Analysis of the Timing of Spoken Korean Using a Classification and Regression Tree (CART) Model

Hyunsong Chung* · Mark Huckvale*

ABSTRACT

This paper investigates the timing of Korean spoken in a news-reading speech style in order to improve the naturalness of durations used in Korean speech synthesis. Each segment in a corpus of 671 read sentences was annotated with 69 segmental and prosodic features so that the measured duration could be correlated with the context in which it occurred. A CART model based on the features showed a correlation coefficient of 0.79 with an RMSE (root mean squared prediction error) of 23 ms between actual and predicted durations in reserved test data. These results are comparable with recent published results in Korean and similar to results found in other languages. An analysis of the classification tree shows that phrasal structure has the greatest effect on the segment duration, followed by syllable structure and the manner features of surrounding segments. The place features of surrounding segments only have small effects. The model has application in Korean speech synthesis systems.

Keywords : timing, duration, prosody, speech synthesis, CART

1. Introduction

The quality of speech recognition and speech synthesis systems for the Korean language has been considered to be behind the standard of such technology in the major western languages, Chinese, and Japanese. It is generally believed that the poor quality of the existing commercial Korean language synthesis systems is due to weaknesses in the implementation of prosody: timing, intonation, and prosodic phrasing.

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 (Lee, 1996; Lee and Oh, 1999) 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

* Dept. of Phonetics and Linguistics, University College London

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 realized 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 Han (1964), Lehiste (1970), and Lee (1990), 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). We downloaded the script of the KBS 9 o'clock news broadcast on January 19, 2000 and that of the MBC 9 o'clock news broadcast on January 20, 2000. The KBS news script contained 412 sentences and the MBC news script contained 338 sentences. From these, 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 (535 sentences), while 10% went into the evaluation data set (68 sentences), and 10% into the test data set (68 sentences).

The sentences were not selected to modify the natural distribution of prosodic contexts in such material. Although there are coverage issues arising from the large number of possible contexts (van Santen, 1995), we felt that it would be better to have a typical sample of contexts.

The distribution of the 42,103 segments in the training data is shown in Table 1. In this

table, segment /a/ is the most frequent vowel with 3,786 occurrences (8.99%) and the segment /ɯi/ is the least frequent with 49 occurrences (0.12%). Among sonorants, /n/ is the most frequent with 4,399 occurrences (10.45%), and [ɾ] is the least frequent with 1,155 occurrences (2.74%). Among obstruents, /k/ is the most frequent with 2,839 occurrences (6.74%), and /p'/ is the least with 57 occurrences (0.14%). There was a very similar pattern of distribution and mean duration in the evaluation and test data sets.

Table 1. Distribution of segments in the training data set.

Phone	Counts	%	Mean (ms)	Phone	Counts	%	Mean (ms)
i	3650	8.67	58	n	4399	10.45	62
u	1223	2.90	52	ŋ	1572	3.73	69
e	1176	2.79	92	l	1363	3.24	67
o	1831	4.35	81	ɾ	1155	2.74	30
ɛ	1021	2.43	75	p^h	287	0.68	88
a	3786	8.99	86	p	1179	2.80	53
ʌ	1725	4.10	75	p'	57	0.14	61
ɨ	2264	5.38	49	t^h	294	0.70	88
wa	339	0.81	94	t	1952	4.64	49
we	291	0.69	71	t'	264	0.63	68
wi	106	0.25	86	k^h	247	0.59	93
wʌ	150	0.36	83	k	2839	6.74	57
ja	84	0.20	101	k'	314	0.75	70
je	86	0.20	87	ts^h	503	1.19	101
jo	188	0.45	82	ts	1458	3.46	68
ju	207	0.49	80	ts'	191	0.45	72
jʌ	895	2.13	78	s	1679	3.99	75
ɯi	49	0.12	111	s'	602	1.43	104
m	1779	4.23	56	h	898	2.13	45

2.2 Subject

The subject was a male speaker of modern standard Korean who had lived in Seoul, Korea for 16 years and had lived in London, England for the last 3 years. He was 20 years old and did not have any experience in this kind of recording. Because he is in the category of younger generation who uses modern standard Korean, he did not make phonological distinctions between long vowels and short vowels, between /we/ and /wɛ/, between /je/ and /jɛ/, between /wi/ and /y/, and between /we/ and /ø/.

