論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Prosody Modeling Based on Gaussian Process Regression for Thai Speech Synthesis
著者(和文)	MoungsriDecha
Author(English)	Decha Moungsri
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第10923号, 授与年月日:2018年6月30日, 学位の種別:課程博士, 審査員:小林 隆夫,奥村 学,山口 雅浩,杉野 暢彦,篠﨑 隆宏
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10923号, Conferred date:2018/6/30, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Prosody Modeling Based on Gaussian Process Regression for Thai Speech Synthesis

Decha Moungsri

May 2018

Summary

This thesis describes techniques to improve prosody modeling for Thai speech synthesis. Prosody modeling is an important issue, since Thai is a tonal language that possesses complicated intonation characteristics. Hidden Markov model (HMM)-based statistical parametric speech synthesis (SPSS) is one of popular speech synthesis frameworks and has been implemented in many languages, including Thai. The conventional HMM-based framework models acoustic features at the state level and uses tree-based context clustering to handle a variety of contextual factors. Although the conventional HMM-based SPSS framework can be utilized for prosody modeling, the synthetic speech is still imperfect. The degradation of synthetic speech quality is mainly caused by limitations of HMM-based SPSS and the suprasegmental level has not been appropriately incorporated into the prosody modeling. To overcome this problem, this thesis focuses on an alternative framework called Gaussian process regression (GPR)-based SPSS and proposes novel techniques to incorporate the suprasegmental level into the prosody modeling.

First, this thesis describes an implementation of GPR-based Thai speech synthesis including the GPR-based framework, Thai linguistics, and definitions of contextual factors, and kernel functions. The GPR-based SPSS was first introduced in Japanese speech synthesis to overcome the limitations of HMM-based SPSS. The GPR-based SPSS uses Gaussian process (GP) to model the relationships between frame-level contextual factors and acoustic features. GP is a nonparametric Bayesian model in which the model complexity grows as the amount of training data increases. The GPR-based SPSS uses a kernel trick that is more flexible than the tree-based context clustering in determining the similarity of complicated contextual factors.

Another advantage is that the GPR-based SPSS framework models acoustic features at the frame level, which is more suitable than state-level modeling for acoustic features that change rapidly within one state. An experiment was conducted to evaluate the performance of the GPR-based framework by comparing it with the HMM- and DNN-based ones. The experimental result showed that the synthetic speech generated by the GPR-based method had more naturalness than the HMM- and DNN-based ones.

Secondly, this thesis describes novel techniques to incorporate suprasegmental features into the GPR-based duration and F0 modeling. The syllable level in the Thai language contains crucial linguistic functions such as stress, tone, and prominence that are primary factors of prosodic features. The conventional single-level model is insufficient for capturing these factors. To overcome the limitations of a single-level model, multi-level-model techniques were proposed for duration and F0 modeling. This thesis proposed two methods of multi-level duration modeling: two-stage prediction and the product of Gaussian process experts. Two-stage prediction uses phone and syllable-level duration models that are trained separately, and the predicted syllable duration from the syllable-duration model is used as an additional context in phone-level duration prediction. In the product of Gaussian process experts, the predictive distributions of phone- and syllable-duration models are combined by the product of experts framework. The mean of the combined predictive distribution is used as the predicted phone duration. This thesis examines multi-level F0 modeling by combining frame- and syllable-level models. F0 contours are generated by jointly maximizing predictive distributions of frame- and syllable-level models. The experimental results showed that the multi-model methods outperformed the single-model one.

Lastly, this thesis describes the use of stress information to improve prosody generation, because stress is a major factor that affects prosody. However, Thai is a non-lexical stress language whose stress cannot be obtained from the input text, and the manual labeling of stress information is time-consuming. To overcome such problems, an unsupervised technique was proposed to annotate the speech corpus with stress information. First, a dimensionality reduction technique, called the Gaussian process latent variable model (GP-LVM), was used to project acoustic features of stress onto

a latent space in which the similarity between syllables can be easily observed. Then, an unsupervised clustering was performed to classify syllables into stressed and unstressed classes. The classification result showed that the use of the latent variables can achieve higher accuracy than the use of the observed acoustic features. In the experiment, the stressed/unstressed classes were used as an additional context in GPR-based prosody generation. Performance comparison of the GPR-based method was done using stress information obtained from manual labeling and unsupervised labeling. The objective evaluation showed that the prosodic features generated by the unsupervised labeling yielded a comparable result to the manual labeling technique. The subjective evaluation confirmed that the manual and unsupervised labeling methods produced similar quality of speech in terms of naturalness.

Acknowledgments

I would like to thank to Professor Takao Kobayashi of Tokyo Institute of Technology for his kindness, support and opportunity for scholarship and studying in Japan. Additionally, I also appreciate Doctor Tomoki Koriyama for his support and encouragement. Without their support and useful advice, this thesis cannot be achieved.

I would like to thank all my colleagues at Kobayashi Laboratory for their suggestion, friendship, and awesome experience during five years, especially Doctor Vataya Chunwijitra. Moreover, I would like to thank my Thai friends at Tokyo Institute of Technology for their help in the evaluations.

Also, I am really grateful to Japanese government (Monbukagakusho) scholarship for financial support during these five years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General background

Voice user interface (VUI) enables human-machine interaction through voice/speech. One of the advantages of using VUI is that it provides hands-free and eyes-free interaction. VUI consists of many processes such as automatic speech recognition, natural language understanding and generation, and text-to-speech synthesis. Recently, various mobile phones and digital assistants such as Apple Siri, Google home, and Amazon Alexa have incorporated VUI as the primary feature.

Speech synthesis is an essential part of VUI that allows a device to generate speech response to the user. The goal of speech synthesis is to generate natural-sounding speech. In the last decade, the statistical parametric speech synthesis (SPSS) approach has become the mainstream technique for speech synthesis. Indeed, hidden Markov model (HMM)-based SPSS [1] is the most successful conventional framework for speech synthesis. The HMM-based SPSS uses HMMs to simultaneously model spectral feature, fundamental frequency (F0), and duration. Then, the trained HMMs are used to generate speech parameters from a given input text. Decision tree-based context clustering [2] is employed to handle the variety of contextual factors and unseen input contexts. The HMM-based approach has been utilized for various applications such as text-to-speech, voice conversion [3], adaptation [4], singing voice synthesis [5], and style control [6]. Although the HMM-based SPSS

can generate a fair quality speech, it has the limitations that degrade the quality of synthetic speech. First, in the HMM-based SPSS, acoustic features of multiple frames are integrated into one state and represented by a state-level feature parameter set. This approach is inappropriate for acoustic features that change rapidly within a state. Secondly, the tree-based context clustering is incapable of handling complicated context. Specifically, it is ineffective in expressing concepts such as XOR, parity, and multiplexer. For such cases, decision tree could become large and eventually leads to an over-fitting problem. Another limitation is that the tree-based context clustering ties the states assigned to a leaf node into one state that decreases the context diversity.

Recently, the use of graphics processing unit (GPU) computing successfully accelerates the training of deep neural networks (DNNs) with a large amount of training data. This allows DNNs to be utilized for various applications. DNN-based SPSS [7] was introduced to overcome the limitations of the HMM-based method. The DNN-based SPSS uses frame-level acoustic features and contextual factors as output and input of DNN, respectively. Then, DNN is trained to learn the relationship of acoustic features to input contexts. Various network architectures have been proposed for DNN-based SPSS, such as recurrent neural network, deep belief networks, and long short-term memory recurrent neural network [8–10].

