



# 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, OH 43210, USA*

Received 15 June 2004; received in revised form 6 November 2004; accepted 18 January 2005

Available online 9 February 2005

---

## 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 (Han, Chunghye, Ko, Eon-Suk, Yi, Heejong, Palmer, M., 2002. Penn Korean Treebank: development and evaluation. In: Proceedings of the 16th Pacific Asian Conference on Language and Computation. Korean Society for Language and Information.) and K-ToBI prosodic labeling conventions (Sun-Ah, J., 2000. K-ToBI (Korean ToBI) labelling conventions. Version 3.1. Available from: URL <<http://www.linguistics.ucla.edu/people/jun/ktobi/K-tobi.html>>.), 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 corpus (6936 words or 21,436 syllables, with 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).

© 2005 Published by Elsevier Ltd.

---

## 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 also

---

\* Tel.: +1 614 459 3677.

E-mail address: [kyoon@ling.osu.edu](mailto:kyoon@ling.osu.edu).

unintelligible. This is partly because the phrasing module in a TTS system affects other modules, such as the intonation and duration modules. The phrase-final lengthening effects of prosodic phrasing are well known, but several recent studies demonstrate that prosodic phrasing also influences segmental properties.

Segmental contrasts in a given language have traditionally been investigated based on word tokens produced either in isolation or embedded in carrier sentences. For example, the voicing contrast of /p/ versus /b/ in English has been accounted for mainly in terms of their voice onset time (VOT) in word-initial position. However, if we look at the same segments in various other utterance positions, we readily see that their acoustic correlates consistently change. For example, the voicing contrast of English /p/ and /b/ that can be accounted for by means of VOT in word initial positions is “neutralized” in word-medial post-stress position, e.g. rapid versus rabid. Although the VOT parameter has been neutralized in that position, other parameters, such as the stop closure duration, which could not manifest itself in word initial position, now “takes over” the contrastive role played by word initial VOT. If a multitude of acoustic parameters constitute a given contrast in a language, the role of each parameter might be different in its acoustical and perceptual magnitude in different positions.

In a study on the relationship of English /h/ and glottal stop with the prosodic structure, [Pierrehumbert and Talkin \(1992\)](#) demonstrated that the pronunciation of /h/ and glottal stop depends on word- and phrase-level prosody. Their experiment showed that being in an accented syllable or adjacent to a phrase boundary in English increases gestural magnitude, making vowels more vocalic and consonants more consonantal.

In another study with American English /z/, [Smith \(1997\)](#) claimed that /z/ is produced with varying amount of devoicing depending on whether it is syllable-final, word-final, or sentence-final. In unstressed syllables and ends of words or phrases, the devoiced /z/s and voiceless /s/s seemed to be distinguished not by the amount of voicing, but by other acoustic cues, such as the duration of the fricatives and of the preceding vowels and aerodynamic cues, such as the air flow of the fricatives. All these phenomena support the idea that the phonetic realization of phonemes is dependent on the prosodic category to which they belong.

Work on Korean also shows that segmental properties are affected by the prosody of an utterance. A brief summary of the prosodic hierarchy of Seoul Korean follows. 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) ([Jun, 1993](#)). An IP is marked by a boundary tone and phrase-final lengthening with a sense of pause, which is optional. An AP, which is smaller than an IP but larger than a PW, is marked by a phrasal tone. 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.

Prosodic categories such as accentual phrases in Korean have been shown to act as the domain of application of such rules as post-obstruent tensing and vowel shortening ([Jun, 1998](#)). In a study on the effect of prosodic domains on segmental properties of three Korean coronal stops /t, t<sup>h</sup>, t<sup>\*l</sup>, [Cho and Keating \(2001\)](#) showed that initial consonants in higher prosodic domains are articulatorily stronger and longer than those in lower domains: the former has more linguopalatal contact and is longer in duration than the latter. Acoustic properties, such as VOT, total voiceless

---

<sup>1</sup> \* stands for tense.

interval, percent voicing during closure, and nasal energy minimum were also found to vary with prosodic position. Keating et al. (1998) looked at four languages, including Korean, and found that speakers of all the languages studied distinguished the IP level from a lower phrase level by the peak linguopalatal contact of the domain-initial consonants.

Prosodic effects on segments have also been found in Korean fricatives. In her study on Korean coronal fricatives, Kim (2001) observed prosodic effects on segmental properties. Linguopalatal contact was greater, acoustic duration was longer, centroid frequency was higher, and  $H_1 - H_2$  (the difference between the values of the first and second harmonics) for /s\*/ was lower in higher domains than in lower domains. Her speakers distinguished at least two levels of prosodic positions, usually between IP-initial and AP-medial, in more than one parameter.