2.3 Recording Procedure

The recordings were made in an anechoic chamber on digital tape using 2 channels at 44,100 samples/sec/channel. Channel 1 was the speech signal from microphone, channel 2 was a Laryngograph signal. They were resampled to 16 kHz and transferred to disk. The recordings were carried out in 12 sessions over a two-month time span. Though fewer sessions would have been ideal, the speaker found it difficult to maintain voice quality after 30 minutes of recording per day. The speaker was prompted with a script displayed on a computer monitor, and sessions were recorded without interruption. The speaker was requested to read each sentence rapidly and fluently to simulate a real news reading style. Sentences containing errors and disfluencies were repeated until a fluent utterance was produced. The naturalness of the speech was monitored based on the perceived consistency of the speech tempo, energy, and pitch range.

2.4 Database Annotation

Each segment was annotated with the following features together with the actual duration:

- phonemic identity of the target segment, that is, the segment name and the phonological 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 Phonological Word (PW), Accentual Phrase (AP), Intonational Phrase (IP), and the Utterance (UTT)

The phonological features of the target segment were decided based on the manner of articulation, rather than the place of articulation. This follows Chung and Huckvale (1999), where the place of articulation did not have significant effect on segmental duration. In order to investigate the effect of prosodic boundaries on the segment duration, we parsed each sentence into prosodic phrases. We set up a hierarchy of four levels: UTT, IP, AP, and PW. The UTT was taken to be the whole sentence, while each IP ended with a clear pause. Each AP had an underlying tonal pattern of LHLH which is sometimes phonetically realized as LH in a short AP (Jun, 1998). 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.

Details of these features are illustrated in Table 2. These 69 features and the segment names were used in this experiment.

Table 2. The 69 features used in the experiment

Feature	Description	Feature	Description
mono	monophthong	_lab	following labial
di	diphthong	_cor	following coronal
stp	plosive	_dor	following dorsal
aff	affricate	_glt	following glottal
fri	fricative	_hiV	following high vowel
nas	nasal	_mdV	following mid vowel
lat	lateral	_loV	following low vowel
fla	flap	CV	CV syllable structure
V_	preceding vowel	CVC	CVC syllable structure
vcl_	preceding voiceless	VC	VC syllable structure
nas_	preceding nasal	V	V syllable structure
lat_	preceding lateral	ON	onset
fla_	preceding flap	NUC	nucleus
stp_	preceding plosive	CODA	coda
aff_	preceding affricate	PW_1	first syllable in PW
fri_	preceding fricative	PW_2	post-initial syllable in PW
asp_	preceding aspiration	PW_m	medial syllable in PW
tns_	preceding tense consonant	2_PW	penultimate syllable in PW
lab_	preceding labial	1_PW	last syllable in PW
cor_	preceding coronal	AP_1	first syllable in AP
dor_	preceding dorsal	AP_2	post-initial syllable in AP
glt_	preceding glottal	AP_m	medial syllable in AP
hiV_	preceding high vowel	2_AP	penultimate syllable in AP
mdV_	preceding mid vowel	1_AP	last syllable in AP
loV_	preceding low vowel	IP_1	first syllable in IP
_V	following vowel	IP_2	post-initial syllable in IP
_vcl	following voiceless	IP_m	medial syllable in IP
_nas	following nasal	2_IP	penultimate syllable in IP
_lat	following lateral	1_IP	last syllable in IP
_fla	following flap	UTT_1	first syllable in sentence
_stp	following plosive	UTT_2	post-initial syllable in sentence
_aff	following affricate	UTT_m	medial syllable in sentence
_fri	following fricative	2_UTT	penultimate syllable in sentence
_asp	following aspiration	1_UTT	last syllable in sentence
_tns	following tense consonant		

3. Analysis

3.1 Classification And Regression Tree (CART) Modelling

CART analysis (Breiman et al., 1998) has become a common method for building statistical models from simple feature data. CART trees partition the data set according to a binary tree of tests on feature values. For duration modelling, the nodes on the tree contain yes/no questions about the context features associated with a segment, while leaves contain the mean duration of all training segments that end up in one partition. When the tree is being built, a set of values within one partition is split according to the available questions, and the split which minimizes the variance of the data across two partitions is chosen. The tree building process terminates when partitions reach a minimum size, or when performance on some held out data reaches a maximum value.