Gaussian process regression (GPR)-based SPSS [11–13] is another framework that has been proposed to overcome the limitations of the HMM-based SPSS. The GPR-based method uses Gaussian process (GP) to model frame-level acoustic features and contextual factors. The regression is performed by calculating the predictive distribution for the given input context. The GPR-based method uses a kernel trick to handle the context diversity. The advantages of the GPR-based SPSS are its flexibility and adaptability to various types of contextual factors. GP is a nonparametric Bayesian model in which the model complexity grows as the amount of training data increases. Furthermore, GP uses a Bayesian inference which is a robust parameter estimation that can avoid over-fitting. A drawback of GPR-based SPSS is its high computational cost. However, the use of GPU and computational optimization techniques [14, 15] can alleviate the problem. The GPR-based

SPSS showed a performance improvement relative to the HMM-based one while it produced a comparable result to the DNN-based one with the limited amount of training data [16]. However, GPR-based SPSS is still imperfect because a frame-level model is insufficient in capturing the prosodic features at suprasegmental level.

1.2 Scope of thesis

This thesis focuses on prosody modeling for GPR-based SPSS. The purpose is to improve the prosody modeling by incorporating suprasegmentals. This study case chose Thai as the target language, which is a tonal language that has a complex F0 movement. Moreover, suprasegmentals play a crucial role in prosody. Therefore, it is expected that incorporating suprasegmental level can show improvement in prosody modeling. Lastly, Thai has clear boundaries of suprasegmental units, such as syllable or word, then an error from incorrect boundary-labeling can be avoided.

Chapter 2 introduces a GPR-based Thai SPSS framework, including an explanation of Thai contextual factors and kernel functions. The experimental section shows a comparison of HMM-, DNN-, and GPR-based Thai speech synthesis. Chapter 3 proposes a multi-level model method, called a two-stage method, for duration prediction. The two-stage method consists of syllable- and phone-duration models. The training of phone-duration model uses syllable duration as an additional context. The syllable-duration model is trained to predict syllable duration to be used as an additional context of input text in phone-duration prediction. The experiments are performed with the two-stage method with DNN- and GPR-based duration prediction, and a comparison with a multi-level HMM-based one. Chapter 4 proposes an alternative multi-level model method for GPR-based duration prediction. This method simultaneously trains phone- and syllable-duration models. For performing phone-duration prediction, predictive distributions of syllable- and phone-level duration models are calculated. Then, a product of expert technique is used to combine these predictive distributions, and the mean of the combined distribution is used as predicted phone-duration sequence. Chapter 5 proposes a multi-level model method for F0 generation.

This method uses frame- and syllable-level models. Acoustic features of the syllable-level model are the discrete cosine transform (DCT) coefficients extracted from log F0 contour. In this method, the generated log F0 contour is obtained by maximizing a log-combined predictive distribution of the frame- and syllable-level models. Chapter 6 proposes an automatic method to annotate a speech corpus with stress information. An experiment was conducted to evaluate the performance of using stress information and the accuracy of the automatic annotation. Finally, Chapter 7 presents the conclusion of this thesis and future work.

Chapter 2

Gaussian Process Regression-Based Thai Speech Synthesis

shows the summary of temporal event context.

2.1 Introduction

In the last decade, HMM-based SPSS has been used for Thai speech synthesis [17]. Thai is a tonal language, like Chinese and Vietnamese, that tone has a role in distinguishing the meaning of words. Therefore, tone modeling is a crucial topic of Thai speech synthesis. Various studies have focused on improving tone modeling in Thai. A tone-separated decision tree structure was proposed to avoid different tone-type assigned to the same leaf-node of a decision tree [18]. However, the tone-separated tree is still insufficient for representing diversity, since tone contours can have multiple shape even in the same tone-type. Several approaches extract additional features from acoustic features and incorporate them as contexts in SPSS. The tone-geometrical features which extracted from F0 contours and syllable duration were proposed to represent the variety of tone contours [19]. T-Additionally, Tilt [20, 21], an extension of Tilt model [22], was another method used for expressing the shape of tone contour. A quantized F0 symbol [23] was proposed to repre-

Table 2.1: Thai phonemes and tone in IPA.

Syllable components	Type	IPA symbol	Total
Initial consonant	Single consonant	p, t, c, k, ʔ, p^h, t^h, c^h, k^h, b, d, m, n, ŋ, f, s, h, r, l, w, j	21
	Cluster consonant	pr, pl, tr, kr, kl ,kw, p^hr, p^hl, t^hr, k^hr, k^hl, k^hw	12
Vowel	Short vowel	i, ɨ, u, e, ɜ, o, æ, a, ʉ, ia, ɨa, ua	12
	Long vowel	iː, ɨː, uː, eː, ɜː, oː, æː, aː, ʉː, iːa, ɨːa, uːa	12
Final consonant		p, t, c, k, m, n, ŋ, w, j	9
Tone		ā (0), à (1), â (2), á (3), ǎ (4)	5

sent the diversity of tone contours by dividing phone-duration interval into equal portions, then calculates a quantized F0 symbol from F0 value that assigns in a portion. These features were used as an additional context in the HMM-based SPSS to improve the quality of speech synthesis.

These techniques could alleviate the tone correctness problem. However, Thai synthetic speech is still imperfect. One of the reasons is that the tree-based context clustering is ineffective for Thai that has complicated contextual factors. Another reason is that a discrete state is insufficient to model the change within a state of acoustic features. These problems are caused by the limitations of the HMM-based method. Therefore, the GPR-based SPSS was investigated to improve the quality of Thai synthetic speech.

This chapter examines GPR- and DNN-based Thai SPSS systems and compare their performance with the HMM-based one. The contextual factors and kernel function are defined to apply the GPR-based SPSS to Thai. The experimental results are shown in both objective and subjective evaluations.

Table 2.2: Thai consonant articulatory.

			Place of articulation				
			Labial	Alveolar	Palatal	Velar	Glottal
Manner of articulation	Stops	Voiceless unaspired	p	t	c	k	ʔ
		Voiceless aspired	p^h	t^h	c^h	k^h	
		Voiced	b	d			
	Non-Stops	Nasal	m	n		ŋ	
		Fricative	f	s			h
		Trill		r			
		Lateral		l			
		Approximant	w		j		

Table 2.3: Thai vowels in the IPA system.

		Vowel advancement		
		Front	Central	Back
Vowel height	High	i, iː	ɨ, ɨː	u, uː
	Mid	e, eː	ɜ, ɜː	o, oː
	Low	æ, æː	a, aː	ʉ, ʉː
Diphthongs		ia, iːa	ɨa, ɨːa	ua, uːa

2.2 Thai characteristics and prosody

The sound system of the Thai language is often described in terms of syllable units [24]. Thai syllables are combinations of four components as follows:

$$\begin{matrix} & T & \\ C_i & - & V - (C_f) \end{matrix}$$

where C_i, V, C_f, and T are initial consonant, vowel, final consonant, and tone, respectively. Table 2.1 is a summary of the syllable components.

2.2.1 Thai phonological system

The initial-consonant can be a single or a cluster consonant. The cluster consonant is pronounced as one of / p, p^h, t, t^h, k, k^h /, followed by one

of / r, l, w /. Table 2.2 lists the manner and place of articulation of Thai consonants. The vowel can be a single vowel or diphthong. The diphthongs are pronounced as one of / i, iː, ɨ, ɨː, u, uː / followed by / a /. The vowels can be classified into short vowels and their counterpart, long vowels. Table 2.3 is an articulation chart of Thai vowels. The final-consonant is a single phone, and it can be absent in a syllable.