Yoon (2003) looked at two Korean voiceless coronal fricatives and found that each fricative in different prosodic positions displayed characteristics that appear to signal its prosodic location by means of durational differences. For /s\*/, fricative noise duration was longest for PW-medial position and shortest for AP positions, with IP-initial position lying in between, whereas aspiration noise duration was nearly the same for all positions. For /s<sup>h</sup>/, however, both fricative and aspiration noise duration seem to play a role in signaling prosodic positions, i.e., both noise segments were longest for IP-initial position, whereas for the rest of the positions, fricative noise duration was about the same while aspiration noise duration was slightly longer for AP positions than for PW-medial position. In all cases, there was a trade-off between the duration of the following vowel and that of the aspiration noise duration.

Motivated by these findings, we have started working on a diphone-based concatenative TTS system within the Festival TTS framework (Taylor et al., 1998). Our aim is to build a prosodically conditioned diphone database. We believe that prosodic effects on segments can be properly modeled by recording different diphones in different prosodic contexts. For example, a *p*–*a* diphone can be collected from four different prosodic positions in Korean: initial to an IP, an AP, a PW and medial to a PW. We are working on building a diphone database from a prosodically labeled (following K-ToBI conventions) database of read speech. In order to select appropriate prosodic diphones at synthesis runtime, it is very important that we establish an accurate prosodic phrasing model. We employed the classification and regression trees (CART) (Breiman et al., 1984) algorithm, one that has proved very successful in building phrase break models in earlier studies, such as Hirschberg and Prieto (1996) and Navas et al. (2002).

This paper is organized as follows. First, Section 2 details the corpora used in this study and the decision tree training. Section 3 then presents the test results and errors produced by the models, and Section 4 discusses the results.

## 2. Methods

### 2.1. Corpora

A 400-sentence subset of the Korean Newswire Text Corpus (Linguistic Data Consortium, LDC) was read by two native speakers of Seoul Korean (one man, the author, and one woman). Recording of an additional 487 sentences from YTN News Channel broadcasts was also included. The first corpus contained 28,666 syllables or 9246 words, with an average of 72 syllables or 23

words per sentence, while the second had 23,615 syllables or 7092 words, with an average of 48 syllables or 14 words per sentence.

The unit *word* was defined as follows. Since Korean lexical items can be inflected with prefixes, suffixes, postpositions, tense morphemes, etc., by *word* we mean a fully inflected lexical item, in Korean, an *eojeol*. In practice, an *eojeol* corresponds to a space-delimited orthographic word unit. The unit syllable was defined as represented in the Korean writing system *hangul*. Hangul is an alphabetic writing system (King, 1996). It has symbols for consonants and vowels, which are grouped into orthographic syllables. The boundaries of the orthographic syllables do not always match those produced by speakers because resyllabifications can occur. The resyllabification process can easily be predicted and the morphophonological processes that trigger it are well known in Korean phonology. However, the number of syllables do not change by the resyllabification process.

One major difference between the two text corpora is the average length of a sentence. Although the two corpora are both news articles, they differ very much in their style because the Newswire articles were intended for readers while the YTN articles were intended for reporters to speak. Newspaper articles that were intended for readers are usually long. Two or more sentences frequently get combined by conjunction or subordination. Nouns get modified by relative clauses. Adverbial phrases are usually longer. By contrast, news articles intended for speakers are usually concise and avoid modifications that would otherwise make them unclear and hard to understand. These differences make the two corpora very different although they can be said to belong to the same category of news articles.

The Korean Newswire text corpus was read by two speakers because we intend to build a diphone-based concatenative TTS system for the reading style of one of the speakers. The YTN corpus was read by many different television reporters. The corpus was recorded from the online video archives of the news channel in order to see how the multi-speaker corpus would compare 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 (Han et al., 2002). 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 numerals to reflect our observation that Korean speakers usually put breaks after multiples of 10 (labeled as NNU) and 10,000 (labeled as NUE). The full POS tags are given in Table 1. The Treebank also uses phrase structure annotation for syntactic bracketing, similar to the annotation schemes used by the Penn English Treebank (Marcus et al., 1993). The bracketing tagset for the Treebank is 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. (2002) at the

