

A prosodic phrasing model for a Korean text-to-speech synthesis system

Kyuchul Yoon

Department of Linguistics
The Ohio State University
1712 Neil Avenue, Columbus Ohio 43210, USA
kyoon@ling.osu.edu

Abstract

This paper presents a prosodic phrasing model for Korean to be used in a text-to-speech synthesis (TTS) system. Read text corpora were morpho-syntactically parsed and prosodically labeled following the Penn Korean Treebank [1] and K-ToBI prosodic labeling conventions [5] respectively. Decision trees were trained with morpho-syntactic and textual distance features to predict locations of accentual and intonational phrase breaks. Our phrasing model cross-validated on a 300-sentence (6,936 words or 21,436 syllables, an average of 72 syllables or 23 words per sentence) predicted non-breaks with $F=92.4\%$ and breaks with $F=88.0\%$ ($F=72.8\%$ for accentual phrase breaks and $F=71.3\%$ for intonational phrase breaks).

1. Introduction

Phrase break assignment is an important task in any text-to-speech synthesis system. Both missing breaks and unnecessary breaks could make TTS systems not only unnatural but unintelligible as well. This is partly because the phrasing module in a TTS system affects other modules such as the intonation and duration module. Several recent studies demonstrate that prosodic phrasings influence segmental properties in addition to the well-known phrase-final lengthening effects. For example, the prosodic effects on segments for English have been studied for /h/ and glottal stop [12] and for American /z/ [13]. Other works have demonstrated that segmental properties of Korean coronal stops [3] and fricatives [16] are affected by different prosodic domains.

Motivated by these findings, we have started working on a diphone-based concatenative TTS system within the Festival TTS framework [15]. Our aim is to build a prosodically conditioned diphone database. We believe that prosodic effects on a segment can be properly modeled by recording different diphones in different prosodic contexts. In order to achieve this, it is very important that we establish an accurate prosodic phrasing model. We employed classification and regression trees (CART) [2] algorithm, one that has proved very successful in building phrase break models in earlier studies.

1.1. Prosodic hierarchy of Seoul Korean

There are two tonally defined prosodic phrases above the level of the prosodic word (PW) in Korean, i.e. the intonational phrase (IP) and the accentual phrase (AP) [4]. An IP is marked by a boundary tone and phrase-final lengthening with an optional sense of pause where as an AP, smaller than an IP but larger than a PW, is marked by a phrasal tone THLH (T = H if the AP initial segment is aspirated or tense, and T = L otherwise). An IP can have one or more APs and the final tone of the last AP within an IP is replaced with the boundary tone of the IP.

2. Methods

2.1. Corpora

A 400-sentence subset of the Korean Newswire Text Corpus [8] was read by two native speakers of Seoul Korean (one man, the author, and one woman) and audio comprising 487 sentences from YTN News Channel broadcasts was additionally recorded. The former corpus contained 28,666 syllables or 9,246 words, an average of 72 syllables or 23 words per sentence while the latter had 23,615 syllables or 7,092 words, an average of 48 syllables or 14 words per sentence. Here, the unit of word needs to be defined. Since Korean lexical items can be inflected with prefixes, suffixes, postpositions, tense morphemes, etc., it would be more appropriate to call a word an *eojeol*, a fully inflected lexical item. In practice, an *eojeol* corresponds to a space-delimited orthographic word unit.

The reason why the former text corpus was read by two speakers was that we intend to build a diphone-based concatenative TTS system for the reading style of one of the speakers. The second speech corpus was recorded from online video archives from the news channel in order to see how the multi-speaker corpus compares with the single-speaker corpus in terms of the predictive power of the models. The three speech corpora were morpho-syntactically parsed and prosodically labeled as shown in the next section.

2.2. Parsing and labeling

The three speech corpora were morpho-syntactically parsed using the Penn Korean Treebank annotation tools [1]. The Treebank uses three major types of part-of-speech (POS) tags - 14 content tags, 15 function tags, and 5 punctuation tags. Of these, we modified and expanded the NNU tag used for ordinal and cardinal numeral to reflect our observation that Korean speakers usually put breaks after multiples of 10 (labeled as NUU) and 10,000 (labeled as NUE). The Treebank also uses phrase structure annotation for syntactic bracketing, similar to the annotation schemes used by the Penn English Treebank [11]. The bracketing tagset for the Treebank are divided into three types: morphological POS tags (e.g., NNC, VV, ADV), syntactic phrasal category tags (e.g., NP, VP, ADVP) and syntactic function tags (e.g., -SBJ for subject, -OBJ for object, -ADV for adverbial).