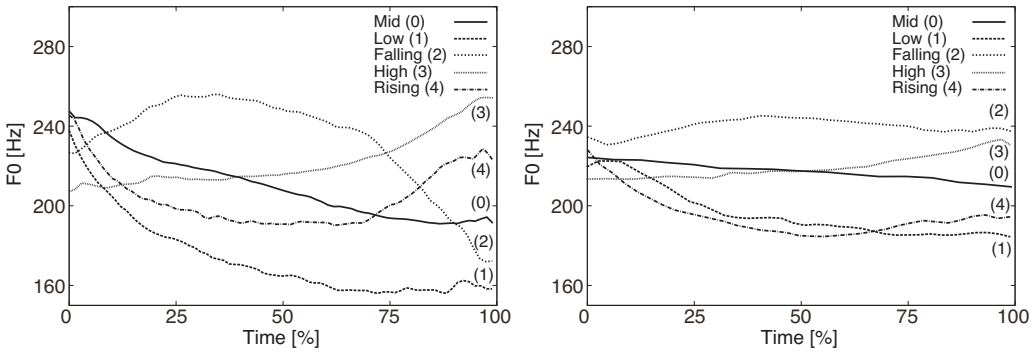
In this study, consonants from loan-words that are not included in native Thai phonemes are included in the speech database used in the experiments. The loan-word initial consonants are / br, bl, fr, fl, dr / and the final consonants are / f, s, c^h, l /.

2.2.2 Thai Prosody

The main factor of Thai prosody is tone-type. Thai syllables are pronounced with one of five tones. Five IPA tone markers or digits are used to indicate tone type as follows: mid (ā or 0), low (à or 1), falling (â or 2), high (á or 3), and rising (ǎ or 4). Characteristics of each tone can be represented by the F0 contour in the syllable unit. Five tones can be classified into two groups, static tones (Tones 0, 1 and 3) and dynamic tones (Tones 2 and 4), in which the F0 contours of the dynamic ones change faster than the static ones. Tone is a crucial factor to distinguish meanings of words having the same phone sequence. For example, "k-a-jˆ-0"[1] means "far" and "k-a-jˆ-2" means "near".

Thai prosodic features are usually studied in terms of syllable units. In continuous speech, the F0 contour and duration in syllable units are influenced by various factors. To investigate the diversity of F0 contours and durations, many studies have classified syllables into simple stressed/unstressed classes. Stressed syllables are similar in shape to the typical contours and have long durations [25]. Unstressed syllables are diverse in F0 contour shape and have short durations. Most studies of stress in Thai conclude that a stressed syllable is an emphasized one and the last syllable of a word or phrase. Many studies described additional rules of stressed syllables. The previous studies [26, 27] showed examples of stress position for various words

[1]Caret (ˆ) and the digit indicate final-consonant and tone-type, respectively.



(a) F0 contours of stressed syllable (b) F0 contours of unstressed syllable

Figure 2.1: Example of F0 contours in (a) stressed and (b) unstressed syllables.

and described the pattern of stress at the word level. A comprehensive study of Thai intonation [28] showed that the position of a stress affects the perception of sentence structure. The speaking rate reduces F0 movement in both stressed and unstressed syllables [29]. In, [30], it showed that F0 variability depends on the tones of neighboring syllables. The lengths of vowels and final consonants are affected by whether the syllable is stressed or unstressed [31]. Figure 2.1 shows an example of F0 contour shapes of each tone in stressed and unstressed syllables which are the same phones, and were extracted from speech samples included in Thai speech database TSynC-1 [32].

2.3 Implementation of GPR-based Thai speech synthesis

2.3.1 Gaussian process regression for speech synthesis

The idea of speech synthesis is to model the relationship between the acoustic features and contextual factors. Gaussian process regression was introduced to statistical parametric speech synthesis by defining the input/output variables of GP and a kernel function to calculate the covariance matrix are defined [11–15].

GPR-based SPSS defines the contextual factors and acoustic feature as

input and output variables of GP, respectively. In GPR, the relation between output variable y_n and input variable $\mathbf{x}_n$ is defined by

$$y_n = f(\mathbf{x}_n) + \epsilon \tag{2.1}$$

where $f(\cdot)$ is a noise-free latent function and ϵ is Gaussian noise with variance σ^2. The matrix forms of the input and output variables of training data are $\mathbf{X}_N = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]^\intercal$ and $\mathbf{y} = [y_1, y_2, ..., y_N]^\intercal$, and $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_N)]^\intercal$, respectively. $\mathbf{X}_T$ and $\mathbf{y}_T$ are denoted those of test data. $\mathbf{y}$ and $\mathbf{y}_T$ are sampled from a GP, then the joint distribution of $\mathbf{y}$ and $\mathbf{y}_T$ is expressed as

$$p(\mathbf{y}, \mathbf{y}_T | \mathbf{X}_N, \mathbf{X}_T) = \mathcal{N}(\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_T \end{bmatrix}; 0, \mathbf{K}_{N+T} + \sigma^2 \mathbf{I}). \tag{2.2}$$

$$\mathbf{K}_{N+T} = \begin{bmatrix} \mathbf{K}_{NN} & \mathbf{K}_{NT} \\ \mathbf{K}_{TN} & \mathbf{K}_{TT} \end{bmatrix} \tag{2.3}$$

where $\mathbf{K}_{NN}$ and $\mathbf{K}_{TT}$ are the covariance matrices of the training and test data, respectively, and $\mathbf{K}_{NT} = \mathbf{K}_{TN}^\intercal$ is the covariance matrix between training and test data. The (m, n) element of the covariance matrix is given by $k_{mn} = \kappa(\mathbf{x}_m, \mathbf{x}_n)$, where $\kappa(\mathbf{x}_m, \mathbf{x}_n)$ is a kernel function for calculating the similarity between input variables $\mathbf{x}_m$ and $\mathbf{x}_n$. To perform a regression, the predictive distribution of $\mathbf{y}_T$ is given by

$$p(\mathbf{y}_T | \mathbf{y}, \mathbf{X}_N, \mathbf{X}_T) = \mathcal{N}(\mathbf{y}_T; \mu_T, \Sigma_T). \tag{2.4}$$

$$\mu_T = \mathbf{K}_{TN}[\mathbf{K}_{NN} + \sigma^2 \mathbf{I}]^{-1} \mathbf{y} \tag{2.5}$$

$$\Sigma_T = \mathbf{K}_{TT} + \sigma^2 \mathbf{I} - \mathbf{K}_{TN}[\mathbf{K}_{NN} + \sigma^2 \mathbf{I}]^{-1} \mathbf{K}_{NT}. \tag{2.6}$$

The parameters of predictive distribution μ_T and Σ_T are used in generating speech parameters.

[11–15] applied GPR into speech synthesis by proposing a frame-level context for spectral features, aperiodicity, and F0, and a phone-level context for phone duration modeling. To apply the GPR-based method into Thai, the frame-level context and a kernel function are defined.

Table 2.4: Examples of temporal events of Thai phonetic features for GPR-based speech synthesis.

