Unit Generation Based on Phrase Break Strength and Pruning for Corpus-Based Text-to-Speech

Sanghun Kim, Youngjik Lee, and Keikichi Hirose

This paper discusses two important issues of corpus-based synthesis: synthesis unit generation based on phrase break strength information and pruning redundant synthesis unit instances. First, the new sentence set for recording was designed to make an efficient synthesis database, reflecting the characteristics of the Korean language. To obtain prosodic context sensitive units, we graded major prosodic phrases into 5 distinctive levels according to pause length and then discriminated intra-word triphones using the levels. Using the synthesis unit with phrase break strength information, synthetic speech was generated and evaluated subjectively. Second, a new pruning method based on weighted vector quantization (WVQ) was proposed to eliminate redundant synthesis unit instances from the synthesis database. WVQ takes the relative importance of each instance into account when clustering 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 the maximum number of instances, 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 45% had almost no perceptible degradation as compared to the synthetic speech without instance reduction.

I. INTRODUCTION

Text-to-Speech (TTS) systems based on small speech databases generate highly intelligible synthetic speech. However, the synthetic speech is still significantly unnatural compared to human speech. The major problems arise from insufficient synthesis units and excessive signal processing to realize natural prosody. To overcome these problems, Hunt and Black [1] proposed an optimum unit selection method, which has been applied successfully to the CHATR system. The CHATR system has two significant features that differ from those of the conventional TTS system: it uses a dynamic unit selection algorithm, and it uses multiple unit instances containing natural human prosody. Since it enables us to generate synthetic speech without prosodic modification, it preserves the voice quality and speaking style of the original speaker. The AT&T TTS system [2] has adopted CHATR's unit selection method. The selected units may be either phones or diphones, and they can be synthesized by a variety of methods, including Pitch Synchronous Overlap and Add (PSOLA), Harmonic+Noise Model (HNM), and simple unit concatenation. In such a kind of corpus-based TTS system, a large speech database is necessary to reflect various prosodic and phonetic contexts. Selecting appropriate units among multiple instances minimizes the artificial prosodic processing so that the original speaker's characteristics can be preserved. Based on recent research results, we started developing a new TTS system based on a large speech corpus. To maximize the phonetic/prosodic context coverage of triphones, we designed a sentence set based on the occurrence of Korean triphones. To reduce the necessity of prosodic modification, the synthesis units should be finely discriminated to fit in the target context as closely as possible. Thus, the synthesis unit that was classified depending on the phonetic con-

Manuscript received May 16, 2001; revised Nov. 6, 2001.

Sanghun Kim (phone: +82 42 860 5141, e-mail: ksh@etri.re.kr) and Youngjik Lee (e-mail: ylee@etri.re.kr) are with Speech Information Technology Research Center, ETRI, Daejeon, Korea.

Keikichi Hirose (e-mail: hirose@gavo.t.u-tokyo.ac.jp) is with the Department of Electrical, Electronic, Information and Communication Engineering, Graduate School of Engineering, University of Tokyo, Japan.

text was further discriminated by phrase break strength. This process enables us to realize the major prosodic events of phrase boundary. This paper will describe the discriminating process according to phrase break strength.

The huge database, which usually reaches to a few hundred megabytes, requires a large memory size and slows down the computational speed. Although a set of specialized sentences has been designed for recording to reduce the duplicated instances, similar instances in terms of prosodic and spectral features still remain in the resulting speech database. Furthermore, due to speaker's condition change, some instances for a triphone deviate largely from others of the same triphone. They cannot contribute to the synthetic speech quality: sometimes they deteriorate it. They should be excluded from the synthesis database. Whistler system by Microsoft [3] selects a small number of instances based on Hidden Markov Model (HMM) matching scores. 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. Black and Taylor [4] clustered phonetic and prosodic contexts using a decision tree. They pruned synthesis units by discarding 1~4 instances located 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 diverse instances from the 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 database design, which is an important issue for corpus-based speech synthesis. In this paper, we tried to reduce the database size by pruning redundant synthesis unit instances based on VQ method.

This paper is organized as follows. In Section II, we give an overview of our baseline system. Section III deals with synthesis unit generation: design of recording sentence set, synthesis unit discrimination, and generalization of unseen triphones. Then, the subjective evaluation result will be presented. In Section IV, we propose a new method for pruning redundant synthesis unit instance based on WVQ (Weighted Vector Quantization) and show the evaluation results. Finally, Section V concludes this paper.

