Modelling Duration In Text-to-Speech Systems*

정현성(대구대)

<차 례>

<Abstract>

Modelling Duration in Text-to-Speech Systems

Hyunsong Chung

The development of the durational component of prosody modelling was overviewed and discussed in text-to-speech conversion of spoken English and Korean, showing the strengths and weaknesses of each approach. The possibility of integrating linguistic feature effects into the duration modelling of TTS systems was also investigated. This paper claims that current approaches to language timing synthesis still require an understanding of how segmental duration is affected by context. Three modelling approaches were discussed: sequential rule systems, Classification and Regression Tree (CART) models and Sums-of-Products (SoP) models. The CART and SoP models show good performance results in predicting segment duration in English, while it is not the case in the SoP modelling of spoken Korean.

* Keywords : duration, prosody, text-to-speech, CART, Sums-of-Products

* This research was supported in part by the Daegu University Research Grant, 2003.

1. Introduction

Thanks to the development of computer technology over the last few decades, linguists and engineers have been able to implement linguistic theories and descriptions within speech technology to create automatic text-to-speech (TTS) conversion systems in some world languages. In the prosody modelling stage of the text-to-speech conversion stages, the phonological representation is converted into a speech waveform. This involves the prediction of suitable intonation and timing, followed by means to realize the phonological units as sound. The prediction of intonation and timing can be made by rule, using knowledge of typical fundamental frequency contours and durations for phonological events in different contexts [1][2][3]. Recently, however, there has been a change whereby the phonological description itself is used to search a database of known contours and timings to extract single known "cases" that match the prosodic pattern [4]. Depending on the choice of signal generation method, the predicted fundamental frequency and duration may be imposed on the generated sound by means of signal processing algorithms for prosody manipulation. However, some success has been made by systems that prefer a larger corpus of cases over signal modification. Even though some signal generation methods avoid modifying the timing of units during signal generation, it is still necessary to understand what phonological features influence timing in context. This paper discusses the development of the durational component of prosody modelling in text-to-speech conversion in English and Korean, showing the strengths and weaknesses of each approach. It also investigates the possibility of integrating linguistic feature effects into the duration modelling of TTS systems.

2. Overview of Text-to-Speech Systems

Three approaches to signal generation have been widely used for TTS: rule-based synthesis, concatenative diphone synthesis, and corpus-based unit selection. Rule-based synthesis tends to be the preference of phoneticians and phonologists, who seek to encode a cognitive, generative model of human speech production. Rule-based synthesis usually uses a formant synthesizer for signal generation. Rules map phonological and phonetic properties to the control signals of a formant synthesizer. The synthesizer includes components to simulate the generation of several different kinds of sound sources and components to simulate the vocal-tract transfer function, and a component to simulate sound radiation from the head [5]. The main problem with rule synthesis is the lack of knowledge of how to map phonetic descriptions to formant parameters in a natural and coherent manner.

As opposed to rule-based synthesis, concatenative synthesis is based on the concatenation of recordings of elementary speech units to make a human-sounding synthetic speech signal. Such units as diphone allow the modelling of some coarticulation effects across phones, and avoid the "targets" and "interpolation" metaphor used in rule synthesis. The concatenation process involves three procedures: (1) stretch or compress acoustic units; (2) attach successive acoustic units to each other; (3) impose an intonation contour. The information about timing imposed on the speech signal in (1) and (3) needs to be calculated using the prosodic component of the TTS system. When stage (2) takes place, there may be audible spectral discontinuities. In order to avoid this, this stage requires various forms of interpolation and smoothing.

The corpus-based unit selection approach to signal generation is a recently developed technique, which is becoming widely used in speech synthesis. This technique describes both the synthesis target and the components of the speech database as phonological trees, and uses a selection algorithm which finds the largest parts of trees in the database which match parts of the target tree. Often this method is used without explicit modification of pitch and timing during synthesis. The technique tries to avoid many of the errors made by prosody prediction modules by incorporating their operation implicitly in the selection process. In diphone synthesis a single diphone is used for every instance of that diphone in a target synthesis and its pitch and duration are modified by signal processing to match its target prosody. In unit selection synthesis, however, there are many instances of each unit type, each with different pitch, durations and prosodic contexts. These are compared to the target and the most appropriate can be chosen. Furthermore the comparisons can be made using phonological features, thereby obviating the need to make explicit models of pitch and duration in Hertz and milliseconds. Unit selection algorithms are often successful at finding units of the appropriate pitch and duration specified in the target description. However, this technology requires an extremely large speech corpus, because it needs to be able to find a sequence of multi-segment units in the corpus that satisfies a number of requirements: (1) phone label match; (2) prosodic label match; (3) spectral match to adjacent units. [6] argued that this causes a problem of coverage and suggested that in corpus based synthesis, it is necessary to restrict synthesis to a single task domain of limited vocabulary and sentence structure to satisfy the above three criteria.

