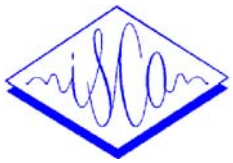

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

Efficient Method of Establishing Words Tone Dictionary for Korean TTS system

Seong-hwan Kim and Jin-young Kim
DSP Laboratory
Dept. of Electronics Engineering
Chonnam National University, 500-757 Kwangju, South Korea
Tel. +82 62 267 0595, Fax: +82 62 514 6472.
E-mail : kseong@dsp.chonnam.ac.kr, kimjin@dsp.chonnam.ac.kr

ABSTRACT

In this paper, we propose an efficient method to establish Word Tone Dictionary(WTD). Vector qantization(VQ) is applied for compressing word tones for compressing word tones, and a phonetic-syntactic distance is adopted for searching the word tone dictionary. Because word tone is a sequence of syllable tones, VQ is used in encoding the syllable tones. As word tones in utterances are specified by their syntactic roles and phonetic features, we propose an adequate distance function to search the appropriate word tone in WTD. It is a combined distance function of syntactic distance and phonetic distance. We tested on a 100-utterance corpus. Preliminary experiments showed that the proposed method could lead to the natural pitch-controlled speech.

1. INTRODUCTION

An adequate prosody control, especially intonation control, is very important for synthetic speech to be natural[1-2]. An intonation pattern can be decomposed into global tones and segmental tones. The global tone is determined by the syntactic structure of the utterance and the segmental tone is represented as a sequence of word tones.

According to the recent researches on the Korean prosody based on large speech corpus, it is observed that there are no fixed word tone patterns. Word tones vary dynamically depending on a variety of phonetic conditions and syntactic roles of words in utterances[3-4]. So it is very difficult to implement word tones appropriately in Korean TTS. Other researchers have adopted neural networks(NN) to get over this problem[5]. As NN method calculate the averaged pitch per syllable, it is inappropriate implementation of word tones varying considerably.

In this study we propose Word Tone Dictionary method for implementing segmental tones. However, the proposed method have two problem, one is to reduce the size of dictionary and the other is to search the dictionary. To overcome these problems, we devise an efficient method to implement WTD. This method adopts the vector quantization and a phonetic-syntactic distance function(PSDF). VQ is introduced for compressing word tones and PSDF is used for searching the dictionary.

2. WORD TONE DICTIONAYR METHOD

In this section, we explain how the corpus is established and how WTD is implemented with VQ and PSDF.

2.1 The corpus

We establish the speech corpus covering the Korean syntactic structures well. 100-utterances were spoken by a female speaker at normal speed. For each utterance, the pitch contour and the syntactic structure were extracted manually. Also, the corpus was labeled with the unit of syllable. The Fujisaki's intonation modeling method is used for analyzing syntactic structure[6]. That is, the right-dependent(RD) parameters was used to represent the depth in the constituent tree. Of course, the syntactic role(a part of speech) was noted. The example of syntactic analysis was shown in the figure 1.

2.2 Structure of Word Tone Dictionary

As explained briefly in the section 1, there is no word-level accent in Korean. That is, primary and secondary accents in English don't exist in Korean. For the given word, its word tone is not fixed. As word tones of two words having similar phonetic and syntactic features are close to each other in the aspect of the pitch pattern, Korean linguists say that the Korean word tone is

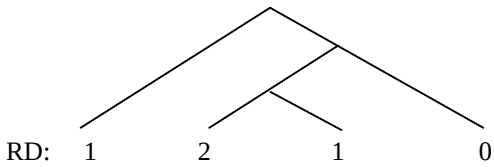

Fig. 1. Example of syntactic analysis.

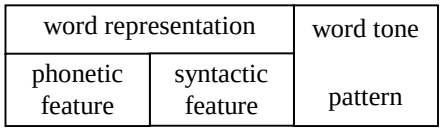

Fig. 2. Structure of WTD.

determined by phonetic features and syntactic roles in the sentence. Thus, a word in the WTD is represented by the phonetic features and syntactic roles. These features are registered in WTD instead of the word itself. Also, the word tone pattern is registered in the WTD. So, the structure of WTD is composed of word features and tone patterns as shown in figure 2.

2.3 Compressing the syllable tones by VQ

A word tone is a sequence of syllable tones. However, the length of syllables in the utterances is not fixed respectively, neither does the length of words tone. Therefore, the length of syllables has to be normalized to compress syllable tones with VQ, and interpolation method can be used for normalization. We thought of the normalized syllable tone as the N-order vector. That is, normalized syllable tone is represented by N-samples. So does VQ codebook .

