Perceptual Evaluation of Duration Models in Spoken Korean

Hyunsong Chung*

ABSTRACT

Perceptual evaluation of duration models of spoken Korean was carried out based on the Classification and Regression Tree (CART) model for text-to-speech conversion. A reference set of durations was produced by a commercial text-to-speech synthesis system for comparison. The duration model which was built in the previous research (Chung & Huckvale, 2001) was applied to a Korean language speech synthesis diphone database, "Hanmal (HN 1.0)". The synthetic speech produced by the CART duration model was preferred in the subjective preference test by a small margin and the synthetic speech from the commercial system was superior in the clarity test. In the course of preparing the experiment, a labeled database of spoken Korean with 670 sentences was constructed. As a result of the experiment, a trained duration model for speech synthesis was obtained. The "Hanmal" diphone database for Korean speech synthesis was also developed as a by-product of the perceptual evaluation.

Keywords: Duration, Prosody, Speech Synthesis, Perceptual Evaluation

1. Introduction

Perceptual evaluation of prosody model is essential to decide the quality of synthesized speech. It is not always the case that improved statistical modeling leads to improved speech quality. This paper is divided into three parts. The first part briefly illustrates the CART duration model and its linguistic implications, some of which are described in Chung & Huckvale (2001). The second part explains the details of a Korean speech signal generation system. The third part describes a perceptual evaluation which was carried out using this system in combination with a duration model developed in the experiment. Perceptual evaluation investigates the clarity and the listener's preference for synthetic speech produced by two duration models of the CART and a commercial Korean text-to-speech (TTS) system.

* Dept. of Computer Science, University College Dublin, Ireland

2. CART Duration Model

2.1 Issues in the timing of spoken Korean

It has been generally considered that the following issues are unsolved or controversial in the research of the timing of spoken Korean:

a. Which type of phrase boundary has the most influence on the duration of adjacent syllables? Utterance, intonational phrase, accentual phrase and phonological word boundaries are all claimed to have lengthening effects (Han, 1964; Kim, 1974; Jun, 1993; Lee, 1996; Chung et al., 1997; Lee & Koo, 1997). However, more information is needed over which boundary is more important, how much the relative size of initial and final boundary effects would be, and whether syllables in post-initial or penultimate positions are also lengthened.
b. How does the structure of a syllable affect its constituents? In Korean, CVC, VC, V and CV syllable structures can be observed. These structures are believed to have an influence on the segment duration. More information is required about how each syllable structure affects segment duration. The different behaviors of onset consonants and coda consonants also need further study.
c. Which segmental features show a systematic effect on duration? In English, following segments have more influence than preceding segments (Peterson & Lehiste, 1960). In Korean, it is claimed that preceding segments are more important than following segments (Han, 1964: Lee, 1996).

The CART duration modeling (Riley, 1992) in this paper does not just try to build the best predictive model of segment duration in context, but also seeks to learn more about which factors and which structures are most important in Korean prosody. It is unfortunate that there are very few published studies on the timing models of Korean. There have been almost no attempts to apply linguistic findings of connected speech of Korean to the duration modeling or to text-to-speech conversion systems. Though some studies (Lee, 1996; Lee & Oh, 1999) have tried to model the duration of Korean using regression tree model such as CART, they usually more concentrated on improving the performance of the models rather than investigating linguistic implication of the models.

2.2 Database

The main corpus consists of 670 sentences with various lengths spoken by one speaker in a news reading style. This choice was made since contextual factors have to be rich enough to capture all aspects that affect timing and so it is necessary to avoid

lists and sentences of repetitive structure. The size of 670 sentences was chosen because this was believed to be big enough to cover the majority of segmental contexts and was comparable to or bigger than those of similar studies (van Santen, 1992; Lee, 1996; Lee & Oh, 1999).