II. OVERVIEW OF BASELINE SYSTEM

The synthesis system is composed of three parts: language processing, prosody processing, and signal (unit selection) processing. Language processing performs text filtering, morphological analysis, text preprocessing, and letter-to-sound conversion. The text filtering module filters undesired symbols (i.e., two byte graphic characters, control characters, and nonsense symbols) out of input texts. It converts typical text forms (i.e., date, telephone number, e-mail address, and URL address) into appropriate reading styles.

In prosody processing, an HMM based Part-of-Speech (POS) sequence model [6] was adopted to predict appropriate phrase break strength from input texts. The performance shows 73.5% in 5-level break strength prediction. In unit selection, a Viterbi search scheme with minimum accumulated distortion criterion was utilized to find the best combination of triphone instances. As for feature vectors, LPC-based cepstrum coefficients, energy, pitch and phoneme duration were extracted and normalized using Z-score. The phase mismatches between units at concatenating boundaries may cause perceptible glitches. To cope with this problem, time domain overlap-and-add process was applied while concatenating synthesis units [7]. The synthesis database size reached 600Mbyte~1Gbyte. To reduce the database size, the original speech (16kHz, 16bits) was compressed using waveform coding, i.e., u-law PCM (8kHz, 8bits) and ADPCM (16kHz or 11kHz, 4bits).

III. SYNTHESIS UNIT GENERATION

1. Synthesis Unit

Recent synthesis systems generate natural synthetic speech by concatenating speech segments such as phone, diphone (i.e., units that begin in the middle of the stable state of a phone and end in the middle of the following one), demisyllable, syllable, and word. With respect to synthesis units, concatenating discontinuity and the number of synthesis unit should be considered. Though longer units (i.e., syllable and word) reflect coarticulation effects of phones well, a large number of synthesis units (about 10,000 syllables in Korean) are necessary. In practice, it is difficult to prepare these synthesis units. Hence, most concatenation synthesis systems have adopted sub-syllabic synthesis units such as phone, diphone (about 1,000 diphones in Korean), and demisyllable (about 2000~3000 demisyllables in Korean). In the case of diphones and demisyllables, concatenating discontinuity can be minimized since they are concatenated at the spectrally stable region. Still, the demisyllable has a problem when concatenated at a syllable boundary. In the case of phone-sized synthesis units, it is difficult to cover all the phonetic contextual variations. Because there is likely to be concatenating distortion, the diphone has been widely applied to many synthesis systems because it has a suitable number of units and minimum concatenating distortion. These days, however, a large amount of phonetic contexts can be taken into account. Therefore, a phone-sized unit is advantageous in utilizing more appropriate context sensitive units, which is likely to generate better synthetic speech. It can be consistently segmented using an automatic speech recognizer. Moreover, it is easier to manipulate than other units with respect to context clustering and prosodic processing. Therefore, we selected a phone-sized synthesis unit (i.e., triphone) that reflects the preceding and following phonetic context of a phone.

2. Design of Recording Sentence Set

To construct a triphone based synthesis database, triphone coverage with respect to phonetic and prosodic contexts should be considered. For the phonetic aspect, there are over fifty thousand triphones in Korean: $[\{v, c_i, \#\}+\underline{v}+\{c_i, c_f, v, \#\}]$, $[\{v, c_f, \#\}+\underline{c_i}+\{v\}]$, and $[\{v\}+\underline{c_f}+\{c, \#\}]$, where v, c_i, c_f, and # stand for 21 vowels, 19 syllable initial consonants, 7 syllable final consonants, and silence, respectively. For the prosodic aspect, each triphone should retain enough instances to cover possible prosodic variations appearing in the utterances as the number of necessary triphone instances may reach over several millions in the end. However, it is impractical to get all the necessary triphone instances [8].

In Korean texts, function words (i.e., particles or inflections) play an important role in demarcating the syntactic boundaries, while in Korean speech, major prosodic variations occur mostly in the function words. Thus, it is better to collect more triphone instances found in the function words. The function words are evenly distributed in the well-formed sentences, so we selected a few hundred thousand well-formed sentences from the text corpus that were extracted mainly from middle/high school textbooks and newspapers. When selecting the sentences, the sentence length was limited to 15~20 words to reduce the speaker's effort. To avoid the duplicated units being included in the resulting speech database, the greedy algorithm was applied [6].