In this experiment, we wanted to establish which context features were most important, and so we built trees 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* (Taylor et al., 1999) as a tool for running this CART tree building process.

An example of the data format for a single segment record input to this process is:

```
100  u  1  0  0  0  0  0  0  0  0  0  0  1  0  0
 0   0  0  0  0  1  0  0  0  0  0  0  0  0  1  0
 0   0  0  0  0  0  0  1  0  0  0  0  0  0  0  1
 0   0  0  1  0  1  0  0  0  0  0  0  1  0  0  0
 0   1  0  0  0  0  1
```

To calculate this data, the annodated transcription was first processed into a hierarchical prosodic structure encoded in XML. The feature string of each segment was then automatically generated from the phonological structure using the ProXML scripting language (Huckvale, 1999). 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. In the above input data, the first column indicates ms duration of the segment, the second column is the identity of the target segment, the other columns are 69 binary features shown in Table 2. For example, the third column has the value 1 for the feature "mono" (monophthong), the next column

the value 0 for the feature "di" (diphthong), and the last column is the value 1 for the feature "1_UTT" (last syllable in sentence).

An example of a CART decision tree is illustrated in Figure 1. The tree below is a part of the actual CART decision tree which is the result of the CART building process described in 3.3 below.

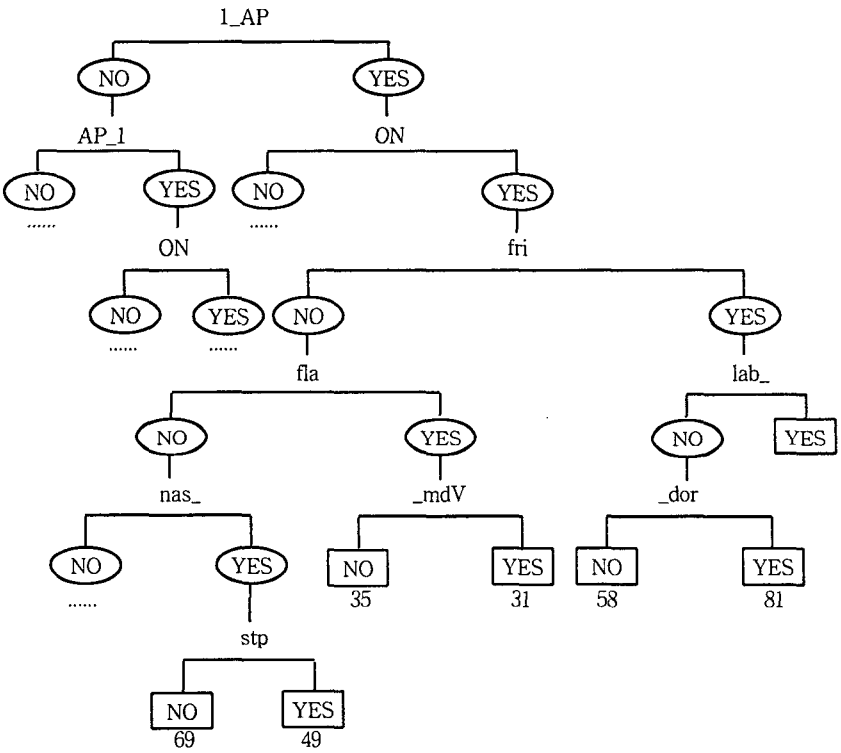

Figure 1. A simplified example of a CART decision tree.

Taking a look at the question nodes (non-terminal nodes), we can see 1_AP (AP- final position), AP_1 (AP-initial position), and ON (onset) are principally used at the question nodes. In particular, 1_AP is used at the root node. The number on the terminal node is the mean duration of the segments identified by the combination of question nodes. This tree clearly shows the interactions among several features in estimating the segment duration. When 1_AP feature combines with ON, fri (fricative), and lab_ (preceding labial) features, it produces a mean duration of 120 ms. But when 1_AP feature and ON feature combine with fla (flap) feature and _mdV (following mid vowel) feature, it shows a mean duration of 31 ms. This simplified tree can be schematized as following seven bundles of features.

