Proc. EuroSpeech 2001 Aalborg Denmark

Linguistic Factors Affecting Timing in Korean With Application to Speech Synthesis

Hyunsong Chung and Mark A. Huckvale

Department of Phonetics and Linguistics
University College London, U.K.
{hchung, mark}@phonetics.ucl.ac.uk

Abstract

This paper describes the results of a study of the phonetic and phonological factors affecting the rhythm and timing of spoken Korean. Stepwise construction of a CART model was used to uncover the contribution and relative importance of phrasal, syllabic, and segmental contexts. The model was trained from a corpus of 671 read sentences, yielding 42,000 segments each annotated with 69 linguistic features. On reserved test data, the best model showed a correlation coefficient of 0.73 with a RMS prediction error of 26 ms. Analysis of the classification tree during and after construction shows that phrasal structure had the greatest influence on segmental duration. Strong lengthening effects were shown for the first and last syllable in the accentual phrase. Syllable structure and the manner features of surrounding segments had smaller effects on segmental duration. The model has application within Korean speech synthesis.

1. Introduction

The aim of this paper is to study the timing of spoken Korean in order to contribute to the improvement of the “naturalness” of Korean speech synthesis. Though many researchers are actively investigating the intonation of Korean for text-to-speech (TTS) systems, the research on the duration of Korean has been limited to the study of phonemic contrasts between vowel segments, segment durations in controlled contexts, and choice of the rhythm unit. Only a few research results [1][2] deal with the rhythmic patterns across sentences required for implementation in TTS systems.

This paper concentrates on duration modelling within a news-reading speech style. We collected 671 read sentences from one speaker of standard Korean. The phonological features of each segment and the context of each segment in the prosodic phrase structure were marked with 69 segmental and phrasal features. Statistical modelling explored the relationships between these features and the realised duration. A CART (Classification And Regression Tree) model was used and evaluated in the material. Objective quality of the modelling was evaluated by root mean squared prediction error (RMSE) and the correlation coefficient between actual and predicted durations in reserved test data.

This paper also explores the linguistic basis for the models. It investigates how the segmental and prosodic contexts combine to best predict the duration.

2. Design of corpus

In duration modelling, annotated speech data is used to establish the statistical relationships between the durations of the segments and the contexts in which they occur. Since these durations tend to be quite variable and the number of contexts tends to be great, a large amount of data is required. Furthermore, as pointed out in [3], [4] and [5], among others, the speech data should be from one individual to obtain a coherent pattern of variation in context. Control over speaking style also helps to reduce variability. In this experiment, we worked only with a news-reading style. News texts seemed most appropriate for speech synthesis applications, because they are factual and dense in information.

2.1. Material

We collected news scripts from two main Korean broadcasting stations: KBS (Korea Broadcasting System) and MBC (Munhwa Broadcasting Corporation). A male speaker of modern standard Korean recorded the scripts and we chose 671 sentences after removing speech errors and those utterances which seemed less grammatical. We divided the sentences among three groups: 80% went into the training data set (42,103 segments in 535 sentences), while 10% went into the evaluation data set (5,299 segments in 68 sentences), and 10% into the test data set (5,438 segments in 68 sentences). The distribution of the 42,103 segments in the training data is shown in Table 1. Phonetic transcription was generated from a dictionary and a set of rules and alignments which were performed automatically and then hand-checked. There was a very similar pattern of distribution, mean duration and standard deviation in the evaluation and test data sets.

Phone	Counts	%	Mn.	sd.
i	3650	8.67	58	37.40
u	1223	2.90	52	30.41
e	1176	2.79	92	50.07
o	1831	4.35	81	49.70
ɛ	1021	2.43	75	37.39
a	3786	8.99	86	44.27
ʌ	1725	4.10	75	40.29
ɯ	2264	5.38	49	27.41
wa	339	0.81	94	58.54
we	291	0.69	71	30.42
wi	106	0.25	86	42.75
wʌ	150	0.36	83	37.08
ja	84	0.20	101	42.28
je	86	0.20	87	32.23
jo	188	0.45	82	40.95
ju	207	0.49	80	38.04
jʌ	895	2.13	78	33.78
ɰi	49	0.12	111	53.03
m	1779	4.23	56	23.62
n	4399	10.45	62	39.66
ŋ	1572	3.73	69	27.95
l	1363	3.24	67	34.00
ɾ	1155	2.74	30	9.20
p^h	287	0.68	88	29.19
p	1179	2.80	53	25.10
p’	57	0.14	61	21.83
t^h	294	0.70	88	28.08
t	1952	4.64	49	22.10
t’	264	0.63	68	16.97
k^h	247	0.59	93	23.59
k	2839	6.74	57	33.19
k’	314	0.75	70	23.44
ts^h	503	1.19	101	30.35
ts	1458	3.46	68	33.47
ts’	191	0.45	72	16.39
s	1679	3.99	75	29.03
s’	602	1.43	104	20.23
h	898	2.13	45	24.53