Despite the success of corpus-based approaches to synthesis, the analysis of language timing is still an important endeavor. We still need an understanding of how segmental duration is affected by context. This understanding will help us to decide what features we need to index speech in the corpus, which features are most important in the unit selection matching function, what contexts need to be incorporated into the sentences recorded for the database, and what features have to be located and specified from the input text.

3. Modelling Duration in English

The underlying linguistic representation in synthesis is symbolic, consisting of entities such as phoneme sequences, in combination with morphological, syntactic and prosodic information. The prosody prediction component computes the timing and pitch contour for the phrase. Prosody modelling refers to the equations involved in these computations, that is using the phonological structure to predict pitch and timing values in Hertz and milliseconds. Prosody modelling is one of the most important factors in determining the naturalness of synthesised speech. This section focuses on the development of the durational component of prosody modelling in text-to-speech conversion in English.

Following [2], [7] categorizes current duration prediction systems into three classes: sequential rule systems, equation systems, and binary prediction trees. Such rule systems as [8] are considered sequential rule systems which could be easily converted to equation systems such as Sums-of-Products (SoP) models in [2]. CART models [9] are considered binary prediction tree systems which have been criticized by [2] as just a collapsed form of lookup table. What these systems share is that they map symbolic input vectors provided by linguistic analysis routines onto acoustic quantities (duration), which may then used by the synthesis component to generate speech with the desired acoustic-prosodic characteristics. In the following sections, sequential rule systems [8, 10], CART decision tree models, and SoP models are overviewed and discussed for segmental duration prediction. Though neural networks have been used for duration modelling, this technique has the problem that its means of operation is not explicit, so that it does not give us any linguistic intuitions.

3.1. Sequential Rule Systems

The duration model in [1] assumes that (1) each phonetic segment type has an inherent duration that is specified as one of its distinctive properties, (2) the effect of each phonological context can be expressed as a percentage increase or decrease in the duration of the segment, but (3) segments cannot be compressed shorter than a certain minimum duration. The duration of a segment can then be written as:

$$\mathrm{DUR} = k(\mathrm{INHDUR} - \mathrm{MINDUR}) + \mathrm{MINDUR}, \qquad (1)$$

where DUR is the output duration in ms, INHDUR is the inherent duration of the segment in ms, MINDUR is the minimum duration of the segment in ms (which for vowels is usually 45% of the inherent duration), and k is the scale factor determined by applying rules in contexts. Combining rules from previous researches, [1] proposed the following duration rules and contexts:

(1) Pause insertion before clause boundaries and before orthographic comma
(2) Clause-final lengthening
(3) Phrase-final lengthening
(4) Non-word-final shortening
(5) Polysyllabic shortening
(6) Non-word-initial consonant shortening
(7) Unstressed segment shortening
(8) Lengthening of emphasized vowels
(9) Shortening of vowels preceding voiceless consonants
(10) Shortening in consonant clusters
(11) Lengthening of stressed vowels or sonorants due to preceding aspirated plosives

[10] proposed a similar duration rule for eight American English vowels as follows:

$$T = T_0 + S(K_1 + K_2 \times C), \qquad (2)$$

where T is the output duration, T_0, K_1, and K_2 are constant values for each vowel, C is the consonantal context factor which depends on which consonant follows the vowel, and S is other factors such as the position of the vowel in a word and in a sentence, the word prominence, stress, speech rate, function word status, etc. This formula allows for interactions between segmental and prosodic features. [10] argued that the duration of a vowel in word-medial position is little affected by its segmental context or its stress, unlike [8] who suggested a non-word-final shortening rule and a polysyllabic shortening rule.