The morpho-syntactic annotation proceeded in two steps. First, we ran the text portion of the three corpora through a morphological tagger, a tool developed by Han *et al.* [1] at the University of Pennsylvania. The output POS tags were then hand-corrected to be fed into the second step of syntactic bracketing. The bracketing was done using the emacs-based interface developed for the Penn English Treebank [11], which was modified for Korean. The parsed corpora provided morpho-syntactic information to be used later in the training phase.

The three read speech corpora that we recorded were prosodically labeled by a trained labeler (the author) following the K-ToBI (Korean tones and break indices) prosodic labeling conventions [5]. The labeling proceeded in two steps. First, each of the recorded sentence was loaded into Praat and manually segmented by the prosodic word. Then a Praat script automatically assigned default tone labels, which were manually corrected by the labeler. The labeled corpora, which we call OSU Talkbank, consist of time-aligned eojeol, AP and IP boundaries, etc. as produced by each of the speakers.

2.3. Building phrasing models

The classification and regression trees (CART) technique was used to train our prosodic phrasing models. Specifically, the Wagon tool provided by the Edinburgh Speech Tools Library was used. The features that we used to train the models can be divided into three categories: morphological features, syntactic features, and non-syntactic textual distance features. Each of them are listed below. The use of the morpheme identity feature was based on our observation that certain postpositions and endings were consistently associated with phrase breaks.

(1) morphological features:(with respect to a potential break)

- POS tags of four preceding tokens
- POS tags of three following tokens
- morpheme identity of the immediately preceding token (a subset of postpositions and endings)

(2) syntactic features:

- terminal phrasal category tags of four preceding and three following tokens
- pre-terminal phrasal category tags of four preceding and three following tokens

(3) non-syntactic textual distance features:

- token length in syllables
- distance in syllables from the previous comma
- distance in syllables to the following comma
- distance in eojeols from the previous comma
- distance in eojeols to the following comma
- distance in syllables from the sentence start
- distance in syllables to the sentence end
- distance in eojeols from the sentence start
- distance in eojeols to the sentence end

The above features were extracted from the text/speech corpora and divided into training and test sets. Two sets of 300 sentences of the Newswire corpus and one set of 365 sentences of the YTN corpus were reserved for training sets. Two sets of 50 sentences from the Newswire corpus and one set of 61 sentences from the YTN corpus were also reserved for testing the actual models. Another set of 50 sentences from the Newswire corpus was used as held-out corpus to find the optimal decision tree parameters and feature sets in the training phase.

With the features extracted from the 300 training sentences and 50 held-out sentences, we found optimal tree parameters and feature combinations. This proceeded in four steps. Firstly, we varied the stop value of the Wagon tool, the minimum number of examples for a leaf node of a decision tree, from 1 to 100 with an interval of 5, built models with the training set, tested them against the held-out set and found that the stop value of 10 performed optimally.

Secondly, for each of the three feature categories, we varied the window length features such as the POS tags and phrasal category tags. For the morphological feature category, for example, we varied the POS window length from three (one POS on one side with respect to the potential boundary and the other two POSs on the other side) to seven (divided into three and four POSs). For the other features such as the distance features, we tried different combinations of these features to find the optimal combination. We repeated the procedures that we did in the first step and found that within each of the feature categories, the morpheme identity, sequences of three POSs (two preceding and one following), sequences of seven terminal phrasal categories (four preceding and three following) and a combination of token length in syllables and all the distance features containing commas performed optimally.

In the third step, we tried different combinations of feature categories obtained in the second step and found that combining all the optimal features discovered in the second step performed best.

In the last step, we varied the size of the tagsets following the work of Taylor and Black [14] by successively collapsing 1) subtypes of POS categories such as nouns, verbs, and adverbs, 2) subtypes of postpositions, 3) subtypes of endings, 4) subtypes of affixes and 5) subtypes of punctuations. This resulted in tagset sizes of 35, 25, 21, 16, 13, and 10. Of these the tagset size of 35 performed best, followed by the size of 21.