2.3 Results of CART modeling

Because details of the CART modeling are illustrated in Chung & Huckvale (2001), this section explains which linguistics features are incorporated in the duration model based on the CART decision tree. In the CART model, a total of 69 features were used which include the prosodic phrase feature, the syllable structure feature and the segmental feature such as major class features of the target segment and surrounding segments. The best performance of the CART model in this experiment was comparable with other published results (Lee, 1996; Lee & Oh, 1999) with the correlation coefficient, 0.79 and the root mean squared prediction error, 24.04 ms. The results show that the accentual phrase boundary has the most influence either on the accentual-initial or on the accentual final syllable. The accentual phrase boundary significantly lengthens the segment duration. Between the phonological word initial position and the phonological word final position, the latter has more lengthening effect. Utterance boundaries and intonational phrase boundaries do not contribute much to the duration once the accentual phrase boundary has been taken into account. This is believed to be partly because each utterance boundary and intonational phrase boundary is also an accentual phrase boundary and a phonological word boundary. Shortening effects are seen in all post-initial positions and in penultimate positions from each boundary except in the post-initial position of the accentual phrase. Though certain types of adjacent consonants affect vowel duration, there is no general effect on the duration that could be attributed to the structure of the syllable without consideration of consonant type. In terms of the effects by surrounding segments, the preceding and following [nasal] features have the most influence in the experiment. Nasals tend to shorten the target vowel. This suggests that nasals seem to have a more shortening effect than homorganic voiced obstruents. This can be explained by the fact that vowel needs a special adjustment of the vocal folds to maintain vibration during voiced plosives. No such adjustment is required for nasals (Lehiste, 1970). Though they are not in the top ten most important factors, [stiff vocal fold] feature which covers aspirated plosives and tense plosives in Korean has a significant shortening effect in the vowel duration. This fact supports the idea that the glottal opening and the tenseness of surrounding segments is the major controller of the vowel duration. In agreement with previous studies of Korean (Chung, Kim & Huckvale, 1999) or English (House, 1961; Crystal & House, 1988), the results show little effect caused by the place features of surrounding segments.

3. Perceptual Evaluation

3.1 TTS evaluation

In November, 1998, a TTS comparison test (van Santen et al., 1998) was conducted by ESCA (European Speech Communication Association)/COCOSDA (International Committee for the Co-ordination and Standardization of Speech Databases and Assesment Techniques). A total of 68 TTS systems participated in the test, 17 of which competed in the English language section. One system was a Korean language TTS system. Three types of text were used in the test: newspaper text (easy vs. difficult), semantically unpredictable sentences and telephone directory listings. Three categories were tested in the English language TTS systems: overall voice quality, intelligibility and overall impression. The test for "overall voice quality" or "naturalness" used a rating scale from 1 to 15 for each of the system and the intelligibility results were expressed in terms of percentage of correctly transcribed words in sentences that were designed to minimize contextual information. Listeners' overall impression was tested on a scale from 1-15. For a further overview of the general TTS evaluation procedures, refer to van Bezooijen and van Heuven (1998). In this experiment, this kind of extensive perception test was not carried out due to the lack of time and resources. So the perceptual study in this paper is only a brief and rather informal investigation.

3.2 Korean language diphone database for speech synthesis

In order to create synthetic speech manipulated by a duration model and to evaluate its perceptual quality, a Korean language diphone database "Hanmal (HN 1.0)" was developed based on the MBROLA synthesis system (Dutoit et al., 1996) in collaboration with Kyongsok Kim (Kim & Chung, 1999). This diphone database has been publicly available since September 17, 1999 from the MBROLA web site so that other researchers could synthesize Korean speech and investigate the relationships between prosody variation and naturalness.

MBROLA is a speech synthesis system based on the concatenation of diphones. It takes a list of phones as input, together with prosodic information (duration of phones and a piecewise linear description of pitch), and produces speech signals, at the sampling frequency of the diphone database used. Diphones are speech units that begin in the middle of the stable state of a phone and end in the middle of the following one. Their main usefulness in synthesis is that they minimize concatenation problems, since they contain most of the transitions and co-articulations between phones. They also require relatively small amounts of memory, as their number remains small (compared to synthesis units such as half-syllables or triphones). To prepare a diphone database capable of satisfying these requirements for Korean,1,986 nonsense words were created to

cover a catalogue of 1,986 diphones.

The diphone recordings were processed by the MBROLA team in Belgium to produce the "Hanmal" diphone database. Applications based on this database are supported on a wide range of computing platforms using the MBROLA signal generation engine. Diphone concatenation and prosody manipulation can be performed using the MBR-PSOLA algorithm (Dutoit et al., 1996).

3.3 Design of evaluation

Nine sentences with various lengths were selected from broadcast news scripts, which were different from the data set used in building duration models. See the appendix for the test sentences. The Romanization in these sentences is based on "the Romanization system for Korean Language" announced by the Ministry of Culture and Tourism of the Republic of Korea on July 4, 2000.