In order to predict consonant duration, [11] used an additive model in which [11] added coefficients specified by various segmental and prosodic contexts to produce an estimated duration value. Thus the model has a large number of arbitrary constants to explain the various contextual effects. [11] argued that consonant duration modelling is so complex that the model in [1] which used a fixed set of constants for all consonants could not predict the complexities of consonant duration.

YorkTalk [3] is a rule-based system that uses a non-linear model of timing. The basic timing unit is the syllable, which is modelled by the temporal interpretation function "overlay". Syllable overlay is calculated by using the distance of overlaid syllable and the distance of syllable in a monosyllabic utterance. The "distance" is a measurement of the separation between the onset and the coda in syllables. The same mechanism is applied to the temporal compression of prosodic feet. This distance has a direct relation with the following structural information: (1) Nucleus property: short or long; diphthong or monophthong; (2) Coda property: simple or branching; (3) Rhyme property: heavy or light; voiced or voiceless; (4) Syllable strength: strong or weak; (5) position in Foot: initial, medial, or final; (6) the weight and strength of adjacent syllable(s). Because in a single syllable, the onset and the coda are "overlaid" on syllable nuclei, the syllable end coincides with the rhyme end, nucleus end, and coda end. In polysyllabic words, the adjacent syllables are overlaid and the temporal compression is expressed as follows:

$$\text{Syllable}_n\text{Start} = \text{Syllable}_{n-1}\text{End} - \text{Overlay} \qquad (3)$$

The YorkTalk model exploits a hierarchical metrical structure to describe the relationships between syllables. The model focuses on the temporal relations between syllables rather than on the durations of individual syllables or segments.

3.2. Classification and Regression Tree (CART) Modelling

The principle of the CART methodology was initially proposed by [12] and it was applied to duration modelling by [9]. CART analysis has become a common method for building classification models from simple feature data. This analysis was suggested by [9] as an alternative to heuristically-derived duration prediction rules for duration modelling in synthesis. CART trees partition a 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 that 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.

[9] argued that CART is a promising method for duration modelling in that (1) the most significant features are selected based on statistics, (2) it provides "honest" estimates of its performance, (3) both categorical and continuous features are permitted, (4) humans can interpret and explore the result. In the analysis of 1,500 hand-segmented and labelled short utterances of English from a single speaker, [9] used the following features in a CART analysis:

(1) Segment identity
(2) Previous segment context (up to three segments to the left)
(3) Following segment context (up to three segments to the right)
(4) Stress (unstressed, primary, secondary)
(5) Distance to the left boundary of the word in segments
(6) Distance to the right boundary of the word in segments
(7) Distance to the left boundary of the word in vowels
(8) Distance to the right boundary of the word in vowels
(9) Distance to the left boundary of the phrase in words
(10) Distance to the right boundary of the phrase in words

[9] described features (5)-(8) as lexical position, and (9)-(10) as phrasing position. In order to make the CART approach more practical, [9] classifies the segment identity of each phone in terms of 4 features: consonant manner, consonant place, vowel manner, and vowel place, with each class having several values. The optimal regression

tree had about 250 nodes and predicted the durations of test data with an error of 23 ms standard deviation. Unfortunately, the quality of the synthesised speech derived from the CART decision tree model was not noticeably better than that which was derived from rule based duration predictions. [9] suggested that this was due to insufficient training data in certain contexts or because of inadequate predictive power of the available features. Nevertheless, [9] argued that CART technique is valuable in that tree building and evaluation is rapid and that the technique maybe easily applied to other feature sets, to other languages, to other speakers, and to other speaking rates. In other applications of CART to duration modelling, [13] showed that the performance of the CART decision tree model in the BT's Laureate TTS system was subjectively better than the rule-based method.

