Duration Models and the Perceptual Evaluation of Spoken Korean

Hyunsong Chung

Department of Computer Science
University College Dublin
hyunsong.chung@ucd.ie

Abstract

This paper builds predictive models of segment duration in context based on the CART models and "additive-multiplicative" models for Korean text-to-speech. It uses a corpus of 670 read sentences collected from one speaker of standard Korean. The best performance was obtained from a CART decision tree model, which shows that the correlation between the observed and the predicted durations is 0.77 and the mean squared error of prediction is 25.11 ms. Linguistic implications of these models are also discussed. The perceptual evaluations of these models are carried out using a Korean language diphone database based on the MBROLA synthesis system in order to investigate the clarity and the listener preference for durations.

1. Introduction

There have been very few studies in the analysis and modelling of the prosody of Korean, particularly in the area of segmental durations. Those that have been conducted are poorly suited to the issues in contemporary speech synthesis systems. This paper sets out to perform a new analysis and modelling of Korean segmental duration. These studies are based on previous work where possible, but extended to take into account the demand of contemporary approaches to duration modelling as used in English and Japanese synthesis. However, this paper does not just try to build the best predictive model of segment duration in context. It also seeks to learn more about which factors and which structures are most important in Korean prosody. The outcome of this work is both a better model of Korean timing for use in synthesis, and a better understanding of the Korean language.

2. Data Corpus

2.1. Text processing

The main corpus consisted of 670 sentences spoken by one speaker in a news reading style. 80% of the sentences went into the training data set (42,103 segments in 535 sentences), while 20% went into the test data set (10,737 segments in 135 sentences). Besides the 670 sentences, an extra evaluation data set (10,609 segments in 135 sentences) was also prepared for evaluating the CART (Classification and Regression Tree) models [1]. The sentences are news broadcasts from the two main Korean broadcasting stations. The phone alignments were performed automatically using the Hvite program from the Hidden Markov Model Toolkit [2] and then hand-checked. Each phonetic transcription was parsed into a hierarchical prosodic structure in which the symbolic transcription is replaced by feature descriptions stored in tree nodes. Each pronunciation was encoded as a metrical structure comprising syllables, onset, rhyme, nucleus and coda nodes as well as the segments, which are described using features. A prosodic hierarchy consisting of utterance (UTT), intonational phrase (IP), accentual phrase (AP), and phonological word (PW) nodes was used. After this process, the hierarchical structures were stored in extensible mark-up language, XML. The hierarchical structures were then aligned with the checked annotations in the speech signal.

2.2. Prosodic structure

In this paper, UTT is always a whole sentence. IP was assumed to be demarcated by a clear pause whether or not it ends with any kind of boundary tone. AP can be demarcated by a phrase final tonal pattern of "LH" in Korean [3]. Because of the lack of a lexical stress system in standard Korean, the metrical foot was not used. A PW can contain one content word with one or more suffixes, case particles or enders.

2.3. Generation of training and test data for modelling

For the modelling process, a feature string for each segment was automatically generated from the phonological structure using the ProXML scripting language [4]. The script looked at each segment in turn and constructed a binary or n-ary 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, e.g. segment name, or phonemic features of the target segment, i.e. major class features of the segment
- phonemic features of the preceding and the following segments
- syllable structure: position and structure of containing syllable
- position of syllables in UTT, IP, AP and PW

Two groups of feature descriptions were prepared: one with general class features and their sub-levels, and the other with binary distinctive features. The first group of features, "compact feature set", treats vowels and consonants separately and was used in CART and "additive-multiplicative" modelling. The second group, "binary feature set", was used for the CART analysis only in order to investigate which distinctive features have most influences on duration.

