UNIT DATABASE PRUNING BASED ON THE COST DEGRADATION CRITERION FOR CONCATENATIVE SPEECH SYNTHESIS

Nobuyuki Nishizawa and Hisashi Kawai

KDDI R&D Laboratories Inc., Japan

{no-nishizawa, Hisashi.Kawai}@kddilabs.jp

ABSTRACT

A novel method of unit database pruning for concatenative speech synthesis is proposed. The proposed method uses sums of the unit preference criterion, which are calculated from cost degradation from the optimal sequence, instead of the appearance frequencies of units, which is used in the conventional method. Therefore, the proposed method is an extension of the conventional method. Since not only the optimal units but also the other candidate units can be taken into account for pruning, unit databases can be pruned with less experimental speech synthesis. The results of a unit selection experiment on 4-hour pruned unit databases built from the original 10.6-hour database indicate that the amount of the experimental speech synthesis can be reduced to 25% of that required for the conventional method without loss of the quality of synthetic speech in terms of average cost.

Index Terms— Speech synthesis, unit selection, database pruning, preselection

1. INTRODUCTION

In concatenative speech synthesis[1], large-scale unit databases are often used because the quality of generated sounds depends on the availability of speech segment units suitable for target speech sounds. In recent years, huge unit databases built from several tens of hours of speech data can be used. Storage for such a database amounts to several tens of gigabytes. However, such a database is not yet acceptable for many computers with limited resources. Therefore, reduction of databases, which is often called database pruning, is still required. Of course, degradation of quality caused by reduction should be as small as possible.

In previous studies, two approaches have been mainly investigated. One is based only on the statistical distribution of units included in the database, and the other is based on results of experimental speech synthesis. In the typical method of the former approach, clustering of units in the database is performed. Then, only representative units for the clusters are kept, and the other units of the clusters are removed[2]. However, since the statistical distribution of units in the database is often different from that of the targets, a unit database may include many useless units for speech synthesis. In the former approach, pruning only such useless units is impossible because the statistical features of targets for speech synthesis are not taken into account.

On the other hand, in the typical method based on the latter approach, the appearance frequencies of selected units by experimental speech synthesis are counted, and units whose appearance frequencies are low are removed[3]. In the method, many executions of experimental speech synthesis efficient to simulate actual speech synthesis are required. If the purpose of the speech synthesis system is reading of text, for example, news or novels, experimental speech synthesis is not difficult because the targets for experimental speech synthesis can be predicted from large-scale text corpora by using modules for TTS (text-to-speech) systems. However, for purposes where large-scale text corpora are not available, experimental speech synthesis is not easy.

Therefore, in this study, a novel pruning method with a small amount of experimental speech synthesis is proposed. Different from the conventional method, not only the optimal units but also the other candidate units are evaluated for pruning in the proposed method. The method is based on a preselection criterion, which has been originally proposed for unit preselection at runtime[4].

The rest of the paper is organized as follows. Section 2 overviews concatenative speech synthesis and our speech synthesizer. Section 3 explains the criterion based on cost degradation from the optimal sequence. Section 4 describes the method of unit database pruning. Section 5 gives an evaluation of the experiments on unit selection with pruned databases by the proposed method is given in Section 5. Finally, section 6 concludes the paper.

2. CONCATENATIVE SPEECH SYNTHESIS

In concatenative speech synthesis, speech segments, each of which is often generalized as a unit, are selected from a database so that a criterion, which is often called cost, is minimized. To find the unit sequence with the minimal cost, a Viterbi search, which is based on the dynamic programming (DP) approach, is basically employed.

In our TTS (text-to-speech) system, which is based on XIMERA[5], cost function C that is calculated by integrating the target and concatenation costs over the entire utterance corresponds to the degradation of naturalness caused by using a unit sequence $\{u_i\}$ to synthesize an utterance for the target information sequences $\{t_i\}$. C is defined by a recurrence equation:

$$\begin{aligned} C(u_1|t_1) &= C_T(u_1|t_1) \\ C(u_1,...,u_i|t_1,...,t_i) &= C(u_1,...u_{i-1}|t_1,...,t_{i-1}) \\ &\quad + C_C(u_{i-1},u_i) + C_T(u_i|t_i) \end{aligned} \tag{1}$$

where C_T, C_C, and t_i denote target cost, concatenation cost, and target information at time i, respectively.