3.3. Sums-of-Products (SoP) Modelling

SoP duration modelling is the third main technique applied to duration modelling. The motivation of SoP models is that certain data which shows the significant interactions among contextual factors cannot be fitted by a simple rule model. [2] argued that previous studies on contextual effects on segmental duration have focussed more on theoretical issues and putative underlying processes rather than completeness of empirical description. [2] said that the first step to construct a duration rule system for a TTS system is to make a list of factors which describe the contexts of a segment. The second step is to produce a duration model to explain complex interactions. In the TTS duration model, [2] tried to show the durational behaviour of a single speaker and produce a simple equation to predict the durations based on contextual factors. [2] also said that duration databases for statistical analysis commonly confound factors in that not all combinations of factors and levels occur with equal frequency. According to [2], the factor confounding results in mean durations that correspond to the levels of the factor of interest (the critical factor) being affected by other factors (confounding factors). One such example is word-final lengthening of unstressed syllables. Because, in word-final position, vowels in English are more likely to be unstressed and stressed vowels are more likely to be longer than unstressed vowels, statistics show that word-final vowels are shorter than non-word final vowels when all vowels are analysed altogether. However, when stressed and non-stressed vowels are analysed separately, word-final vowels are longer than non-word final vowels. When a pair of interacting factors such as the vowel and stress factors needs to be described, the quasi-minimal pairs technique can be used. Segment durations occurring with a combination of levels on confounding factors and with several levels on the critical factor are divided into "quasi-minimal" sets. If there are not enough duration events for all sets, a piecewise multiplicative correction method can be introduced which assumes that the effect of the critical factor and the joint effects of the remaining factors combine multiplicatively. [2] gives the example of the interaction between the syllabic position factor and the stress factor. [2] argued that these interactions are better described by a multiplicative rule than an additive rule. However, such interactions are not necessarily completely multiplicative, so he uses the term "piecewise." Where the quasi-minimal sets and multiplicative correction methods have difficulties with factors that have many levels, [2] introduces SoP models, which is called "a special case of an additive-multiplicative models, consisting of the sum of a single product term and any number of single-factor terms."

According to SoP models, the duration for a unit in the context combination described by the feature vector d is given by:

$$\mathrm{DUR(d)} = \sum_{i \in K} \prod_{j \in I_i} S_{i,j}(d_j) \qquad (4)$$

Here, K is a set of indices, each corresponding to a product term. I_i is the set of indices of factors occurring in the i-th product term. [2] suggested that major interactions between factors could be described as a complete multiplicative rule (a single product term) or a piecewise multiplicative rule (more than two terms) in a "Sums-of-Products" model. Otherwise, other interactions are described as additive in the model. The multiplicative interactions predict constancy when effect size is measured as a percentage, the additive interactions do it when it is measured in milliseconds.

In the experiment on vowel duration in American English, [2] used training data of 18,000 vowel segments and test data of 6,000 vowel segments was used with the following context factors:

(1) Vowel identity (V), 9 levels
(2) Accent (A), 3 levels
(3) Syllabic stress (S), 3 levels
(4) Prevocalic consonants (Cpre), 3 levels
(5) Postvocalic consonants (Cpost), 6 levels

(6) Within-word position: Preceding syllable/segment counts (Wpre), 3 levels

(7) Within-word position: Following syllable/segment counts (Wpost), 5 levels

(8) Utterance position (U): 4 levels

Based on the observations of the data, [2] suggested the following seven-term SoP model for English vowel duration.

$$\begin{aligned}&\log[\mathrm{Dur}(A,\ S,\ V,\ C_{pre},\ C_{post},\ W_{prc},\ W_{post},\ U)] =\\&\quad [S_{1,1}(A)\ \mathrm{x}\ S_{1,2}(S)] + S_{2,1}(V) + S_{3,1}(C_{pre}) + S_{4,1}(W_{pre}) + S_{5,1}(W_{post}) +\\&\quad [S_{6,1}(C_{post})\ \mathrm{x}\ S_{6,2}(U)]\end{aligned} \tag{5}$$

This model shows two multiplicative interactions: the first between pitch accent and syllabic stress and the second between post-vocalic consonant and utterance position. The first interaction is described as one term $[S_{1,1}(A)\ \mathrm{x}\ S_{1,2}(S)]$ and the last two terms $S_{5,1}(W_{post}) + [S_{6,1}(C_{post})\ \mathrm{x}\ S_{6,2}(U)]$. The deviation of this formula was based on the "covariance analysis method" [14]. Also note the fact that the model as a whole operates in the log duration domain. The overall correlation between observed and predicted durations based on this SoP model on test data reached 0.908.

4. Modelling Duration in Korean

There are very few published studies on the modelling of Korean prosody for TTS. In this section, a couple of timing models of Korean are reviewed.