3. Analysis

3.1. CART models for the "compact feature set"

CART duration analysis has become a common method for building classification models from simple feature data,

suggested by [1] as an alternative to heuristically derived duration prediction rules for duration modelling in synthesis. The strengths of CART modelling come from the ease with which trees may be built from duration data and from the speed of classification of new data. It also shows good performance in subjective terms. CART models cope with complex interactions because it makes very few assumptions about the structure of the data. The weakness of CART models lies in the fact that it cannot interpolate between known context to find values for unknown contexts. Another weakness of this model is that it relies on objective function for partitioning that may not be the best in a perceptual sense.

The "Wagon" CART building program [5] was used as a tool for running this CART tree building process. Firstly, CART analysis directly predicted the duration. Two separate stepwise CART models for vowels and for consonants were trained. These used 19,071 vowels and 23,032 consonants in the training data set described by the name and major class features of each segment and the segmental and prosodic phrasal features describing the context. Training ended when additional features made no significant improvement in performance. This tree was "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 then tested on 4,829 vowels and 5,908 consonants. Finally the correlation between actual and predicted durations and the mean squared error of prediction was found for the test set.

In the second CART model, the tree predicts z-scores rather than the duration directly. Each segment duration was first converted into log ms. Then each log duration was transformed to a z-score using the mean and standard deviation for each phoneme type. The log transformation was used to create more normal probability distributions for duration. Because z-scores encode the inherent properties of each segment, the names and the major class features of the target segment were not used in this model. After prediction, the segmental duration can be calculated by the formula:

$$\text{Duration} = \text{mean} + (\text{z-score} \times \text{standard deviation}) \quad (1)$$

The CART performance results for vowels and consonants using the "compact feature set" are summarised in Table 1.

Table 1: CART performance results for vowels and consonants using "compact feature set".

	Vowels		Consonants	
	RMSE	Corr.	RMSE	Corr.
Duration	27.51 ms	0.78	24.20 ms	0.71
z-score	26.01 ms	0.77	25.21 ms	0.70

3.2. Additive-multiplicative models for "compact feature set"

In [6], van Santen claimed that "additive-multiplicative" models capture the "directional invariance" of the segment duration and the factor interactions are better described by a multiplicative rule than an additive rule. The strength of this model is that with relatively few parameters, durations can be well estimated from training data. Unlike CART models, they naturally interpolate to unseen contexts. A formula is small so it is easy to apply and understand. The weakness of this approach to modelling is that it is difficult to unravel all interactions in training data and it needs a large corpus with wide variety of contexts. In this paper, firstly, the model formula was specified in terms of which factors were to be incorporated and how the parameters associated with each factor were to be combined in the model. Each parameter was then initialised to a value specified by the experimenter. Limits on the allowed range of values were established. A function optimisation strategy was then employed whereby perturbations in the values of the parameters were investigated in terms of their effects on the model performance. The simulated annealing method [7] was used for the optimisation. This paper has not attempted to derive a new "additive-multiplicative" model for Korean vowels but instead it has adapted the model reported in [8] of Japanese timing. Because feature classes of the model in [8] are different to those that have been used for Korean, they were modified as follows:

$$\begin{aligned}\text{DUR}(\text{id, man, prev, foll, syll, left\_pos, right\_pos}) = \\ S_{1,1}(\text{id}) + [S_{2,1}(\text{man}) \times S_{2,2}(\text{prev})] + [S_{3,1}(\text{man}) \times S_{3,2}(\text{foll})] + \\ [S_{4,1}(\text{foll}) \times S_{4,2}(\text{syll})] + S_{5,1}(\text{left\_pos}) + S_{6,1}(\text{right\_pos}) \quad (2)\end{aligned}$$

where "id" is the identity of the vowel, "man" is the manner feature of the target vowel, "prev" the manner of the preceding vowel, "foll" the manner of the following vowel, "syll" the syllable structure, "left_pos" the syllable distance to the left phrase boundary, "right_pos" the syllable distance to the right phrase boundary. The correlation of this model trained by the simulated annealing method [7] was 0.68 and the prediction error was 32.13 ms.

