Analysis of Segmental Duration for Thai Speech Synthesis

Chatchawarn Hansakunbuntheung[1] and Yoshinori Sagisaka[2]

[1]Information R&D Division, National Electronics and Computer Technology Center, Thailand
{chatchawarnh, chatchawarn.hansakunbuntheung}@nectec.or.th

[2]Global Information and Telecommunication Institute, Waseda University, Tokyo
sagisaka@giti.waseda.ac.jp

Abstract

This paper presents a characteristic study of Thai segmental duration and adapts the analysis results to construct a Thai phone duration model for Thai speech synthesis. The study uses Hayashi's categorized linear regression model to analyze the effects of various factors including current phonemes themselves, surrounding phonemes, phone positions in word, phone positions in phrase, part-of-speeches and Thai tones. These factors have combined to form a Thai phone duration model. The model gives rather high correlation of 0.788. Thought, it has fairly high RMS error of 33.14 ms, a evaluation shows the high consistency of the model on unknown data.

1. Introduction

Due to the requirements on speech applications and speech-to-speech research in recent years, many researches on speech synthesis extend their speech domain to new languages. Accordingly, new speech and language models are necessarily studied on. For Thai language, we need these models. Like others new languages, the primary research topics that need to be done for speech synthesis are finding appropriate speech units and building natural prosodic models. Since the present capacity of storage is large, the numerous speech utterances can be stored and retrieved to re-synthesize desired utterances. However, the quality of speech unit is just a factor of naturalness of synthetic speech. The prosodic model is another essential research topic of naturalness of synthetic speech. Many conventional methods were proposed for this topic. Nevertheless, the methods cannot be directly applied to a new language. Due to the characteristic distinction of each language, these characteristics need to be studied before applying the methods to those new languages. One of those fundamental characteristics of prosody is speech unit duration.

In the present, researches on speech unit duration model have many proposed methods. Several methods are based on neural networks [1], Classification and Regression tree (CART) [2], and linear regression model [3][4][5]. Neural-network-based methods often give appropriate results. However, the results are hard to interpret their meaning such as their weights and bias values, and the relationship between the input and the output of the networks. In case of CART, the whole model is not tied hence the effect of an input factor cannot be compared to the others. Unlike linear regression model, the whole model is fully tied.

In Thai, the number of research on speech duration is quite small. Most researches are specific to sets of phones [6][7], or are rule-based [8][9][10]. Hence, the whole structure of the duration model has not been studied.

In previous work [5], the primary study of duration model was carried out. The syllable-based duration model is studied and predicted by the multiple linear regression models. The model uses phone identities, articulation features, syllable position, tones and number of syllable in phrase as the input factors. The result shows significant of some factors on syllable duration. However, more analyses of underlying model are required to have more understanding of Thai duration model.

This paper aims to (1) analyze segmental duration effects on Thai speech, and (2) model Thai segmental duration prediction using linear regression analysis. To reach these aims, first, we select a speech corpus as the analyzing data. The information from the corpus is collected and grouped as input factors. Next, linear regression models are applied to the factors. Finally, the models and the factors are analyzed and evaluated.

2. Experimental data

In these experiments, the NECTEC's phonetically balanced Thai speech corpus [11] for speech synthesis (TSynC) is selected for analyzing. The corpus contains about 436,700 phones extracted from 5,200 sentences read by one female speaker. The speaker read the sentences with Thai central accent (Thai standard accent). The information that tagged on this corpus consists of:

- Phone, syllable, word, phrase and sentences boundaries
- Tone marks
- Syllable position in word and phrase
- F0, Energy
- Voiced/unvoiced, toned/toneless part

The corpus is separated into two sets: a training set and a test set. The training set contains about 90% of the corpus and the rest (about 10%) is the test set.

3. Experiments

In this step, a set of experiments is established to study effects of phonetic information, lexical information, syntactic information and prosodic information on phone duration.
First, the analyzing information is gathered from the corpus mentioned above and has details as below.