The obtained sentence set consists of about 3,600 sentences and contains 14,882 unique triphones. The total number of triphone instances exceeds 410,000. Source utterances are produced by a female announcer during normal text reading. To obtain phoneme segmentation results, we utilized an HMM-based continuous speech recognizer. In order to adapt to the target speaker, the speech recognizer has been calibrated using the target speaker's database. With adapted distribution and codebook weight parameters, the speech recognizer conducted Viterbi alignment to segment an utterance into its corresponding phonetic symbols. Figures 1(a), (c), and (d) show an utterance waveform, its spectrum, and its automatic segmentation results, respectively. To detect the pitch period, we utilized a laryngograph. It converts physical vocal cord activity to electric signals so that the voicing/epoch can be accurately detected without invoking a manual process. Figure 1(b) shows the laryngograph signal.

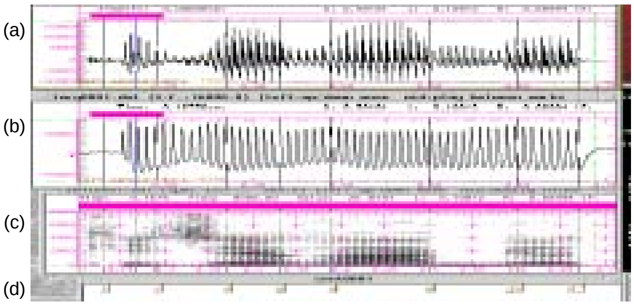

Fig. 1. The automatic phoneme segmentation result: (a) speech signal, (b) laryngograph signal, (c) spectrum, and (d) phoneme labels.

Although the phoneme boundary was located automatically, we still needed a manual process for correcting severe segmentation errors such as "vowels+vowels", "vowels+final consonants (or coda)", etc. The synthesis unit database was constructed so that synthesis unit cost becomes zero when the two adjacent synthesis units are consecutively placed in the source utterances. By doing so, the spectral mismatches can be reduced as much as possible.

3. Synthesis Unit Discrimination

In corpus-based synthesis, synthesis units should be as finely discriminated as possible in terms of phonetic and prosodic contexts. To accomplish this, first, each phone was discriminated depending on preceding and following phonetic context of a phone. Second, the phone was further discriminated depending on phrase break strength, which can reflect major prosodic events. This section describes the discriminating process of triphone to reflect prosodic context.

In general, an utterance is prosodically grouped with several chunks of unit. These units, called prosodic phrases, are formed by the syntactic and semantic structure of an utterance. Prosodic phrase is usually signaled by significant prosodic events such as preboundary lengthening, intonation falling/rising, and pause insertion. It occurs in several levels corresponding to the perceptually distinct phrase break strength. Price [9] assigned a 7-level break index to express the degree of decoupling between words. In Tone and Break Indices (ToBI) system proposed by Beckman [10], there is a 5-level break index. It merged Price's break indices 4~6 into the same break strength. In K-ToBI (Korean ToBI version 3.1) system proposed by Jun [11], there are 4 different levels: 0 for clitic group, 1 for phrase internal word boundary, 2 for accentual phrase, and 3 for intonation phrase. To grade the phrase break strength, Wightman [12] utilized the segmental duration in the vicinity of prosodic phrase boundary. In the results, 7-level break strength could be statistically re-graded to 4 levels (i.e., 0~1, 2, 3, 4~6) and also the durational acoustic cue was useful for grading minor prosodic phrase such as break indices 0~3. On the other hand, intonation and pause were effective to grade major prosodic phrases (i.e., break indices 4~6).

