CART modeling of Korean Prosodic Phrases

Hyo Sook Kim
(Eoneo Inc., Korea)

1. INTRODUCTION

To produce highly qualified synthetic speech, controlling prosodic features are necessary. Primary prosodic features are duration, pitch and loudness. Based on those features, several prosodic phrases are defined. Concerning Korean Prosodic phrases, binary classification is mainly suggested[2,3,5,6,7,10]. Maltomak and Malmadi are well known units defined upon the existence of stressed syllable and pause for breath[6,7]. In K-ToBI system two prosodic phrases are also used, accentual phrases(AP, hereafter) and intonational phrases(IP, hereafter). IPs are usually cued by long pauses which appear between Eojeols. Maltomak and Malmadi could be mapped onto AP and IP respectively. Differ from Maltomak, AP doesnt contain the phrases which have phrase final lengthening. Pause duration contributes to the classification of prosodic boundary degrees[1].

In this paper we used K-ToBI transcription system[10]. In determining phrases, acoustic features such as phrase final lengthening, pitch difference, pause duration and change of segmental quality are considered. In K-ToBI, AP and IP are defined by tonal markings, but the break index value indicates the labelers subjective sense of disjuncture. So, K-ToBI also has flag m for the cases of mismatch between tones and disjuncture type. To avoid mismatch cases and to get consistent data, we modified original K-ToBI. We concluded that break index 1 and 2 could be merged into AP. And the cases which have break index 3 but the pause duration is shorter than 200ms are also classified as AP. Consequently, break index is modified as Zero, AP and IP. Maintaining above criteria, one of the co-authors tagged 400 sentences.

2. DATA AND CART

2.1 Data

We used 400 sentences made up of editorials, essays, novels and news scripts. Professional radio actress read those sentences in normal speed. All 400 sentences end with declarative endings. Eojeol(words differentiated from others with spaces in text) is basic units to be tagged break index. But many Korean compound nouns are long to be spoken without break. Text was corrected to represent the speech of a speaker in that case. For example, there is a Korean compound noun Noryangjin#susan#jusik#hwesa (#means noun boundary). If a speaker pronounced it as Noryangjin/ susan/ jusikhwesa,(/ means AP boundary) we modified text as Noryangjin [space] susan [space] jusikhwesa.
Punctuation marks play important role determining prosodic boundaries. So we regard punctuation mark as independent morpheme. Original sentences(representing speakers speech) have 5620 Eojeols. Punctuation marks made additional Eojeols. Finally we have 6207 Eojeols. 987 Eojeols were tagged as Zero, 3427 Eojeols were tagged as AP and 1793 Eojeols were tagged as IP. Prosodic boundary may not occur between morpho-syntactically related Eojeols[9]. In read speech speakers process text as flat structure rather than hierarchical structure[8]. So morphologic analysis alone may produce a good result in prosodic phrasing without syntactic analysis[8,9].
In this paper, morphological analysis was carried. Eojeol is consist of content word part and optional(but mostly) functional word part. So we divided POS(part-of-speech) into left(content word) POS and right(functional word) POS[4,9]. If there is no functional part in Eojeol, left POS was used as right POS. Referring to previous studies we made following POS list[1-4]. POS tagging was done automatically and corrected manually.

Table 1.1 Part-of-Speech list

number	Part-of-Speech
1.	Genitive case marker
2.	Adnoun
3.	Adnominal ending
4.	Noun
5.	Verb
6.	Noun
7.	Nominalizing ending
8.	Accusative case marker
9.	Delimiter
10.	Auxiliary connecting ending
11.	Adverb
12.	Adverbial case marker
13.	Mark (excluding , and .)
14.	Number
15.	,
16.	Bound noun
17.	Conjunctive adverb
18.	Conjunctive marker
19.	.
20.	Sentence-terminating ending
21.	Subordinating ending
22.	Nominative case marker
23.	Topicalized case marker

2.2. CART

We trained CART classification tree predicting prosodic boundary under 0se(standard error) rule and 1se rules. Adjacent Eojeols POS was considered[5]. Besides POS, the number of syllables in observed Eojeol and next Eojeol and the location of observed Eojeol in sentences were also used as variables explaining prosodic boundary. POS-related parameters are categorical variables and others are real-valued variables.

Table 2.1 Explaining variable list

variable name	content
syl	Number of syllables of observed Eojeol
n_syl	Number of syllables of following Eojeol
loc	Location of observed Eojeol in sentence
ppL	Left POS of previous previous Eojeol
ppR	Right POS of previous previous Eojeol
pL	Left POS of previous Eojeol
pR	Right POS of previous Eojeol
L	Left POS of observed Eojeol
R	Right POS of observed Eojeol
nL	Next left POS of observed Eojeol
nR	Next right POS of observed Eojeol
nnL	Next next left POS of observed Eojeol
nnR	Next next right POS of observed Eojeol

When target variables were three(Zero, AP, IP), the average prediction accuracy was 77.5% on training data and 75.2% on test data. This study is part of prosodic module of TTS system. In TTS, IP boundary detection is more important. So we divided target variables. NonIP(Zero, AP) and IP classification tree and Zero and AP classification tree. We trained on 4188 Eojeols and tested on 2109 Eojeols in NonIP-IP classification tree. Except 1793(which is tagged as IP)Eojeols, 4144 Eojeos were used as training data and 1423 Eojeols were as test data in classifying Zero and AP.

