Automatic Detection of Korean Accentual Phrase Boundaries

*Ki-Yeong Lee, **Min-Suck Song

※ This work was supported by GRANT No. KOSEF 981-0913-064-2.

Abstract

Recent linguistic researches have brought into focus the relations between prosodic structures and syntactic, semantic or phonological structures. Most of them prove that prosodic information is available for understanding syntactic, semantic and discourse structures. But this result has not been integrated yet into recent Korean speech recognition or understanding systems.

This study, as a part of integrating prosodic information into the speech recognition system, proposes an automatic detection technique of Korean accentual phrase boundaries by using one-stage DP, and the normalized pitch pattern. For making the normalized pitch pattern, this study proposes a method of modified normalization for Korean spoken language.

For the experiment, this study employs 192 sentential speech data of 12 men' s voice spoken in standard Korean, in which 720 accentual phrases are included, and 74.4% of the accentual phrase boundaries are correctly detected while 14.7% are the false detection rate.

I. Introduction

In many recent linguistic studies, the relations between prosodic structures and syntactic or phonological structures are examined[1-3], and the usefulness of prosodic information is proved for understanding semantic, syntactic or discourse structures. But the results of these studies have little been integrated yet into current speech recognition or understanding systems. Especially, into the systems for Korean, they have never been integrated. The main reason for this is that it is difficult to develop the proper method for drawing out these prosodic features, because they are varied according to speakers.

Nevertheless, the studies have been continually made in the 1990' s on the possibility of integrating prosodic features into speech recognition or understanding systems, and as the result, Ostendorf for English[4-5], Verbmobil project for German[6], Shimodaira for Japanese[7-8], etc. propose the methods of detecting and labelling the boundaries of prosodic units, as a part of the study for recognizing and understanding the spoken language. For Korean, the studies are made by Kiyeong Lee and Minsuck Song[9-10], and a method is proposed for understanding the meanings of Korean ambiguous sentences by prosodic information.

The purpose of this study is to propose a method of detecting Korean accentual phrase boundaries automatically, as the prearrangement process for recognizing and understanding continuous speech. The accentual phrase is one of Korean prosodic units proposed by Sun-Ah Jun[11]. We observe speech signals with the naked eye at first, and draw out distinctive features proper to the accentual phrase(AP). The distinctive features are marked with pitch contour. To compose the standard patterns of AP, we adopt a method of modified normalization, because the pitch contours of APs show the down-step phenomenon. By using these standard patterns, we segment the input of continuous speech automatically into AP units by using one-stage DP[12]. Chapter two describes the characteristics of Korean AP, and chapter three proposes the method of modified normalization and segmentation algorithm. The experiment and the conclusion are given in chapter four and five, respectively.

II. Accentual phrases of korean

Most of current speech recognition systems in Korea are designed for recognizing word units. These systems show

*Dep. of Electronic Communication Engineering in Kwandong University
**Dep. of English Language and Literature in Kwandong University
Manuscript Received : December 17, 1998.

the good recognition rate of 95-98%. But, for human beings and computers to communicate each other freely by natural spoken language, the system must be developed so as to understand human utterance. With the current recognition techniques, it is very difficult to develop this system and we come up against a wall. It is prosodic information that can give a clue for breaking this wall.

Nespor and Vogel[2] proposes that human languages universally consist of the hierarchical structure with seven prosodic units, such as syllables, feet, phonological words, clitic groups, phonologicalphrases, intonational phrases and phonological utterance. These units are closely related with the prosodic and phonological rules proper to each language. As Minsuck Song, et al.[12], on the basis of this hierarchical prosodic structure, explains that if we can parse continuous speech above sentence unit into smaller prosodic units only by the prosodic features(not syntactic or semantic features), then we can easily develop the speech understanding system by using the current technique of recognizing independent words, because we can employ the parsed smaller units as the input of recognition systems.

Sun-Ah Jun[11] proposes that not all the seven prosodic units of Nespor and Vogel[2] are necessary for each language, but only a few units proper to each language are linguistically significant. And for Korean, she suggests, accentual and intonational phrases are linguistically significant. Kook Chung, et al.[13] linguistically proves her suggestion to be valid for analyzing continuous speech with the evidence of natural speech.

The intonational phrase(IP) of Sun-Ah Jun is a prosodic unit which corresponds to the intonational phrase of Nespor and Vogel[2], and is characterized by an intonational contour made up of two tonal levels of H(igh) and L(ow).