In practice, it is not easy to grade distinctive levels of minor prosodic phrases (i.e., break indices 0~3). Moreover, intra-word triphones in minor prosodic phrases are phonetically influenced by each other. In this case, we have to consider not only the break strength but also the phonetic context of the following word. These triphones are scarcely used in the synthesis process. Thus, discriminating minor prosodic phrases may not contribute to synthetic speech quality. For that reason, we have focused on the major prosodic phrases (i.e., break indices 4~6). The minor prosodic phrases (i.e., break indices 0~3) were assigned to break index 1. As explained previously, the major prosodic phrase boundaries are marked by complicated prosodic events. In the K-ToBI system, break indices 2 and 3 are discriminated according to whether a strong subjective sense of pause is perceived or not. The pause must be a simple and decisive feature among prosodic features. On examining the data, the pause length was utilized to grade the major prosodic phrase. Even if no silence was found between words, we put pause marks if there was a perceptible break and no strong assimilation across the words. The discriminating process is straightforward. First, we investigated pause length distribution in all the utterances to determine the degree of decoupling words. As shown in Fig. 2, which is a histogram of pause length, the pause length can be divided into three parts: short, medium, and long pause. We marked distinctive levels of major prosodic phrase depending on it. Each part was assigned to break indices 2~4, respectively.

The resulting break index consists of 5 levels, i.e., 0 for word internal triphone, 1 for word juncture, 2 for short pause, 3 for a little longer pause than break index 2, and 4 for long pause. The strong phrase break strength (i.e., break index 4) is allocated to the end of sentence or clause. Finally, triphones at word boundaries were discriminated depending on the phrase break strength. Furthermore, Wightman said that segmental lengthening in the vicinity of prosodic boundaries is found to be restricted to the rhyme of the syllable preceding the phrase boundary. On examining the data, we also discriminated the vowel of final syllable rhyme.

For example, the Korean word "학생" ("student" in English) is converted into phonetic symbols /#-h-a-K-S-E-O-#/ (IPA code corresponding Roman characters is shown in Appendix). Symbol '#' indicates the pause. Each phone is discriminated by its preceding and following phonetic contexts. Then, the word "학생" consists of 6 triphones: /${}_{\#}h_{a}$/, /${}_{h}a_{K}$/, /${}_{a}K_{S}$/, /${}_{K}S_{E}$/, /${}_{S}E_{O}$/, and /${}_{E}O_{\#}$/. The subscript indicates the phonetic context. To reflect the prosodic context, the word initial (i.e., /${}_{\#}h_{a}$/) and the word final triphone (i.e., /${}_{E}O_{\#}$/) are further discriminated as the phrase break strength. If the break strength of the pre-word and post-word boundary is 3 and 4, they become ${}_{\#3}h_{1a}$ and ${}_{E0}O_{4\#}$, respectively. The triphones of an internal word have break strength 0. The syllable rhyme triphones /${}_{S}E_{O}$/ reflecting Wightman's scheme becomes /${}_{S0}E_{1O}$/. The right break index of the phone /E/ is not 0 but 1. By doing so, the prosodic variations due to the phrase boundaries can be reflected. Finally, the prosodic context sensitive triphones become /${}_{\#3}h_{0a}$/, /${}_{h0}a_{0K}$/, /${}_{a0}K_{0S}$/, /${}_{K0}S_{0E}$/, /${}_{S0}E_{1O}$/, and /${}_{E0}O_{4\#}$/.

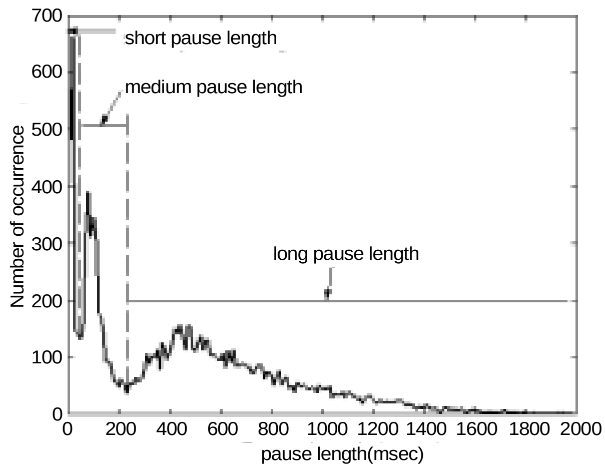

Fig. 2. Histogram of pause length.

4. Generalization of Unseen Triphones

While synthesizing unlimited texts, an unseen triphone that doesn't exist in the synthesis database may occur. Usually, over 1% of unseen triphones occurred when a large amount of sentences was synthesized. It means that an unseen triphone occurs once for every sentence of 15~20 word length. It causes spectral mismatch at the concatenating point, and has a bad effect on the synthetic speech quality. Therefore, it should be substituted for the proper triphone with a similar phonetic context. This is the generalization process of unseen triphones. Similar phonetic context means that the acoustic realization of the phonetic context is similar with that of the desired triphone. To generalize unseen triphones, we defined the similar phonetic context of all phones based on human expert knowledge.