3. PROSODIC BOUNDARY PREDICTION

3-1. NonIP and IP Prediction

Prediction rate was 90.4% on training data and 88.5% on test data under 0se rule. 87.9% on training data and 88.8% on test data under 1se rule.

Table 3.1 Confusion matrix of NonIP and IP under 0se rule.

Actual	Predicted Class		Actual total
Class	NonIP	IP	
NonIP	1283.00	157.00	1440.00
IP	76.00	503.00	579.00
Pred. Tot	1359.00	660.00	2019.00
Correct	0.891	0.869	
Tot. Correct	0.885		

<R>, <nR>, <nL>, <nnR>, <nnL> were important variable in NonIP and IP classification.

Table 3.2. Variable Importance in NonIP and IP classification. (under 0se rule)

Variable	Importance
R	100.000
nR	77.317
nL	71.470
nnR	68.841
nnL	66.316
L	38.006
pR	10.187
n_syl	7.084
pL	5.561
ppR	5.321
loc	3.217
ppL	3.041
syl	1.310

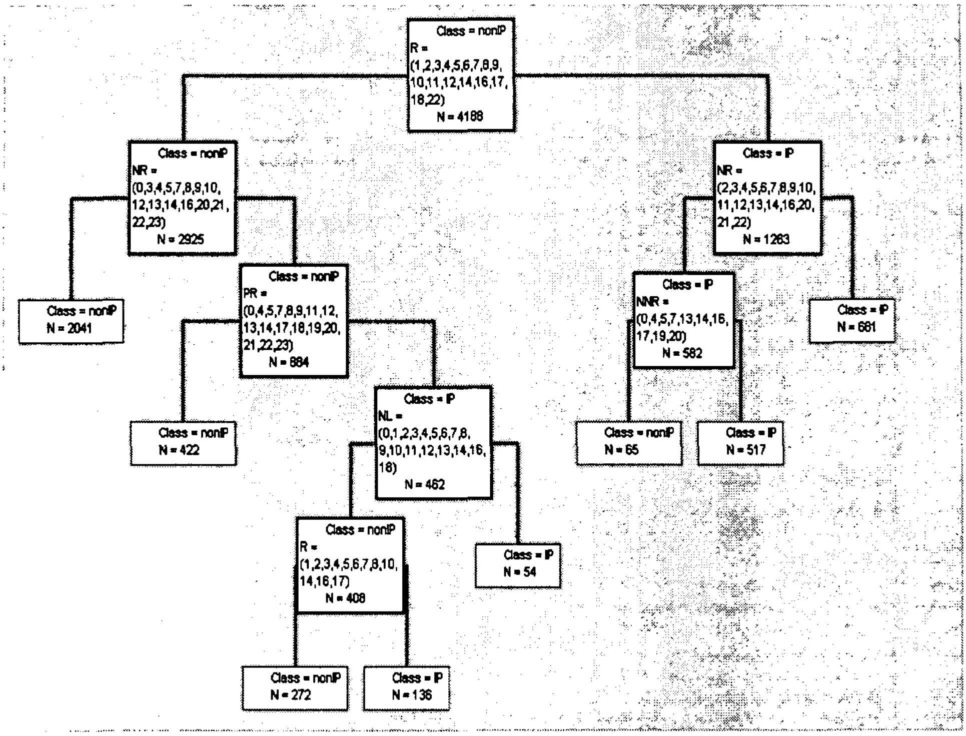

Figure 3.1 NonIP and IP classification tree(under 1se rule)

Firstly, 4188 Eojeols are divided by the variable <R>. If the condition is satisfied, the tree will branch on left side. For example, the rightmost terminal node will be classified as IP through following branching rule. If <R> is 13(mark excluding , and .), 15(,), 19(.), 20(sentence-terminating ending), 21(subordinating ending) and 23(topicalized case marker), then Eojeols go to right branch. If <NR> is 1(genitive case marker), 15(,), 17(conjunctive adverb), 18(conjunctive marker), 19(.) and 23(topicalized case marker), then Eojeols are classified as IP.

3.2 Zero and AP prediction

Prediction rate was 79.7% on training data and 78.7 on test data under 0se rule. 77.2% on training data and 78.6% on test data under 1se rule.

Table 3.3 Confusion matrix of Zero and AP under 0se rule.

Actual Class	Predicted Class Zero	AP	Actual total
Zero	259.00	70.00	329.00
AP	233.00	861.00	1094.00
Pred. Tot.	492.00	931.00	1423.00
Correct	0.787	0.787	
Tot. Correct	0.787		

In determining Zero and AP, <R>, <L> and <n_syl> were important.

