Pruning of Redundant Synthesis Instances Based on Weighted Vector Quantization

Sanghun Kim*, Youngjik Lee*, and Keikichi Hirose**

*Spoken Language Processing Team, Human Interface Department, Electronics and Telecommunications Research Institute
**Dept. of Frontier Informatics, School of Frontier Sciences, University of Tokyo

ksh@etri.re.kr

Abstract

A new method of pruning redundant synthesis unit instances in a large-scale synthesis database was proposed based on weighted vector quantization (WVQ). WVQ takes relative importance of each instance into account when clustering the similar instances using vector quantization (VQ) technique. The proposed method was compared with two conventional pruning methods through objective and subjective evaluations of synthetic speech quality: one to simply limit maximum number of instance, and the other based on normal VQ-based clustering. For the same reduction rate of instance number, the proposed method showed the best performance. The synthetic speech with reduction rate 50% sounded with no perceptible degradation as compared to that without instance reduction.

1. Introduction

A large corpus-based synthesis method without prosodic modifications, originated in CHATR [1], can generate highly natural synthetic speech. It selects the most appropriate instance out of multiple tri-phone instances for a synthesis unit in terms of minimizing spectral and prosodic distortions and concatenates it to other selected instances. To reduce the necessity of prosodic modification, a large-scale database with instances of various prosodic variations is requested. The quality of obtained synthetic speech is roughly proportional to database size, and necessary size usually reaches to a few hundred mega-byte. The huge database requires a large memory size and slows down the computational speed. Although a set of phoneme balanced sentences has been designed for recording to reduce the computational cost of unit pronunciation, the overlap-and-add process was utilized to find out the best combination of tri-phone instances for a synthesis unit. Furthermore, due to speaker’s condition change, some instances for a tri-phone deviate largely from others of the same tri-phone [2]. They cannot contribute to the synthetic speech quality: sometimes deteriorate. They should be excluded from the database.

Whistler system by Microsoft selects a small number of instances based on HMM matching scores [3]. It was reported that very high concatenating quality was achieved by choosing instances with the highest HMM score. However, they only considered phonetic contexts without taking prosodic contexts into account [4]. Black and Taylor [2] clustered phonetic and prosodic contexts using a decision tree. They pruned synthesis units by discarding 1~4 instances locating furthest from each cluster center. Reduction rates of 20% to 50% were realized without serious degradation in synthetic speech quality. In CHATR, Campbell and Black [5] selected the most from prosodic viewpoint for each unit using VQ clustering technique. The cluster number (i.e. codebook size) was determined according to the number of instances for each unit.

In fact, pruning methods have not been investigated well for synthesis corpus design, which is an important issue for corpus-based speech database. In the current paper, we tackle the pruning problem and propose a new method, where the relative importance of the unit instances is taken account.

2. Synthesis processing

A Korean corpus-based synthesis method has been developed and successfully utilized in text-to-speech conversion system [6]. To realize the major prosodic variations, the phrase boundaries were classified into 4 levels of break strength (no break, minor, medium and major break) as pause length [7][8]. Each synthesis unit instance was classified into one of 4 phrase break strengths depending on where it was locating in a sentence. Since the major prosodic phenomena were reflected in the synthesis units, we can realize highly natural prosody only by selecting tri-phone instances out of the database depending on the following concatenation cost:

\[
C_{c}(u_{i-1}, u_{i}) = \min_{\text{path}} \left\{ C_{j}(u_{i-1}, u_{i}) + \sum_{j} w_{j} C_{j}(u_{i-1}, u_{i}) \right\}
\]

where \( Q \) is the number of sub-cost \( (j) \). \( C_{j} \) denotes concatenation cost of unit \( u_{i-1} \) and \( u_{i} \) (phonetic and prosodic feature). \( w_{j} \) indicates the weight of \( j^{th} \) sub-cost function of \( C_{j} \). As for feature vectors, LPC-based cepstrum coefficients, energy, pitch and phoneme duration were extracted and normalized using Z-score (i.e. \( z = \frac{\mu - x}{\sigma} \)). Viterbi search scheme with minimum accumulated distortion criterion was utilized to find out the best combination of tri-phone instances. The phase mismatches between units at concatenating boundaries may cause perceptible glitches. To cope with this problem, the overlap-and-add process was introduced during concatenation.