- Phonetic information: current phone (Ph_0), first preceding phone (Ph_{-1}), second preceding phone (Ph_{-2}), first succeeding phone (Ph_{+1}), second succeeding phone (Ph_{+2})
- Lexical information: part of speech (POS)
- Syntactic information: position in word (PosWrd), position in phrase (PosPhr)
- Prosodic information: tone (T)

Next, these factors are organized into several groups to study their effects on target phone duration. The groups of factors are listed below.

Table 1*: Analysis Factors.*

Individual factor								
Ph_0	Ph_{-1}	Ph_{+1}	Ph_{-2}	Ph_{+2}	T	POS	PosWrd	PosPhr
Factor group								
Ph_0 + T								
Ph_0 + T + POS + PosWrd + PosPhr								
Ph_{-1} + Ph_0 + Ph_{+1}								
Ph_{-2} + Ph_{-1} + Ph_0 + Ph_{+1} + Ph_{+2}								
Ph_{-2} + Ph_{-1} + Ph_0 + Ph_{+1} + Ph_{+2} + T + POS + PosWrd + PosPhr								

4. Segmental duration analysis method

To analyze the factors of temporal control factors, we adopted an equation from the Hayashi's quantification theory (Type I) [12]. The theory statistically predicts the relationship between a response value and categorical values using the multiple linear regression method as the following equation:

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

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 value of all data, x_{fc} represents the regression coefficient, and $\delta_{fc}(i)$ represents the characteristic function:

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

$$\sum_i (\hat{y}_i - y_i)^2 \qquad (3)$$

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

In this paper, we use the linear regression model and contextual information to predict phone duration. The coefficients from the regression model can represent the tied relationship between the whole information and the phone duration. From this point of view, the model in equation (1) can be interpreted as follow:

Predicted phone duration = Mean phone duration + summation of duration offsets (effects) of current phone, surrounding phones, tone, part-of-speech, and position in the word and phrase.

In addition, the regression coefficients and the root-mean-squared (RMS) error of the models at evaluation part can be used to analyze the effects.

5. Analysis results

After analyzing each factor group using the linear regression, the regression coefficients of each group are used for calculating their correlation coefficients and RMS errors as shown on Table 2 and 3. The correlation coefficients express the contribution of the factors on the segmental duration of the target phone while the RMS errors represent the fitting coefficient of the factors.

Table 2: *Correlation coefficients and RMS errors of the considering factors.*

Factors	Partial correlation coefficient	RMS error (ms)
Ph_0	0.442	48.60
Ph_{-2}	0.141	53.63
Ph_{-1}	0.246	52.50
Ph_{+1}	0.559	44.91
Ph_{+2}	0.472	47.74
T	0.133	53.67
POS	0.216	52.89
PosWrd	0.308	51.53
PosPhr	0.535	45.77
Factor group	**Correlation coefficient**	**RMS error (ms)**
Ph_0 + T + POS + PosWrd + PosPhr	0.701	38.67
Ph_{-1} + Ph_0 + Ph_{+1}	0.694	39.02
Ph_{-2} + Ph_{-1} + Ph_0 + Ph_{+1} + Ph_{+2}	0.775	34.25
Ph_{-2} + Ph_{-1} + Ph_0 + Ph_{+1} + Ph_{+2} + T + POS + PosWrd + PosPhr	0.791	33.20

Table 3: *Correlation coefficients and RMS errors of current phonemes grouped by their functions.*

Factors	Partial correlation coefficient	RMS error (ms)
Initial consonant phoneme	0.521	21.94
Vowel phoneme	0.440	64.54
Final consonant phoneme	0.241	49.80
Factors	**Correlation coefficient**	**RMS error (ms)**
Initial consonant phoneme + Tone	0.525	21.88
Vowel phoneme + Tone	0.447	64.29
Final consonant phoneme + Tone	0.253	49.67

5.1. Effects of current phones and surrounding phones

The correlation values in Table 2 and Figure 1 clearly show that phone duration highly depends on current phonemes, first succeeding phonemes and second succeeding phonemes. In addition, both succeeding phonemes also have higher correlation than current phonemes.