Durations were calculated by using the best CART model. For CART modelling, both the name and manner features of segments were used. Also, durations were extracted from one of the best Korean commercial TTS systems. In the experiments, the CART model was named "model 1" and the model by the commercial TTS was named "model 2". F_0 contours for the sentences were copied from natural read versions and were used for both systems. The duration and F_0 contour information of these models were then applied to the MBROLA Korean diphone database "Hanmal". The synthesized speech by the two models were played to subjects for perceptual evaluation:

a. Model 1: Durations directly predicted from CART decision tree model using names and major class features of the target segment and the segmental and prosodic phrasal features describing the context
b. Model 2: Durations from the commercial TTS system

Ten subjects participated in the perceptual evaluation, all of whom are native Korean speakers using modern Seoul and Gyeonggi dialect. They had lived in Korea for more than 15 years and had lived in London, England for the last couple of years. The subjects were aged between 18 and 20. They were given the selected nine sentences produced using each of the two different models in the two judgement tasks. Thus each subject listened to nine pairs of sentences twice. In order to avoid any judgement bias derived from the order in which the models were used, the order of the models within the pairs was randomized. The random ordering was done in a way so that, overall, each model had the same distribution with respect to position in the pairs. After each pair, the subjects were given 10 seconds to make a ranking decision on the quality of the synthesized speech. Two aspects of the quality were evaluated: clarity and preference.

Subjects were asked to make a judgement on the clarity of the sentences the first time they listened to the pair and to make a judgement on their general preference in a second listening. The better one was graded "1" and the worse "0".

3.4 Results of the perceptual evaluation

The 90 judgements (9 sentences×10 subjects) for each of the two models were obtained. Each was converted to pairwise preferences and are summarized as shown in Table 1 and Table 2. "Sign test" (Press et al., 2002) was used to check whether any differences between these models were simply the result of chance. Subjects were also encouraged to discuss their subjective impression of the synthetic speech.

Table 1. Pairwise preference summaries for clarity level and general preference.

	Clarity	Preference
Model 1	29	53
Model 2	61	37

Table 2. Likelihood of results occurring by chance for each pair of models (sign test).

Clarity
Model 1 vs. Model 2
$p < 0.01$

Preference
Model 1 vs. Model 2
$p = 0.1$

In terms of clarity, subjects preferred the commercial model to the CART model. The difference was statistically significant at $p<0.01$. In terms of general preference, the CART model was more preferred by subjects, though the difference was not statistically significant.

During a final discussion, the sentences were played again to obtain their subjective impressions. Eight subjects suggested that the speech produced by a commercial TTS model was relatively slow. One of the subjects' complaints was that nasal consonants were perceived relatively long in many instances from the CART model. All subjects agreed that in most cases vowel durations were satisfactory in both models. Five subjects indicated that in the case of the CART model, bilabial stops sounded tense in some cases. Overall, subjects seemed more sensitive to consonant duration than vowel duration. Comments about discontinuities between concatenated diphones in the synthesized speech were made for both models. The fact that the general preference for the CART and the commercial TTS durations has similar scores shows that CART duration prediction in this experiment is at least as good as that of the commercial one.

4. Conclusion

This paper described a small-scale subjective perceptual evaluation of the quality of the durations produced by the CART duration model of spoken Korean. Durations calculated by the best CART model were applied to the Korean diphone database "Hanmal (HN 1.0)". A reference set of durations was produced by a commercial TTS system. Clarity and preference were used as criteria in the evaluation of the naturalness of the synthesized speech from the two models. It was found that the subjects were more sensitive to consonant durations than to vowel durations. The CART model was preferred in the preference test by a small margin, while the synthetic speech from the commercial TTS system was significantly preferred in the clarity test.

In the course of preparing the experiment, a labeled 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 diphone database for Korean speech synthesis was developed as a by-product of the perceptual testing, which is now publicly available and currently in use by other researchers. In the future, more research will be done on the design of the TTS evaluation procedure.

References

Chung, H. & M. Huckvale. 2001. "Analysis of the timing of spoken Korean using a Classification and Regression Tree (CART) model." *The Korean Journal of Speech Sciences*, The Korean Association of Speech Sciences, 8(1), 77-91.

Chung, H., K. Kim & M. Huckvale. 1999. "Consonantal and prosodic influences on Korean vowel duration," *Proceedings of 6th Eurospeech*, 2, 707-710, Budapest, Hungary.

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. 1997. *A Study of Korean Prosody and Discourse for the Development of Speech Synthesis/Recognition System.* Korea Telecom Research & Development Group Technical Report.

Crystal, T. H. & A. S. House. 1988. "Segmental durations in connected speech signals: Preliminary results," *Journal of Acoustical Society of America*, 72(3), 705-716.

Dutoit, T., V. Pagel, N. Pierret, F. Bataille & O. van der Vreken. 1996. "The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes." *Proceedings of 4th International Conference on Spoken Language Processing*, 3, 1393-1396.

Han, M. 1964. *Duration of Korean Vowels, Studies in the Phonology of Asian Language II*, Acoustic Phonetics Research Laboratory, Los Angeles: University of Southern California.

House, A. S. 1961. "On vowel duration in English." *Journal of Acoustical Society of America*, 33(9), 1174-1178.