Table 1. Distribution of segments in the training data set. Mn. = mean duration in ms; sd. = standard deviation in ms.

2.2. Database processing

We annotated the transcription for prosodic boundaries. Four prosodic boundaries are assumed in this paper: utterance (UTT), intonational phrase (IP), accentual phrase (AP) and phonological word (PW). Each sentence has the default UTT boundary in the starting and the end point of the sentence. When we found a clear pause in the actual utterance, we not only marked the pause in the annotation file in the speech data, but also put the IP boundary in the pronunciation string. Based on the fundamental frequency contour in the speech data, AP boundary was also marked in the pronunciation. Each AP has an underlying tonal pattern of LHLH which is sometimes phonetically realised as LH in a short AP [6]. Then PW boundary was also indicated in the pronunciation. The PW is a morphological and syntactic unit which is demarcated by one content or functional word with one or more suffixes, case particles, or endings. When more than two prosodic phrase structure occur in the same place, we only marked the higher prosodic structure. For example, each IP-initial position was also AP-initial and PW-initial position.

The annotated transcription was processed into a hierarchical prosodic structure encoded in XML comprising UTT, IP, AP, PW, syllable, onset, rhyme, nucleus and coda nodes as well as segments, which are described using features. The feature string of each segment was then automatically generated from the phonological structure using the ProXML scripting language [7]. The script looked at each segment in turn and constructed the binary feature string from the properties of the target segment, the properties of its neighbours and its position in the prosodic structure. Each segment was annotated with the following features together with the actual duration:

- phonemic identity of the target segment, i.e. segment name, or phonemic features of the target segment, i.e. major class features of the segment
- phonological features of the preceding and the following segments
- syllable structure: position and structure of containing syllable
- position of syllables in the Utterance (UTT), Intonational Phrase (IP), Accentual Phrase (AP) and Phonological Word (PW)

Overall, each segment was annotated with total of 69 features.

3. Analysis of corpus

3.1. Classification And Regression Tree (CART) Model

In this experiment, we wanted to establish which context features were most important, and so we built CART trees [8] in a *stepwise* fashion. In this approach each single feature is taken in turn and a tree consisting of nodes only asking questions of that feature is built. The single best tree is then kept and each remaining feature is taken in turn and added to the tree to find the best tree possible with just two features. The procedure is then repeated for a third, fourth, fifth feature and so on. This process continues until no significant gain in accuracy is obtained by adding more features. We used the *Wagon CART building program* [9] as a tool for running this CART tree building process.

Using the binary feature string of each segment, we present below three approaches to CART modelling of the news database. The first allows the tree to ask questions based on the name of the segment; this gives good performance but a tree which is less easy to interpret. The second restricts questions on the tree to features of the target segment but not its name; it is hoped that this will force generalisations across segment types. The third replaces the millisecond duration values with durations calculated in z-scores of the log duration value of each segment type. The idea is to remove from the tree any influences caused by differences in inherent duration and variability of segment type.

3.2. CART using segment names

A stepwise CART model was trained using all 42,103 segments in the training data set described by the name of each segment and 61 segmental and prosodic phrasal features describing the context. Training ended when additional features made no significant improvement in performance; this was after 46 features were incorporated. This tree was then 'pruned' by removing questions and pooling leaf nodes so that the performance of the tree on the evaluation data set was maximised. The tree was pruned back to 26 features in this process. Finally the correlation between actual and predicted durations and the mean squared error of prediction was found for the training set, the evaluation set and the test set as shown in Table 2.

Data set	**RMSE (ms)**	**Correlation**	**Feature numbers**
Training	24.23	0.77	46
Evaluation	24.04	0.79	26
Test	26.48	0.73	26

Table 2. CART model performance using segment names.

