

# Analysis and Modeling of Syllable Duration for Thai Speech Synthesis

Chatchawarn Hansakunbuntheung<sup>1</sup>, Virongrong Tesprasit<sup>1</sup>,  
Rungkarn Siricharoenchai<sup>1</sup>, Yoshinori Sagisaka<sup>2</sup>

<sup>1</sup>Information R&D Division  
National Electronics and Computer Technology Center, Thailand  
{chatchawarnh, virongr, rungakarn}@nectec.or.th

<sup>2</sup>Global Information and Telecommunication Institute  
Waseda University, Tokyo  
sagisaka@giti.waseda.ac.jp

## Abstract

This paper describes the analysis results on the control factors of Thai syllable duration, and a statistical control model using linear regression technique. The analyses have been carried out both at a syllable level and at a phrase level. In a syllable level duration control, the effects of five Thai tones and syllable structures are investigated. To analyze syllable structure effects statistically, we applied the quantification theory with two linguistic factors: (1) phone categories by themselves, and (2) the categories grouped by articulatory similarities. In a phrase level, the effects of position in a phrase and syllable counts in a phrase were analyzed. The experimental results showed that tones, syllable structures, and position in a phrase play significant roles on syllable duration control. Syllable counts in a phrase slightly affects the syllable duration. These analysis results have been integrated into a statistical control model. The duration assignment precision of the proposed model is evaluated using 2480-word speech data. Total correlation 0.73 between predicted values and observed values for test set samples shows the fair precision of the proposed control model.

## 1. Introduction

In speech synthesis, speech naturalness has been widely studied. Many studies have been conducted to explore which factors contribute to the naturalness of speech and how dominant they are. One of the most crucial factors that determine the naturalness is segmental duration. It is well-known within this field that segmental durations are controlled by many universal language factors such as phonetic composition, and position in a phrase. Though these control factors are quite similar, the same factor could play different roles across languages.

In Thai, there is little research that relates to duration control for speech synthesis. Most studies on segmental duration were carried out for purely linguistic interests. For examples, Luangthongkum [1] studied rhythm in Thai polysyllabic words and measured syllable duration in the words. The duration ratios of the syllables were roughly estimated. Referring to Luangthongkum's research, Surinpi boon's research limited its study to the rhythm of Thai accent in trisyllabic and tetrasyllabic words [2]. At the segmental level, Trongdee studied the phone duration of the non-stop consonants when occurring with three vowels (ii, aa, and uu) [3]. Similarly, Tarnsakun studied the phone duration of the stop consonants with the same vowels [4]. For the modeling of segmental duration, only a few rule-based models (e.g. Mittrapiyanurak [5]) have been proposed.

Contrastingly, statistical and modeling techniques have been widely applied in other languages. Regression tree approach [6] has been widely used in many languages. An artificial neural network is also used as an optimizer for the syllable duration calculation of British English by Nick Campbell [7]. Takeda [8] and Kaiki [9] applied the linear regression method to analyze and model segmental duration. Iwahashi [10] has proposed a constrained tree regression model that interpolates tree regression and linear regression. Although these statistical models are quite powerful by themselves, we need awareness of temporal control characteristics of the target language for the efficient modeling.

This study aims at: (1) the analyses of duration control factors on Thai syllable duration using the quantification theory and statistical linear regression method and (2) an interpretable duration modeling for Thai speech synthesis. (The speech domain of study is specific to reading speech style.) First, the control factors and the design of speech data are explained in Section 2. Next, the four analyses of duration effects using multiple linear regression method are employed to analyze duration effects in syllable level and phrase level. In the syllable level, the analyses consist of syllable structure and tone effect. Another level consists of analyses of phrase-position and phrase length effect. Finally, the four models of duration effects are integrated and evaluated.

## 2. Factors and speech data design for analyses

For the analyses of control factors of Thai syllable duration, we considered the following: the structural design of Thai syllables, Thai Tones, positional difference in a phrase, and syllable length of a phrase. The details of each factor are described below respectively.

### (a) The structural design of Thai syllables

As there are too many to collect all the Thai syllables in different contexts, we restricted the number of syllables based upon Thai syllable structure and their phonetic features. Thai has syllable structure written as  $C_i(C_d)V(C_f)$ , where  $C_i$  represents the initial consonants,  $C_d$  represents the initial double consonants,  $V$  represents the vowels, and  $C_f$  represents the final consonants. The phones were grouped by their phonetic features such as places, manner, and tongue position, and were selected to span the defined groups. The total number of distinct syllables used here is 1,735. In recording step, these syllables were recorded with a carrier sentence and located in the mid-sentence to avoid phrase-position effect. Each syllable was read 5 times.

