A Preselection Method Based on Cost Degradation from the Optimal Sequence for Concatenative Speech Synthesis

Nobuyuki Nishizawa and Hisashi Kawai

KDDI R&D Laboratories Inc., Japan

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

10.21437/Interspeech.2007-538

Abstract

A novel unit preselection criterion for concatenative speech synthesis is proposed. To reduce the computational cost for unit selection, units that are unlikely to be selected should be pruned as preselection before Viterbi search. Since the criterion is defined as the difference between the cost of the locally optimal sequence where a unit is fixed and that of the globally optimal sequence, not only the target cost but also the concatenation cost can be taken into account in preselection. For real-time speech synthesis, a preselection method using decision trees, where a unit can be bound to multiple nodes of a tree, is also introduced. Results of a unit selection experiment show that the proposed method using decision trees built from 8-hour training data is superior in the costs of the selected units to the conventional online preselection based on target costs. The experimental results also show that the method is more effective where the computational cost is strongly limited.

1. Introduction

In concatenative speech synthesis, large-scale unit databases are often used since the quality of generated sounds depends on the availability of speech segment units suitable for the targets. The unit suitability is evaluated not only (a) on the target cost, which corresponds to the similarity between a synthesis target and a selected unit, but also (b) on the concatenation cost, which corresponds to smoothness between adjacent units[1]. Consequently, the computational cost of the search for the optimal unit sequence is proportional to the square of the number of unit candidates at each time instant even where Viterbi search is adopted. Since the number of candidates for a unit often amounts to tens of thousands the computational cost becomes impractically large. Therefore, unit candidates should be pruned before Viterbi search is conducted. This process is called *preselection* in this study.

If prosodic (F_0 and duration) modification of speech segments is applied, for example, by the PSOLA (pitch-synchronous overlap and add) method[2], the computational cost for unit selection can be drastically reduced by rough preselection since small differences in selected units from the targets and discontinuities to the adjacent units caused by the preselection can be modified. However, in practice, existing prosody modification methods degrade naturalness to some extent. Therefore, we adopt concatenative speech synthesis without prosodic modification, e.g. CHATR[3], to preserve naturalness of the original segments. In this case, accurate preselection is required.

The target cost is often used as a criterion for preselection[1][4][5] since units far from the target, which have large target costs, are unlikely to be selected in the optimal sequence of units. However, preselection based on the target costs is disadvantageous in many cases because not all units that are close to the targets can be smoothly concatenated to each other. To avoid pruning of truly suitable units, the reduction of computational cost by the preselection is practically limited. For that reason, real-time concatenative speech synthesis without prosodic modification has been often regarded as impractical.

Therefore, we propose a novel criterion where both the target cost and concatenation cost are taken into account. The criterion is based on the difference in cost between the locally optimal sequence of units where the unit is temporarily fixed and the globally optimal sequence that is determined without restriction. Since the computation cost of the above proposed criterion is too large for real-time speech synthesis, decision trees are prepared off-line in order to predict preselected units from targets. A method for preselection using decision trees based on context-clustering of units has already been proposed[4]. However, in this method, possible substitution of context for a target is directly restricted by the structure of the built tree. Since a unit in our synthesizer can be flexibly used for targets in contexts that are different from the context of the unit, building of precise decision trees by the method may be difficult. Therefore, another method based on context-clustering of targets from training data, not units, is introduced. Since a unit can be bound to more than one leaf node of a decision tree built by the method, the tree can flexibly represent the possible substitution of context. While the clustering method is similar to the context-clustering algorithm in HMM-based speech synthesis[6], a different class-division algorithm is used because the algorithm in HMM-based speech synthesis does not directly predict units but predicts the physical quantities of targets.

2. Concatenative speech synthesis

2.1. Cost functions

In concatenative speech synthesis, speech segments are selected from a database so that a criterion, which is often called cost, is minimized. In this study, each of the speech segments in the database is generalized as a unit. In our TTS (text-to-speech) system, which is based on XIMERA[5], the 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:

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

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

The 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). All of these features except for the phonetic environment are predicted by using HMM-based speech synthesis techniques[6]. On the other hand, the 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[7].

2.2. Unit selection algorithm

To find the unit sequence with the minimal cost, a Viterbi search, which is based on the dynamic programming (DP) approach, is employed. Subsequences that are not the local optimum are pruned as soon as possible because they never form the globally optimal sequence. The search is achieved by iterative operations as follows: (1) for each unit at a given time, all of the combinations between each of the sequence hypotheses from the initial target to the previous time and the unit at the given time are evaluated; (2) all of the combinations except for the best one are discarded. Consequently, the computational count of concatenation costs is equal to the sum of the products of the number of units at each time and the number of units at the previous time. In practice, to reduce the computational cost, the beam search method is often adopted as in the following operation: (3) unit sequences with small possibilities are discarded at the given time. In this study, the decision to discard a sequence is controlled by the number of preserved sequences, which is called *beam width*.