Fig. 3 shows the procedure compressing syllable tones. After all of the syllable tone vectors is gathered, the codebook is trained by a clustering algorithm. LBG-algorithm is used in our study[7]. The syllable tone vectors in the above processing are gathered independently on the number of syllables in word. So there is only one codebook regardless of the number of syllables in words. The encoding process of syllable tones is represented in the figure 4. Normalized word tones are described by a sequence of codebook index.

2.4 Word representation in WTD

As explained in the section 2.2, a word is represented by its phonetic and syntactic features in the WTD. We used such features as articulation place, articulation method and intensity. In the case of consonants the following

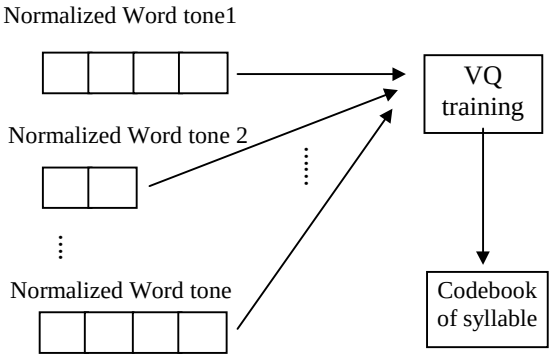

Fig. 3. VQ traing of syllable tones.

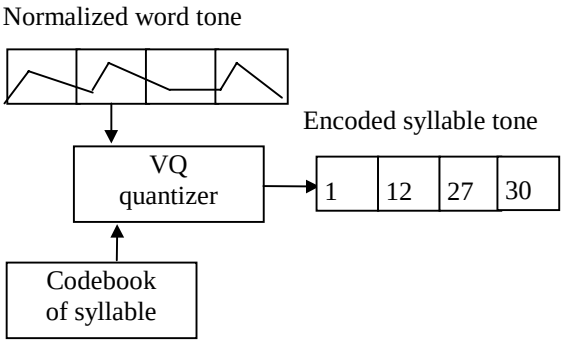

Fig. 4. Codebook of normalized word tone.

feature parameters are used.

- CF1(Articulation place) : palatal, dental, labial, alveolar, glottal
- CF2(Articulation method) : plosive, nasal, lateral, fricative, aspiration
- CF3(Intensity) : lenis, affricate, tense

Also, we assume that vowels are characterized by the following parameters.

- VF1(Front/Back) : front, center1, center2, back
- VF2(High/Low) : high, low, middle
- VF3(Roundness) : round, not_round
- VF4(diphthong) : not_glide, yglide, wglide

Using the above features, we can represent each word in the point of articulation features. For example, ‘cham’ in the sentence referenced in the figure 1 is represented as table 1.

On the other hand, words in sentences are basically characterized by RD in syntactic trees and their syntactic role. The syntactic role means a part of speech. Also, we used the syntactic pause as one of syntactic features, as prior and posterior syntactic pauses affect word tones. Thus, the following four parameters are adopted to represent syntactic features.

- TF1 : forward right dependent
- TF2 : backward right dependent
- SF1 : part of speech
- SF2 : not_pause, pause

Table 1. Phonetic features of ‘cham’.

‘ch’	CF1 = ALVEOLAR CF2 = FRICATIVE CF3 = LENIS
‘a’	VF1= CENTER2 VF2 = LOW VF3 = not_ROUND VF4 = not_GLIDE
‘m’	CF1 = LABIAL CF2 = NASAL CF3 = LENIS

2.5 Phonetic-Syntactic distance function

When we want to synthesize a given word, the matched word tone is searched in the established WTD. Thus, the searching criteria must be defined to select the adequate word tone. In other words, the distance function must be given. By the way, phonetic and syntactic features represent a word registered in WTD. Hence, we propose a combined distance function of phonetic distance and syntactic distances function as follows.