Based on the results of the preceding experiments, we have decided to use the stop value of 10, the features of morpheme identity, sequences of three POSs (two preceding and one following), sequences of seven terminal phrasal categories (four preceding and three following) and a combination of token length and distance features containing commas, all with the POS tagset size of 35.

We have trained our prosodic phrasing models using the two sets of 300-sentence Newswire and one set of 365-sentence YTN training sentences and tested against the two sets of 50-sentence Newswire and one set of 61-sentence YTN test sentences respectively.

Of the three phrasing models, we chose the best one for cross-validation, which was the model trained on the Newswire corpus. We split our 350 sentences (excluding from the 400 sentences the 50-sentence held-out set used in parameter estimation) into 50 sets of 7 sentences, drawing equally from all parts of the corpus. We then carried out 7-fold cross-validation. Each instance was trained on 300 sentences and tested on 50. The results are as follows.

3. Results

Our cross-validated prosodic phrasing model (Newswire corpus read by the author) predicted non-breaks with $F=92.4\%$ and breaks with $F=88.0\%$ ($F=72.8\%$ for accentual phrase breaks and $F=71.3\%$ for intonational phrase breaks). The averaged confusion matrix and the recall/precision values are given in Tables 1 and 2.

3.1. Error analysis

The types of errors made by our model are compared with those from other studies in Table 3. Noting the fact that each *eojeol* roughly corresponds to an AP in carefully read Korean, the errors of any type involving an IP (shaded in Table 3) should be classified as potentially more critical than others. These errors occupy 6.54% of all the breaks/non-breaks in the prediction of our model. The model built by Kwon [7] performs better in predicting AP breaks (Table 2). However, he used IP breaks in building their AP prediction model. Thus, a direct comparison is not appropriate.

3.2. Model analysis

All of the cross-validation models had their first split on the feature of the POS of the immediately following token and the second split on the POS of the immediately preceding token, followed by the distance in syllables from either the

Table 1: Averaged confusion matrix from our cross-validated phrasing models.

	actual	predicted
1459	non-break	non-break
6	non-break	HL% (IP)
144	non-break	LHa (AP)
6	HL% (IP)	non-break
145	HL% (IP)	HL% (IP)
1	HL% (IP)	L% (IP)
69	HL% (IP)	LHa (AP)
1	L% (IP)	non-break
58	L% (IP)	HL% (IP)
6	L% (IP)	L% (IP)
2	X?% (IP)	LHa (AP)
7	L% (IP)	LHa (AP)
82	LHa (AP)	non-break
78	LHa (AP)	HL% (IP)
511	LHa (AP)	LHa (AP)

Table 2: Comparison of our cross-validated recall/precision values with other studies

break type	current	Kim [6]	Lee [10]	Kwon [7]
non-break	90.7/94.3	75.5/76.1	90.8/85.1	80.1/91.6
H% (IP)	0.0/N/A			
HL% (IP)	65.6/50.6			
L% (IP)	7.6/88.1			
LHa (AP)	76.2/69.7	43.4/54.6		96.1/79.7
IP (total)	71.2/71.3	80.7/70.4		78.9/58.0
breaks(total)	90.9/85.4	78.4/77.8	77.1/85.4	90.6/72.2

preceding or the following comma. Thus, it appears that the POS information plays an important role in the phrase break prediction. The token length in syllables was also ranked high in the decision trees and sometimes competed with terminal phrasal categories. With performance improvement of automatic parsers, more use of syntactic category information could further enhance the predictive power of phrasing models.

4. Discussion

We have described the procedures for building a Korean prosodic phrasing model and presented its performance. Our prosodic phrasing model cross-validated on a 300-sentence corpus (6,936 *eojeols* or 21,436 syllables, an average of 72 syllables or 23 words per sentence) and tested against a 50-sentence corpus (1,091 *eojeols* or 3,320 syllables) predicted non-breaks with 90.7%/94.3% (recall/precision) and breaks with 90.9%/85.4% (76.2%/69.7% for accentual phrases and 71.2%/71.3% for intonational phrases). Compared to earlier similar attempts to build phrasing models, our study used a larger training corpus and showed improvement in over-

Table 3: *Types of errors. Numbers in parentheses indicate the total counts of (non-)breaks for evaluation.*

type of errors		current (2481)	Kim (2044)	Lee (1438)	Kwon (1084)
insertion	AP	5.60%	5.43%		13.4%
	IP	0.24%	0.06%	5.42%	14.8%
deletion	AP	3.19%	7.19%		2.12%
	IP	0.24%	0.04%	9.39%	5.44%
substitution	AP \Rightarrow IP	3.02%	4.50%		
	IP \Rightarrow AP	3.04%	2.01%		

all recall/precision values. However, it is not possible to compare the relevant works in a direct way because non-break/break decisions were made following different theories and the composition of the corpora was different.