Although the beam search is an efficient algorithm, when a large unit database is used, the computational cost may still be too much. For example, if a database of waveform segments is built from a 10-hour corpus, the number of candidate units at a given time can amount to tens of thousands. Therefore, the number of candidate units should be reduced by a preselection technique before the Viterbi search is conducted. In our system, the effect of the preselection is controlled by the number of candidates after this reduction, the upper limit of which is called *preselection width*.

Units connected in the original corpus tend to consist of the optimal sequences for utterances because concatenation of the units does not cause audible discontinuities at the boundary. Therefore, in our system, pairs of units that are connected in the original corpus are always evaluated in the Viterbi search regardless of the result of preselection. By this method, discontinuities caused by inappropriate pruning of units in preselection can be reduced.

3. Preselection based on cost degradation from the optimal sequence

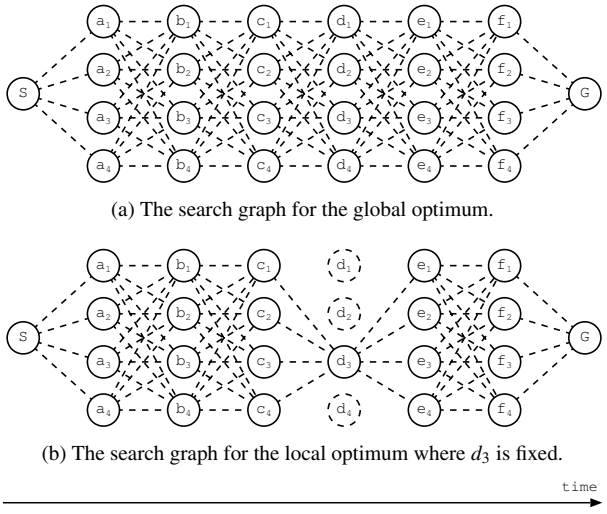
In the conventional preselection, the target cost is used as a criterion because units far from the targets rarely constitute the optimal sequence. However, since it does not take into account of concatenation cost, units that are close to the target but cannot be smoothly concatenated to the adjacent units in a sequence may be kept in preselection. If a preselection method taking into account of concatenation cost is introduced, further reduction of computational cost without degradation will be enabled. In this section, a novel criterion for preselection based on degradation in cost from the optimal sequence is proposed.

3.1. Cost degradation from the optimal sequence

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

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

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



(a) The search graph for the global optimum.

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

Figure 1: 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.

Figure 1 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).

3.2. Forward-backward Viterbi search to compute cost degradation

In Viterbi search, the best scores from the start node to all nodes in the search graph are preserved in the stage of forward search. Similarly, when backward search is performed, the best scores from the goal node to all nodes can be also obtained. Therefore, $D(u|t)$ at each node can be computed by a forward Viterbi search and a backward Viterbi search as in:

$$\begin{aligned}
&\min C(\mathbf{u}_{S\to u\to G}|\mathbf{t}) \\
&\quad = \min C(\mathbf{u}_{S\to u}|t_1,t_2,...,t_i) \\
&\qquad + \min C(\mathbf{u}_{u\to G}|t_i,t_{i+1},...,t_L) - C_T(u|t_i) \\
&\quad = \min C(\mathbf{u}_{S\to u}|t_1,t_2,...,t_i) \\
&\qquad + \min C(\mathbf{u}_{G\to u}|t_i,t_{i+1},...,t_L) - C_T(u|t_i) \qquad (3)
\end{aligned}$$

where the first and second terms of the right-hand side are obtained from the forward and backward search, respectively. Therefore, the computational cost to compute D for all of the units is twice as large as that for conventional unit selection algorithms, which perform only the forward Viterbi search.

4. Building decision trees for preselection

In this study, preselection is conducted by using decision trees whose input and output are target information and a set of units,

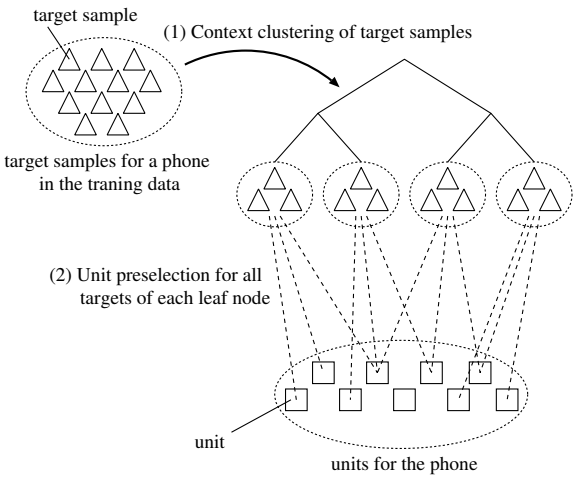


Figure 2: The method for building decision trees for preselection.

respectively. The use of decision trees eliminates the computation for preselection at synthesis time. Therefore, a method based on the top-down context-clustering of targets is introduced. The targets as the training data are generated from sentences that are not contained in the corpora for unit selection.