### (b) Thai Tones

As Thai is a tonal language, the tone of syllables require analyzing. In Thai, tone alternation is constrained by two syllable types: (1) sonorant-ending syllables (called live syllables), and (2) obstruent-ending syllables (called dead syllables). Thai syllables can be alternated with 5 tones in live syllables and 4 tones in dead syllables. Therefore, the experimental data was constructed for both syllable types. A candidate of each syllable type was recorded with the carrier sentence and alternated with tones. Each syllable was read 5 times for each tones.

### (c) Positional difference in a phrase

To study the phrase-position effect, a carrier sentence was composed to carry the monosyllabic target words at the beginning, middle, and ending. The carrier sentence is composed of 6 syllables in this pattern, "XCCCXCCCX", where X represents the target words and C represents the syllables of the carrier sentences. Each sentence was read 5 times.

### (d) Syllable length of a phrase

To investigate the duration effects of the number of syllables in a phrase, we collected the sets of text, which randomly selected from radio-broadcasted scripts. The selected text had 397 phrases, which contains 2,840 words. The topic of the text related to general science knowledge for public. Next, the speech utterances of the text were recorded in reading style and tagged with phone identities.

## 3. Statistical analysis using linear regression

The speech data was analyzed to ascertain the contribution of above control factors on syllable duration using the Hayashi's quantitative theory. The analyses of temporal control factors were divided into two levels: syllable-level and phrase-level analysis.

### 3.1 Quantification Theory (Type I)

To analyze the factors of temporal control factors, we adopted an equation from the Hayashi's quantification theory (Type I) [11]. The theory statistically predicts the relationship between a response value and categorical values using the multiple linear regression method as the following equation:

$$\hat{y}_i = \bar{y} + \sum_f \sum_c x_{fc} \delta_{fc}(i) \quad i = 1, 2, 3, \dots, N \quad (1)$$

Where N represents the total number of data,  $\hat{y}_i$  represents the predicted syllable duration of the i-th sample,  $\bar{y}$  represents the average duration of all syllables,  $x_{fc}$  represents the regression coefficient, and  $\delta_{fc}(i)$  represents the characteristic function:

$$\delta_{fc}(i) = \begin{cases} 1 & : \text{if } i^{\text{th}} \text{ data is in} \\ & \text{the category c of the factor f} \\ 0 & : \text{otherwise} \end{cases} \quad (2)$$

$$\sum_i (\hat{y}_i - \bar{y})^2 \quad (3)$$

The regression coefficients  $x_{fc}$  can be calculated by minimizing equation (3) using a conventional multiple linear regression method

### 3.2 Syllable-level analysis

#### 3.2.1 Syllabic structure effect on syllable duration

In order to study the duration effect of syllabic structure, the collection of Thai syllables in section 2 (a) was used as the experimental data. In addition, two sets of phonetic categories were defined as the factors for the characteristic function. The first set determined the syllabic structures by phonemic identities as shown in Table 1. For another set, it determined the structure by articulatory categories such as places, manners, and tongue position. The categories are shown in Table 2.

Phone Class	Phonemic Identity Category
Initial consonant (Ci)	z, ch, b, ng, s, r, l, j
Initial double consonant (Cd)	r, l, w
Vowel (V)	i, ii, v, vv, u, uu, e, ee, q, qq, o, oo, x, xx, a, aa, @, @@, ia, iia, va, vva, ua, uua
Final consonant (Cf)	p <sup>^</sup> , t <sup>^</sup> , k <sup>^</sup> , m <sup>^</sup> , n <sup>^</sup> , ng <sup>^</sup> , w <sup>^</sup> , j <sup>^</sup> , z <sup>^</sup>

Table 1: The categories of phonemic identities.

Phone Class	Articulation Category
Initial consonant (Ci)	CiBilabial, CiAlveolar, CiPalatal, CiVelar, CiGlottal, CiPlosive, CiNasal, CiTrill, CiFricative, CiLateral, CiApproximant, CiVoiced, CiAspirated
Initial double consonant (Cd)	CdLabial, CdAlveolar, CdTrill, CdLateral, CdApproximant
Vowel (V)	VFront, VCentral, VBack, VHigh, VMid, VLow, VShort, VLong, VDiphthong
Final consonant (Cf)	CfBilabial, CfAlveolar, CfPalatal, CfVelar, CfNasal, CfApproximant

Table 2: The categories of articulation

In the first experiment, the categories of phonemic identities were applied with the quantification theory and the result is shown in Figure 1. In the next experiment, the theory was applied with the same data using the articulatory categories, and the result is shown in Figure 2. The relationship between the predicted duration and the control factors was expressed by the regression coefficients.