The generalization process is conducted as follows. First, the phonetic context of the unseen triphone is replaced by the similar phonetic context. Then, we try to find the new triphone if it exists in the synthesis database. Second, if there is no matched triphone in the synthesis database, the broad phonetic contexts (i.e., voiced/unvoiced, tense/lax, aspirated/unaspirated, rounded/unrounded, etc) are applied [6]. If an unseen triphone differs in phrase break strength, it is replaced with another break strength. For example, the unseen triphone $/_{\#4}g_{1a}/$ may be substituted for $/_{\#3}g_{1a}/$ or $/_{\#2}g_{1a}/$. In the result, the synthesis database is able to cover about 285,000 triphones: 28,000 for exactly matched triphones and 257,000 for similarly matched triphones.

5. Subjective Evaluation

To evaluate the synthesis system, an informal listening test was performed. In the experiment, 12 people participated and received scores ranging from 1 (worst) to 5 (best). We selected 50 sentences from 589 phonetically balanced sentences (PBS), which was extracted from a million words text corpus based on entropy maximization. The synthetic speech was played twice with a 5 second interval. The resulting score was 3.0 (minimum=2.62, maximum=3.42). The proposed TTS system was compared with the old one that was based on demi-syllable units [13]. The number of demi-syllable units was 1228. The old TTS system used the Fujisaki model for intonation and the Klatt model for duration. Most participants preferred the synthetic speech of the proposed TTS system to that of the old TTS system. Although the new synthesis system still generates some unstable sound, it creates human-like synthetic speech without prosodic modification.

IV. PRUNING REDUNDANT SYNTHESIS UNIT INSTANCE

1. Weighted VQ

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. Furthermore, 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 occurrences of the selected instance ($freq_m$) 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^c_j) of feature parameters (i.e., cepstrum, pitch, power and duration) was adjusted experimentally [14]. 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 [15] was modified and used.

In this experiment, we extracted 12 dimensional training feature vectors (x_m): 2 pitch values (begin and end of phone), 5 orders of cepstrum coefficients for unit u_i (extracted from left and right phone boundaries). Since the mismatches of duration and power in the synthetic speech are less audibly perceptible than those of pitch and cepstrum. The duration and power were excluded from training feature vectors.

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.

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

where $\|e\|^2 = e_1^2 + e_2^2 + \ldots + e_n^2$ and M is the number of training vectors.

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

$$c_n^{(i+1)} = \sum_{Q(x_m)=c_n^{(i)}} x_m w_m, \quad n=1,2,\ldots,N \qquad (2)$$

$$w_m = \frac{freq_m}{\sum_{Q(x_m)=c_n^{(i)}} freq_m} \qquad (3)$$

where $freq_m$ is the frequency of the selected m^{th} instance.

Step 4: Set $i=i+1$ and calculate the average distance.

$$Dist^{(i)} = \frac{\sum_{m=1}^{M} \left\| x_m - Q(x_m) \right\|^2 \times freq_m}{K \times \sum_{m=1}^{M} freq_m} \qquad (4)$$

where K is the dimension of the vector.

Step 5: Repeat steps 2 ~5 until the decreasing rate of the average distance is less than a given threshold (ε).

$$\text{if } \left(\frac{Dist^{(i-1)} - Dist^{(i)}}{Dist^{(i-1)}} < \varepsilon \right) \text{ Stop;} \qquad (5)$$

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). Figure 3(a) shows the distribution of $freq_m$, the frequency of selection, in the triphone $/_{m0}\mathbf{o}_{0d}/$.

Figure 3(b) shows all the training vectors and VQ clustering results of triphone $/_{m0}\ \mathbf{o}_{0d}/$ in the pitch-cepstrum (1st coefficient) two-dimensional space. In Fig. 3(c), 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 frequently selected instances. The result shows that two of the VQ codewords are moved to the densely distributed regions due to weighting, indicating that the WVQ reflects well the relative importance of instances.