Figure 2 shows the method schematically. The method consists of two steps: (1) a tree structure is constructed by context-clustering of target samples, (2) preselected units for targets at each leaf node of the tree are bound to the node. A unit can be bound to multiple nodes of a tree. Since a straightforward distance measure between targets for clustering based on the proposed criterion cannot be defined, a class-division algorithm is introduced.

4.1. Warping of degradation values

In the clustering, the difference in D should be examined only for units with small D because unit with large D will pruned by the preselection. Therefore, before building trees, values of D are warped by the following equation:

$$p(u|t) = \exp(-\lambda D(u|t)) \tag{4}$$

If u is the optimal unit in the training data, p is equal to 1. As D becomes larger, u converges to 0.

This warping can be controlled by a parameter λ. As λ becomes smaller, units with larger D also contribute to the clustering. On the other hand, if λ is set to a large value, the preselection method emulates a conventional method based on the appearance frequency of a unit selected for training data.

4.2. Division algorithm for building tree structure

In the method, different trees are constructed for each phone. In the following discussion, T and U denote a set of target samples for a given phone, and a set of candidate units for the phone, respectively.

To build a tree structure, a target set T is recursively divided into two subsets T_1 and T_2 by asking questions regarding the targets. The question that maximizes the sum of the interclass variances of p for all units is selected to divide T. This method assumes that the targets where the values of p are large at the same units resemble each other. The sum of the interclass variances of p between T_1 and T_2 is defined as follows:

$$\sum_{u\in U}\left\{|T_1|(\mu_{u,T_1}-\mu_{u,T})^2+|T_2|(\mu_{u,T_2}-\mu_{u,T})^2\right\} \tag{5}$$

$$\text{where}\quad \mu_{u,T}=\frac{1}{|T|}\sum_{t\in T}p(u|t) \tag{6}$$

where $|T|$ denotes the number of elements in T.

In this study, target information consists of quinphone context, log duration and log F_0 of the unit. The total number of questions is 376 including 248 questions for phonetic environment, 64 questions for phone duration, and 64 questions for F_0.

To avoid overfitting, the division is terminated when the number of targets belonging to a node reaches a preset minimum.

4.3. Unit preselection for each respective subspace

To build decision trees for preselection, sets of units must be bound to all of the leaf nodes as the preselection results. The set of units for each node should be a balanced result for all target samples of the node. Such a preselection can be conducted by the following equations:

$$U_n=\underset{\substack{U'\subseteq U\\ |U'|<K}}{\operatorname{argmax}}\sum_{t\in T_n}\left\{\sum_{u\in U'}p(u|t)\right\}^{\frac{1}{\gamma}} \tag{7}$$

where K and T_n denote the preselection width limit and the target set at tree node n, respectively. For appropriate preselection for multiple target samples, γ must be set greater than 1. In this study, γ is empirically set to 2.

Since computation for obtaining U_n by Equation (7) is extremely complex, U_n is approximately obtained by a greedy search instead of directly solving Equation (7) as $U_n^{(K)}$ defined by the following recurrence equations:

$$u_{n_{i+1}}=\underset{u\in U-U_n^{(i)}}{\operatorname{argmax}}\sum_{t\in T_n}\left\{p(u|t)+\sum_{u'\in U_n^{(i)}}p(u'|t)\right\}^{\frac{1}{\gamma}} \tag{8}$$

$$\text{where}\quad U_n^{(i)}=\{u_{n_1},u_{n_2},...,u_{n_i}\} \tag{9}$$

5. Experiment

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

5.1. Basic configuration of the concatenative speech synthesis system

In the experiment, a TTS system based on XIMERA[5] was used. The unit database for the experiment was built from a Japanese speech corpus of approximately 11.9 hours pronounced by a female speaker. The size of each unit in the database was a half-phoneme for vowels and unvoiced fricatives, or a phoneme for the other consonants.

5.2. Building decision trees

First, cost difference values were computed for 8-hour training data that consisted of 2452, 2347, and 1388 sentences from novels, news, and travel dialogues. Target information of sentences was predicted by using the text processing module of the TTS. Reduced data, viz. 0.5-hour, 1-hour, 2-hour, and 4-hour data were also composed from the full training data for comparison with different conditions in the size of training data.