Phonetic features	p	pl	a	uːa	kˆ
Labial	+	+	−	−	−
Alveolar	−	−	−	−	−
Velar	−	−	−	−	+
Glottal	−	−	−	−	−
Voiceless unaspired	+	+	−	−	+
Lateral	−	+	−	−	−
Cluster	−	+	−	−	−
Vowel Short	−	−	+	−	−
Vowel High	−	−	−	−	−
Vowel Low	−	−	+	+	−
Vowel Front	−	−	−	−	−
Vowel Central	−	−	+	+	−
Vowel Back	−	−	−	−	−
Dipthong	−	−	−	+	−
Initial consonant	+	+	−	−	−
Vowel	−	−	+	+	−
Final consonant	−	−	−	−	+
⋮	⋮	⋮	⋮	⋮	⋮

2.3.2 Frame-level context for Thai language

A context of a partial frame $\mathbf{x}_n = (x_{n,1}, \ldots, x_{n,K})$ is an array of temporal event contexts. $x_{n,k} = (\mathrm{p}_{n,k}, c_{n,k})$ is the context of k-th event contains the relative position context $\mathrm{p}_{n,k}$ and the k-th event context $c_{n,k}$. $\mathrm{p}_{n,k} = (\mathrm{p}_{n,k}^{(-1)}, \mathrm{p}_{n,k}^{(0)}, \mathrm{p}_{n,k}^{(+1)})$ and $c_{n,k} = (c_{n,k}^{(-1)}, c_{n,k}^{(0)}, c_{n,k}^{(+1)})$ are included preceding (-1), current (0), and succeeding $(+1)$ temporal events for corresponding speech unit. The event context $c_{n,k}^{(u)}$ is a vector that represents a particular linguistic information. The relative position context $p_{n,k}^{(u)}$ is a vector that represents a distance from a frame to an event in corresponding unit-scales.

Thai frame-level context is derived from linguistic information that available in T-Sync-1, a Thai speech corpus. T-Sync-1 corpus provides four layers of speech units, phone, syllable, word, and utterance. The temporal event

Table 2.5: Temporal context for Thai GPR-based speech synthesis.

Unit:	phone
Type:	{beginning, end} of each phonetic feature
Scale:	phone-normalized scale, time*[a]
Unit:	syllable
Type:	{beginning, end} of tone-type
Scale:	{syllable, word}-normalized scale, time*
Unit:	word
Type:	{beginning, end} of part of speech
Scale:	{syllable, word}-normalized scale, time*
Unit:	utterance
Type:	{beginning, end} of utterance
Scale:	{syllable, word, utterance}-normalized scale, time*

[a]The scales marked with * are not used for the duration model.

contexts consisted of linguistic information of these layers. In phone layer, the temporal event contexts were the beginning and end of phonetic features. Each phonetic feature was represented by a binary value (+1 or -1) as an event context. Table 2.4 shows an example of Thai phonetic features. In syllable layer, the temporal event contexts were the beginning and end of tone-type which was represented by a one-hot vector. In word layer, the temporal event contexts were the beginning and end of part-of-speech (POS) which was represented by a one-hot vector. In utterance layer, no linguistic information is provided in T-Sync-1, then the temporal event contexts were the beginning and end of an utterance. Each temporal event context included a vector of relative position context. Table 2.5 shows the summary of temporal event context.

Figure 2.2 illustrates an example of frame-level context defined for the beginning of tone-type event $x_{n,k}$. $\mathrm{p}_{n,k}^{(-1)}$, $\mathrm{p}_{n,k}^{(0)}$, and $\mathrm{p}_{n,k}^{(+1)}$ denotes the relative position from the beginning of preceding, current, and succeeding tone-type temporal events to the n-th frame position, respectively. The relative position $\mathrm{p}_{n,k}^{(u)}$ was represented by a vector which the elements was the distance from the frame to an event with different scales as described in Table 2.5. Then,

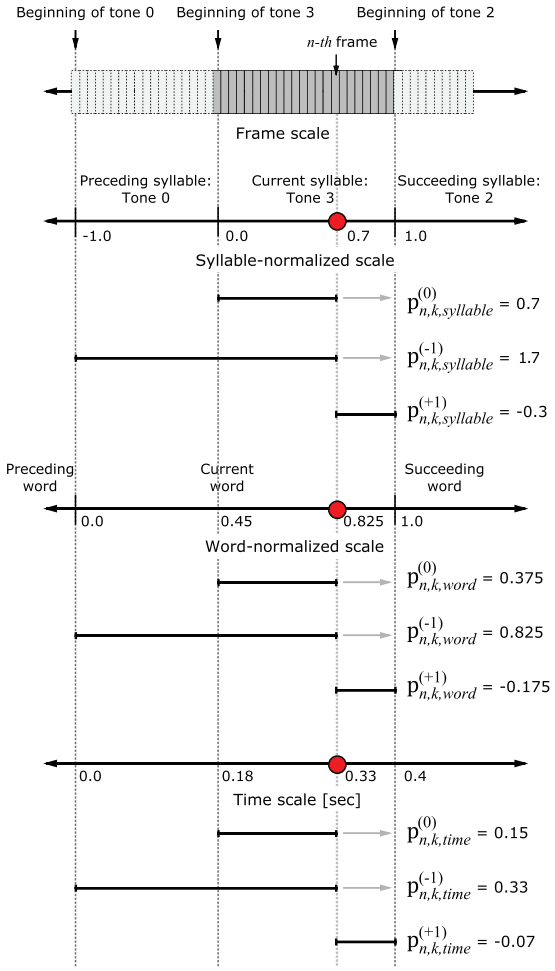


Figure 2.2: Illustrative example of contextual factor of beginning of tone-type temporal event $x_{n,k} = (\mathrm{p}_{n,k}, c_{n,k})$.

$\mathrm{p}_{n,k}^{(u)}$ was three elements which express as

$$\mathrm{p}_{n,k}^{(u)} = [\mathrm{p}_{n,k,syllable}^{(u)}, \mathrm{p}_{n,k,word}^{(u)}, \mathrm{p}_{n,k,time}^{(u)}], \quad u = -1, 0, +1$$

where the actual value of $\mathrm{p}_{n,k}^{(u)}$ from Figure 2.2 is

$$\begin{aligned}\mathrm{p}_{n,k} &= (\mathrm{p}_{n,k}^{(-1)}, \mathrm{p}_{n,k}^{(0)}, \mathrm{p}_{n,k}^{(+1)}) \\ &= ([1.7, 0.825, 0.33], [0.7, 0.375, 0.15], [-0.3, -0.175, -0.07]).\end{aligned}$$

$c_{n,k}^{(-1)}$, $c_{n,k}^{(0)}$, and $c_{n,k}^{(+1)}$ denote tone-types of preceding, current, and succeeding

syllables, respectively. The temporal event context $c_{n,k}$ is expressed as

$$\begin{aligned} c_{n,k} &= (c_{n,k}^{(-1)}, c_{n,k}^{(0)}, c_{n,k}^{(+1)}) \\ &= ([1,0,0,0,0], [0,0,0,1,0], [0,0,1,0,0]) \end{aligned}$$

where $[1,0,0,0,0]$, $[0,0,0,1,0]$, and $[0,0,1,0,0]$ represent tones 0, 3, and 2, respectively.