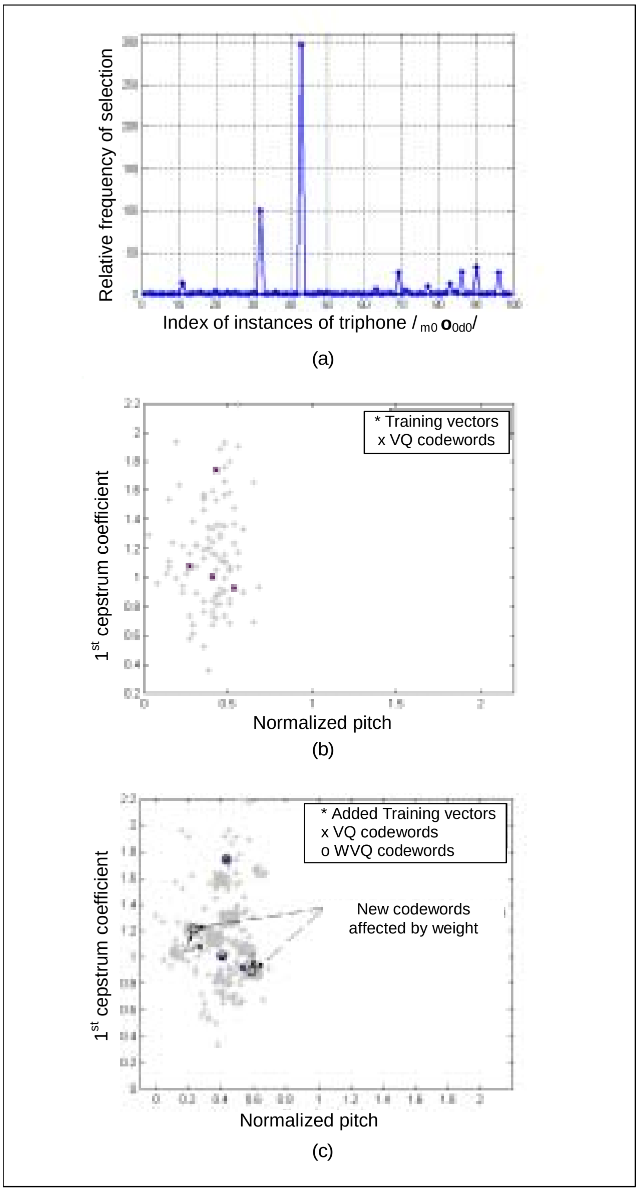

Fig. 3. VQ/WVQ clustering result of triphone $/_{m0}\mathbf{o}_{0d}/$: (a) Relative frequency $freq_m$ in triphone $/_{m0}\ \mathbf{o}_{0d}/$, (b) VQ, (c) WVQ.

2. 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.

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 and pitch) are roughly constant under the 45% reduction rate and start to in-

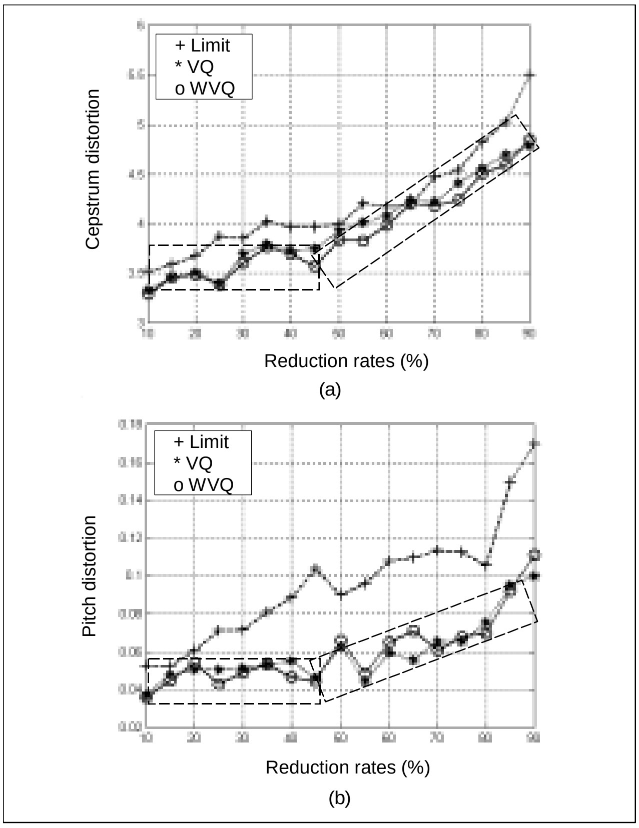

Fig. 4. Objective evaluation results (Distortion vs. reduction rate): (a) cepstrum distortion, (b) pitch distortion.

crease over that rate. It means that the synthetic speech quality can be preserved even if 45% reduction rate is used. Reduction rate is the degree of pruning of the synthesis database.