Secondly, target trees were built for 376 kinds of units contained in the sentences of the 0.5-hour training data. The sizes of trees were controlled by the lower limit of the number of target samples at a node. The limit was empirically set to 4. λ in Equation (4) was set to 0.1, 0.2, 0.5, and 1.0 according to a preliminary experiment.

For comparison, decision trees based on target cost were also built. In this case, degradation of target cost:

$$D_T(u|t) = C_T(u|t) - \min_{u' \in U} C_T(u'|t) \quad (10)$$

was used instead of D. λ was set to 0.01, 0.02, 0.05, and 0.1.

5.3. Unit selection experiment

For targets of unit selection, 476 sentences were selected from ATR's phonetically balanced 503 sentences[8], the pronunciations of which were not included in the corpus for the unit database. Unit selection was conducted for various preselection widths. The beam width of the beam search was fixed to the same value of the preselection width. For comparison, unit selection with the conventional online preselection based on the target cost was also examined.

Figure 3 shows the mean cost per unit for (A) the proposed method (preselection using decision trees based on cost degradation from the optimal sequence), (B) preselection using decision trees based on the target cost, and (C) the conventional online preselection based on the target cost. In the figure, the results of (A) and (B) indicate the minimal costs in various λ settings. In most cases of the proposed method, the optimal setting of λ is 0.2. The results indicate that (A) is superior to (C) in terms of cost when 8-hour data is used for the training of the decision trees. Note that preselection using decision trees is considerably superior in terms of computational complexity to the online preselection. The results of (A) for several sizes of training data suggest that the 8-hour training data is not large enough to converge the performance. Therefore, a larger amount of training data may improve the performance.

If we compare (A) of 8-hour training data to (B), the same cost is achieved by half or a smaller preselection width. For a detailed comparison, Figure 4 shows the differences in costs between (A) and (B). When the same preselection widths are set to (A) and (B), the difference in cost between (A) and (B) becomes larger as the preselection width is set smaller. In other words, the proposed method is more effective where the computation cost for the unit selection is strongly limited.

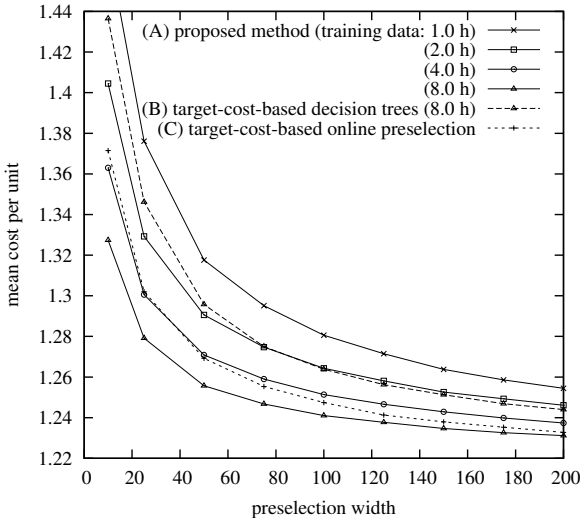


Figure 3: Mean integrated cost per unit of selected units in 476 test sentences by methods with preselection using decision trees and the conventional method.

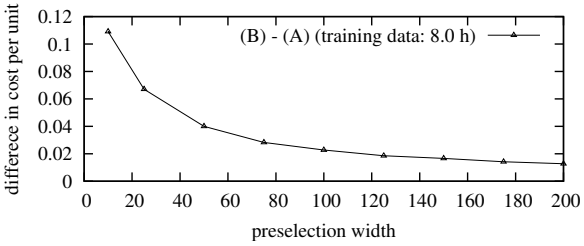


Figure 4: Differences in costs per unit between (A) and (B) of the 8-hour training data in Figure 3.

6. Conclusion

To reduce the computation cost for unit selection, a novel preselection criterion based on cost difference from the optimal sequence was proposed. Since online computation of the proposed criterion is impractical in view of computational complexity, a method for building decision trees for predicting the results of preselection was introduced. To evaluate the proposed methods, a unit selection experiment was conducted. The results showed that the proposed method was superior in terms of cost to the conventional online preselection based on target costs when the decision trees were built from 8-hour training data. In a comparison between preselection methods using decision trees, the results showed that the proposed method was effective especially where the computation cost for the unit selection was strongly limited.

Future work includes training of decision trees from larger training data where a larger unit database is used, and evaluation of computational cost in practical conditions.

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] Moulines, E. and Charpentier, F., "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Communication, vol. 9, pp. 453–467, 1990

[3] Black, A. and Taylor, P., "CHATR: a generic speech synthesis system," Proc. COLING 94, vol. II, pp. 983–986, Kyoto, Japan, Aug. 1994.

[4] 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.

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

[7] 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, Geneva, Switzerland, pp. 297–300, Sept. 2003.

[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).