When determining current phonemes in detail, as shown in Figure 1, vowel phone duration and final consonant phone duration, also, have characteristics as the effects of current phone. Differently, initial consonant phone duration is significantly affected by current phonemes, first preceding phonemes and second succeeding phonemes, consequently. In addition, the correlation coefficients of vowel phonemes and final consonant phonemes themselves are obviously less than both the succeeding phonemes, that is, both highly control current phone duration and final consonant phone duration.

Accordingly, the RMS errors of the correlation coefficients of vowel phonemes and final consonantal phonemes are fairly high, as shown in Table 3.

In Japanese, Kaiki [4] shows that current phonemes, preceding phonemes and following phonemes are the significant factors on vowel phone duration. Nevertheless, the correlation values of the significant surrounding phonemes of both languages are higher than the correlation values of the vowel phonemes**.**

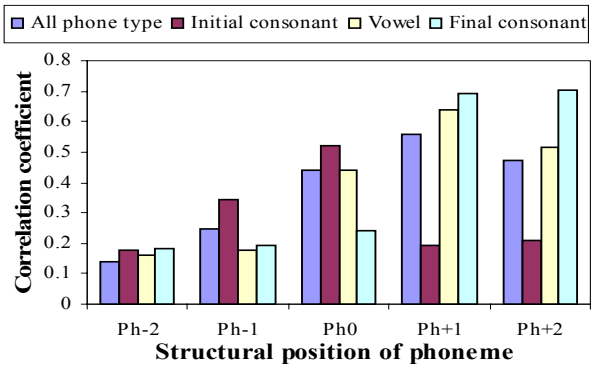

Figure 1: *Correlation coefficients of structural position of phoneme.*

5.2. Effects of tones

Table 2 and Table 3 show that tones have the least correlation to phone duration. Theirs regression coefficients and RMS errors show insignificant effects to phone duration.

However, the regression coefficients of all Thai tones, as shown in Figure 2, reveal an interesting result, that is, all static tones lengthen phone duration while all dynamic tones shorten it. In this case, the static tones – including low tone, mid tone and high tone – are tones that have incline-like or slight-movement F0 contours. Unlike the static tones, the dynamic tones – including falling tone and rising tone – are tones that have sharp-movement F0 contours.

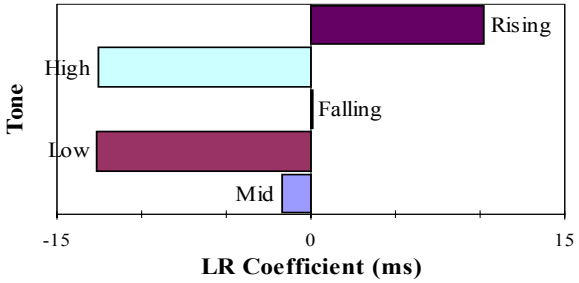

Figure 2: The *regression coefficients of tone factors.*

5.3. Effects of part-of-speech

The correlation coefficients of Part-of-speech, as shown in Table 2, have low influence on phone duration. However, the range of regression coefficients is wide and it reveals several interesting characteristics, that is, the directions of duration effects moderately correspond to types of part-of-speech as shown in Figure 3. Part-of-speeches that tend to lengthen phone duration are endings, adjectives, most of determiners, most of classifiers and most of adverbs. On the other hand, prefixes, conjunctions, negative, preposition, verbs and most of auxiliaries generally shorten phone duration. In noun, label noun and proper noun, which are uncommon noun, have longer duration than common noun.