In order to make a duration model, [15] suggested using the following factors for Korean segments:

(1) Syllable structure

(2) Surrounding context: 13 features for consonants, 4 features for vowels

(3) Position of the syllable in the word: initial, medial, final

(4) Number of syllables in the word

(5) Position of the syllable in the phrase: initial, medial, final

(6) Number of syllables in the phrase

[15] used a regression tree model for the statistical processing of the segment duration data set. For training data, [15] used 15 sentences spoken by three males and four females, each sentence spoken in three different tempos: fast, slow, and normal. Another sentence with three different tempos was used for test data. [15] calculated the correlation between predicted and observed duration of syllables and that of segments. More than 75% of segments had actual durations within 25 ms of the predicted value. The quality of the prediction was as in Table 1.

<Table 1> Results of regression tree models in [15].

Tempo	Normal	Fast	Slow
Correlation (segment)	0.74	0.69	0.74
Correlation (syllable)	0.86	0.83	0.88

On the other hand, the tree-based modelling in [16] used the following features for the prediction of segmental duration:

(1) Preceding segment, observed segment, and following segment
(45 segment categories)
(2) The part-of-speech context corresponding to segmental context
(23 word classes)
(3) Location of the syllable in PW (phonological word) and AP
(accent phrase) (initial, medial, and final)
(4) The length of PW and AP in syllables

[16] trained on 240 sentences (15,037 segments) and tested on 160 sentences (9,494 segments). [16] carried out two separate performance tests: one with part-of-speech information and the other without it. The performance of tree-based modelling of segmental duration is shown in Table 2.

<Table 2> Performance of tree-based modelling of segmental duration in [16].

	Correlation	RMSE
Including part-of-speech	0.820	22.06 ms
Excluding part-of-speech	0.823	21.88 ms

RMSE = root mean squared prediction error

[16] concluded that part-of-speech information did not contribute to improving the performance. The number of nodes used in the tree was 73.

[17] carried out an analysis of contextual effects of linguistic features on segment duration of spoken Korean in a news-reading speech using CART and SoP models. The two models were applied and evaluated on the corpus of 670 read sentences (52,840 segments) collected from one speaker of standard Korean. The duration of each segment and its phonological context were extracted from the corpus and the statistical modelling explored the relationship between the context features and the realised duration. The 69 phonological features were used in the analysis as shown below:

(1) 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
(2) Phonemic features of the preceding and the following segments
(3) Syllable structure: position and structure of containing syllable
(4) Position of syllables in UTT (utterance), IP (intonational phrase), AP (accent phrase) and PW (phonological word)

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. The SoP models used in the duration prediction of the vowels and consonants in [17] are as follows:

SoP model for vowels in [17]:

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

SoP model for consonants in [17]:

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

where “id” is the identity of the segment, “man” the manner of the target segment, "prev" the manner of the preceding segment, "foll" the manner of the following segment, "left_pos" the syllable distance to the left phrase boundary, "right_pos" the syllable distance to the right phrase boundary, "syll" the syllable structure, "syllpo" the segment position of the target segment in the syllable, i.e. onset or coda.

The best performance results of the CART and SoP models from [17] are summarized as follows:

<Table 3> CART and SoP performance results for vowels and consonants in [17].

	Vowels		Consonants	
	RMSE	Correlation	RMSE	Correlation
CART	25.51 ms	0.78	24.20 ms	0.71
SoP	36.89 ms	0.61	30.08 ms	0.51

5. Discussion

The strength of the sequential rule systems lies in the fact that the rules are derived directly from linguistic analysis and phonological structure so that they are easy to understand and to use. Rules might be common across languages or at least make explicit the differences between languages. Rule systems also have weaknesses, however. They are incomplete in that they only cover some phenomena. Rules tend not to be tested on varied material such as sentences of different lengths. Though YorkTalk tries to create a declarative formulation of knowledge, generally, rule interactions occur in the rule systems, which make them difficult to develop and extend. It is also not easy to adapt rule systems to changes in speaker, style, tempo, or genre.

On the other hand, 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. Also, in theory, it is possible to interpret the results of modelling. The weakness of CART models lies in the fact that it cannot interpolate between known contexts to find values for unknown contexts. This is particularly a problem when the data set is small or when the number of factors is large. Another weakness of this model is that it relies on objective function for partitioning that may not be the best in a perceptual sense. We also need to find ad hoc means to terminate tree growth. Despite the claim in [9], the interpretation of models can actually be quite difficult for large trees.