Table 1  
Summary of tagset used

Category	Tag description	Tag label
Noun	Proper noun	NPR
	Common noun	NNC
	Dependent noun	NNX
	Personal/demonstrative pronoun	NPN
	Ordinal/cardinal/numeral	NNU
	Numeral (multiples of 10)	NUU <sup>a</sup>
	Numeral (multiples of 10,000)	NUE <sup>a</sup>
	Words in foreign characters	NFW
Predicate	Verb	VV
	Adjective	VJ
	Auxiliary predicate	VX
Adverb	Constituent/clausal adverb	ADV
	Conjunctive adverb	ADC
Adnominal	Configurative/demonstrative	DAN
Postposition	Case	PCA
	Possessive	PAN
	Adverbial	PAD
	Conjunctive	PCJ
	Auxiliary	PAU
Copula		CO
Ending	Final	EFN
	Co-/sub-ordinate/adverbial/complementizer	ECS
	Auxiliary	EAU
	Adnominal	EAN
	Nominal	ENM
	Pre-final (tense, honorific)	EPF
Affix	Suffix	XSF
	Prefix	XPF
	Verbalization	XSV
	Adjectivization	XSJ
Comma		SCM
Termination	Sentence end markers	SFN
Left quotation mark		SLQ
Right quotation mark		SRQ
Symbols	Others	SSY

<sup>a</sup> Extended tags.

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 (Marcus et al., 1993), 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 (Sun-Ah, 2000). The labeling proceeded in two steps. First, each of the recorded sentence was loaded into Praat (Boersma, 2001) and manually segmented by the prosodic word. Then a Praat script automatically assigned default tone labels, which were manually corrected by the human 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 (Taylor et al., 1999) 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 these features is 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 beginning,
  - distance in syllables to the sentence end,
  - distance in *eojeols* from the sentence beginning,
  - distance in *eojeols* to the sentence end.

The two syntactic features correspond to the lowermost phrasal category and the phrasal category of its mother node, respectively, in a syntactic tree diagram. For example, the terminal (i.e., the lowermost in the syntactic tree) phrasal category of the English phrase “in the park” would be an NP whereas the pre-terminal (i.e., the mother node of the lowermost phrasal category, therefore higher in the syntactic tree) phrasal category would be a PP. Commas in Korean texts are rather sparsely used, especially in news articles. However, our informal

observation shows that when they are present, they strongly indicate the presence of phrase breaks or pauses.

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. We proceeded in four steps. First, 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. We built models with the training set, tested them against the held-out set, and found that the stop value of 10 performed optimally.

Second, 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 performed 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 final step, we varied the size of the tagsets following [Taylor and Black \(1998\)](#) 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 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 2 and 3.

We also tested the performance of our cross-validated models on a 60-sentence subset of the YTN corpus and the averaged recall/precision values are given in Table 4. We did this because we think it is also important to test the performance of our cross-validated models on a different text of the same genre. As explained earlier, the YTN corpus was a very good candidate for this evaluation because it was readily available to us and its style was very different from the Newswire corpus, on which we built our cross-validated models. Performance drops are observed overall but not as much for the precision values of non-breaks and intonational breaks, and the recall values of the total breaks. It would be ideal to test our models on different text genres as well,

Table 2  
Averaged confusion table 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 3  
Comparison of our cross-validated recall/precision values with other studies

Break type	Current	Kim et al. (1999)	Lee (2000)	Kwon et al. (2002)
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

Table 4  
Averaged recall/precision values of our cross-validated model on a 60-sentence subset of the YTN corpus

Break type	Recall	Precision
Non-break	85.9	94.4
H% (IP)	0.0	N/A
HL% (IP)	49.5	27.7
HLH% (IP)	0.0	N/A
L% (IP)	9.3	95.3
LH% (IP)	0.0	N/A
LHa (AP)	72.8	50.9
IP (total)	60.1	96.0
Breaks (total)	90.7	77.7

but unfortunately we do not have any corpus available to us at the moment. We wish this evaluation could be done in the near future.

### 3.1. Error analysis

The types of errors made by our model are compared with those from other studies in Table 5. Noting the fact that IP breaks are frequently associated with a pause, the errors of any type involving an IP 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, an improvement over other studies, except for Kim et al. (1999). However, the model of Kim et al. (1999) performed worse than our model in predicting breaks/non-breaks (Table 3). The model built by Kwon et al. (2002) performs better in predicting AP breaks (Table 3). However, he used IP breaks predicted from his model 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,

Table 5  
Types of errors

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%
Deletion	AP	3.19%	7.19%	2.12%
	IP	0.24%	0.04%	9.39%
Substitution	AP ⇒ IP	3.02%	4.50%	
	IP ⇒ AP	3.04%	2.01%	

Numbers in parentheses indicate the total counts of (non-)breaks for evaluation.

followed by the distance in syllables from either the preceding or the following comma. It thus appears that the POS information plays an important role in 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 (6936 eojeols or 21,436 syllables, an average of 72 syllables or 23 words per sentence) and tested against a 50-sentence corpus (1091 eojeols or 3320 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 overall 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 et al. (1999) built a prosodic phrasing model for Korean by training it on a 300 sentence corpus uttered by three speakers. In another study, Lee (2000) trained his model on a 240 sentence corpus (2286 eojeols, collected from different genres and uttered by a trained professional) and tested against a 160 sentence corpus (1438 eojeols). Their recall/precision values are given in Table 3. Lee followed a theoretical framework proposed by Lee (1989) 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 on which we trained our models were from a single 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. Another significant aspect of our study is that we followed the K-ToBI prosodic labeling conventions (Sun-Ah, 2000) for annotation of the corpora, which will facilitate performance comparisons of similar subsequent studies. Also, with increasing availability of syntactically parsed and K-ToBI labeled corpora, further performance enhancement can be expected.