Target cost function C_T represents the degradation of naturalness caused by the disagreement between a target and a selected unit in the phonetic environment, phone duration, log F_0 (fundamental frequency), and MFCC (mel-frequency cepstral coefficients). On the other hand, concatenation cost function C_C represents the degradation of naturalness caused by discontinuity at the unit boundary in F_0 and MFCC. To accurately emulate the human perception of naturalness, the target cost function and the concatenation cost function were optimized by extensive perceptual experiments[6].

All of the target features except for the phonetic environment are predicted by using HMM-based speech synthesis techniques[7]. Figure 1 schematically shows the structure of the TTS system.

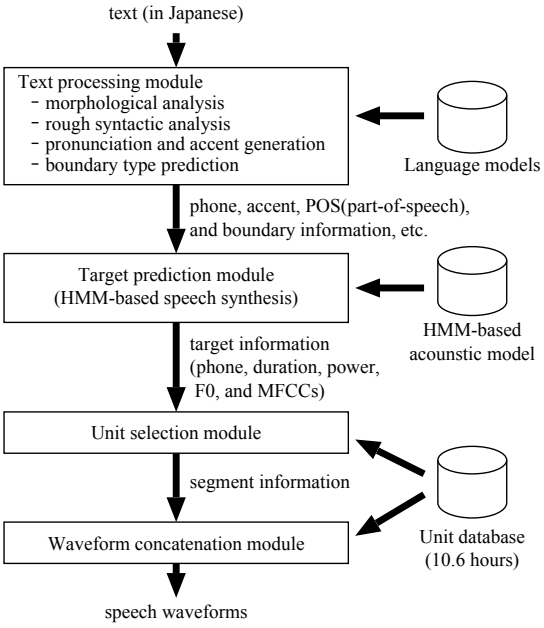


Fig. 1. Structure of the target TTS system.

3. PRESELECTION CRITERION BASED ON COST DEGRADATION FROM THE OPTIMAL SEQUENCE

In this section, a preselection criterion based on cost degradation from the optimal sequence is introduced for later discussion.

In conventional preselection, the target cost is often used as a criterion[1] because units far from the targets rarely constitute the optimal sequence. However, since it does not take concatenation cost into account, units that are close to the target but cannot be smoothly concatenated to the adjacent units in a sequence may also be kept in preselection. Therefore, we proposed a criterion on degradation in cost from the optimal sequence[4].

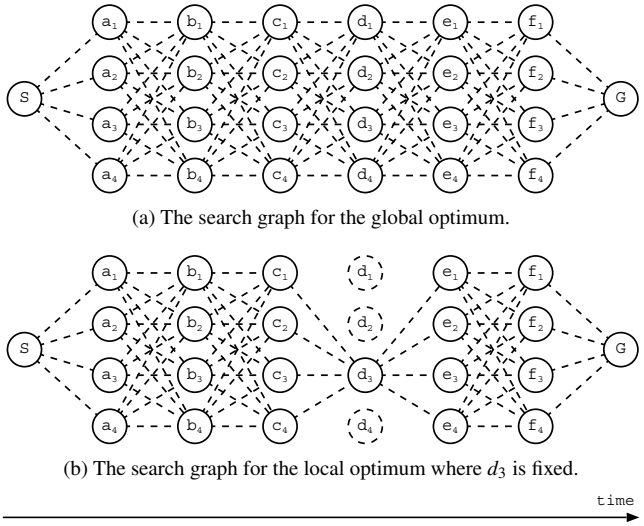
The criterion stands on the following idea: If unit u_a in the optimal unit sequence is forcibly replaced by another unit u_b, the sequence may not be optimum in all possible sequences where u_b is fixed. For search of the optimal sequence in such sequences, another unit selection where u_b is temporally fixed must be performed. Similarly, the inappropriateness of a preselected unit should be globally evaluated regarding the difference between the local optimum where the unit is fixed, and the global optimum.