The intonational contour of the IP is derived from two constituents : the pitch accent and the phrase tone. The pitch accent is a pitch event phonologically linked to a particular stressed syllable and phonetically realized at or around the designated syllable in an utterance. In other words, it is the realization of a lexical stress. The phrase tone is an autosegment which exists independently of lexical entries, and consists of phrase accents and a boundary tone. The phrase accents occur after the rightmost pitch accent, and a boundary tone occurs at the right edge and (optionally) at the left edge of the IP. Thus, the phase accent marks the boundary of intermediate phrases which are smaller units than the IP, and the boundary tone marks the boundary of the IP. The smaller units than the IP are accentual phrases(AP) which are subunits of the IP. In sum, the natural utterance is composed of the hierarchical structure which has APs and IPs as its constituents.[11:33]

The AP is marked by F0 contours. Although the F0 contour has various patterns according to pragmatic meanings such as focus, topic, etc. and to dialects in Korean, Seoul dialect has the basic pattern of LH or LHLH according to the number of syllables which are contained in an AP, and if two H tones appear in an AP the second one is higher than the first one. As another characteristic, APs' contours show the down-step or declination phenomenon in an IP. That is, the L toneof an AP is lower than that of the previous AP and the H tone is lower than that of the previous AP, in turn. The last AP' s contour of an IP is overridden by the boundary tone of the IP. Using these characteristics, we can set up the basic pitch pattern of APs, as follows:

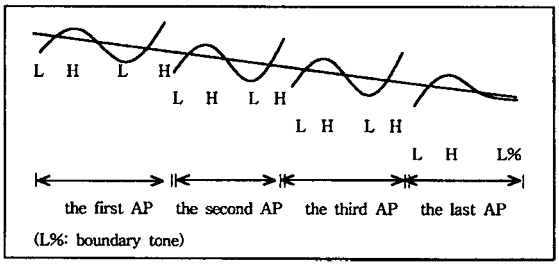


Figure 1. The basic pitch pattern of APs.

Ⅲ. Segmentation algorithm of accentual phrases

The segmentation algorithm of APs is to segment a presegmented IP into APs. We employ the one-stage

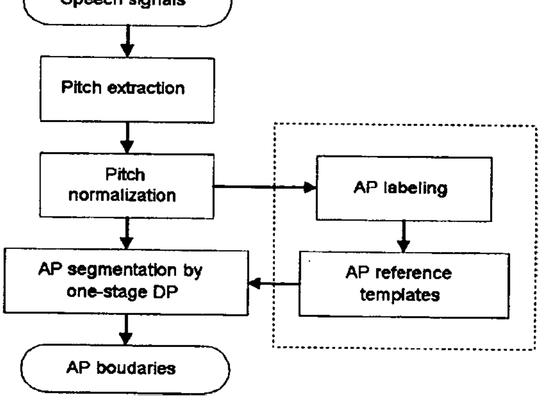


Figure 2. The system of segmenting APs.

DP[14] and the standard patterns made by normalization. Figure 2 shows the block diagram of our segmentation system. The first step is to extract pitch contours from inputted speech signals. And then, pitch contours are normalized so that we can make standard patterns and test patterns for detecting the boundaries of APs by using one-stage DP.

3.1. The Method of Pitch Normalization

We adopt the center-clipped autocorrelation method[15] for the algorithm of drawing out the pitch contour. For autosegmenting APs, we compose multi-templates with pitch contours of hand-labelled APs, and employ the algorithm of one stage DP. But we must normalize the slant pitch contours because the pitch contour of speech is sloped by the down-step effect, as mentioned in the previous chapter. We propose the modified normalization as the method of pitch normalization. The algorithm is as follows : N is the number of frames of speech, and p(n) is the pitch value of each frame.

(1) Get the average pitch-value, p_{savg}, of the first quarter of speech.

$$P_{savg} = \frac{4}{N} \sum_{n=0}^{\frac{N}{4}-1} p(n) \qquad (1)$$

(2) Get the average pitch-value, p_{favg}, of the last quarter of speech.

$$P_{favg} = \frac{4}{N} \sum_{n=0}^{\frac{N}{4}-1} p(N-n) \qquad (2)$$

(3) Get the slope of down-step, $l(n)$, by linear-interpolating the gap between p_{savg} and p_{favg}.

$$l(n) = \frac{n}{N}(P_{favg} - P_{savg}) \qquad (3)$$

(4) Get the normalized pitch value, $p(n)$, by using the slope, $l(n)$.

$$p\ (n) = p(n) - l(n) \qquad (4)$$

(5) Normalize each value of the equation (4) by dividing it by the maximal value.

Figure 3 shows the normalized pitch contour extracted from the speech by using the above method.

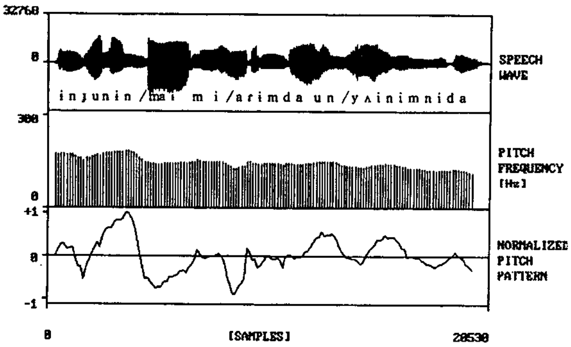


Figure 3. Normalized Pitch Pattern.