Figure 1 shows the duration effects as follows: (1) Ci and CiCd are the most dominant factors on the duration whereas Cf are the less ones, and (2) CiCd lengthen syllable duration whereas Ci shorten the duration.

Figure 2 shows the duration effects as follows: (1) most articulatory factors, especially the factors of Ci, shorten the duration, and (2) the short vowel category is clearly the most shortening factors in vowel categories.

In addition, we found the correspondences between the first experiment and the second experiment as the followings: (1) the order of V ranked by their dominance on the duration are short-diphthong, short, long-diphthong and long vowels respectively, and (2) the nasal Cf are the most dominant factors on the duration whereas the stop Cf are the less dominant ones.

The correlation value of the first experiment are 0.797 for the training set and 0.79 for the test set. The correlation value of the second experiment are 0.692 for the training set and 0.686 for the test set.

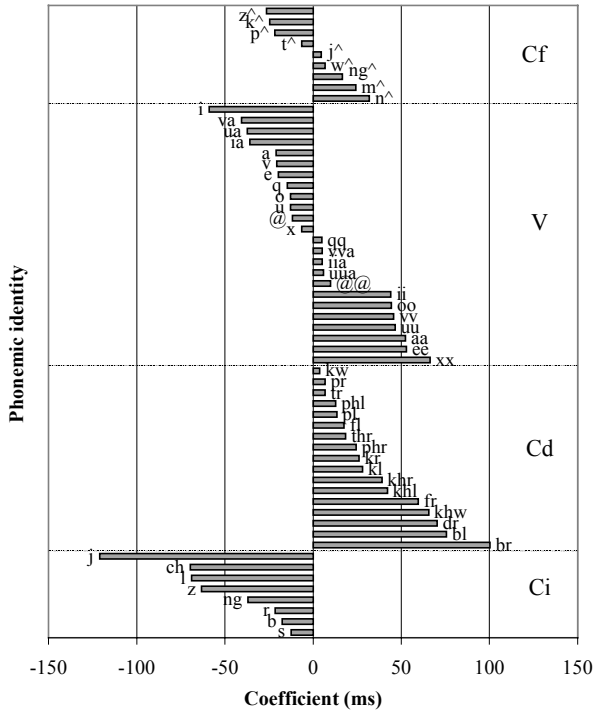


Figure 1: The duration effects of phonemic identities on syllable duration.

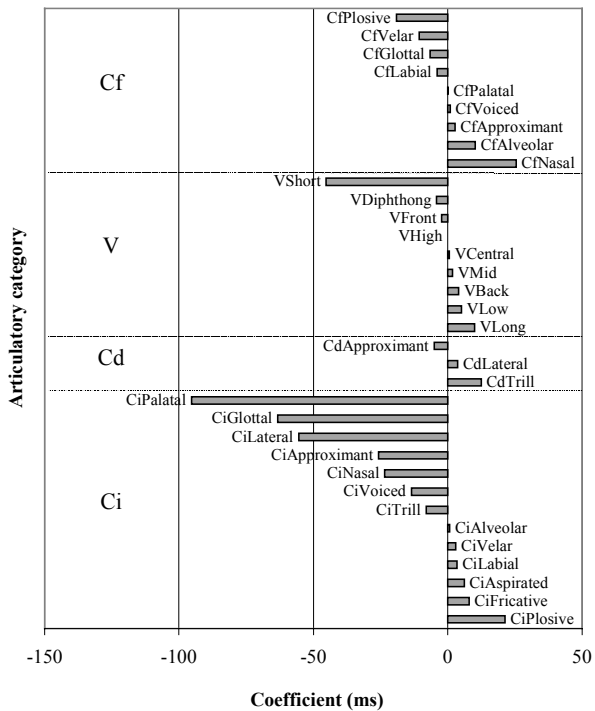


Figure 2: The duration effects of articulatory categories on syllable duration.