Therefore, as a criterion of preselection for unit u when the target is t_i, cost degradation D is defined by:

$$D(u|t_i) = \min C(\mathbf{u}_{S\to u\to G}|\mathbf{t}) - \min C(\mathbf{u}_{S\to G}|\mathbf{t}) \quad (2)$$

where $\mathbf{u}_{S\to u\to G}$ and $\mathbf{u}_{S\to G}$ denote the sequences of units that correspond to paths from start node S to goal node G through the node for unit u and from start node S to goal node G in the search graph for the unit selection, respectively, and $\mathbf{t}$ denotes a sequence of targets. When unit u for target t_i is a component of the optimal unit sequence, the value of D is equal to 0.

Figure 2 schematically shows an example of search graphs for preselection. In this figure, D for d_3 is equal to the difference between the cost of the optimal path shown in graph (b) and the cost of the optimal path in graph (a). All of D for all unit u and all target t of an utterance can be computed by a forward Viterbi search and a backward Viterbi search[4]. Consequently, D of all candidate units for a utterance can be computed with double computational costs of conventional unit selection for searching the optimal sequence, where only the forward Viterbi search is performed.



(a) The search graph for the global optimum.

(b) The search graph for the local optimum where d_3 is fixed.

Fig. 2. An example of search graphs. In this example, the difference in cost between the optimal sequences of two graphs is evaluated for d_3 in preselection.

4. PRUNING OF UNITS

To reduce size of the unit database without degradation of synthetic sounds, units that are less preferred for the target sentences of the TTS system should be removed. Since the target sentences are not usually accurately known, less preferred units must be predicted, for example, by experimental unit selection for similar sentences of the target sentences. In the conventional method, the appearance frequencies of selected units were used as the criterion of pruning. On the other hand, in this study, instead of appearance frequency, cost degradation criterion D is used.

In the previous study for unit preselection at runtime[4], D is converted into unit preference p by the following equation at first:

$$p(u|t) = \exp(-\lambda D(u|t)) \quad (3)$$

If u is the optimal unit for target t, p is equal to 1. As D becomes larger, u converges to 0. This is because the difference in D should be focused only on units with small D. Unit with large D will be pruned by the preselection. In the proposed method, the same conversion is adopted at first.

Secondly, the sum of p is calculated for each unit u. The sum denotes $P(u)$:

$$P(u) = \sum_{t\in T} p(u|t) \quad (4)$$

where T is the set of all target information for experimental speech synthesis. In this study, $p(u|t)$ is defined as zero when unit u is not a candidate for target t. Finally, units where $P(u)$ are small are removed from the unit database. However, to avoid pruning all candidate units for a certain phone, a lower limit of the number of units for each phone is also adopted. In this study, the limit is set to 1.

The pruning can be controlled by parameter λ in equation (3). As λ becomes smaller, units with larger D will be also preserved in the pruning. On the other hand, if λ is set to a large value, the preselection method emulates the conventional method based on the appearance frequency of a unit selected for training data. i.e., the proposed method is an extension of the conventional pruning method based on the appearance frequencies of selected units.

5. EVALUATION

To evaluate the proposed pruning method, several sets of pruned unit databases were built and a unit selection experiment using the pruned unit databases was conducted.

In the evaluation, the original unit database was built from a Japanese speech corpus of approximately 10.6 hours pronounced by a female speaker. The corpus consisted of novels, news, travel dialogues, names, words, etc. The size of each unit in the database was a half-phoneme for vowels and unvoiced fricatives, or a phoneme for the other consonants.

5.1. Pruning of the unit database

First, cost difference values were computed for 16-hour target information that was predicted from 4907, 4664, and 2779 sentences from novels, news, and travel dialogues that were not included in the corpora for the unit database. For prediction, the text-processing module and the target prediction module of the TTS system were used. For comparison of different sizes of experimental speech synthesis, 0.5-hour, 1-hour, 2-hour, 4-hour and 8-hour data were reduced from the full data.