3.2. Segmentation algorithm of APs

There is a strong correlation(but no 100% agreement) between syntactic and prosodic phrases. In Korean, syntactic phrases are divided in orthography by a space, and are in accordance with accentual phrases although there are some exceptions(e.g., defective nouns).

Because the pitch contour marks the AP, we segment the speech signal of a sentence into APs by hand labelling, as follows, and then use the pitch contours of the segmented units as the standard patterns:

[inju-nin]$_{\text{TOPIC}}$ / [maim-i]$_{\text{SUBJECT}}$ / [arimdaun]$_{\text{ADJECTIVE}}$ / [yʌin-inmida]$_{\text{NOUN PREDICATE}}$
"Eunju is a woman whose heart is beautiful."

"Eunju is a woman whose heart is beautiful."

The real speech signal of this example is shown in Figure 3. The pitch contour of the signal appears as the form of LH. The normalized pitch contour extracted by the proposed method in the previous section has the leveled slope of the down-step effect, and its form of LH is more evident. We employ the comparison of pattern in the algorithm of segmenting APs. The method of pattern comparison is one stage DP which is able to compare continuous patterns, and the standard pattern is made of the normalized pitch contour extracted from the real speech signal and marked manually.

The equation of the distance measure used in one stage DP is as follows :

$$\begin{aligned} d_{AB}(n,m) &= (1-\alpha)\cdot\{\bar{p}_A(n) - \bar{p}_B(m)\} \\ &+ \alpha\cdot\{\Delta\bar{p}_A(n) - \Delta\bar{p}_B(m)\}, \quad 0 \le \alpha \le 1 \end{aligned} \qquad (5)$$

$$\Delta\bar{p}(n) = \bar{p}(n) - \bar{p}(n-1) \qquad (6)$$

where α $(0 \leq \alpha \leq 1)$ is a weighting ratio of the distance for the normalized pitch value of the equation (4) to the distance for the difference pitch value of the equation (6) which is the pitch variation.

Ⅳ. Experiment and Results

4.1. Experiment

We use speech signals of 10 kHz sampling rate, and 25.6 msec per frame, 12.8 msec for scanning interval. The script for data collection is composed of 16 sentences(all sentences are declarative), of which 4 sentences consists of 4 APs, 4 sentences of 5 APs, and 8 sentences of 3 APs. The total number of APs is sixty. Twelve male speakers of standard Korean read the script without any guideline, and record with their own equipment in order to keep the quality of data as natural as they can be. This experiment employs 192 sentential speech data of 12 men' s voice spoken in standard Korean, in which 720 APs are included.

In the experiment, two cases are compared : one is the case in which the single-pattern of one sentence pronounced by one speaker is used as the standard pattern, and the other is the case in which the multi-pattern of one sentence(the same sentence of the case one) pronounced by three speakers is used as the standard pattern. The standard pattern of the single-pattern is made up of the normalized pitch contours of 4 APs which are extracted from a sentence spoken by a speaker, and that of the multi-pattern is made up of the normalized pitch contours of 16 APs which are extracted from the same sentences spoken respectively by three speakers.

4.2. Results

Figure 4 shows the correct detection rates according to weights, α in the equation (5).

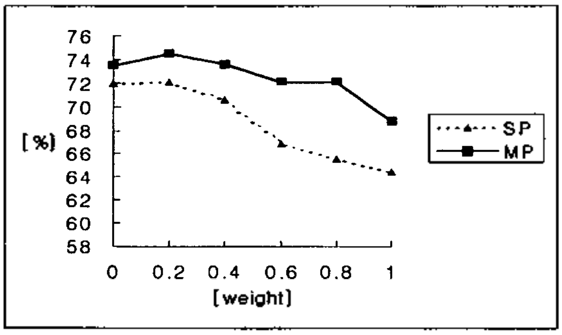


Figure 4. Correct detection rates according to weights.

SP indicates the case of the single pattern, and MP indicates the case of the multi-pattern. We can see the highest correct detection rate, 72.1% for the SP and 74.4% for the MP, when the weight, α is 0.2 in both cases. This result reveals that there is a prosodic unit of AP in Korean and that the pitch contour of LH tones plays a significant role in segmenting or marking APs in Seoul dialect of Korean. We can also see the fact that the multi-pattern is superior to the single pattern as the standard pattern.

Table 1. False detection rates according to weights,α [%].