To obtain an "additive-multiplicative" model for consonants, we adapted the model used for vowels, having no specific information which might guide an alternative design. The model is:

$$\begin{aligned}\text{DUR}(\text{id, man, prev, foll, syll, syllpo, left\_pos, right\_pos}) = \\ S_{1,1}(\text{id}) + [S_{2,1}(\text{man}) \times S_{2,2}(\text{prev})] + [S_{3,1}(\text{man}) \times S_{3,2}(\text{foll})] + \\ S_{4,1}(\text{syll}) + S_{5,1}(\text{syllpo}) + [S_{6,1}(\text{man}) \times S_{6,2}(\text{left\_pos})] + \\ [S_{7,1}(\text{man}) \times S_{7,2}(\text{right\_pos})] \quad (3)\end{aligned}$$

where "syllpo" is the segment position in the syllable, i.e. onset or coda. The performance results of these models can be summarised as follows.

Table 2: Performance results summary for vowels and consonants using "additive-multiplicative" model.

Vowels		Consonants	
RMSE	Correlation	RMSE	Correlation
32.13 ms	0.68	28.86 ms	0.54

3.3. CART models for "binary feature set"

In order to explore the importance of individual levels of each factor, the feature set was extended so that each n-ary feature was replaced with a number of binary features. For example, "left_pos" and "right_pos" features in "compact feature set" were replaced by "first", "post-initial", "medial", "penultimate", and "last" phrase position features with binary values assigned for each feature. A total of 69 features were available in the data set. The procedure for this modelling is similar to that of 3.1 except for the feature set. When the duration was directly predicted from the tree, the tree was pruned back to 35 features. When the tree predicted z-scores rather than duration directly, the tree was pruned back to 40

features after the evaluation process. The performance results of these models can be summarised as follows:

Table 3: CART performance results summary for "binary feature set".

	RMSE	Correlation
Duration	25.11 ms	0.77
z-score	26.44 ms	0.74

3.4. Mean feature effect

Based on the CART decision tree in 3.3, the mean z-score changes arising from each selected feature acting on its own were calculated. We call this "mean feature effect" analysis. The objective of this analysis is to obtain from the CART tree information about the relative size of the effect of each feature on the segment duration. We know from the stepwise building of the tree which features were most important and in which order they were applied in the tree from root towards the leaf nodes. We can use this information to re-analyse the training data to establish the mean effect of each feature. The procedure is as follows: firstly the data is partitioned into two groups according to the value of the most important feature (here, 1_AP: AP-final position) and the means of each partition are calculated (1_AP=0: -0.12, and 1_AP=1: 0.87, values in z-scores). The difference between these means (0.99) is called the mean effect of feature "1_AP". Next the mean duration value of each partition is then subtracted from the individual segment durations in that partition. In effect this "takes into account" the mean operation of feature "1_AP". The data can then be partitioned according to the value of the second most important feature in the CART analysis (here, ON: onset position). This gives us two further means (ON=0: 0.04, ON=1: -0.06) and the mean effect of feature ON (-0.09). The mean values from the two partitions can be subtracted as before to take into account feature "ON", and the process repeated for the third most important feature and so on. The top 10 changes are given in Table 4.

Table 4: Mean feature effect caused by selected features in the training data.

Ranking	Feature	Partition 0		Partition 1		Diff.
		Mean	Size	Mean	Size	
1	1_AP	-0.12	37170	0.87	4933	0.99
2	ON	0.04	25704	-0.06	16399	-0.09
3	AP_1	-0.06	37041	0.47	5062	0.53
4	nas_	0.03	34478	-0.12	7625	-0.15
5	_nas	0.07	34467	-0.30	7636	-0.36
6	PW_1	-0.03	28198	0.06	13905	0.09
7	vce_	-0.12	13783	0.06	28320	0.18
8	1_PW	-0.06	27915	0.11	14188	0.16
9	CVC	0.05	24978	-0.08	17125	-0.13
10	cor_	-0.03	22082	0.03	20021	0.06

Partition 0 = mean and size of partition when feature is 0.
Partition 1 = means and size of partition when feature is 1.

