshiba. In this system, prosody information and synthesis units are automatically derived from the speech corpus and synthetic speech is produced through pitch synchronous overlap-and-add process on the synthesis units. Naturally, speech quality depends on the quality of employed synthesis unit.

The conventional TTS system uses a context-oriented clustering method (COC) [3] or decision-tree-based clustering of context-dependent phonetic units [1] to generate synthesis units, where phonemic clustering with context dependence is carried out on the basis of inner-cluster variance. These methods try to minimize distortion within each cluster. However, synthesized speech suffers from the distortion caused by prosodic modification, which is not taken into consideration in clustering.

A closed-loop training method for automatic generation of synthesis units minimizes the distortion in synthetic speech [4]. The basic idea is to select the optimal unit that minimizes the distortion caused by prosodic modification from the candidates extracted from speech corpus. It was shown that the closed-loop training method improves synthetic speech quality over the conventional method.

To enhance the quality of the synthesis unit even further, an analytic approach can be employed to automatically generate the optimal unit. In this approach, the optimal unit is not selected from speech corpus but produced by solving an equation.

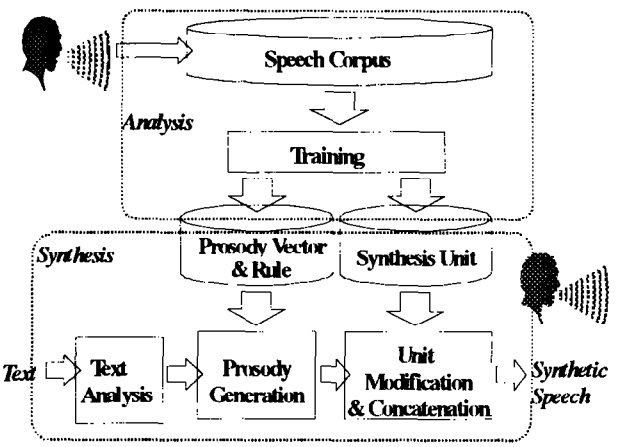

Figure 1. Block diagram of Totally Speaker Driven TTS System

This paper proposes an analytic method to generate synthesis units automatically. It can be applied to various kinds of synthesizers such as LSP-based and waveform-based synthesizers. It also can be applied to any language. In this paper, it is applied to generating CV/VC-type synthesis units for a Japanese synthesizer that is based on waveform concatenation. First, a distortion is defined as squared errors between the synthesized vector and corresponding training vectors from speech corpus. The optimal unit is obtained by solving an equation where the differential of the distortion with respect to the unit equals zero. The analytic method was compared with the selective method [4] in terms of speech quality and flexibility in measuring distortion. The analytic method improved smoothness in synthesized speech. However, it showed limitation in measuring distortion because the distortion function must be differentiable with respect to the synthesis unit. 302 CV/VC units were generated by the proposed method in five hours by SUN Ultra2 from a 40 minute long speech corpus. It was confirmed that the proposed approach produces clear and smooth synthetic speech on listening tests, even with a small storage size for the synthesis units.

Text analysis and prosodic generation from the speech corpus are described in other paper [5][6].

2. CLOSED-LOOP TRAINING

2.1. Selective Method

This method selects the optimal unit vector that minimizes the distortion in synthesized speech from candidate unit vectors extracted from a speech corpus. The procedure of the selective method is shown in Figure 2. First, speech segments are prepared as the candidate vectors for synthesis units and the training vectors by extracting them from the speech corpus. Then we calculate a distortion between the training vectors and the synthesized speech vector from the candidate unit-vector by modifying its pitch period and duration so that they are identical with those of the training vectors. The optimal unit vector is selected from the candidate so as to minimize the distortion. The modification of the pitch period and duration is performed by the pitch-synchronous overlap-add (PSOLA) method [7].