Not surprisingly, the single most important feature in the model is the identity of the segment being predicted. Other features were dominated by the prosodic phrase features and syllable structure features. The second most important feature, AP-final position feature had a large effect, followed by the AP-initial position feature, onset position feature, CVC syllable structure feature, and preceding voiceless feature. Subsequent features had much less effect, the 9th feature only improving the correlation coefficient by 0.01.

3.3. CART using segment class features

The same stepwise CART model training procedure was carried out using class features instead of names. From 52 features found from the training data, the tree was pruned back to 36 features using the evaluation data.

Data set	**RMSE (ms)**	**Correlation**	**Feature numbers**
Training	25.67	0.74	52
Evaluation	26.04	0.76	36
Test	27.98	0.70	36

Table 3. CART model performance using segment class features.

The most important feature was the AP-final feature, followed by onset position, AP-initial, CVC syllable structure, preceding voiceless, utterance-medial position, following nasal, flap consonant, nasal consonant, and plosive features. No place of articulation features appear in the top 10. The growth in the correlation coefficient levels off rapidly after 5 features. When the 9th feature was added to the CART, the correlation coefficient improved only by 0.01. Many parallels can be drawn with the previous CART model, where the most important features were dominated by the prosodic phrase features and syllable structure features.

3.4. CART using z-scores of segments

In this CART modelling, we first converted each duration into log duration in ms. Then we transformed each log duration to a z-score using the mean and standard deviation log ms for each segment type. The log transform was used to create more normal probability distributions for duration. In the CART model, a positive z-score corresponds to durations longer than the mean and a negative z-score corresponds to durations shorter than the mean. Because z-scores encode the inherent properties of each segment, the segment names were not used in this model.

Data set	**RMSE (z-score)**	**Correlation**	**Feature numbers**
Training	0.73	0.67	54
Evaluation	0.74	0.66	36
Test	0.77	0.63	36

Table 4. CART model performance using z-scores.

Among the most influential 10 features, five of them were prosodic phrase features and syllable structure features; four of them were manner features; one was a place feature. The most important feature was the AP-final position feature, followed by onset position feature, AP-initial position feature, preceding aspiration feature, preceding nasal feature and following nasal feature. Subsequent features had less effect, only improving the correlation coefficient by 0.01 or less.

Using the results on the training data, we calculated the mean z-score changes arising from each selected feature acting on its own. These are given in Table 5.

Feature	1_AP	ON	AP_1	asp_	nas_
z-score	0.86	-0.04	0.35	-0.44	-0.17
Feature	_nas	PW_1	CVC	nas	_cor
z-score	-0.29	0.07	-0.09	-0.01	-0.04

Table 5. Mean z-score changes of selected features in

the training data.

When the segment is in AP-final position (1_AP), the segment has the positive z-score 0.86, so it has a large lengthening effect. The lengthening effects of the sentence-final feature (z-score 0.85) and the IP-final feature (z-score 0.98) can be seen to be largely due to the fact that these boundaries are also marked by the AP-final feature. This explains why the sentence-final and the IP-final feature do not appear in the top 10 features. Also in this table, the AP-initial position feature (AP_1) and the PW-initial position feature (PW_1) have a lengthening effect. The onset position feature (ON), preceding aspiration feature (asp_), preceding nasal feature (nas_), following nasal feature (_nas), CVC syllable structure feature (CVC), nasal segment feature (nas) and following coronal feature (_cor) all have a shortening effect.

4. Discussion

These results can be compared to previous analyses of the rhythmic pattern of spoken Korean.

- In all three CART models, the AP boundary has a significant effect on the segment duration. The feature AP-final has a lengthening effect. Neither the sentence-final feature nor the IP-final feature has a significant effect. All final lengthening can be interpreted as AP-final lengthening. Other studies found that a vowel in sentence final position [3][10] or a vowel in IP-final position [5][11][12] is longer than in other positions. It is possible that because their data was restricted to constrained carrier phrase sentences, they failed to find a generalization of duration patterns over different sizes of prosodic units.

- In this CART model, a CVC syllable indeed has a shortening effect with a mean z-score of -0.09. [3] and [13] found that vowels in CVC syllable structure are much shorter than those in CV or V syllable structures.

- This CART model shows that the presence of a preceding aspirated segment has a shortening effect, as found in [1], [3] and [14]. It is believed that the wide opening of the glottis in the articulation of aspirated consonants shortens the following segments. We made aspiration part of stop, whereas this result is evidence that it might have been better to label it as part of the vowel (syllable nucleus).

