Extraction of Prosodic Information from Spoken Word for Synthesizing Korean Speech

Jin Young Kim and Koeng-Mo Sung

This paper describes a study on prosodic information extracted from spoken Korean words, as a preliminary study to synthesize natural Korean speech. We investigated prosodic features-duration, amplitude and pitch-of some 500 words, of which selection was based on appearing frequency in Korean elementary school textbooks. Some results on word-based prosodic information are deduced.

1. Introduction

Researches on speech-synthesis in Korea were started from the beginning of 80' and some speech-synthesis systems were developed (Lee, J. C. & Y. J. Lee (1990), Lee, J. H. (1990)). The criteria of the evaluation of synthesized speech are intelligibility and naturalness (Bristow (1984), Allen et al. (1987)). But, until now the area of most researches were synthesis-unit and method. Now, we need to study prosodic information of Korean speech for enhancing naturalness of synthesized speech. Some Korean linguists have studied Korean accent (Lee, H. B. (1989)), but most of them are not useful for practically synthesising Korean speech, for quantitative studies have not been performed. Therefore quantitative studies on Korean prosodic is needed for the purpose of practical speech synthesis. We must extract prosodic information from a wide variety of sentences and make rules of it for unlimited vocabulary speech synthesis. But, this requires a great meanss experiments and analyses. Thus we investigated word-based prosodic information as a preliminary study. Prosodic information, such as length, amplitude and pitch, was measured and analysed from spoken words instead of sentences. As a result, We got some word-based prosodic characteristics.

2. Method of Experiment

About 500 words for extracting prosodic features were selected according to their appearing frequencies in Korean elementary school textbook. The selected words are composed of 1−6 syllables. Each word spoken by the speaker who was 29-years old and used Korean standard language was recorded by Kay-Soangraph 5500. And then, prosodic features-duration, amplitude and pitch-of each syllable of words were calculated for analysis. Fig. 1 shows the example of word '가계비'(cost of living). In this figure (a) shows waveform and amplitude of this word and(b) shows the spectrogram of it.

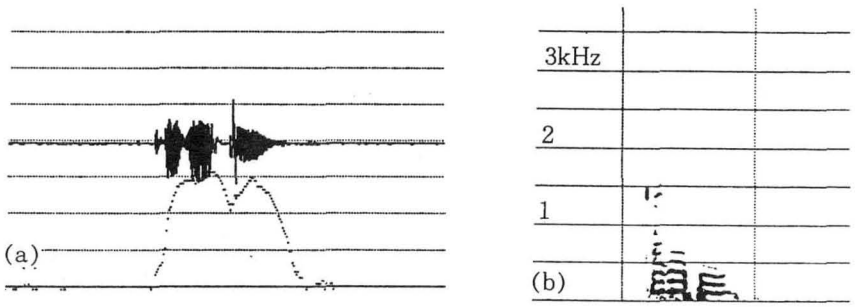

Fig. 1. (a) Waveform and Amplitude (b) Spectrogram

3. Result

3.1. Duration

3.1.1. Total Duration of Spoken Word

At first, total length of each spoken word was investigated to observe the variation of word durational depending on number of syllables composing each words. The results are shown in Fig. 2. It can be deduced from the figure that the increment of word duration depending on the number of syllables decrease as the number increases and that the increment is almost constant as number is above 3.

3.1.2. Duration of Initial Consonants

A Korean syllable is generally composed of three components, which com-

prise the initial consonant, vowel and final consonant (CVC pattern). It often happens that the duration of syllable is slightly changed by that of the vowel and final consonant. Thus, the analysis of initial consonants was formerly performed to study duration-change of syllables. Korean initial consonants have been classified according to their duration.

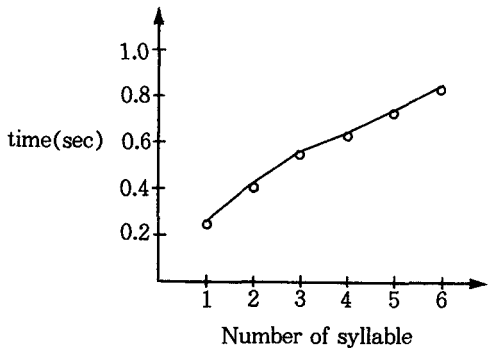

Fig. 2. Duration of Spoken Words

The results is as follows.