In pruning, λ was selected from 1, 2, 5, 10, 20, 50, 100, 200, 500, or 1000. For comparison, unit databases pruned by the conventional method based on frequencies were built. In our implementation of the conventional method, units to be removed were randomly selected when their appearance frequencies were the same.

5.2. Unit selection experiment

In this evaluation, the ATR's 503 phonetically balanced sentences[8], which were not included in either the corpus for the unit database or the text for pruning, were used as targets of unit selection. Similarly to experimental speech synthesis, the targets for unit selection were predicted from the sentences by using the TTS system. For comparison, unit selection with pruned databases by the conventional method based on the appearance frequency of selected units was also conducted.

Figure 3 shows the mean and the standard deviation of cost per unit of selected units for each evaluation sentence with pruned databases by the proposed method and the conventional method. As a typical condition, the size of the reduced database is 4-hour (the reduction rate is 62.3%); the TTS system with an uncompressed 4-hour, 16-bit, 16-kHz sampling database can be stored in a CD-ROM disk. The horizontal axis corresponds to the total duration of targets for pruning. The horizontal dotted line indicates the result when the original unit database is used. Therefore, the lines are the lower limit of results when pruned unit databases are used. In the figure, the results of the proposed method indicate the minimal costs in several λ settings. The figure indicates that the results using pruned unit databases by the proposed method are superior to those by the conventional method. In this experiment, the costs of selected units with pruned databases by the conventional method are achieved with those by the proposed method with less than 25% of the experimental speech synthesis required for the conventional method.

Secondly, the relationship between a setting of λ and the results of unit selection was examined. Figure 4 shows the mean costs of selected units. The figure shows that a large λ is not good especially where the data size for pruning is small. This is comprehensible result because a large λ emulates the conventional method. On the other hand, as the data size for pruning grows, the optimal λ becomes larger. i.e., the difference between the results of the proposed method and those of the conventional method becomes smaller. However, the difference between the results at optimal λ and those where λ is the optimum for small data for pruning is slight. This implies that setting of λ is not sensitive for the results of unit selection.

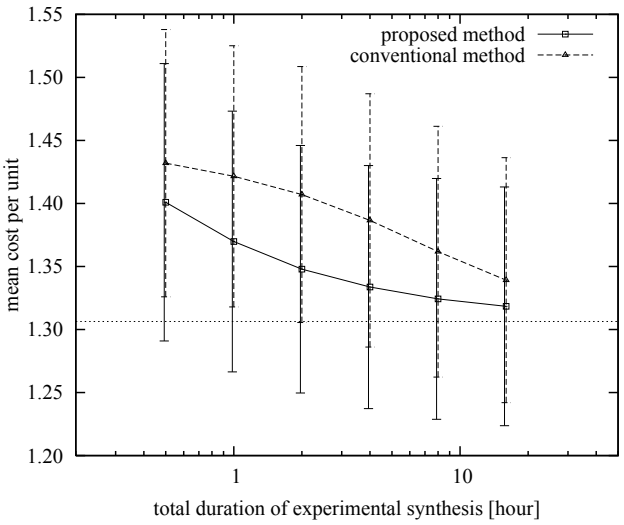


Fig. 3. Mean and standard deviation of cost per unit of the 503 sentences for the pruned 4-hour database. Each error bar represents the standard deviation from the mean value. The dotted line indicates the mean cost per unit for the original 10.6-hour unit database.