α	0.0	0.2	0.4	0.6	0.8	1.0
SP	18.5	16.3	16.7	17.4	16.9	17.5
MP	16.0	14.7	14.7	15.3	14.0	16.5

Table 1 shows false detection rates according to weights, which mean boundaries detected the wrong position by automatic segmentation. The rates are also lower when the multi-pattern is used as the standard pattern, and are lowest wheń the weight, α is 0.2. We consider that these errors might occur because the LH pattern can appear within the AP, not in the boundary of AP. We expect that the correct detection rate will increase or the false detection rate will decrease if other prosodic parameters such as the duration of AP' s final syllable, energy, etc. are used, or if Markof chain with large amount of data is employed as the segmentation algorithm.

Ⅴ. Conclusion

This study proposes a method of segmenting Korean speech signals automatically into accentual phrases(AP). We employ one stage DP as the segmentation algorithm, and single-pattern(SP) and multi-pattern(MP) templates of normalized pitch contour made by hand. For experiment, we use speech data of 12 speakers. The results of the experiment show the correct detection rate of 74.4% and the false detection rate is 14.7%, when the multi-template is used as the standard pattern.

From the experiment we can conclude that :

(1) for recognizing AP boundaries, the normalized normalization is useful, which is proposed in this study.

(2) as the standard pattern, the multi-pattern template is superior to the single-pattern template.

(3) the difference pitch value, which is the pitch variation, is helpful for improving segmentation accuracy.

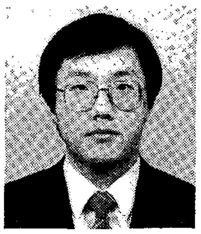
References

1. E. O. Selkirk, Phonology and Syntax, The MIT Press, Cambridge, Massachusetts, 1984.
2. M. Nespor and I. Vogel, Prosodic Phonology, Foris Publications, Dordrecht, 1986.
3. M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," Phonology Yearbook 3, ed. J. Ohala, pp. 255-309, 1986.
4. C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech, Audio Processing,* Vol. 2, No. 4, pp. 469-481, 1994.
5. C. W. Wightman and M. Ostendorf, "Automatic recognition of intonational features," in *Proc. of IEEE Int. Conf. ASSP*, pp. 1-221, 1992.
6. A. Kipp, M. Wesenick, and F. Schiel, "Automatic detection and segmentation of pronunciation variants in German speech corpora," in *Proc. of ICSLP96*, 1996.
7. H. Shimodaira and M. Kimura, "Accent phrase segmentation using pitch pattern clustering," in *Proc. of IEEE Int. Conf. ASSP*, pp. I-217, 1992.
8. H. Shimodaira and M. Nakai, "Prosodic phrase segmentation by pitch pattern clustering," in *Proc. of IEEE Int. Conf. ASSP*, pp. II-185, 1994.
9. Kiyoung Lee and Minsuck Song, "Automatic recognition of sentence-final intonational patterns for Korean predicates," in Proc. of the 12th Workshop for Speech Communication and Signal Processing, Acoustic Society of Korea, pp.131-134, 1995.
10. Kiyoung Lee, et. al. "Intonation conversion using the other speaker' s excitation signal," *The Journal of the Acoustic Society of Korea*, Vol. 14, No. 4, pp.27-34, 1995.
11. Sun-Ah Jun, The Phonetics and Phonology of Korean Prosody, Doctoral Dissertation, The Ohio State University, 1993.
12. Minsuck Song, et al., "A theoretical model of spoken language processing," in Proc. of the 1994 Conference of HCI(Human and Computer Interface), Korea Information Science Society, pp. 1-9, 1994.
13. Kook Chung, et al., A Study of Korean Prosody and Discourse for the Development of Speech Synthesis/Recognition System, Annual Report of Korea Telecom Research and Development Group, 1996.
14. Hermann Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol.ASSP-32, No.2, pp.263-271, 1984.
15. L.R. Rabiner, R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Inc., Englewood Cliffs, N.J., pp.150-158, 1978

▲ Ki-Yeong Lee

Ki-Yeong Lee received the B.S. the M.S, and the Ph. D. degrees in Electronic Engineering from Myong Ji University, Seoul, Korea, in 1984, 1986, and 1992, respectively. Since 1993, he is a professor of Dept. of Information Communication Engineering at Kwandong University His major research areas are fuzzy analysis, speech coding, speech synthesis/recognition, and digital signal processing.

▲ Min-Suck Song



Min-Suck Song received the B. A. the M. A. and the Ph. D. degrees in Linguistics from Hankuk University of Foreign Studies, Seoul, Korea, in 1983, 1987, and 1992, respectively. Since 1994, he is a professor of Dept. of English Language and Literature at Kwandong University. His main research areas are phonology and phonetics. Especially, he is interested in experimental phonology, computational phonology, speech synthesis, and speech recognition.