### 3.2.2 Tone effect on syllable duration

To analyze the effect of Thai tones on duration of live and dead syllables, the speech data in section 2 (b) was applied with the theory in section 3.1. Four-fifth of each speech data was used as a training set and the rest was used as a test set. After analyzing tone models, the result was illustrated in Figure 3 and the correlation values of the analysis model for live and dead syllables are shown in Table 3.

The result shows that tones moderately affect syllable duration and most of them, especially the falling tone, lengthen the syllable duration.

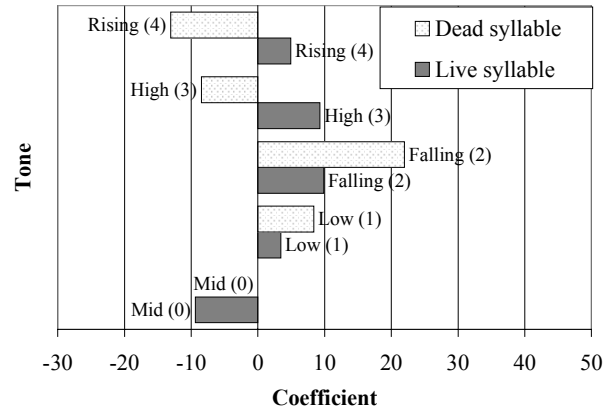


Figure 3: The duration effects of Thai tones on syllable duration.

Syllable Type	Correlation Coefficients	
	Training Set	Test Set
Live syllable	0.694	0.470
Dead syllable	0.862	0.623

Table 3: The correlation coefficients of tone effects analysis

### 3.3 Phrase-level analysis

At phrase level, the effects of phrase position and phrase length were investigated.

#### 3.3.1 Phrase position effect on syllable duration

To analyze this effect, the speech data in section 2 (c) was used for the experiment. The data was equally divided into a training set and a test set. Using the quantification theory, the regression coefficients of the phrase positions were analyzed and the result is shown in the Figure below.

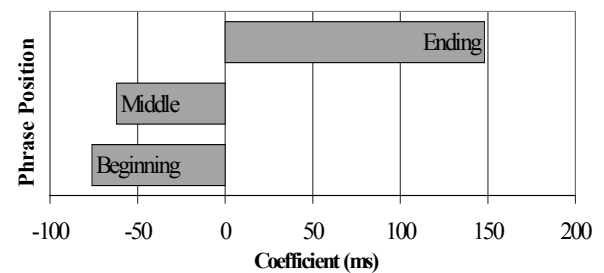


Figure 4: The duration effects of phrase position

The Figure 4 shows that phrase positions clearly affect the syllable duration, especially at the ending position of the phrases. The ending phrase position exhibits the lengthening effect. Meanwhile, the beginning position and the middle position exhibit the shortening effect on the duration.

The fitting of the analysis model were assessed. The assessment shows that the correlation values were 0.994 for the training set and 0.985 for the test set.

### 3.3.2 Analysis of phrase length effect on syllable duration

In this case, the length of a phrase was determined by the number of syllable in the phrase. We used the speech data in section 2 (d) and grouped the phrase utterances by phrase length, then calculated average syllable duration and standard deviation values of each group. Afterward, the values were plotted versus the number of syllables in the phrases as shown in Figure 5.

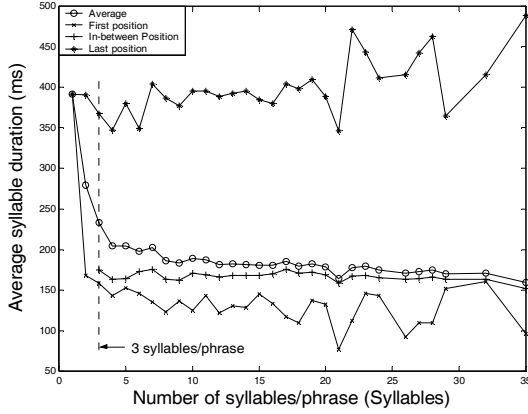


Figure 5: The relationship between the average syllable duration and the number of syllables

The Figure 5 shows that the average of syllable duration exponentially decreases with the number of syllables per phrase. This is especially true for the phrases with less than 3 syllables/phrase, the duration shows the rapid decreasing in rate. Meanwhile, the rate of the phrases, which have greater than 2 syllables/phrase, is slightly changed. The phenomenon of the duration of the phrases which have less than 3 syllables/phrase, can be described by determining the duration curves of the first position and the last position of the phrases. Both duration curves are effected by the phrase position effects as described in section 3.3.1. Furthermore, the number of syllables is not large enough to compensate for the effects. On the other hand, the duration curves of the phrases, which have greater than 2 syllables/phrase, is apparently flat. Hence, we express the curves by linear model.