Distance(iWord, rWord)
=(1-*x*)Phonetic_Distance(iWord, rWord)
+ *x* Syntactic_Distance(iWword, rWord),

where *x* is the weighting factor, iWord is a input word and rWord is a registered word. In the above equation. each of phonetic and syntactic distance functions is calculated by the phonetic feature variables(PFV) and the syntactic feature variables(SFV) defined in section 2.4. CF1, CF2, CF3, VF1, VF2, VF3 and VF3 are the PFVs. And, TF1, TF2, SF1 and SF2 are SFVs. Using the PFVs and SFVs, the each distance function is calculated as follows.

$$\text{Phonetic\_Distance} = \sum \delta(\text{CFi}_{\text{iWord}} - \text{CFi}_{\text{rWord}}) + \sum \delta(\text{VFi}_{\text{iWord}} - \text{VFi}_{\text{rWord}})$$

$$\text{Syntactic\_Distance} = \sum \delta(\text{SFi}_{\text{iWord}} - \text{SFi}_{\text{rWord}}) + \sum \left| TFi_{iWord} - TFi_{rWord} \right|$$

, where $\delta(x)$ is the delta function.

3. EXPERIMENTS AND EVALUATIONS

In this section, We explain how WTD is implemented and practically applied for. We have performed experiments to determine the vector dimension of codewords, codebook size and the weighting factor in the proposed distance function. Of course, the quality of pitch-controlled speech is evaluated.

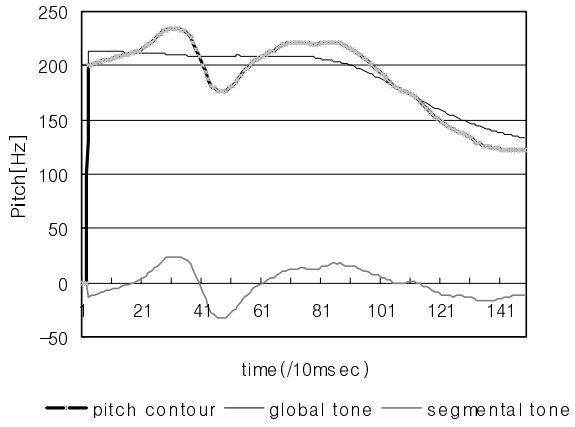

Fig. 5. Examples of the analysis results of intonation.

3.1 Examples of intonation decomposition

The intonation is composed of global tones and the segmental tones. In our experiment, the global tone is obtained with the 101-points hamming window LPF followed by median filter. The original pitch is sampled with 100 Hz sampling rates. And the Segmental tone is the difference of the original pitch contour and the obtained global tone. Fig. 5 shows the analysis results of the utterance "Oneulen cham aleumdaun nalimnida".

3.2 Vector dimension and codebook size

In these experiments, it is assumed that the global tone is well tuned by Fujisaki's filter and does not affected to the segmental tone. Thus, the naturalness of synthetic speech is dependent absolutely on segmental tones. By the way, the distortion of synthetic segmental tone is dependent on the vector dimension and codebook size. If the proposed method produce the synthetic speech with the least computation and memory size, We can say that the method is very efficient. Therefore, the determination of vector dimension and codebook size is very important. We have performed experiments to determine the appropriate vector dimension and efficient codebook size.

Firstly, we measure the average distortion per syllable versus the vector dimension, which is the mean squared errors between the original syllable tone and the reproduced syllable tone. The segmental tone computed by the linear piecewise interpolation of the selected codewords. Figure 6 shows the average distortion per syllable versus the vector dimension. The distortion is calculated with the vector dimensions from 2 to 10. From the results, it is observed that the distortion is saturated at the point of the vector dimension 6.

Secondly, we calculated the distortion with the codebook size from 2 to 256. The average distortion per syllable is displayed in figure 7. As in the figure, the average distortion per syllable is reduced inversely to the codebook size. By the way the distortion of subtraction the distortion at codebook size 128 from distortion at coebook size 256 is about 0.5 Hz. As the frequency of the difference is very small, that is be ignoble.

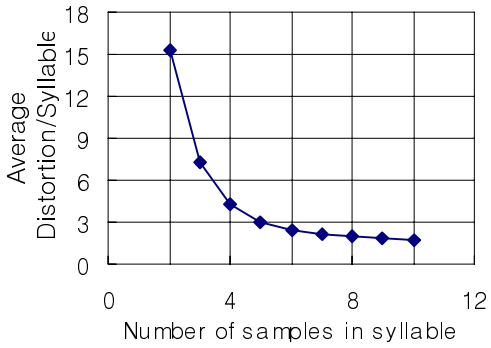

Fig. 6. Plot of average distortion/syllable versus vector dimension

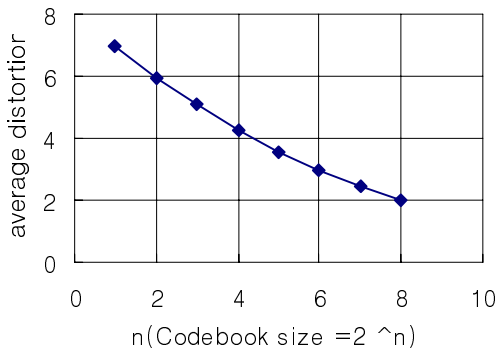

Fig. 7. Plot of average distortion versus codebook size.