3. Database preparation

To construct tri-phone based synthesis database, tri-phone coverage with respect to the phonetic/prosodic contexts should be considered. For the phonetic aspect, there are over fifty thousand tri-phones in Korean: \([\text{V, Ci, #}] * [\text{Cn, Ci, V,} ...\]
The unit selection process determines the best instance sequence by minimizing the accumulated distance within a word or a phrase. Selecting the best instance is usually affected by the preceding and following unit instances. Further, the frequently selected instances are more important and contributive to the synthetic speech quality than other instances. In the pruning process, those facts should be considered. Thus, to reflect the importance of the frequently selected instances, i.e. the relative importance of instances, we propose the weighted vector quantization (WVQ), which considers the relative frequency of selection. The weight \( w_m \) can be obtained in advance by counting the number of occurrence of the selected instance \( \text{freq}_m \) after synthesizing a large text corpus. Then, the weight is directly incorporated into the VQ algorithm. In the current experiment, the unit selection module utilizes the Euclidean distortion measure as concatenating cost. To generate as highly natural synthetic speech as possible, the weight \( w' \) of feature parameters (i.e. cepstrum, pitch, power and duration) was adjusted experimentally. Especially, if the distortion between two units exceeds a given threshold, Viterbi search excludes the path going into that instance. To realize the WVQ algorithm, the Lloyd algorithm [9] was modified and used.

### 4. Weighted vector quantization

The unit selection process determines the best instance sequence by minimizing the accumulated distance within a word or a phrase. Selecting the best instance is usually affected by the preceding and following unit instances. Further, the frequently selected instances are more important and contributive to the synthetic speech quality than other instances. In the pruning process, those facts should be considered. Thus, to reflect the importance of the frequently selected instances, i.e. the relative importance of instances, we propose the weighted vector quantization (WVQ), which considers the relative frequency of selection. The weight \( w_m \) can be obtained in advance by counting the number of occurrence of the selected instance \( \text{freq}_m \) after synthesizing a large text corpus. Then, the weight is directly incorporated into the VQ algorithm. In the current experiment, the unit selection module utilizes the Euclidean distortion measure as concatenating cost. To generate as highly natural synthetic speech as possible, the weight \( w' \) of feature parameters (i.e. cepstrum, pitch, power and duration) was adjusted experimentally. Especially, if the distortion between two units exceeds a given threshold, Viterbi search excludes the path going into that instance. To realize the WVQ algorithm, the Lloyd algorithm [9] was modified and used.

**Weighted VQ algorithm**

Step 1: Choose randomly initial \( N \) codewords \( c_n^{(i)} \) (\( i=0 \)).

Step 2: For each training vector \( x_m \), find the nearest codeword and assign training vector to the corresponding centroid.

\[
Q(x_m) = \arg \min_{c_n^{(i)}} \left\| x_m - c_n^{(i)} \right\|^2 \quad m=1,2,\ldots,M
\]

where \( \left\| \cdot \right\|^2 = e_1^2 + e_2^2 + \ldots + e_K^2 \) and \( M \) is the number of training vectors.

Step 3: Update the centroid vector that the frequency of the selected instances in each cluster should be reflected.

\[
c_n^{(i+1)} = \sum\{Q(x_m)=c_n^{(i)}\} x_m w_m
\quad n=1,2,\ldots,N
\]

\[
\text{freq}_m = \frac{\sum\{Q(x_m)=c_n^{(i)}\} \text{freq}_m}{\text{freq}_m}
\]

where \( \text{freq}_m \) is the frequency of the selected \( m \)th instance.

Step 4: Set \( i=i+1 \) and calculate the average distance.

\[
\text{Dist}^{(i)}(x) = \frac{\sum_{m=1}^{M} \left\| x - Q(x_m) \right\|^2 \times \text{freq}_m}{K \times \sum_{m=1}^{M} \text{freq}_m}
\]

where \( K \) is the dimension of vector.

Step 5: Repeat step 2 ~5 until the decreasing rate of the average distance is less than a given threshold (\( \varepsilon \)).
if($\frac{\text{Dist}(i-1) - \text{Dist}(i)}{\text{Dist}(i-1)} < \epsilon$) Stop; \hspace{1cm} (7)
else go to Step 2.