2.3.3 Kernel function

The kernel function is essential to determine the similarity between two frame contexts. Since the frame-level context is an array of temporal event contexts, then the kernel function $\kappa(\mathbf{x}_m, \mathbf{x}_n)$ is the sum of the similarities of each temporal event $\kappa_k(x_{m,k}, x_{n,k})$ which is expressed as

$$\kappa(\mathbf{x}_m, \mathbf{x}_n) = \sum_{k=1}^{K} \theta_{r,k}^2 \kappa_k(x_{m,k}, x_{n,k}) + \delta_{mn}\theta_{floor}^2 \tag{2.7}$$

where $\theta_{r,k}^2$ and θ_{floor}^2 are kernel parameters. The kernel function $\kappa_k(x_{m,k}, x_{n,k})$ of a temporal event is defined particularly for each temporal event [2]. The kernel function $\kappa_k(x_{m,k}, x_{n,k})$ is expressed as

$$\kappa_k(x_{m,k}, x_{n,k}) = \sum_{u=-1}^{+1} \sum_{v=-1}^{+1} [w(\mathrm{p}_{m,k}^{(u)}) w(\mathrm{p}_{n,k}^{(v)}) \cdot \kappa_p(\mathrm{p}_{m,k}^{(u)}, \mathrm{p}_{n,k}^{(v)}) \cdot \kappa_c(c_{m,k}^{(u)}, c_{n,k}^{(v)})]. \tag{2.8}$$

where $w(\cdot)$, $\kappa_c(\cdot)$, and $\kappa_p(\cdot)$ are a weight function, the event kernel, and the position kernel. The event kernel is a linear kernel which is defined by

$$\kappa_c(c_{m,k}^{(u)}, c_{n,k}^{(v)}) = c_{m,k}^{(u)} \cdot c_{n,k}^{(v)}. \tag{2.9}$$

The position kernel is a squared exponential (SE) kernel given by

$$\kappa_p(\mathrm{p}_{m,k}^{(u)}, \mathrm{p}_{n,k}^{(v)}) = \exp\left(-\frac{(\mathrm{p}_{m,k}^{(u)} - \mathrm{p}_{n,k}^{(v)})^2}{l_k^2}\right) \tag{2.10}$$

where l_k denotes a length-scale hyper-parameter. The SE kernel is a standard approach to calculate a distance between two input that defined by a real number.

[2] This study defines all temporal event context in the same manner, then all context are used the same kernel.

2.4 Experiments

This section conducts experiments to compare the performance of HMM-, DNN-, and GPR-based methods in speech synthesis.

2.4.1 Experimental conditions

This experiments used a set of phonetically balanced sentences from the Thai speech database T-Sync-1 developed by NECTEC [32]. One professional female speaker uttered the sentences in the reading style of standard Thai accent. The training set varied from 250 to 950 utterances. The test set had 50 utterances. Speech signals were sampled at a rate of 16 kHz. Spectral features, aperiodicity, and F0 were extracted by STRAIGHT [33] with a 5-ms frame shift. The acoustic feature vector consisted of the 0th to 39th mel-cepstral coefficients, five-band aperiodicity, log F0, and their delta and delta-delta coefficients.

The HMM-based Thai SPSS [18] used context-dependent tri-phone hidden semi-Markov models (HSMMs) having five-state, left-to-right, no-skip model topology [34]. Decision-tree-based context clustering performed with the minimum description length (MDL) criterion [35]. The DNN-based SPSS was the framework that was proposed by [7]. The DNN-based method was performed by having 3 and 6 hidden-layers. Each hidden-layer had 1024 nodes and the *tanh* function was used as the activation function. The input and output features used in network training were normalized to zero-mean and unit variance. GPR-based model training was conducted with using the partially independent conditional approximation [15] and the kernel function parameters are optimized using the expectation-maximization (EM)-based method [14].

2.4.2 Objective evaluation

Figure 2.3 shows the mel-ceptral distance. The GPR-based method had the lowest distortion, and the DNN-based ones had lower distortions than the HMM-based one. The DNN-based with 6 hidden-layers had lower distortion than that of 3 hidden-layers.

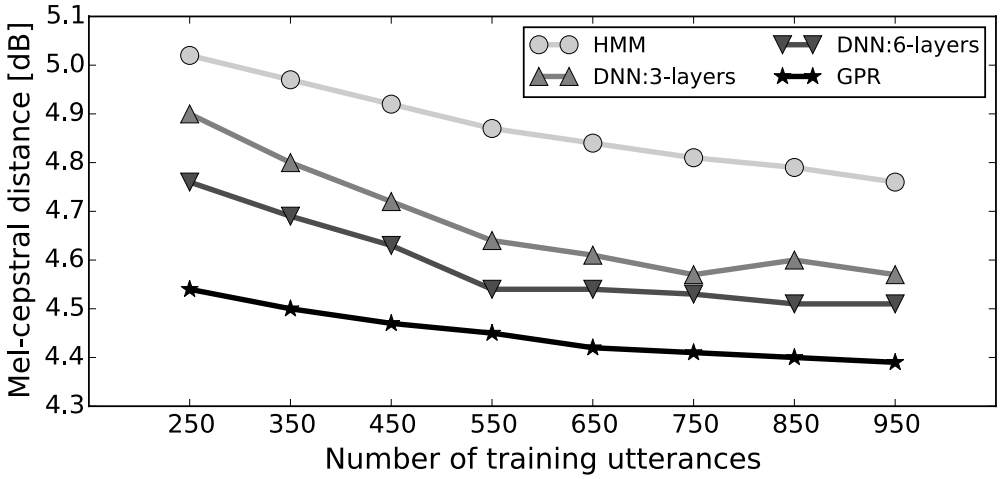


Figure 2.3: Mel-cepstrum distortions.

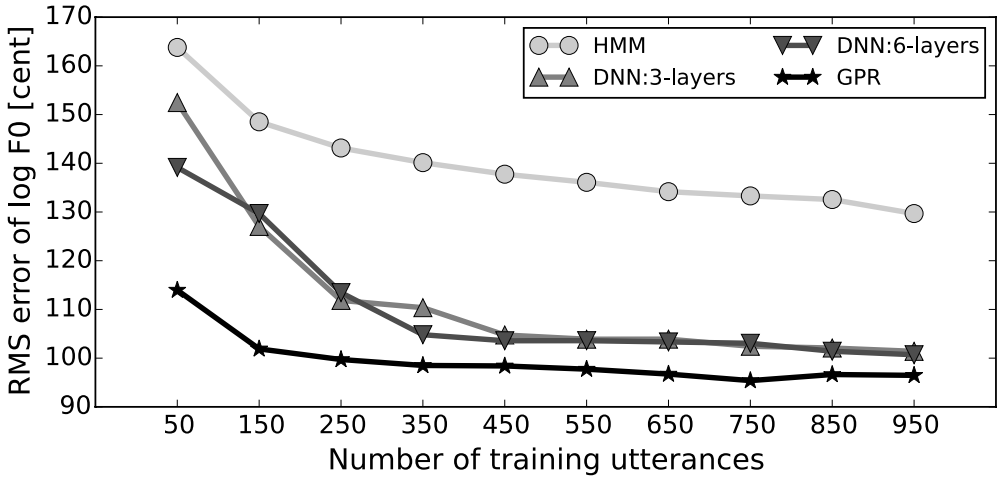


Figure 2.4: Log F0 distortions.

Figure 2.4 shows the RMS errors of log F0. The GPR-based method also showed the lowest distortion in log F0. The DNN-based method was lower distortion than the HMM-based method. Both DNN-based methods had a similar result. Figure 2.5 shows a comparison of generated F0 contours in which the GPR-based method was closer to the original one than the HMM- and DNN-based ones.