Mean effect analysis gives an overall picture of the effect of the most important features, but it doesn't accurately reflect the actual operation of the tree, since it ignores interactions between features. Thus it could be that feature "ON" has a very different effect in "AP_1" positions than elsewhere. However we have found no evidence of strong interactions in the top 10 most important features. In this table, a positive mean feature effect in z-score corresponds to a lengthening effect of duration and a negative z-score is a shortening effect of duration. When the segment is in AP-final position (1_AP), the segment has the positive mean feature effect of 0.99, so it has a large lengthening effect. Also in this table, the AP-initial position feature (AP_1), the PW-initial position feature (PW_1), the PW-final position feature (1_PW), the preceding voicing feature (vce_), and the preceding coronal feature (cor_) had lengthening effects. On the other hand, the onset position feature (ON), the preceding nasal feature (nas_), the following nasal feature (_nas), and the CVC syllable structure feature (CVC) had shortening effects.

4. Discussion of the Analysis

4.1. Performance of the models

The results in the paper showed that the CART models had overall better performance than the "additive-multiplicative" models. The performance was best when the duration was directly predicted from the CART tree, where the names and the major class features of the target segment were used in the tree. The prediction error and correlation coefficients of the best CART model in this paper were comparable with the best published results in Korean [9]. The results from "additive-multiplicative" model for Korean in this analysis were worse than those for English or Japanese found in other studies. The calculation of the values for the same feature in different product terms should be investigated to improve the performance of this model.

4.2. Linguistic implication

By using a CART decision tree model with segment durations as z-scores, the linguistic implications of the model were investigated. The AP boundary had the most influence either to AP-initial or to AP-final syllables. It significantly lengthened the segment duration in the AP-final syllable. Though both the PW-initial position and the PW-final position were important in duration prediction, the PW-final position feature had more lengthening effect. UTT boundaries and IP boundaries did not contribute much to the duration once the AP boundary had been taken into account. This is believed to be partly because each UTT boundary and IP boundary is also an AP boundary and a PW boundary in the phonetic transcriptions. Though it is not described in Table 4, the results showed that shortening effect were seen in the syllables in all post-initial positions and in penultimate positions from boundaries. It shows that in Korean, the lengthening effect of the phrase does not penetrate into the syllables in these positions. The CART analysis did not find that syllable structure had a general effect on vowel duration except in the case of CVC syllable structure.

Nasals seem to have a shortening effect than homorganic post-vocalic voiced obstruents. This is explained that vowel needs a special adjustment of the vocal folds to maintain vibrations during voiced plosives [10]. Though they are not in

the top ten most important factors, aspiration and tenseness features of surrounding segments show significant shortening effect. This fact supports the idea that the glottal opening is the major controller of the vowel duration. This may also indicate that probably [stiff vocal cord] feature which covers aspirated plosives and tense plosives in Korean language has a significant shortening effect in the vowel duration. In agreement with previous studies of English and Korean, the models showed little effect caused by the place features of surrounding segments.

5. Perceptual Evaluation

Perceptual evaluation is essential to gauge the quality of synthesised speech. It is not always the case that improved statistical modelling leads to improved speech quality. The perceptual evaluation in the paper investigates the clarity and the listener preference for durations calculated by the best CART and "additive-multiplicative" models in this paper and durations calculated by a commercial Korean TTS system.