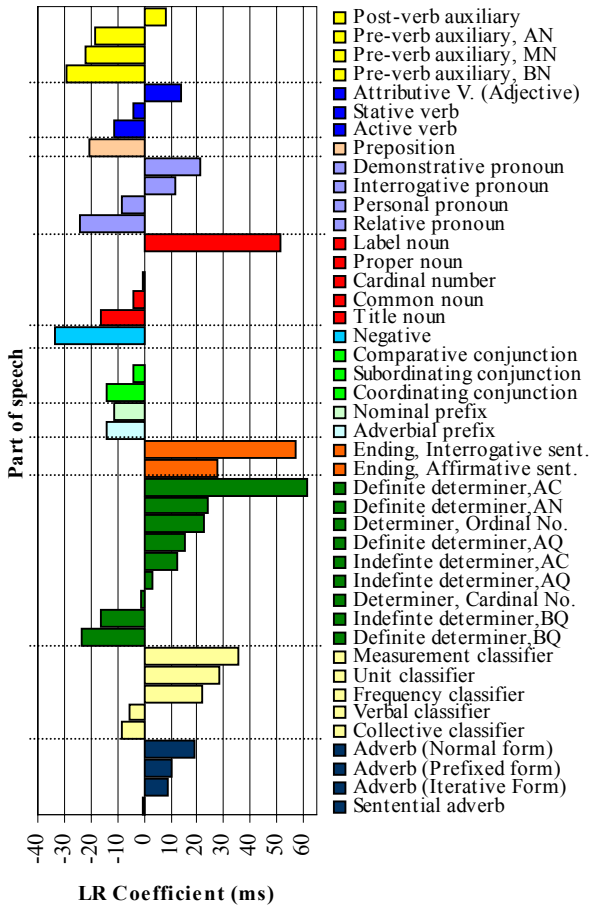

Figure 3: *The regression coefficients of part-of-speech factors.*

5.4. Effects of position in word

Figure 4 shows that duration effects of places in word can be observed. The phone duration is lengthened when placing phones at the end of word. In contrast, phones that occur at the beginning or in the middle of word, shorten the phone duration. This shortening effect similarly occurs with phones in monosyllabic word.

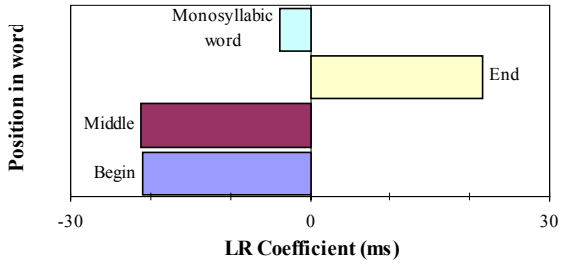

Figure 4: *The regression coefficients of position-in-word factors.*

5.5. Effects of position in phrase

Like position-in-word effects, the shortening effects happen when placing phones at the beginning or in the middle of phrase, while phones at the ending of phrase strongly lengthen the duration. These results can, also, be observed in the previous research [5] and in Japanese [4].

In contrast to position-in-word effects, phone duration in monosyllabic phrase is lengthened, not shortened duration.

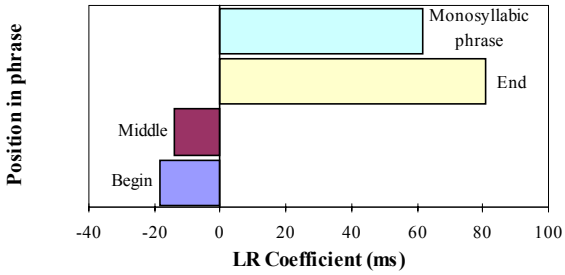

Figure 5: *The regression coefficients of position-in-phrase factors*

5.6. Effects of Factor Groups

After analyzing each factor, the results show positive correlation with estimated duration. Instead of studying each individual factor, several groups of the factors are analyzed. The analysis results in Table 3 show that the more factors are included in the model, the more accuracy of the model increases.