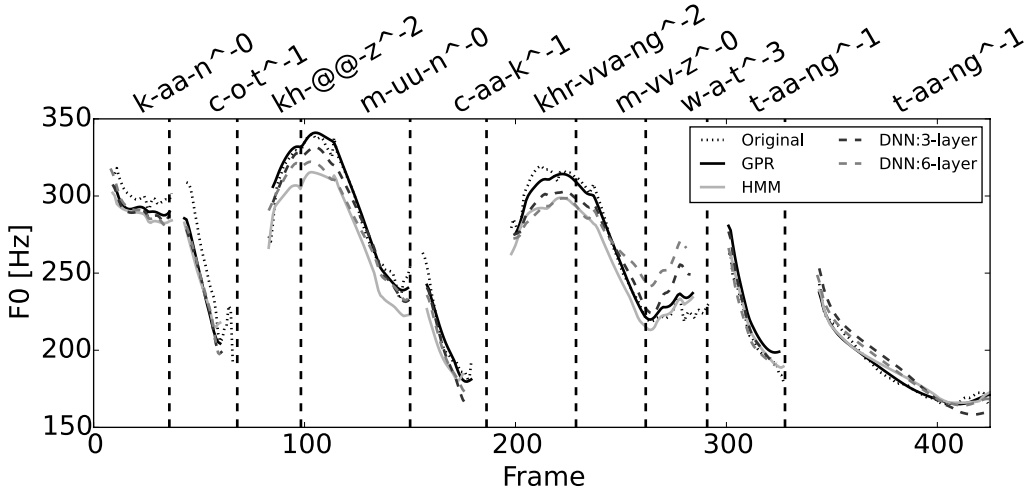


Figure 2.5: Example of generated F0 contours using 950 training utterances.

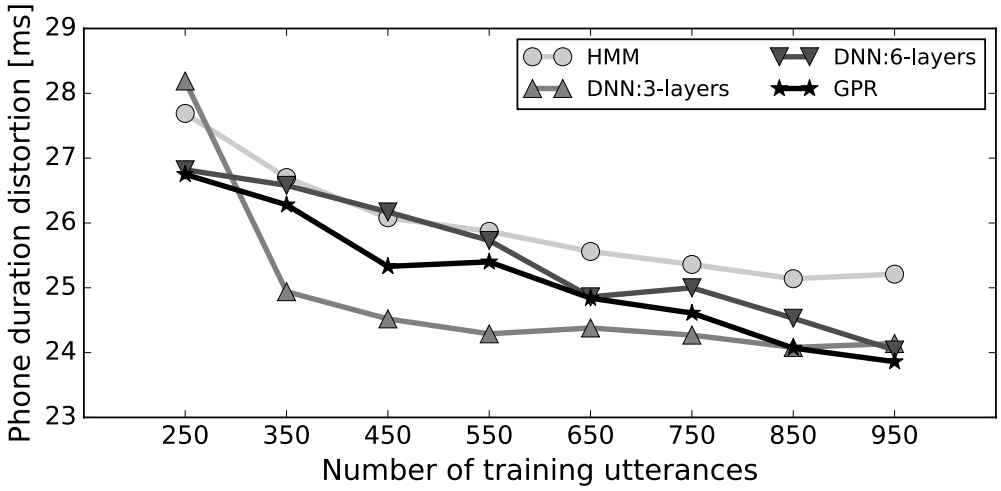


Figure 2.6: Phone-duration distortions.

Figure 2.6 shows the phone duration distortions. The DNN-based method with 3 hidden-layers had the lowest distortion in most case, except at 950 utterances which the GPR-based method had the lowest distortion. The GPR-based method had lower distortion than the DNN-based method with 6 hidden-layers. The GPR- and DNN-based methods had lower distortions than the HMM-based one.

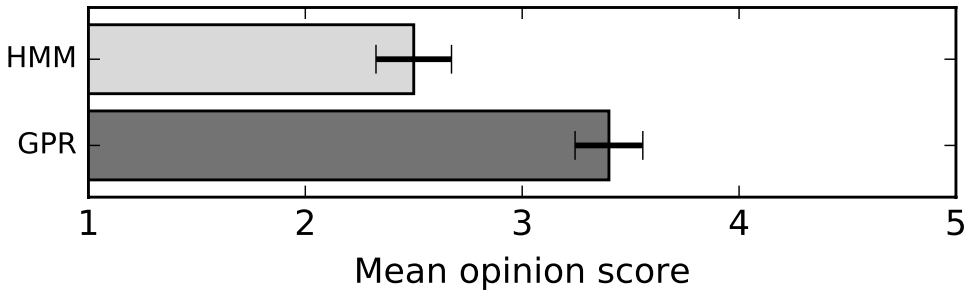


Figure 2.7: Comparison of mean opinion scores (MOSs) of naturalness between HMM-based and GPR-based SPSS.

2.4.3 Subjective evaluation

The subjective evaluation involved with mean opinion score (MOS). The participants evaluated each speech sample on a five-point scale from 1 to 5 according to their satisfaction regarding naturalness. The definition of the rating was 1: bad, 2: poor, 3: fair, 4: good, and 5: excellent. The participants could repeat playback as many times as they required. The speech parameters were generated from 950 utterances of training data. The speech samples were randomly selected from the test set. The evaluation conducted two listening tests with two different groups of ten Thai-native speakers.

The first listening test compared HMM- and GPR-based SPSS. The result of MOS tests is shown in Figure 2.7. The GPR-based SPSS received significantly higher scores than the HMM-based one with p-value is 0.0. The second listening test compared DNN- and GPR-based SPSS. In the DNN-based method, spectral features, aperiodicity, and F0 were generated by the 6-hidden-layer network, and phone durations were predicted by the 3-hidden-layer one. The result of MOS tests is shown in Figure 2.8. The GPR-based method outperformed DNN-based one with p-value is 0.002.

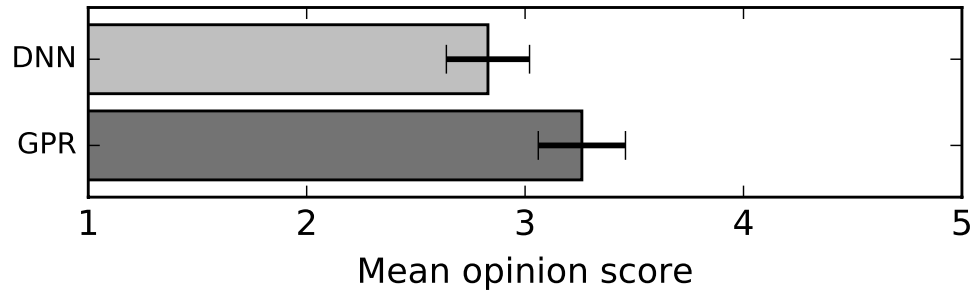


Figure 2.8: Comparison of mean opinion scores (MOSs) of naturalness between DNN-based and GPR-based SPSS.

2.5 Conclusion

This chapter describes the GPR-based Thai SPSS and shows a comparison with the HMM- and DNN-based SPSS. The contextual factors and kernel function were proposed to introduce the GPR-based SPSS into Thai. The result showed that the GPR-based SPSS outperformed the HMM- and DNN-based ones. In the next chapters, this thesis describes several methods to improve the GPR-based SPSS for prosody generation.

Chapter 3

Two-Stage GPR-Based Duration Prediction

This chapter describes a technique to use multi-level models for GPR-based duration prediction, called two-stage. This technique trains a syllable-duration model to predict syllable durations. Then the predicted syllable durations are used as an additional context for phone-duration model. The two-stage method can apply to not only the GPR-based framework but also the DNN-based one. The objective and subjective evaluations showed that the two-stage method could improve the accuracy of phone-duration prediction and naturalness of synthetic speech.