Kim [6] built a prosodic phrasing model for Korean by training it on a 300 sentence corpus uttered by three speakers. In another study, Lee [10] trained his model on a 240 sentence corpus (2,286 eojeols, collected from different genres and uttered by a trained professional) and tested against a 160 sentence corpus (1,438 eojeols). Their recall/precision values are given in Table 2. Lee followed a theoretical framework proposed by Lee [9] and labeled his corpora with only prosodic phrase boundaries which are believed to correspond to our IP boundaries.

Compared to these studies, the three corpora that we trained our models on were from one genre, i.e. newspaper articles and two corpora were read by one man and one woman respectively. Thus, the homogeneity of our text and speech corpora may have contributed to the overall performance. With respect to the homogeneity, it appears from our experiments that a phrasing model trained on a single-speaker corpus performs better than the one trained on a multi-speaker corpus in terms of phrase model performance. Since we intend to build a prosodically conditioned concatenative TTS system for a single speaker, the results of our study seem promising. Also, with increasing availability of syntactically parsed and K-ToBI labeled corpora, further performance enhancement can be expected.

5. Acknowledgements

The author would like to thank Professors Chris Brew, Mary Beckman and Martha Palmer for their help with the study, Professor Hyunsook Kang and Kirk Baker for their advice, and Eunjong Kong, Hyunsook Shin, Shijong Ryu, and Na-Rae Han for their help on parsing and labeling.

6. References

[1] Han, Chung-hye, Han, Na-Rae, Ko, Eon-Suk, and Palmer, Martha, Development and evaluation of a Korean Treebank and its application to NLP, LREC-2002.

[2] Breiman, L., Friedman, J.H., Olsen, R.A. and Stone, C.J., Classification and regression trees, Chapman & Hall, 1984.

[3] Cho, Taehong and Keating, Patricia A., Articulatory and acoustic studies of domain-initial strengthening in Korean, Journal of Phonetics, 29, 2001.

[4] Jun, Sun-Ah, The phonetics and phonology of Korean prosody, PhD thesis, The Ohio State University, 1993.

[5] Jun, Sun-Ah, K-ToBI (Korean ToBI) labelling conventions (version 3.1), 2000.

[6] Kim, Yeon-Jun, Byeon, Heo-Jin and Oh, Yung-Hwan, Prosodic phrasing in Korean; determine governor, and then split or not, Eurospeech99.

[7] Kwon, Ohil, Hong, Munki, Kang, SunMee and Shin, Jiyoung, AP, IP prediction for corpus-based Korean text-to-speech, Journal of Speech Sciences, 9(3):25-34, 2002.

[8] Linguistic Data Consortium (LDC), Korean Newswire Text Corpus, catalog number LDC2000T45, ISBN 1-58563-168-X, 2000.

[9] Lee, H.B., Standard Korean Pronunciation, Educational Science Press, 1989.

[10] Lee, Sangho, Tree-based modeling of prosody for Korean TTS systems, PhD thesis, Korea Advanced Institute of Science and Technology (KAIST), 2000.

[11] Marcus, Mitch, Santorini, Beatrice and Marcinkiewicz, M., Building a large annotated corpus of English, Computational Linguistics, 19(2):313-330, 1993.

[12] Pierrehumbert, Janet and Talkin, David, Lenition of /h/ and glottal stop, Papers in Laboratory Phonology II, 90-116, 1992.

[13] Smith, C.L, The devoicing of /z/ in American English: effects of local and prosodic context, Journal of Phonetics, 25:471-500, 1997.

[14] Taylor, Paul and Black, Alan W., Assigning phrase breaks from part-of-speech sequences, Computer Speech and Language, 12:99-117, 1998.

[15] Taylor, Paul, Black, Alan W. and Caley, Richard, The architecture of the festival speech synthesis system, 3rd ESCA Workshop on Speech Synthesis, 147-151, 1998.

[16] Yoon, Kyuchul, The effects of prosody on segmental variations, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003), 523-527, 2003.