#### Acknowledgments

The author thanks 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.

## References

- Boersma, P., 2001. Praat, a system for doing phonetics by computer. *Glott International* 5, 9–10.
- Breiman, L., Friedman, J.H., Olsen, R.A., Stone, C.J., 1984. *Classification and Regression Trees*. Chapman & Hall.
- Cho, Taehong, Keating, P.A., 2001. Articulatory and acoustic studies of domain-initial strengthening in Korean. *Journal of Phonetics* 29, 25–34.
- Han, Chunghye, Ko, Eon-Suk, Yi, Heejong, Palmer, M., 2002. Penn Korean Treebank: development and evaluation. In: *Proceedings of the 16th Pacific Asian Conference on Language and Computation*. Korean Society for Language and Information.
- Hirschberg, J., Prieto, P., 1996. Training intonational phrasing rules automatically from English and Spanish text-to-speech. *Speech Communication* 18, 281–290.
- Jun, Sun-Ah, 1993. The phonetics and phonology of Korean prosody: intonational phonology and prosodic structure. Ph.D. Thesis, The Ohio State University.
- Jun, Sun-Ah, 1998. The accentual phrase in the Korean prosodic hierarchy. *Phonology* 15 (2), 189–226.
- Jun, Sun-Ah, 2000. K-ToBI (Korean ToBI) labelling conventions. Version 3.1. Available from: URL <<http://www.linguistics.ucla.edu/people/jun/ktobi/K-tobi.html>>.
- Keating, P.A., Cho, Taehong, Fougeron, C., Hsu, Chai shune, 1998. Domain-initial articulatory strengthening in four languages. *LabPhon* 6, 145–163.
- Kim, Sahyang, 2001. The interaction between prosodic domain and segmental properties: domain-initial strengthening of fricatives and post obstruent tensing rule in Korean. Master's thesis, UCLA.
- Kim, Yeon-Jun, Byeon, Heo-Jin, Oh, Yung-Hwan, 1999. Prosodic phrasing in Korean; determine governor, and then split or not. In: *Proceedings of Eurospeech'99*, pp. 539–542.
- King, R., 1996. Korean writing. In: Daniels, P.T., Bright, W. (Eds.), *The World's Writing Systems*. Oxford University Press, pp. 218–227.
- Kwon, Ohil, Hong, Munki, Kang, SunMee, Shin, Jiyoung, 2002. AP, IP prediction for corpus-based Korean text-to-speech. *Journal of Speech Sciences* 9 (3), 25–34.
- Lee, H.B., 1989. *Standard Korean Pronunciation*. Educational Science Press.
- Lee, Sangho, 2000. Tree-based modeling of prosody for Korean TTS systems. Ph.D. Thesis, Korea Advanced Institute of Science and Technology (KAIST).
- Linguistic Data Consortium (LDC), 2000. Korean Newswire Text Corpus. Available from: URL <<http://www ldc.upenn.edu/>>. Catalog number LDC2000T45, ISBN 1-58563-168-X.
- Marcus, M., Santorini, B., Marcinkiewicz, M., 1993. Building a large annotated corpus of English. *Computational Linguistics* 19 (2), 313–330.
- Navas, E., Hernaez, I., Sanchez, J.M., 2002. Assigning phrase breaks using CART's in Basque TTS. In: *Proceedings of the 1st International Conference on Speech Prosody, Aix-en-Provence*, Plenum Press, pp. 527–531.
- Pierrehumbert, J., Talkin, D., 1992. Lenition of /h/ and glottal stop. *Papers in Laboratory Phonology* 2, 90–117.
- Smith, C.L., 1997. The devoicing of /z/ in American English: effects of local and prosodic context. *Journal of Phonetics* 25, 471–500.
- Taylor, P., Black, A.W., 1998. Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language* 12, 99–117.
- Taylor, P., Black, A.W., Caley, R., 1998. The architecture of the festival speech synthesis system. In: *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pp. 147–151.
- Taylor, P., Caley, R., Black, A.W., King, S., 1999. *Edinburgh Speech Tools Library*. Available from: URL <<http://www.cstr.ed.ac.uk/projects/speechtools/>>.
- Yoon, Kyuchul, 2003. The effects of prosody on segmental variation. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, Borovets, Bulgaria.