3.1 Introduction

Duration is one of the prosodic features which contributes to various linguistic functions such as stress, accent, and intonation. For Thai, the duration is the most dominant factor of stress in which significantly influences intonation [28]. For statistical parametric speech synthesis, various duration-modeling techniques have been proposed. In an HMM-based method, state durations of a phoneme HMM are modeled by a multi-dimensional Gaussian distribution, and a decision tree was used to cluster the variation of contextual factors [36]. A constrained tree regression method [37] incorporates linear and tree regressions for duration prediction. A gradient tree boosting [38],

a meta-algorithm of regression trees, iteratively constructs a regression tree in phone-duration modeling. A multi-level model method for HMM-based SPSS [39] are performed by maximizing the joint probability of multiple duration models. In a DNN-based method, a neural network was examined for segmental duration modeling [40]. An architecture of unidirectional long short-term memory recurrent neural network was performed for duration prediction [8]. A deep neural network was examined for duration prediction in short sentences [41]. A robust DNN-based duration prediction was proposed to alleviate problems of dubious and unhelpful data points [42]. Variety of machine learning techniques were examined for duration prediction such as a multiplicative model [43], support vector regression (SVR) [44], sums-of-product (SoP) models [45], and Bayesian networks [46]. Furthermore, many techniques combines multiple techniques for duration prediction [47–53].

Duration is a major prosodic feature that contributes to intonation. For Thai, linguistic functions of syllable level play a crucial role in duration, for example, stress, accent, and co-articulation. The conventional duration modeling uses a single phone-duration model. However, it is ineffective to capture linguistic functions in longer units than phone level. This chapter describes a multi-level model method for GPR-based duration prediction. Specifically, this chapter examines syllable- and phone-duration models for Thai duration prediction. In this technique, called a two-stage method, the syllable-duration model is trained and uses to predict syllable duration. Then, the predicted syllable duration is used as an additional context in phone-duration modeling. This method can apply to both GPR- and DNN-based methods since these frameworks are capable of utilizing a real number (syllable duration) as a context. In the experiment, this chapter conducted a duration prediction by a multi-level model of the HMM-based one to compare with the two-stage ones. The multi-level HMM-based duration prediction was performed with a method proposed by [39] in which jointly maximizes multiple duration models for prediction. The objective and subjective evaluations were conducted to measure the performance of the proposed method.

3.2 Duration prediction using multi-level model

This section describes the two-stage duration prediction in which using multi-level models. The two-stage method was applied to GPR- and DNN-based framework. This sections also describes a multi-level model method for HMM-based framework proposed by [39].

3.2.1 Multi-level GPR-based duration prediction

This technique utilizes syllable- and phone-duration models for duration prediction. The syllable-duration model uses syllable-level context and syllable duration as input and output variables of Gaussian process, respectively. A kernel function is defined for corresponding syllable-level context.

The syllable-level context consists of syllable, word, and utterance layers. The syllable layer contains phonetic features of initial consonant, vowel, and final-consonant of a syllable, and a tone-type. The phonetic features are defined in the same manner as the phone-level context as shown in Table 2.4. The word and utterance layers are the same as phone-level context.

The phone-duration model uses the similar context as the baseline GPR-based duration modeling. The difference is that it included syllable duration as an additional context. The phone- and syllable-level contexts are summarized in Table 3.1 and 3.2, respectively. Figure 3.1 is an overview of a speech synthesis system with the multi-level model for duration prediction.

- Training part
  - Phone-level model

    Ph.1 Extract acoustic features including mel-cepstral coefficients, aperiodicity, and F0.

    Ph.2 Convert labels into phone and frame-level context including the additional context, *syllable-duration context* (see Table 3.1).

    Ph.3 Train phone-level duration model.

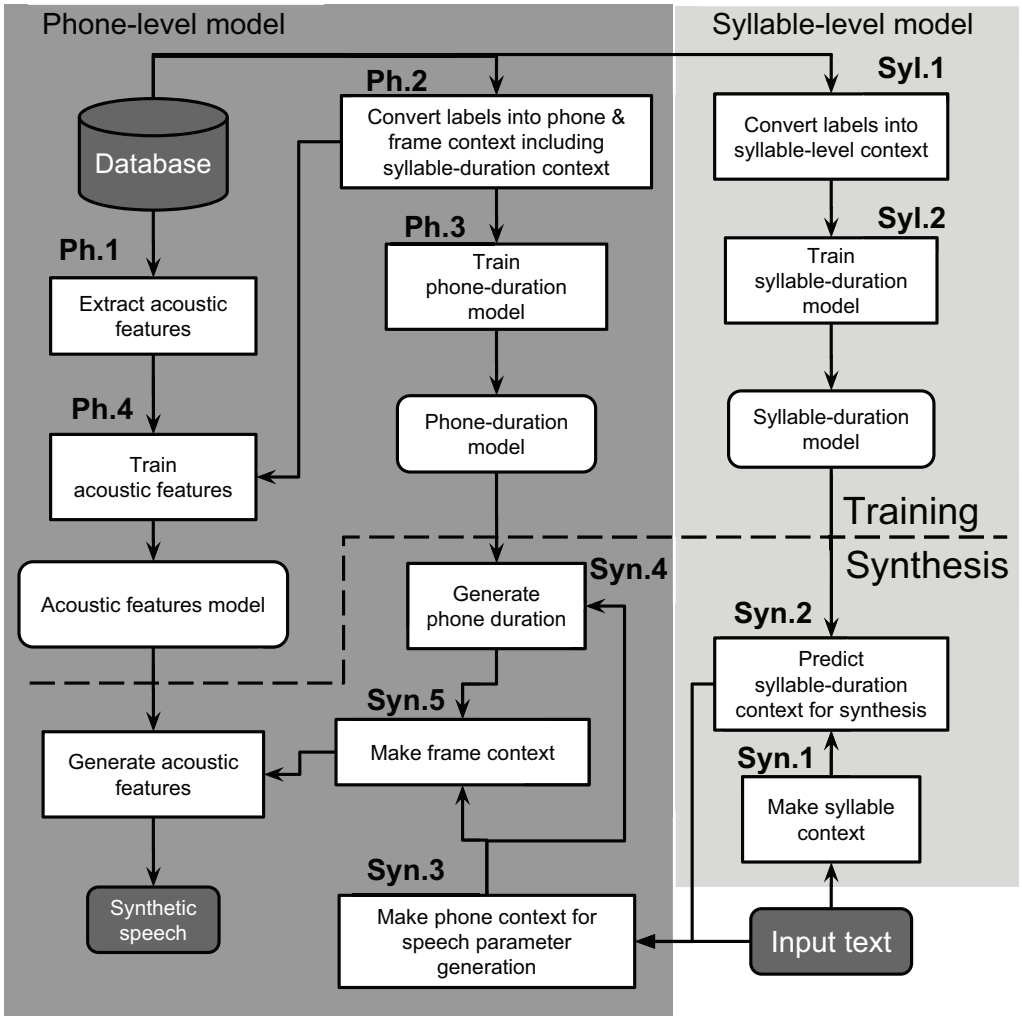


Figure 3.1: Block diagram of GPR-based speech synthesis system with multi-level model for duration prediction.

Ph.4 Train mel-cepstral coefficients, aperiodicity, and F0 models based on GPR-based framework described in Chapter 2.

– Syllable-level model

Syl.1 Convert labels into syllable-level context (see Table 3.2) for model training.

Syl.2 Train syllable-level duration model.

Table 3.1: Temporal context for Thai GPR-based speech synthesis.

Unit:	phone
Type:	{beginning, end} of each phonetic feature
Scale:	phone-normalized scale, time
Unit:	syllable
Type:	{beginning, end} of tone-type
	{beginning, end} of syllable duration
Scale:	{syllable, word}-normalized scale
Unit:	word
Type:	{beginning, end} of part of speech
Scale:	{syllable, word}-normalized scale
Unit:	utterance
Type:	{beginning, end} of utterance
Scale:	{syllable, word, utterance}-normalized scale

- Synthesis part

Syn.1 Make syllable-level context of input text.