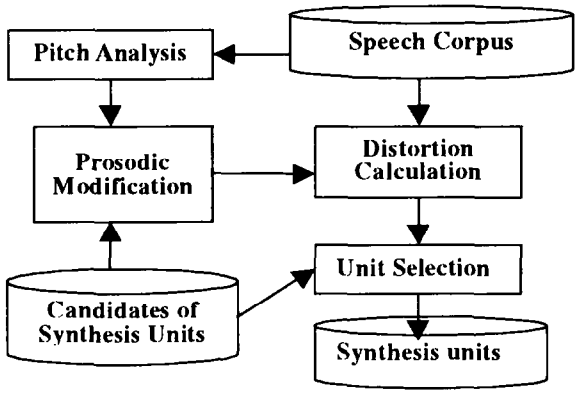

Figure 2. Procedure of synthesis unit selection in closed-loop training

2.2. Analytic Method

Unlike the selective method mentioned above, the analytic method creates the synthesis units in an analytic way in order to reduce the distortion much further. The procedure for generating the synthesis units is as follows:

Step 1: Prepare speech segments as training vectors. The training vectors are given from a speech corpus in the same way as in the selective method.

Step 2: Set initial synthesis unit vectors. We use synthesis units obtained by the selective method as them.

Step 3: Partition the training vectors into cluster sets based on the nearest neighbor condition using a distance between the synthetic speech vector and the training vectors.

Let u_i and r_j be the synthesis unit vector and the training vector, respectively. Then, we partition the training vector r_j into cluster set G_i as follows:

$$G_i = \{r_j : d(r_j, y_{j,i}) < d(r_j, y_{j,k}); \; all \; k \neq i\}, \qquad (1)$$

where, $d(r_j, y_{j,i})$ and $y_{j,i}$ is a distance measure and the synthesized vector from u_i.

Prosodic modification: $y_{j,i}$ is produced from u_i by modifying its pitch period and duration so that they are identical with those of the training vector r_j. We can change the pitch period and duration based on the pitch-synchronous overlap-add process.

Figure 3 shows an example of pitch-synchronous overlap-add process. The synthesis unit vector is decomposed into a sequence of short-term overlapping vectors (ST-vectors) v_m:

$$v_m(n) = w(n)u(n + t_m); \; n = 1,2,\cdots,L, \qquad (2)$$

where $w(n)$ is the Hanning window whose length L is twice as long as the local-pitch period. The successive instances t_m indicate pitch-marks. We obtain the synthetic speech vector by overlap-adding the ST-vectors at pitch marks of the training vector. Elimination or duplication of the ST-vectors is carried out due to changes in the prosody.

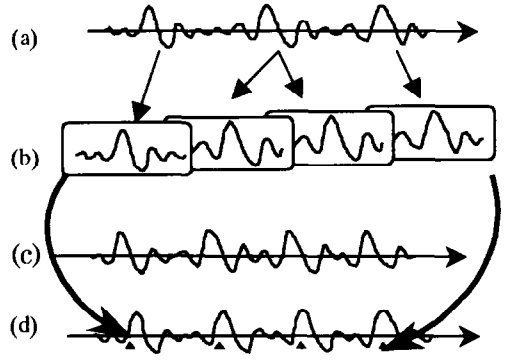

Figure 3. Example of pitch-synchronous overlap-add process: (a) Unit vector, (b) ST-vector, (c) Training vector, (d) Modified vector

The prosody modification can be elegantly described through matrix operations. The synthesis unit vector u is given as a super vector of the ST-vectors v_m as follows:

$$u = (v_1^T, v_2^T, \cdots, v_P^T)^T. \qquad (3)$$

We specify a matrix $I_{t,k}$ as the operator overlapping the ST-vector v_k at the pitch mark t:

$$I_{t,k}(m,n) = \begin{cases} 1; t < m \le t+L, (t-1)L < n \le tL \\ 0; otherwise. \end{cases} \quad (4)$$