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 ranging from 1 (worst) to 5 (best). In the results shown in Fig. 5, 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.

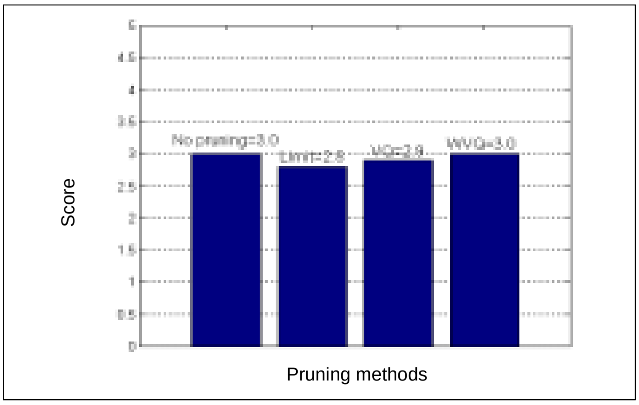

Fig. 5. Subjective evaluation result.

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

V. CONCLUSION

In this paper, we discussed two important issues: synthesis unit generation and pruning redundant synthesis unit instances. To make an efficient synthesis database, the new sentence set for recording was designed. It reflected the characteristics of the Korean language. The synthesis unit that was classified depending on the phonetic context was further discriminated by phrase break strength. This process enables us to realize the major prosodic events of phrase boundary. In the case of unseen triphones, we defined the similar phonetic context of all phones based on human expert knowledge. By doing so, we could substitute it for a similar triphone as closely as possible. To prune redundant synthesis unit instances, we proposed the weighted VQ pruning method. By considering the relative importance of instances during pruning process, we could efficiently reduce the synthesis 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.

REFERENCES

[1] A. Hunt and A.W. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," *Proc. of the Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, 1996, pp. 373-376.

[2] M. Beutnagel, A. Conkie, and A. Syrdal, "Diphone Synthesis Using Unit Selection," *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia:ESCA, 1998, Paper F.2(R5t2).

[3] H. Hon, A. Acero, X. Huang, J. Liu, and M. Plumpe, "Automatic Generation of Synthesis Units for Trainable Text-to-Speech Systems," *Proc. of the Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, 1998, pp. 293-296.

[4] A.W. Black and P.A. Taylor, "Automatically Clustering Similar Units for Units Selection in Speech Synthesis," *Proc. of Eurospeech97*, vol. 2, 1997, pp. 601-604.

[5] N. Campbell, and A.W. Black, "Prosody and Selection of Source Units for Concatenative Synthesis," *A Collection of Technical Publications*, ATR-ITL, 1996, pp. 45-58.

[6] S.H. Kim, Y.J. Lee, and K. Hirose, "A New Korean Corpus-Based Text-to-Speech System," to be published on *Int'l J. of Speech Technology*.

[7] E. Moulines and F. Charpentier, "Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones," *Speech Comm.*, vol. 9, 1990, pp. 453-467.

[8] D. H. Yang, "An Algorithm for Predicting the Relation between Linguistic Items and Corpus Sizes," *ETRI J.*, vol. 22, no. 2, June 2000, pp.20-31.

[9] P.J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The Use of Prosody in Syntactic Disambiguation," *J. Acoust. Soc. America*, vol. 90, 1991, pp. 2956-2970.

[10] K. Silverman, M. Beckman, J. Pierrhumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg, "ToBI: a Standard Scheme for Labeling Prosody," ICSLP, Oct. 1992, pp. 867-879.

[11] M. Beckman and S.A. Jun, "K-ToBI(Korean ToBI) Labelling Conventions," *Speech Sciences*, vol. 7, no. 1, 2000, pp. 143-170.

[12] C.W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P.J. Price, "Segmental Durations in the Vicinity of Prosodic Phrase Boundaries," *J. Acoust. Soc. America*, vol. 91, no.3, Mar. 1992, pp. 1707-1717.