To verify WVQ algorithm, we compared the result of WVQ with that of VQ. The $freq_m$ was obtained by synthesizing about 20,000 sentences (i.e. textbooks, news, dialogues, scenarios and so on) and utilized for calculating the weight ($w_m$). Fig. 2 shows the distribution of $freq_m$, the frequency of selection, in tri-phone /məd/.

Figure 2. Relative frequency $freq_m$ in tri-phone /məd/

The mismatches of pitch and cepstrum in the synthetic speech are more audibly perceptible than those of duration and power. In this experiment, we extracted training vectors of 12 dimensions (pitch values and 1st to 5th orders of cepstrum coefficients for unit $u_i$ and $u_{i-1}$) and performed the VQ and WVQ clustering process.

Fig. 3 (a) shows all the training vectors and VQ clustering results of tri-phone /məd/ in the pitch-cepstrum (1st coefficient) two-dimensional space. In Fig. (b), WVQ codewords and $freq_m$ of additional vectors, which was calculated by adding small random values to original training vectors, were presented. The densely distributed regions indicate that there exist the frequently selected instances. The result shows that two of VQ codewords are moved to the densely distributed regions due to weighting. It means that the WVQ reflects well the relative importance of instances.

5. Experimental results

We carried out several experiments using three kinds of pruning methods: ‘Limit’ (the number of maximum instances is simply limited), VQ, and WVQ. At present, our synthesis system restricts the maximum number of instances for real-time synthesis. ‘Limit’ is a baseline of our current synthesis system. To evaluate the performance of the pruning methods, we chose 20 test sentences from 589 phonetically balanced sentences. Then, the synthesis process was performed to compute the accumulated concatenating distortion, which was used for objective evaluation.

As shown in Fig. 4, the WVQ shows better performance than the other methods with regard to cepstrum distortion.

Figure 4. Objective evaluation results (Distortion vs. reduction rate): (a) cepstrum, (b) pitch
In pitch distortion, the WVQ and VQ outperform ‘Limit’ but they are roughly the same. A large pitch distortion in ‘Limit’ might be simply caused by the confined prosody due to limited number of available instance. The objective evaluation results show that the two distortions (i.e. cepstrum, pitch) are roughly constant under the 45% reduction rate and start to increase over that rate. It means that the synthetic speech quality can be preserved even if 45% reduction rate is used.

To evaluate subjectively the performance of pruning methods, we also carried out the informal listening test with 45% reduction rate. The test material is the same as that of the above objective measurement. Four people participated in the experiment. All the participants scored a number ranged from 1(worst) to 5(best). In the results shown in Fig. 5-(a), the WVQ is superior to the other pruning methods. In addition, the WVQ doesn’t degrade the synthetic speech quality when compared to the full search (i.e. no-pruning) even if a large reduction rate is used. The ‘Limit’ results in the worst among the three pruning methods. Fig. 5-(b) shows how 4 kinds of synthetic speech (no-pruning, Limit, VQ and WVQ) are ranked. Top 1 means the score is the highest among 4 methods. The WVQ is even slightly higher than no-pruning and VQ with regard to Top 1. The ‘Limit’ shows the worst performance. Moreover, the perceptual experiments shows that the synthetic speech quality was not seriously deteriorated over 50% reduction rates.

6. Conclusion

In this paper, we proposed the weighted VQ pruning method. By considering the relative importance of instances during pruning process, we can efficiently reduce the database size without degrading the synthetic speech quality. The subjective/objective evaluation results showed that the proposed method outperformed the conventional pruning methods in terms of synthetic speech quality in the case of 45% reduction rate. Moreover, the proposed method generates nearly indistinguishable synthetic speech from that of no-pruning. Even over 50% reduction rate, the new pruning method doesn’t seriously deteriorate the synthetic speech quality.

7. References


Figure 5. Subjective evaluation results:
(a) score; (b) number of rank

The computational load mainly comes from the unit selection process, which finds the best instance sequence among lots of combinations of instances. Pruning process reduces the number of combinations, and therefore accelerates the synthesis speed.