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

Kim, K. & H. Chung. 1999. URL: http://tcts.fpms.ac.be/synthesis/mbrola.html

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

Lee, S. & Y. Oh. 1999. "Tree-based modeling of prosodic phrasing and segmental duration for Koran TTS systems." *Speech Communication*, 28, 283-300.

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

Lee, Y. 1996. "Modeling of segment duration in Korean speech synthesis." *Phonetics and Linguistics in Honor of Professor Hyun Bok Lee*. Seoul: Seoul National University Press, 249-274.

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

Peterson, G. E. & I. Lehiste. 1960. "Duration of syllable nuclei in English." *Journal of Acoustical Society of America*, 32(6), 693-703.

Press, W. H., S. A. Teukolsky, W. T. Vetterling & B. P. Flannery. 2002. *Numerical Recipes in C++*. 2nd edition. Cambridge: Cambridge University Press.

Riley, M. 1992. "Tree-based modeling of segmental duration." *Talking Machines: Theories, Models and Designs*, G. Bailly & C. Benôit (eds.), Amsterdam: North-Holland, 265-273.

van Santen, J. P. H. 1992. "Contextual effects on vowel duration." *Speech Communication*, 11, 513-546.

van Santen, J. P. H., L. C. W. Pols, M. Abe, D. Kahn, E. Keller & J. Vonwiller. 1998. "Report on the third ESCA TTS workshop evaluation procedure." *Proceedings of 3rd ESCA/COCOSDA Speech Synthesis Workshop*.

van Bezooijen, R. & V. van Heuven. "Assessment of synthesis systems." *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, R. Moore & R. Winski (eds.), Berlin: Mouton de Gruyter, 481-563.

Received: Jan. 25, 2002.
Accepted: Mar. 8, 2002.

▲ Hyunsong Chung
Department of Computer Science
University College Dublin
Belfield, Dublin 4 Ireland
Tel: +353-1-716-2923
E-mail: hyunsong.chung@ucd.ie

Appendix

Nine Test Sentences for Perceptual Evaluation

1. *uri-neun minjok jungheung-ui yeogsajeok samyeong-ul tti-go i ttang-e tae-eo nass-da.*
   "We were born in this country with a historical duty to promote national prosperity."
2. *baram-gwa haesnim-i seoro him-i deo seda-go datu-go iss-eoss-seupnida.*
   "The wind and the sun were disputing who was the stronger."
3. *oneul haru-mando jeonnam-gwa gyeongnam naeryukjibang-euroneun baek-millimiteo -ga neomneun manh-eun bi-ga nae-ryeoss-seupnida.*
   "Today, there was also heavy rain, over 100 mm in the inland areas of Jeonnam and Gyeongnam provinces."
4. *beolsseo saheul-jjae nambu jibang-ui hou-neun gyesok-dwaego iss-seupnida.*
   "It has already been three days since heavy rain started in the southern area."
5. *jigeum-do yeojeonhi gyeongnam-gwa jeonnam namhaean jibang-euroneun hou gyeongbo-ga naeryoejeo iss-go jeonbuk-gwa jeonnam naeryukjibang-euroneun houjuuibo-ga naeryeojin gaunde oneulbam sai-edo cheondung beongae-ga chimyeonseo jipjunghou-ga naeril ganeungseong-i nop-seupnida.*
   "There are still warnings of severe rain in the southern coastal areas of Gyeongnam and Jeonnam provinces and forecasts of heavy rain in the inland areas of Jeonbuk and Jeonnam provinces, and there is a high chance of heavy rain with thunder and lightning tonight."
6. *Seoul sigaji-ui gyotongjeongbo ipnida.*
   "This is the traffic bulletin for central Seoul."
7. *nambu sunhwandoro-neun sadangsageori-ui gongsa ttaemune yangbanghyang-ui gyotongheureum-i modu eoryup-seupnida.*
   "There is heavy congestion for both inbound and outbound traffic on the southern circular highway due to road works at the Sadang intersection."
8. *namtaeryeong gogae-eseo isu gyocharo panghyang-euroneun hwamulcha-ui jeungga-ro sisok isipkillo-ui sokdoreul nael su iss-go yaesul-ui jeondang-eseo bongcheon sageori banghyang-euroneun jiche-ga gyesokdwaego iss-seupnida.*
   "The speed of traffic in the area from the Namtaereong Hill to the Isu intersection is around 20 km per hour and delays are continuing in the area from the Seoul Arts Center to the Bongcheon intersection."
9. *Olympic daero-neun yeoido-eseo gimpogonghang banghyang-i sisok sipkillo jeongdo -ui geobugi georeum-eul gyesok-hago iss-seupnida.*
   "On the Olympic Highway, cars are moving slowly at a speed of 10 km per hour in the direction from Yeoido to Gimpo Airport."