[13] S.H. Kim and J.C. Lee, "Korean Text-to-Speech System Using TD-PSOLA," *Australian Int'l Conf. on Speech Science and Technology (SST'94) Proc.*, Perth, Australia, 1994, pp. 587-592.

[14] I.S. Lee, "Encoding of Speech Spectral Parameters Using Adaptive Quantization Range Method," *ETRI J.*, vol. 23, no.1, Mar. 2001, pp.16-22.

[15] R.M. Gray, "Vector Quantization," *IEEE ASSP Magazine*, 1984, pp. 4-28.

Sanghun Kim received the B.S. degree in electrical engineering from Yonsei University, Seoul, Korea in 1990 and the M.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, Korea in 1992. Since 1992, he has been with Research Department and Spoken Language Processing Section of ETRI, Daejeon, Korea. Currently, he is a Senior Researcher in Speech Database Development Team, Speech Information Technology Research Center of ETRI. His interests include speech synthesis, speech recognition, and speech signal processing.

Youngjik Lee received the B.S. degree in electronics engineering from Seoul National University, Seoul, Korea in 1979, the M.S. degree in electrical engineering from Korea Advanced Institute of Science, Seoul, Korea in 1981, and the Ph.D. degree in electrical engineering from Polytechnic University, Brookly, New York, USA. From 1981 to 1985, he was with Samsung Electronics Company, Suwon, Korea where he was involved in the development of video display terminal. From 1985 to 1988, his research topic was concentrated on the theories and applications of sensor array signal processing. Since 1989, he has been with Research Department and Spoken Language Processing Section of ETRI, Daejeon, Korea pursuing interests in theories, implementations, and applications of spoken language translation, speech recognition and synthesis, and neural network.

Keikichi Hirose received the B.E. degree in electrical engineering in 1972, and the M.E. and Ph.D. degrees in electronic engineering, respectively in 1974 and 1977 from the University of Tokyo. From 1977, he is a faculty member at the University of Tokyo: professor of the Department of Electronic Engineering from 1994 and Professor of the graduate course for Information and Communication Engineering from April 1995. From March 1987 until January 1988, he was a Visiting Scientist of the Research Laboratory of Electronics at the Massachusetts Institute of Technology. Although his research interests widely cover the field of speech information processing, such as analysis, synthesis, perception, and recognition, he has major interest on prosody. He is a member of the Institute of Electrical and Electronics Engineers, the Acoustical Society of America, the European Speech Communication Association, the Institute of Electronics, Information and Communication Engineers, the acoustical Society of Japan, the Japan Society of Applied Physics, the Information Processing Society of Japan, and the Japanese Society for Artificial Intelligence.

Appendix

IPA and Roman characters corresponding to Korean characters:

(a) syllable initial consonants

Korean	IPA	Roman
ㄱ	g	g
ㄴ	n	n
ㄷ	d	d
ㄹ	r	r
ㅁ	m	m
ㅂ	b	b
ㅅ	s	s
ㅈ	z	z
ㅊ	ç	c
ㅋ	k^h	k
ㅌ	t^h	t
ㅍ	p^h	p
ㅎ	h	h
ㄲ	g⌝	G
ㄸ	d⌝	D
ㅃ	b⌝	B
ㅆ	s⌝	S
ㅉ	z⌝	Z

(b) vowels

Korean	IPA	Roman
ㅏ	a	a
ㅓ	ʌ	v
ㅗ	o	o
ㅜ	u	u
ㅡ	ɯ	U
ㅣ	i	i
ㅔ	e	e
ㅐ	ɛ	E
ㅑ	ja	ja
ㅕ	jʌ	jv
ㅛ	jo	jo
ㅠ	ju	ju
ㅖ	je	je
ㅒ	jɛ	jE
ㅢ	ɯi	wi
ㅘ	wa	wa
ㅝ	wʌ	wv
ㅚ	we	we
ㅟ	wi	wi
ㅙ	wɛ	wE
ㅞ	we	we

(c) syllable final consonants

Korean	IPA	Roman
ㄱ	g	K
ㄴ	n	N
ㄷ	d	T
ㄹ	l	L
ㅁ	m	M
ㅂ	b	P
ㅇ	ŋ	O