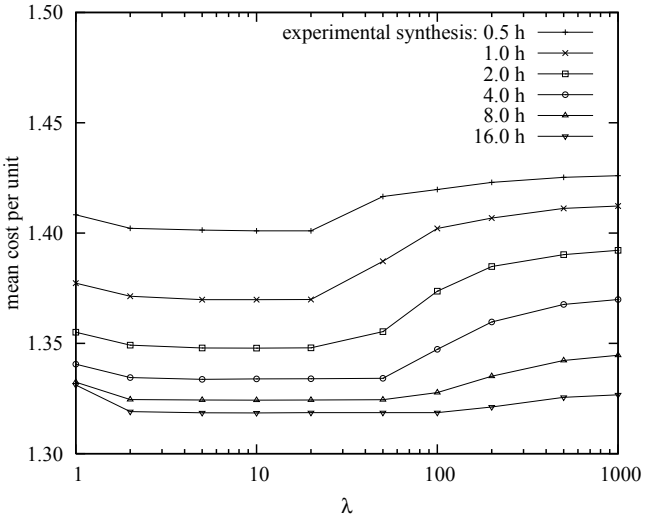


Fig. 4. Mean cost per unit of 503 sentences in various λ settings.

The costs of selected units at various reduction rates were also examined. Figure 5 shows the mean costs where λ in the proposed method was fixed to 10. Where the reduction rate is less than approximately 70%, the proposed method with 4-hour data for pruning is superior to the conventional method with 16-hour data in the cost of selected units. Unlike the conventional method, degradation of cost in the proposed method is slight where the reduction rate is low.

5.3. Discussion

Figure 6 shows the cumulative ratio of selected units by the conventional method of experimental speech synthesis in section 5.1. The figure shows that the selected units are a small part of the units included in the database. Even when 16-hour experimental speech synthesis is conducted, only approximately 23% of the total units are selected.

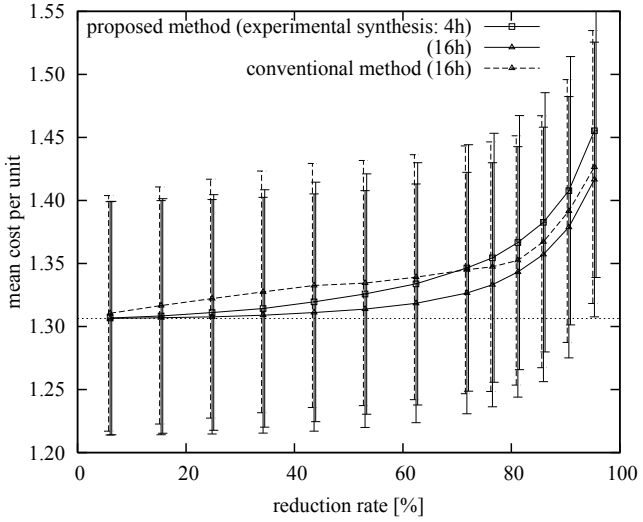


Fig. 5. Mean and standard deviation of cost per unit of the 503 sentences. The error bars and the dotted line are similar to Fig. 3. λ in the proposed method is fixed to 10.

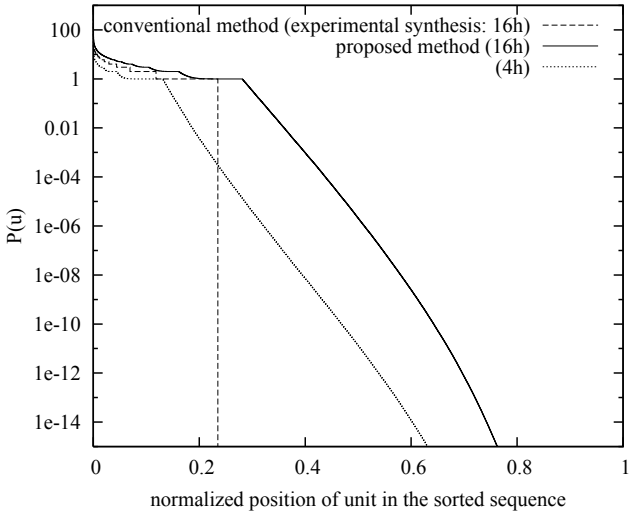


Fig. 7. $P(u)$ that are sorted in descending order. λ in the proposed method is fixed to 10. Normalized position 1 corresponds to the last of the sorted unit sequence.