6. Evaluation

In this evaluation, all studied factors are integrated to form a Thai segmental duration model. The model consists of the following factors:

- Current phonemes
- First preceding phonemes
- Second preceding phonemes
- First succeeding phonemes
- Second succeeding phonemes
- Tones
- Part-of-speeches
- Positions in word
- Positions in phrase

To evaluate the model, the test corpus mentioned in section 2 is used. The evaluation results are shown in Table 4.

Table 4: *The evaluation result of the Thai segmental duration model.*

Data	Correlation coefficient	Average duration (ms)	RMS error (ms)
Training set	0.791	82.45	33.20 (40.27%)
Test set	0.788	84.54	33.14 (39.20%)

The testing results in Table 4 show that the model has rather high correlation with both training and test sets. Though, the RMS errors of the both data sets are quite high, the correlation value and the RMS error of the test set are both really close to the values of the training set (0.38% correlation coefficient difference and 0.18% RMS error difference comparing to the training set). In other words, the model is consistent to both known and unknown data.

7. Conclusion

This paper presents a characteristic study of Thai segmental duration and adapts the analysis results to construct a Thai phone duration model for Thai speech synthesis. The study shows that the significant factors consist of succeeding phonemes, current phonemes, preceding phonemes, positions in phrase, positions in word, part-of-speeches and tones, consequently. These factors have combined to form the Thai phone duration model. The model gives rather high correlation. Though, it has fairly high RMS error, the evaluation shows the high consistency of the model on unknown data.

In future works, the duration model for Thai speech needs more studies on duration factors extending from the factors in this papers, and it requires more suitable analysis method that can represent the duration characteristics of Thai speech.

8. References

[1] Campbell, W. N.; Isard, S. D., 1991. Segment Durations in a Syllable Frame, *Journal of Phonetics*, Special Issue on Speech Synthesis, Vol. 19(1), 37-48.

[2] Riley, M.D., 1992. Tree-based modeling of segmental durations, *Talking Machines*, G. Bailly et. al. (ed.), North-Holland.

[3] Takeda, K.; Sagisaka, Y.; Kuwabara, H., 1989. On Sentence-level Factors Governing Segmental Duration in Japanese, *Journal of the Acoustical Society of America*, Vol. 86 (6), 2081-2087.

[4] Kaiki, N.; Takeda, N.; Sagisaka, Y., 1992. Linguistic Properties in the Control of Segmental Duration for Speech Synthesis, *Talking Machines: Theories Models, and Designs*, Elsevier Science Publishers.

[5] Hansakunbuntheung, C.; Tesprasit, V.; Siricharoenchai, R.; Sagisaka, Y., 2003. Analysis and Modeling of Syllable Duration for Thai Speech Synthesis, *8th Eurospeech 2003*.

[6] Trongdee, T., 1987. *An Acoustic Analysis of Non-stop Consonants in Thai*, Master Thesis, Department of Linguistics, Chulalongkorn University.

[7] Tarnsakun, W., 1988. *An Acoustic Analysis of Stop Consonants in Thai*, Master Thesis, Department of Linguistics, Chulalongkorn University.

[8] Luangthongkum, T., 1977. *Rhythm in Standard Thai*, Ph.D. Thesis, University of Edinburgh.

[9] Surinpiboon, S., 1985. *The Accentual System of Polysyllabic Words in Thai*, Master Thesis, Department of Linguistics, Chulalongkorn University.

[10] Mittrapiyanuruk, P.; Hansakunbuntheung, C.; Tesprasit, V.; Sornlertlamvanich, V., 2000. Improving Naturalness of Thai Text-to-Speech Synthesis by Prosodic Rule, *Proceeding of the 6th ICSLP*, Vol. 3, 334-337.

[11] Hansakunbuntheung, C.; Tesprasit, V.; Sornlertlamvanich , V., 2003. Thai Tagged Speech Corpus for Speech Synthesis, *The Oriental COCOSDA 2003*, 97-104.

[12] Hayashi, C., 1950. On the Quantification of Qualitative Data from the Mathematico-Statistical Point of view, *Annals of the Institute of Statistical Mathematics*, Vol. 2.