Table 3. Estimated mean durations (ms) for various feature bundles in the CART decision tree.

1_AP	1_AP	1_AP	1_AP	1_AP	1_AP	1_AP
ON	ON	ON	ON	ON	ON	ON
fri	fri	fri	fla	fla	nas_	nas_
lab_	_dor		_mdV		stp	
120	81	58	31	35	49	69

An advantage of CART analysis is that it is easy to see the relative importance of each feature and the interactions of features in estimating the duration within the decision tree itself. However a disadvantage is that CART trees can grow quite large.

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 generalizations 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 maximized. 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 4.

Table 4. CART model performance using segment names.

Data set	RMSE (ms)	Correlation	Number of features
Training data	24.23	0.77	46
Evaluation data	24.04	0.79	26
Test data	26.48	0.73	26

The best 10 features and the growth in the correlation coefficients for each stepwise

refinement of the CART model on the training data set alone are illustrated in Table 5.

Table 5. Accumulated correlation coefficients of the best 10 features selected by the stepwise CART trained on the training data alone using segment names.

Feature	name	1_AP	AP_1	ON	CVC	vcl_	IP_m	PW_1	nas_	_stp
Correlation	0.40	0.61	0.65	0.67	0.70	0.71	0.72	0.73	0.73	0.74

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 Analysis Using Segment Class Features

A stepwise CART model was trained using all 42,103 segments in the training data set described by 69 segmental and prosodic phrasal features. Training ended when additional features made no significant improvement in performance; this was after 52 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 maximized. The tree was pruned back to 36 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 6.

Table 6. CART model performance using segment class features.

Data Set	RMSE (ms)	Correlation	Number of features
Training data	25.67	0.74	52
Evaluation data	26.04	0.76	36
Test data	27.98	0.70	36

The best 10 features and the growth in the correlation coefficients for each stepwise refinement of the CART model on the training data set alone are illustrated in Table 7.

Table 7. Accumulated correlation coefficients of the best 10 features selected by the stepwise CART trained on the training data alone using phonemic class features of each segment.

Feature	1_AP	ON	AP_1	CVC	vcl_	UTT_m	_nas	fla	nas	stp
Correlation	0.33	0.47	0.53	0.58	0.60	0.61	0.63	0.64	0.65	0.66

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 longer than mean duration and a negative z-score is shorter than mean duration. Because z-scores encode the inherent properties of each segment, the segment names were not used in this model.

A stepwise CART model was trained using all 42,103 segments in the training data set described by 69 segmental and prosodic phrasal features describing the context. Training ended when additional features made no significant improvement in performance; this was after 54 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 maximized. The tree was pruned back to 36 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 8.

Table 8. CART model performance using z-scores.

Data set	RMSE (z-score)	Correlation	Number of features
Training data	0.73	0.67	54
Evaluation data	0.74	0.66	36
Test data	0.77	0.63	36

The best 10 features and the growth in the correlation coefficients for each stepwise refinement of the CART model on the training data set alone are illustrated in Table 9.

Table 9. Accumulated correlation coefficients of the best 10 features selected by the stepwise CART trained on the training data using z-scores.

Feature	1_AP	ON	AP_1	asp_	nas_	_nas	PW_1	CVC	nas	_cor
Correlation	0.31	0.38	0.46	0.49	0.51	0.54	0.55	0.56	0.57	0.58

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 10.

Table 10. Mean z-score changes of selected features in the training data.

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

When the segment is in AP-final position, 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, and the PW-initial position feature have a lengthening effect. The onset position feature, preceding aspiration feature, preceding nasal feature, following nasal feature, CVC syllable structure feature, nasal segment feature, and following coronal feature all have a shortening effect. The preceding aspiration feature has a significant shortening effect. 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 also 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 an evidence of segment shortening both before and after nasal consonants. This is in partial agreement with Lee (1996) for Korean, where shortening after nasals was observed; and also with Lehiste (1970) for English where

shortening of vowels before nasals was seen.