The above matrix includes (L x L) unit matrix. Let $t_q, (q = 1,2,\cdots,R)$ be the pitch marks of training vector r, then we describe the synthetic speech vector y by matrix operation as follows:

$$y = Au, \quad (5)$$

$$A = \sum_{q=1}^{R} I_{t_q,k_q}, \quad (6)$$

where k_q indicates that the ST-vector v_{k_q} is put at the pitch mark t_q. Matrix A plays a role of overlap-add operator. The following equation shows an example of matrix A, where $P = 3, R = 4, k_1 = 1, k_2 = 2, k_3 = 2, K_4 = 3$.

$$A = \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & 1 & & & 0 & \\ & & & & \ddots & & & \\ & & & 1 & & 1 & & \\ & & & & \ddots & & & \\ & & 0 & & & 1 & 1 & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{bmatrix} \quad (7)$$

Distance measure: We define the distance measure $d(r_j, y_{j,i})$ as a squared error $e_{j,i}$ between the training vector r_j and the synthetic speech vector $y_{j,i}$:

$$e_{j,i} = (r_j - g_{j,i} y_{j,i})^T (r_j - g_{j,i} y_{j,i}), \quad (8)$$

where, $g_{j,i}$ is a gain factor that adjusts the difference in signal level between the training vector and synthetic speech vector. We can optimize the gain factor to minimize the error given in Equation (8) by solving the following equation:

$$\frac{\partial e_{j,i}}{\partial g_{j,i}} = 0, \quad (9)$$

$$g_{j,i} = \frac{y_{j,i}^T r_j}{y_{j,i}^T y_{j,i}}. \quad (10)$$

Step 4: Generate the optimal unit-vector that minimizes distortion in each cluster. At first, we define a distortion E_i in a cluster G_i as the sum of squared errors between the training vectors and synthetic speech vectors in the cluster:

$$\begin{aligned} E_i &= \sum_{r_j \in G_i} (r_j - g_{j,i} y_{j,i})^T (r_j - g_{j,i} y_{j,i}) \\ &= \sum_{r_j \in G_i} (r_j - g_{j,i} A_{j,i} u_i)^T (r_j - g_{j,i} A_{j,i} u_i), \end{aligned} \quad (11)$$

where, matrix $A_{j,i}$ is the overlap-add operator that modifies the pitch period and duration of the synthesis-unit vector u_i so as to be those of the training vector r_j.

Let the differential of the distortion E_i with respect to the synthesis-unit vector u_i be zero:

$$\frac{\partial E_i}{\partial u_i} = 0. \quad (12)$$

From the above equation, we derive the following equation:

$$\left(\sum_{r_j \in G_i} g_{j,i}^2 A_{j,i}^T A_{j,i} \right) u_i = \sum_{r_j \in G_i} g_{j,i} A_{j,i}^T r_j. \quad (13)$$

The optimal synthesis-unit vector u_i is obtained by solving the above equation.

Step 5: Update the synthesis unit-vectors by replacing the old unit by the new one obtained in Step 4.

Step 6: Repeat Step 3 to Step 5 until the sum of distortions for each cluster converges. We compute the sum of distortions for all clusters, $\sum_i E_i$. If it has changed by a small enough amount since the last iteration, we stop. Otherwise, we repeat Step 3 to Step 5.

3. EXPERINMENTAL RESULTS

The proposed method has been applied to generating a set of synthesis units of a Japanese text-to-speech system, which uses waveform concatenation-type synthesizer. 302 CV/VC units were generated by each of the selective method [4] and the analytic method presented in this paper. A hand-labeled speech corpus was used for the training data. Its sampling frequency was 11.025 kHz and its size was about 40 minute long. Synthesis units generated by the selective method were used as the initial units of the proposed closed-loop training.

Figure 4 shows the relation between the total distortion $\sum_i E_i$ and the number of synthesis units of a Japanese diphone "Na", where the number of the training vectors was 100. In Figure 4, the distortions were normalized by that of the selective method when the number of synthesis units was one. The figure shows that the distortion decreases by 25 to 30 % in comparison with that of the selective method.

An informal listening test was conducted for female and male synthetic voices, using 6 subjects, comparing the selective method and the analytic method for synthesis unit generation.