We could confirm that the quality of synthetic speech was almost the same through the listening test regardless of the vector dimension if it is greater than 128.

3.3 Distance function and evaluation

We performed, under condition that vector dimension and codebook size is set up to the above results, the MOS test to the synthetic speech with varying the weighting factor. The quality range is limited from 1 to 5(1:bad, 2:poor, 3:fair, 4:good, 5: excellent). In this evaluation, the materials are synthesized with varying the weighting factor. TD-PSOLA is used to synthesize the pitch-controlled speech.

Figure 8 shows the naturalness score of synthetic speech versus various weighting factors. From the figure, it is observed that the quality of synthetic speech is very good at the weighting factor 0.5. That is, as the syntactic and phonological distance is considered equally, the naturalness of synthetic speech is high. Figure 9 shows the contours of the synthetic speech and natural utterance. As in the figure, the curve of the synthetic speech is similar to the original contour.

4. CONCLUDING REMARKS s

In this study, we proposed a WTD method to synthesize

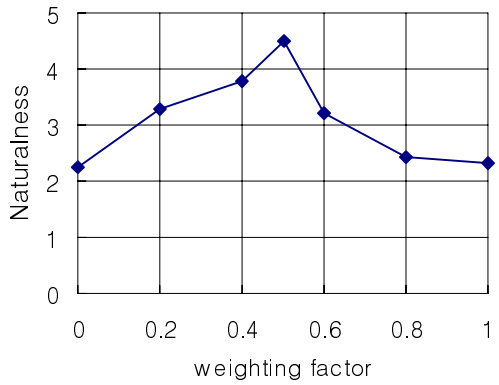

Fig. 8. MOS test results: naturalness verus weighting factor.

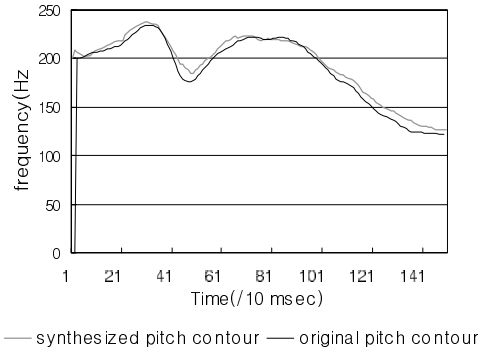

Fig. 9. Example of the synthetic speech.

segmental tones, and devised efficient methods to implement WTD. Vector quantization and phonetic-syntactic distance function are successfully applied to implementing WTD. Through the listening tests we evaluated the performance of the proposed method. The listening test showed that the pitch-controlled speech is very natural.

For further studies the corpus size has to be enlarged. And this method must be evaluated on the practical TTS system. Also, the study on the optimal distance function used in the search of WTD will be continued. It is possible that the distance function used in this paper is not optimal. The last study topic is how we control global tones. We assume that global tones are obtained perfectly. Anyway, our proposed WTD method will be recently ported on the Korean TTS system 'HANSORI'.

REFERENCES

[1] Pierrehumbert, "Synthesizing Intonation,". J. Acoust. Soc. Am. ,Vol70, pp.985-995, 1981

[2] D. Hirst, "Structure and Categories in Prosodic Representation," in Prosody: Models and Measurements, Sp ringer-Verlag, pp.93-109, 1983

[3] Koo, H.S., *An Experimental Acoustic Stusdy of the Phonetics of Intonation in Korean*, Hanshin Publishing co. 1987

[4] Lee, H.Y., *The structure of Korean Prosody*, Hanshin Publishing co. 1990

[5] J.C. Lee, Y.J Lee, S.H. Kim and M.S. Hahn, Intonation Processing for TTS Using Stylization and Neural Network Learning Method," ICSLP'96, pp. 1381-13 84. 1996

[6] Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Jpn., 5(4), pp.233-242, 1984

[7] Y. Linde, A. Buzo, and R. Gray. "An algorithm for vector quantiser design," IEEE Trans. On Communications, pp.84-95, January 1980.