3.5 Summary

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. Han (1964) and Kim (1974) found that a vowel in sentence final position is longer than in other positions. Lee and Koo (1997) found that the syllable before a sentence boundary is longest, and at normal speed, the syllables at IP, AP, and PW boundaries have similar duration. On the other hand, Chung et al. (1996), Jun (1993), and Lee (1990) argued that when an IP is followed by a pause, the IP-final position has a greater lengthening effect than does the AP-final position. 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. Han (1964) and Koo (1998) found that vowels in CVC syllable structure are much shorter than those in CV or V syllable structures.

- This CART model showed that the presence of a preceding aspirated segment has a shortening effect, as found in Han (1964), Kim (1974), Lee (1996) and Kang (2000).

- Though the result of this analysis agrees with that of Lee (1996) in terms of the shortening effect of preceding nasals, the shortening effect of following nasals needs to be investigated. Although our results are not in complete agreement with those reported in previous studies which argue that sonorants generally tend to have a lengthening effect, we have found that there is a tendency for the duration of segments to be shortened by the presence of following nasals, which was not reported in any previous study of Korean.

The prediction error and correlation coefficients found are comparable with recent published results in Korean (Lee and Oh, 1999). In their CART modelling of spoken Korean on segmental duration, Lee and Oh (1999) trained on 240 sentences (15,037 segments) and tested on 160 sentences (9,494 segments). Their RMSE was about 22 ms, and the corre-

lation 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, Lee (1996) showed correlations between 0.74 and 0.69 and an RMSE of less than 25 ms. In Fackrell et al. (2001), we find some modelling performance for some European languages; Dutch, English, French, German, Italian and Spanish. The correlation comparisons between the best results of our modelling and the others are shown in Table 11. Though all results are based on regression tree models, the details of the statistical analysis were slightly different across experiments.

Table 11. Comparisons between best CART model and other results.

Language	Experiment	Correlation
Korean	*this paper*	0.79
Korean	Lee & Oh (1999)	0.82
Korean	Lee (1996)	0.74
Dutch	Fackrell (2001)	0.80
English	Fackrell (2001)	0.78
French	Fackrell (2001)	0.73
German	Fackrell (2001)	0.78
Italian	Fackrell (2001)	0.84
Spanish	Fackrell (2001)	0.75

4. Conclusion

This paper used stepwise CART modelling to analyze the timing of spoken Korean in a connected news-reading speech style. The advantages 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 or in European languages. 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 (Chung, Huckvale

and Kim, 1999). In a future study, we hope to investigate other features proposed by Lee and Oh (1999) not incorporated in the present model. These are the number of syllables in the prosodic phrase and the grammatical class of each word. We could then determine the importance of these relative to the current feature set. It would also be interesting to study a more spontaneous speech style using the same approach.

References

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

Chung, H. & M. Huckvale. 1999. "Modelling of temporal compression in Korean."*Harvard Studies in Korean Linguistics VIII*, S. Kuno et al. eds, 102-116.

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

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

Fackrell, J., H. Vereecken, C. Grover, J. Martens & B. Coile. 2001."Corpus-based development of prosodic models across six languages." *Improvements in Speech Synthesis*, E. Keller ed. Wiley. (in press)

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

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

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

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

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

Kim, K. O. 1974. *Temporal Structure of Spoken Korean: an Acoustic Phonetics Study.* Ph.D. Thesis, University of Southern California.

Koo, Hee-San. 1998. "The influence of consonant environment upon the vowel duration." *Korean Journal of Speech Sciences*, 4(1), 7-18.

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

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

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

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

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

Taylor, P., R. Caley, A. Black & S. King. 1999. *Edinburgh Speech Tools Library* http://

www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/
van Santen, J. P. H. 1995. "Timing in text-to-speech system." *Eurospeech* '95, 1397-1404.

Received: Jan. 23, 2001.
Accepted: Mar. 3, 2001.

▲ Hyunsong Chung
Department of Phonetics and Linguistics
University College London
Gower Street
London WC1E 6BT
United Kingdom
Tel: +44-20-7679-5104
Fax: +44-20-7383-0752
E-mail: hchung@phonetics.ucl.ac.uk

▲ Mark Huckvale
Department of Phonetics and Linguistics
University College London
Gower Street
London WC1E 6BT
United Kingdom
Tel: +44-20-7679-5002
Fax: +44-20-7383-0752
E-mail: M.Huckvale@ucl.ac.uk