20 sentences were processed to generate prosodic data by our Japanese TTS system. 302 CV/VC units with single unit for each diphone were generated by both methods. The storage size of the synthesis units was about 1.4 Mbytes. It took two hours for the selective method and another three hours for the analytic method on SUN Ultra 2. Figure 5 shows the result of the comparative tests. The preference scores of the proposed method were 80 % and 76 % for male voices and female voices, respectively. From interviewing the test subjects, it was confirmed that speech quality was improved in clearness and smoothness.

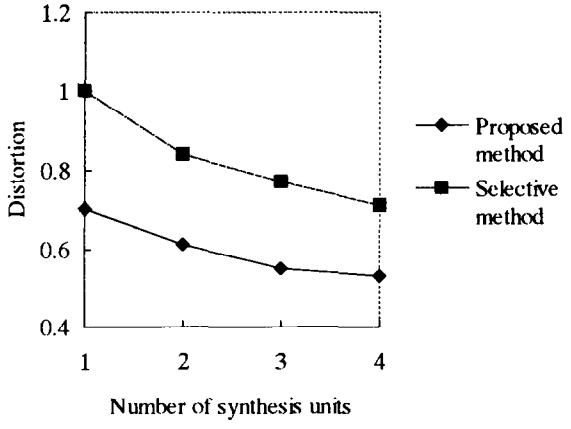

Figure 4. Relation between total distortion and the number of synthesis units for proposed method and selective method.

The analytic method improves smoothness in synthesized speech. However, it has limitation in measuring distortion because the distortion function must be differentiable with respect to the synthesis unit. In this paper, the squared error between the synthesized vector and training vectors was utilized as the distortion because it was differentiable with the synthesis unit, while it was not a perfect measure representing human perception. We need a better distortion measure. It may not be differentiable. The selective method can be applied even that case.

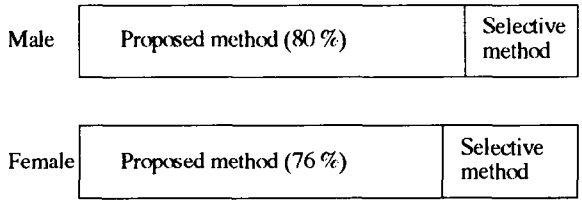

Figure 5. Preference scores for proposed method and selective method

4. CONCLUSION

Closed-Loop Training minimizes distortion caused by synthesis process. The distortion is described in a numeric equation in terms of synthesis units. The optimal unit is obtained by solving a numeric equation derived by setting the differential of the distortion with respect to the synthesis unit zero. It was confirmed that the proposed method produces clear and smooth synthetic speech on listening test. The proposed method can be incorporated with context dependent clustering, such as the prosodic clustering and neighboring phoneme dependent clustering, to improve speech quality even further. These are future works.

5. REFERENCES

1. X. Huang et al, "Recent improvements on Microsoft's trainable text to speech system-Whistler," *Proc. ICASSP97, pp.959-962, Apr. 1997.*
2. E. Lopez-Gonzalo et al, "Automatic prosodic modeling for speaker and task adaptation in text-to-speech", *Proc. ICASSP97, pp.927-930, Apr. 1997.*
3. K. Ito, S. Nakajima and T. Hirokawa, "A new waveform speech synthesis approach based on the COC speech spectrum", *Proc. ICASSP94, pp.577-580, Apr. 1994.*
4. T. Kagoshima and M. Akamine, "Automatic generation of speech synthesis units based on closed loop training", *Proc. ICASSP97, pp.963-966, Apr. 1997.*
5. S. Seto, M. Morita, T. Kagoshima and M. Akamine, "Automatic rule generation for linguistic features analysis using inductive learning technique", to be appeared in *Proc. ICSLP98, Nov. 1998.*
6. T. Kagoshima, M. Morita, S. Seto and M. Akamine, "An F0 contour control model for totally speaker driven text to speech system", to be appeared in *Proc. ICSLP98, Nov. 1998.*
7. C. Hamon, E. Moulines and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modifications of speech", *Proc. ICASSP89, pp.238-241, May 1989.*