Table 3.4 Variable Importance in Zero and AP classification. (under 0se rule)

Variable	Importance
R	100.000
L	57.112
n_syl	54.539
nR	45.252
nL	37.302
syl	35.689
pR	27.183
nnR	13.456
nnL	10.004
ppR	4.210
loc	2.010
ppL	1.328
pL	0.370

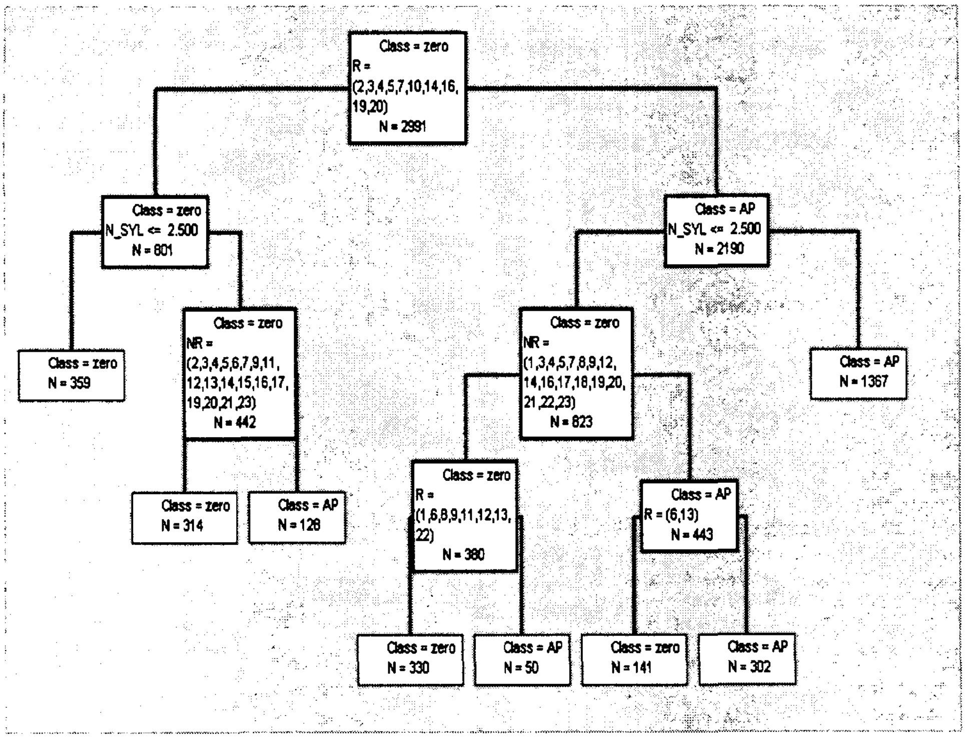

Figure 3.2 Zero and AP classification tree(under 1se rule)

If <R> is 2(adnoun), 3(adnominal), 4(noun), 5(verb), 7(nominalizing ending), 10(auxiliary connecting ending), 14(number), 16(bound noun), 19(.) and 20(sentence-terminating ending), Eojeols go to left branch. And if <n_syl> is below 2.5(syllables), then Eojeols are classified as Zero.

4. CONCLUSION

Prediction accuracy of NonIP and IP was 88.5% and the accuracy of Zero and AP was 78.7%. The low accuracy of Zero and AP represents the difficulty in differentiating Zero from AP. Additional experiment will be carried on specifying characteristics of AP boundary. That will promote the Zero and AP prediction accuracy.

REFERENCES

[1] Kim, S. et al., Corpus-based Korean Text-to-Speech Conversion System , The journal of the acoustical society of Korea, Vol. 20, No. 3, 2001.

[2] Kim, J. et al., A Method of Intonation Modeling for Corpus-based Korean Speech Synthesizer , Korean journal of Speech Sciences, Vol. 7, No.2,2000.

[3] Eom, K. et al., Prediction of Prosodic Boundary strength by means of Three POS(Part of Speech) sets , MALSORI, Journal of the phonetic society of Korea, No. 35-36, 1998.

[4] Lee, S. et al., The Modelling of Prosodic Phrasing and Pause Duration using CART , KSCSP 15. 1998.

[5] Lee, Y. et al., A computational algorithm for F0 contour generation in Korean developed with prosodically labeled databases using K-ToBI system , MALSORI, Journal of the phonetic society of Korea, No. 35-36, 1998.

[6] Lee, H. B., Korean Standard Pronunciation, Gyoyukgwahaksa, 1993(3rd.)

[7] Lee, H. Y., Korean Phonetics, Taehaksa, 1996.

[8] Seong, C. , The Experimental Phonetic Study of the Standard Current Korean Speech Rhythm , ph.D dissertation., SNU., 1995

[9] Jeong, H., On automatic boundary labeling of Korean read speech , 2nd Conference on Phonetics, The Phonetic Society of Korea, 1996.

[10] Jun, Sun-Ah, K-ToBI Labelling Conventions ,