To analysis the result statistically, we applied the data with the linear regression models as shown in the equation below:

$$\hat{y} = \bar{y} + ax \quad (4)$$

where  $\hat{y}$  is the predicted duration,  $\bar{y}$  is the average of all the syllable duration,  $x$  is the number of syllables in each phrase, and  $a$  is the regression coefficient.

After fitting the curves, the regression coefficients of the curves are shown in the table below.

Phrase Position	Coefficient (ms/syllable)
All	-0.374
First	-0.238
In-between	-0.006
Last	0.525

Table 4: The regression coefficients of the average syllable duration at different phrase positions

From Table 4, the coefficient values clearly show that the number of syllables in the phrases slightly affects the syllable duration, especially on syllables, which occur at in-between phrase position.

## 4. Evaluation of the duration models

To evaluate the fitness of the established duration models, we used the analysis coefficients form the experiment except the coefficients of the phrase length effects since they showed insignificant effect on syllable duration. The rest of coefficients are grouped into two models: (1) phonemic identities, tones, and phrase position, and (2) articulatory categories, tones, and phrase position. The speech data used here was the data in the section 2 (d). The result shows that the correlation values are 0.73 for the model using phonemic identities and 0.70 for the latter. The correlation values shows that the best model for this study is the former model.

## 5. Conclusion

This paper has statistically analyzed the duration effects of the control factors on the Thai syllables at the syllable level and at the phrase level. The factors used here are: (1) syllable structures using the phonemic identities and the articulatory categories, (2) tones, (3) phrase position in a phrase, and (4) phrase length. The control factors were applied with the quantitative theory and analyzed by using linear regression method. The analyses explored the significant information for improving the duration model for speech synthesis and related fields. Afterward, the models were evaluated and the result shows the fair precision of the proposed control models.

Finally, we need further research, such as the additional models in the phone level and word level, and the effects of part-of-speech, to improve the model.

## 6. References

- [1] Luangthongkum, T., *Rhythm in Standard Thai*, Ph.D. Thesis, University of Edinburgh, 1977.
- [2] Surinpiboon, S., *The Accentual System of Polysyllabic Words in Thai*, Master Thesis, Department of Linguistics, Chulalongkorn University, 1985.
- [3] Trongdee, T., *An Acoustic Analysis of Non-stop Consonants in Thai*, Master Thesis, Department of Linguistics, Chulalongkorn University, 1987.
- [4] Tarnsakun, W., *An Acoustic Analysis of Stop Consonants in Thai*, Master Thesis, Department of Linguistics, Chulalongkorn University, 1988.
- [5] Mittrapiyanuruk, P., Hansakunbuntheung, C., Tesprasit, V. and Somlertlamvanich, V., "Improving Naturalness of Thai Text-to-Speech Synthesis by Prosodic Rule", *Proceeding of the 6<sup>th</sup> ICSLP*, Vol. 3, pp. 334-337, 2000.
- [6] Riley, M.D., "Tree-based modeling of segmental durations", *Talking Machines* (edited by G. Bailly et. al.), North-Holland, 1992.
- [7] Campbell, W. N. and Isard, S. D., "Segment Durations in a Syllable Frame", *Journal of Phonetics, Special Issue on Speech Synthesis*, Vol. 19(1), pp. 37-48, 1991.
- [8] Takeda, K., Sagisaka, Y. and Kuwabara, H., "On Sentence-level Factors Governing Segmental Duration in Japanese", *Journal of the Acoustical Society of America*, Vol. 86 (6), pp. 2081-2087, 1989.
- [9] Kaiki, N., Takeda, N. and Sagisaka, Y., "Linguistic Properties in the Control of Segmental Duration for Speech Synthesis", *Talking Machines: Theories Models, and Designs*, Elsevier Science Publishers, 1992.
- [10] Iwahashi, N. and Sagisaka, Y., "Statistical Modeling of Speech Segment Duration by Constrained Tree Regression", *Transaction IEICE*, Vol. E83-D, pp. 1550-1559, 2000.
- [11] Hayashi, C., "On the Quantification of Qualitative Data from the Mathematico-Statistical Point of view", *Annals of the Institute of Statistical Mathematics*, Vol. 2, 1950.