5.1. Test procedure

Nine sentences with various lengths were selected from broadcast news scripts, which were different from the data set used in the experiment. Durations were calculated by using the best CART model (model 1) and "additive-multiplicative" model (model 2). To compare the quality of the duration modelling with a commercial Korean TTS system, durations were also extracted from the ETRI (Korean Electronics and Telecommunications Research Institute) TTS demonstration system (model 3). F_0 contours for the sentences were copied from natural read versions. The duration and F_0 contour information of these models were then applied to the MBROLA Korean language diphone database "Hanmal (HN 1.4)" [11]. The synthesised speeches by the three models were played to 10 subjects for perceptual evaluation.

5.2. Results

In terms of clarity, subjects preferred model 3, where ETRI durations were used, to the other models. The CART model followed the ETRI model and the "additive-multiplicative" model was the least preferred. All differences were statistically significant at $p < 0.01$. In terms of general preference, subjects' preferences were more balanced, though "additive-multiplicative" model was still the least preferred. The CART model was most preferred by subjects, though the difference was not statistically significant. In an informal discussion among subjects, it was suggested that ETRI model was slower than other synthetic speech. ETRI had a distinctly slower speaking rate than other models. This suggests that the difference in tempo is significant perceptually, because it could explain why duration obtained by ETRI were preferred for clarity.

The fact that the general preference for the CART and ETRI durations has similar scores shows that CART duration prediction in this paper is at least as good as that of ETRI's. The "additive-multiplicative" model performed worse in both subjective and objective tests.

6. Conclusions

The analyses of this paper are believed to contribute to the study of spoken Korean in the following aspects. Firstly, it showed how much prosodic phrase features influenced duration and which of these were more important. Secondly, it showed how phonological distinctive features could be used for modelling in such a way as to allow a linguistic interpretation of the model. Thirdly, these observations allowed us to determine which factors and which structures are most important in Korean prosody.

In the course of preparing the experiments, a labelled database of spoken Korean was constructed. As a result of the experiments, a trained CART model for synthesis was obtained. Durations of segments in a new text can be rapidly predicted from this model. The "Hanmal" diphone database for Korean speech synthesis was also developed as a by-product of the perceptual testing. This database is now publicly available and currently in use by other researchers.

7. Acknowledgements

The author would like to thank Dr Mark Huckvale and Gordon Hunter at UCL (University College London) for their contribution to the text processing of the data corpus. The author also thanks Dr Tae-Yeoub Jang at Hankuk University of Foreign Studies in Korea and Weonhee Yun at CSTR for their help in the letter-to-phoneme conversion and the phone alignment. Professor Gyeongseog Gim in Busan National University in Korea is the co-author of the MBROLA Korean language diphone database, "Hanmal (HN 1.4)".

8. References

[1] Riley, M., 1992. Tree-based modelling of segmental durations, in *Talking Machines: Theories, Models and Designs*, G Bailly, C. Benôit (eds.) Amsterdam: North-Holland, 265-273.

[2] Young, S.; Jansen, J.; Ollason, J.; Woodland, P., 1996. *HTK Book*, Entropic.

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

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

[5] Black, A.; Taylor, P.; Caley, R., 1999. *The Festival Speech Synthesis System: system documentation, edition 1.4, for Festival Version 1.4.0.*, CSTR web page, University of Edinburgh.

[6] Van Santen, J.P.H., 1997. Timing in text-to-speech system. *Proceedings of Eurospeech '95*, 1397-1404.

[7] Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B., 1992. *Numerical Recipes in C*, 2nd edition. Cambridge: Cambridge University Press.

[8] Venditti, J.; Van Santen, J.P.H., 1998. Modelling segmental durations for Japanese text-to-speech synthesis. *Proceedings of the Third ESCA/COCOSDA International Workshop on Speech Synthesis*, 31-36.

[9] Lee, S.; Oh, Y., 1999. Tree-based modelling of prosodic phrasing and segmental duration for Korean TTS systems. *Speech Communication*, 28, 283-300.

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

[11] Chung, H.; Huckvale, M.; Gim, G., 1999. A new Korean speech synthesis system and temporal model. *Proceedings of 16th ICSP*, 1, 203-208.