- Nasals seem to have an interesting influence on the durations of adjacent segments. Although sonorants are generally thought to have a lengthening effect, we have found evidence of segment shortening both before and after nasal consonants. This is in partial agreement with [1] for Korean, where shortening after nasals was observed; and also with [4] for English, where shorter vowels before nasals was seen.

The prediction error and correlation coefficients found are comparable with recent published results in Korean [2]. In their CART modelling of spoken Korean on segmental duration, [2] trained on 240 sentences (15,037 segments) and tested on 160 sentences (9,494 segments). Their RMSE was about 22 ms, and the correlation coefficients was about 0.82. They used the segment names of surrounding segments and of the observed segment in question, the part-of-speech features of the word, and the position features of the segment in the prosodic phrase, and the length of the prosodic phrases in syllables. In another regression tree modelling of spoken Korean using 15 sentences by three male and four female speakers in three different tempos, [1] showed correlations between 0.74 and 0.69 and an RMSE of less than 25 ms.

Language	Experiment	Correlation
Korean	*this paper*	0.73
Korean	[2]	0.82
Korean	[1]	0.74

Table 6. Comparisons between CART model result in this paper and other results.

5. Conclusion

This paper used stepwise CART modelling to analyse the timing of spoken Korean in a connected news-reading speech style. The advantage of the stepwise approach is that the relative importance of contextual features to the duration of segments can be quantified. It was found that prosodic phrase features had the most influence, among them AP final and AP initial features. The syllable structure and the manner features of surrounding segments were less important. The place features of surrounding segments had little influence. Although the stepwise approach puts more constraints on the CART model, the overall performance of duration estimation was comparable with other results in Korean.

We believe the results in this paper are more representative of the timing of spoken Korean of a news-reading speech style, because the data set constructed for this experiment is larger and with more variability in sentence length than earlier studies. The results can be directly applied within the duration prediction component of a Korean speech synthesis system [15].

Acknowledgements

We thank Tae-yeoub Jang and Weonhee Yun for help with the generation of phonetic transcriptions from a pronunciation dictionary.

References

[1] Lee, Y. H., “Modelling of segmental duration in Korean speech synthesis”, *Phonetics and Linguistics in honour of Professor Hyun Bok Lee*, 249-274, 1996.

[2] Lee, S. H. and Oh, Y. H., "Tree-based modelling of prosodic phrasing and segmental duration for Korean TTS systems", *Speech Communication* 28(4): 283-300, Elsevier Science, 1999.

[3] Han, M., *Studies in the Phonology of Asian Languages II—Duration of Korean Vowels*, University of Southern California, 1964.

[4] Lehiste, I., *Suprasegmentals*, The MIT Press, Cambridge, 1970.

[5] Lee, H. Y., *The Structure of Korean Prosody*, Ph.D. Thesis, University of London, 1990.

[6] Jun, S. A., "The accentual phrase in the Korean prosodic hierarchy", *Phonology*, 15: 189-226, 1998.

[7] Huckvale, M. A., "Representation and processing of linguistic structures for an all-prosodic synthesis system using XML", *Proceedings of Eurospeech '99*, 4: 1847-1850, 1999.

[8] Breiman, L., Freidman, J., Olshen, R. and Stone, C., *Classification and Regression Trees*, Chapman and Hall/CRC, 1998.

[9] Taylor, P., Caley, R., Black, A. and King, S., *Edinburgh Speech Tools Library*, University of Edinburgh, 1999.

[10] Lee, S. H. and Koo, H. S., "The effects of the speaking rate on the duration of syllable before boundary", *Korean Journal of Speech Sciences*, 1: 103-111, 1997.

[11] Chung, K. et al., *A Study of Korean Prosody and Discourse for the Development of Speech Synthesis/Recognition System*, Korea Telecom Research and Development Group, 1996.

[12] Jun, S. A., *The Phonetics and Phonology of Korean Prosody*, Ph.D. Thesis, The Ohio State University, 1993.

[13] Koo, H. S., "The influence of consonant environment upon the vowel duration", *Korean Journal of Speech Sciences*, 4(1): 7-18, 1998.

[14] Kang, K. S., "On Korean fricatives", *Korean Journal of Speech Sciences*, 7(3): 53-68, 2000.

[15] Chung, H., Huckvale, M. A. and Kim, K., "A new Korean speech synthesis system and temporal model", *Proceedings of 16th International Conference on Speech Processing*, 1: 203-208, 1999.