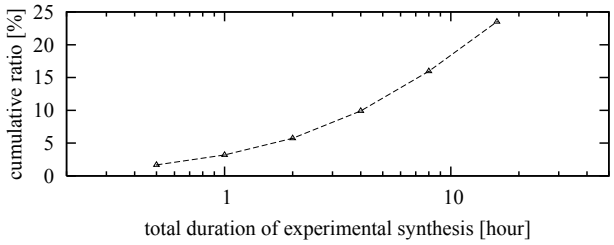


Fig. 6. Cumulative ratio of selected units in units included in the database.

On the other hand, Figure 7 shows the distribution of $P(u)$ that are sorted in descending order. λ in the proposed method is fixed to 10. For the conventional method, $P(u)$ is equal to the frequency of unit u. In the figure, units on horizontally flat parts of the results are randomly pruned in this study. Therefore, the figure demonstrates that most units are not randomly pruned in the proposed method even where the size of experimental speech synthesis is 4 hours. The difference between the result of the conventional method and that of the proposed method especially where the reduction rate is low in Figure 5 implies that units should not be randomly removed.

In the proposed method, not only the optimal units but also the other candidate units are evaluated for pruning. Since $P(u)$ become larger where cost degradation is smaller for more targets, units that are close to many targets on the cost degradation criterion tend to be kept when pruning. In other words, the appearance probability density of targets indirectly affects pruning through the distribution of the cost degradation criterion for all targets.

6. CONCLUSION

To reduce the size of a unit database by small data for pruning, a novel pruning method based on a cost degradation criterion was proposed. Unlike the conventional method based on the frequencies of selected units of the experimental speech synthesis, all candidate units included in the database are evaluated for pruning. To evaluate the proposed method, unit selection experiments were conducted. The results showed that the proposed method can achieve pruning for a 4-hour unit database from the 10.6-hour database by less than 25% of the experimental speech synthesis required for the conventional method at least where the size of the experimental speech synthesis is less than 16 hours. Analysis of the results of speech synthesis by the conventional method shows that units are randomly removed from units that are not tested in the experimental speech synthesis. On the other hand, in the proposed method, since units are removed with the appearance probability density of targets through the distribution of the cost degradation criterion for all targets, more accurate pruning by less experimental speech synthesis is achieved.

7. REFERENCES

[1] Black, A. and Campbell, N., "Optimising selection of units from speech databases for concatenative synthesis," EUROSPEECH '95, vol. 1, pp. 581–584, Madrid, Spain, Sept. 1995.

[2] Black, A. and Taylor, P., "Automatically clustering similar units for unit selection in speech synthesis," Proc. EUROSPEECH '97, vol. 2, pp. 601–604, Rhodes, Greece, Sept. 1997.

[3] Conkie, A., Beutnagel, C. B., Syrdal, A. K., and Brown, P. E., "Preselection of candidate units in a unit selection-based text-to-speech synthesis system," Proc. ICSLP 2000, vol. 3, pp. 314–317, Beijing, China, Oct. 2000.

[4] Nishizawa, N. and Kawai, H., "A preselection method based on cost degradation from the optimal sequence for concatenative speech synthesis," Proc. Interspeech 2007, pp. 2869–2872, Antwarp, Belgium, Aug. 2007.

[5] Kawai, H., Toda, T., Ni, J., Tsuzaki, M., and Tokuda, K., "XIMERA: A New TTS from ATR Based on Corpus-Based Technologies," Proc. 5th ISCA Speech Synthesis Workshop, pp. 179–184, Pittsburgh, Pennsylvania, U.S.A., June 2004.

[6] Toda T., Kawai H., and Tsuzaki, M., "Optimizing Integrated Cost Function for Segment Selection in Concatenative Speech Synthesis Based on Perceptual Evaluations," Proc. EUROSPEECH '03, pp. 297–300, Geneva, Switzerland, Sept. 2003.

[7] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi T., and Kitamura, T., "Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis," Proc. ICASSP 2000, vol.3, pp. 1315–1318, Istanbul, Turkey, June 2000.

[8] Abe, M., Sagisaka, Y., Umeda, T., and Kuwabara, H.. Speech Database User's Manual, ATR Interpreting Telephony Research Laboratories Technical Report, TR-I-0166, Japan, Aug. 1990 (in Japanese).