Finally, the strength of SoP models is that with relatively few parameters, durations can be well estimated from training data. Unlike CART models, they naturally interpolate to unseen contexts. A SoP 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 a wide variety of contexts.

6. Conclusion

This paper has overviewed and discussed three approaches to text-to-speech conversion and three approaches to duration modelling. Rule systems are now seen as a special kind of SoP model, so they are no longer taken seriously because of the difficulty in developing and maintaining them. On the other hand, CART models are simple to build and use, with good performance. SoP models have shown excellent performance, at least in English, but seem rather tricky to build. This might lead to the very poor performance in Korean SoP models. They require complex data analysis to unravel the interactions between factors. Among the synthesis methods, formant and diphone-style synthesis require a numerical model that predicts durations in context; while corpus-based synthesis needs to know which factors are most important for unit-selection to get good prosody. So the discussion in this paper is expected to contribute to building durational models of Korean using CART and SoP approaches to get reasonable performance and to uncover the most important contextual factors. It is hoped that this will make a contribution to the understanding of Korean timing which is still rather undeveloped and under-researched compared to English.

참 고 문 헌

[1] D. H. Klatt, "Interaction between two factors that influence vowel duration", *Journal of the Acoustical Society of America,* Vol. 54, No. 4, pp.1102-1104, 1973.

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

[3] J. Local, R. Ogden, "A model of timing for nonsegmental phonological structure", *Progress in Speech Synthesis,* J. P. H. van Santen, R. W. Sproat, J. P. Olive, J. Hirschberg (eds.), New York: Springer, pp.109-122, 1997.

[4] A. J. Hunt, A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", *Proc. ICASSP,* pp.373-376, 1996.

[5] D. H. Klatt, "Software for a cascade/parallel formant synthesizer", *Journal of the Acoustical Society of America,* Vol. 67, No. 3, pp.971-995, 1980.

[6] J. P. H. van Santen, "Prosodic modelling in Text-to-Speech synthesis", *Proc. Eurospeech,* 1997.

[7] W. N. Campbell, "Timing in speech: a multi-level process", *Prosody: Theory and Experiment,* Studies Presented to Gösta Bruce, Merle Horne (ed.), Dordrecht, the Netherlands: Kluwer Academic Publisher, pp.281-334, 2001.

[8] D. H. Klatt, "Review of text-to-speech conversion for English", *Journal of the Acoustical Society of America,* Vol. 82, No. 3, pp.737-793, 1987.

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

[10] N. Umeda, "Vowel duration in American English", *Journal of the Acoustical Society of America,* Vol. 58, No. 2, pp.434-445, 1975.

[11] N. Umeda, "Consonant duration in American English", *Journal of the Acoustical Society of America,* Vol. 61, No. 3, pp.846-858, 1977.

[12] L. Breiman, J. Friedman et al., *Classification and Regression Trees,* Monterey, Chapman & Hall/CRC, 1984.

[13] P. Deans, A. P. Breen, P. Jackson. "Cart-based duration modeling using a novel method of extracting prosodic features", *Proc. Eurospeech,* Vol. 4, pp.1823-1826, 1999.

[14] J. P. H. van Santen, J. P. Olive, "The analysis of contextual effects on segmental duration", *Computer Speech and Language,* Vol. 4, pp.359-391, 1990.

[15] Y. H. Lee, "Modelling of segment duration in Korean speech synthesis", *Phonetics and Linguistics in Honor of Professor Hyun Bok Lee,* Seoul National University Press, pp.249-274, 1996.

[16] S. Lee, Y.-H. Oh. "Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems", *Speech Communication,* Vol. 28, pp.283-300, 1999.

[17] H. Chung, *Analysis of the Timing of Spoken Korean with Application to Speech Synthesis,* PhD thesis, University College London, University of London, 2002.

접수일자 : 2003년 12월 22일
게재결정 : 2004년 3월 9일

▶ 정현성(Hyunsong Chung)
주소: 경북 경산시 진량읍 내리리 15
소속: 대구대학교 사범대학 영어교육과
전화: 053) 850-4125
FAX: 053) 850-4121
E-mail: hchung@daegu.ac.kr