1) ㄱ, ㄷ, ㅂ : 39.7 msec
2) ㄴ, ㅁ : 42.1 msec
3) ㅅ : 56.3 msec
   ㅆ : 102.0 msec
4) ㅈ, ㅉ : 53.4 msec
5) ㅊ : 86.4 msec
6) ㅋ, ㅌ, ㅍ : 51.1 msec

3.1.3. Pause-Time between Syllables

When we speak words, the pause interval exist between the syllables according to phonical conditions of later syllable. For example, a pause time of some 70msec always exists when the later syllable begins with a strong sound(ㄲ, ㄸ, ㅃ etc.). Let two syllables be [CVC1-C2VC]
The results of the analysis are

1) voiced+C2[strong sound-ㄲ ,ㄸ, ㅃ, ㅈ] : 71.1 msec
   C2[싸] : 51 msec
2) C1[stop sound]+C2[strong sound] : 106.3 msec
3) voiced+C2[aspirate-ㅋ, ㅌ, ㅍ, ㅊ] : 56.4 msec
4) voiced+C2[ㄱ, ㄷ, ㅂ, ㅈ] : 32.4 msec

3.1.4. Duration of [Vowel+Final Consonant]

The Duration of the part of VC in CVC syllable were investigated. Of course, the pause interval explained in the above chapter is excluded from it. The three cases were considered in this paper. They are

1) case 1 : final consonant does not exist
2) case 2 : final consonant is a stop sound
3) case 3 : final consonant is a nasal or liquid

Fig. 3−4 show the duration of VC part of each syllable depending on the number of syllables and order of syllables. From these figures it can be deduced that the speaker of this experiment accented the second syllable. In later chapter it is discussed that this speaker stressed mainly the second syllable. In any case three facts can be observed from this experiment. they are

1) As the number of syllables increases, the syllable of which the duration decrease is accented syllable−2nd syllable in this speaker−and the last syllable. The duration of the others are almost unchanged.

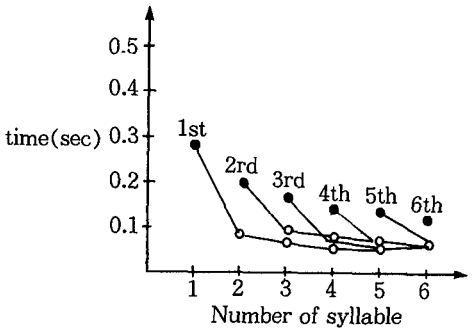

Fig. 3. Duration of Case 1

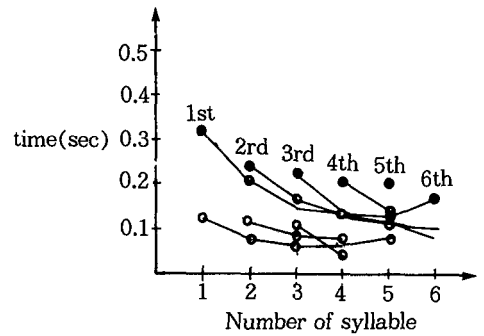

Fig. 4. Duration of Case 2 and Case 3

2) When the number of syllables composing the words is 1-3, the change of duration of syllbales is relatively large.
3) When the final consonants are nasal or liquid, the increment of duration depending on the accent is not distinguished.
Now, some duration rules of [VC] part can be made in accoding to the above three reasonings.

1) case 1:
   Dur[accented syllable]=Max[a1(n-3)+b1, DMINA1]
   Dur[last syllable]=Max[a2(n−4)+b2, DMINL1]
   Dur[other syllable]=DCON1
2) case 2:
   Dur[accented syllable]=Max[a3(n−3)+b3, DMINA2]
   Dur[last syllable]=Max[a4(n−1)+b4, DMINL2]
   Dur[other syllable]=DCON2
3) case 3:
   Dur[last syllable]=Max[a5(n−3)+b5, DMINL3]
   Dur[other syllable]=DCON
      where, DMINA1, DMINA2, DMIN
      DMINL1, DMINL2, DMINL3: Min value

3.2. Stress

As discussed above, the speakers in this experiment have usually put the accent on second syllable. Therefore the amplitude of the second syllable is the biggest through 2−6 syllable words. There are some different accent theories among Korean linguists. We must investigate further whether the second syllable is accented in Korean standard speaking. Experiments with more speakers will be needed for that. Thus we only analyzed the difference of amplitude between the accented syllable and the others. The amplitude of accented syllable was normalized as 0 dB. The result is as follows.

1) 2−syllable word

S1	S2′
−2.58	0.00 dB

2) 3−syllable word

S1	S2′	S3
−3.50	0.00	−3.85dB

3) 4−syllable word

S1	S2′	S3	S4	
−3.63	0.00	−4.06	−4.54dB	

4) 5−syllable word

S1	S2′	S3	S4	S5	
−4.20	0.00	−4.35	−4.94	−4.96dB	

5) 6−syllable word

S1	S2′	S3	S4	S5	S6
−4.20	0.00	−2.94	−5.00	−8.00	−8.21dB

As it turns out the amplitude of the syllables is changed by their phonetical conditions. In this paper, the following three cases are described.

1) When the former syllable hasn't got a final consonant and the latter syllable hasn't an initial consonant, they have equal amplitude.
2) The amplitude of the syllable which has the strong consonant in final position decreases.
3) The amplitude of the syllable which has the aspirate consonant in initial position and a closed vowel /ㅇ], ㅜ, ㅗ / decreases.

In case of 2) and 3) some 9dB decrement was generally observed.

3.3. Pitch

In unlimited vocabulary speech synthesis, the global pitch contour through the sentence is more important than the local pitch variation. Thus we only investigate the initial, the maximum and the minimum(final) value of pitch througth the spoken word. The reference value of pitch is the initial value.

1) 2 syllable word: 0 −25.2 Hz
2) 3−6 syllable word: 0 +10.2 −25.0 Hz

The maximum pitch was on the accented second syllable except 2−syllable word. Table 1 shows the maximum pitch value depending on the number of syllables. From the table two facts can be deduced. They are

1) The maximum pitch value generally increases as the number of syllables increases.
2) The maximum pitch value of which syllable has strong or aspirate as initial consonant is greater than that of other cases.

Table 1. Maximum Pitch Value

(unit: Hz)

# of syllables	C: strong/aspriate	C: else
3	+14	+ 7
4	+15	+ 8
5	+14	+10

4. Conclusion

Korean prosodic information extracted from spoken words is described in this paper. Three prosodic features, such as duration, amplitude and pitch, were investigated toward some 500 words selected according to frequency of occurence and then some results about word-based prosodic characteristics were reported. In the future a study on applying these results to sentence synthesis and experiments with more speakers will be needed.

References

Allen, J., M. S. Hunnicutt and D. Klatt (1987) *From Text to Speech: The MITalk System*, Cambridge Univ. Press.

Bristow, G. (1984) *Electronic Speech Synthesis*, McGraw Hill Company.

Lee, H. B. (1989) *Korean Standard Pronunciation*, Kyoyuk-Kwahak Sa.

Lee, J. C. and Y. J. Lee (1990) 'A Study of Korean Text-to Speech Synthesis Based on LSP,' Speech Communication and Signal Processing Workshop 1990, *Proceeding*, pp. 95-98.

Lee, J. H. (1990) *Design of Natural Language Interface in Searching Database-Korean Text-to-Speech Synthesis System*, Seoul Nat'l Univ.

Applied Electronics Lab.
Department of Electronics Eng.
Seoul National University
56-1 Shillim-dong Kwanak-ku
Seoul 151-742
Korea