Syn.2 Predict syllable durations corresponding to input text by using syllable-level duration model.

Syn.3 Make phone-level context of input text including *syllable-duration context* predicted with syllable-duration model.

Syn.4 Generate phone durations by using phone-level duration model.

Syn.5 Make frame context, generate speech parameters, and synthesize speech.

Step **Ph.3** calculates the similarity between two syllable-duration contexts in the phone-level model by using the SE kernel as follows:

$$\kappa_c(c_{m,k}^{(u)}, c_{n,k}^{(v)}) = \exp\left(-\frac{(c_{m,k}^{(u)} - c_{n,k}^{(v)})^2}{l_{ck}^2}\right), \tag{3.1}$$

where $c_{m,k}^{(u)}$ and $c_{n,k}^{(v)}$ are syllable-duration contexts, and l_{ck} denotes a length-scale hyper-parameter.

Table 3.2: Syllable-level temporal context for Thai GPR-based speech synthesis.

Unit:	syllable
Type:	beginning of each initial-consonant's phonetic feature
	beginning of each vowel's phonetic feature
	beginning of each final-consonant's phonetic feature
	beginning of tone-type
Scale:	{syllable, word}-normalized scale
Unit:	word
Type:	{beginning, end} of part of speech
Scale:	{syllable, word}-normalized scale
Unit:	utterance
Type:	{beginning, end} of utterance
Scale:	{syllable, word, utterance}-normalized scale

3.2.2 Multi-level DNN-based duration prediction

This section briefly explains the two-stage duration prediction for a DNN-based framework. In the two-stage method of the DNN-based framework, the syllable-duration model is trained and predicts syllable duration for the test data. Then, the phone-duration model is trained by incorporating the predicted syllable duration as an additional context. The input features of the DNN-based duration model are defined using the same approach described by [7] and are binary and numerical features including linguistic and positional information. The input features are derived from the GPR-based method as shown in Table 3.1 and 3.2. The architecture of the DNN-based multi-level model for duration prediction is shown in Figure 3.2.

3.2.3 Multi-level HMM-based duration prediction using joint maximizing probability

This section briefly explains an approach of multi-level HMM-based technique to duration prediction [39, 54]. State, phone, and syllable duration models are separately trained. The question sets used in the tree-based clustering

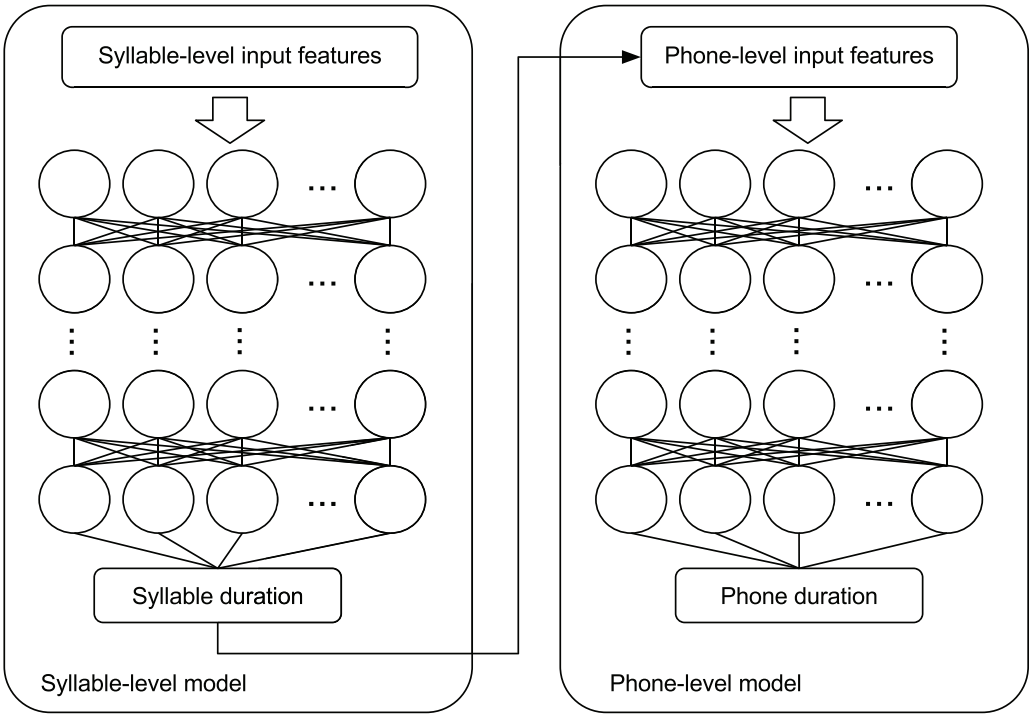


Figure 3.2: Overview of DNN-based multi-level model for duration prediction.

were based on the conventional HMM-based Thai speech synthesis [17], which include those related to phoneme, tone-type, part-of-speech, and position of units. In the syllable-level model, the question set included three phonetic components of a syllable, which was the same as the multi-level GPR-based approach. The question sets of phone- and syllable-levels are summarized in Table 3.3. Duration of each leaf node is modeled by a Gaussian distribution.

For duration prediction, the likelihood of state durations is jointly maximized with the weighted likelihoods of phone and syllable durations. For a given duration sequence $D = [d_1, d_2, \ldots, d_J]$ of J syllables, the log likelihood $L(D)$ of duration is defined as

$$L(D) = \sum_j \Big[\sum_n \Big[\sum_k \log p_{j,n,k}(d_{j,n,k}) + \alpha \log p_{j,n}(d_{j,n}) \Big] + \beta \log p_j(d_j) \Big] \quad (3.2)$$

where $d_{j,n,k}$ is the duration of state k in phone n and syllable j, and $p_{j,n,k}(\cdot)$ is the corresponding probability density function. The pdfs of phone and

Figure 3.3: Question set of phone- and syllable-level models for multi-level HMM-based method for duration prediction.

Level	Question set	
	Phone-duration model	Syllable-duration model
Phoneme	• Phonetic features of phone • Position of phone in syllable	—
Syllable	—	• Phonetic features of consonant • Phonetic features of vowel • Phonetic features of final-consonant
	• Tone-type of syllable • Position of syllable in word • Number of phones in syllable	
Word	• POS of word • Number of syllables in word	
Utterance	• Number of syllables in utterance • Number of words in utterance	

syllable are likewise defined by $p_{j,n}(\cdot)$ and $p_j(\cdot)$, respectively. The durations of syllable and phone are constrained as follows:

$$\sum_k d_{j,n,k} = d_{j,n} \tag{3.3}$$

$$\sum_n d_{j,n} = d_j. \tag{3.4}$$

By maximizing $L(D)$, the solution of state duration $d_{j,n,k}$ is given by

$$d_{j,n,k} = \mu_{j,n,k} + \left[-\alpha \frac{d_{j,n} - \mu_{j,n}}{\sigma^2_{j,n}} - \beta \frac{d_j - \mu_j}{\sigma^2_j} \right] \sigma^2_{j,n,k} \tag{3.5}$$

where $d_{j,n}$ and d_j are obtained by applying the constraints of Eqs. (3.3) and (3.4).

3.3 Experiments

Experiments were conducted to evaluate the performance of the proposed method, two-stage duration prediction. The proposed multi-level GPR- and DNN-based methods were compared with the multi-level HMM-based method in duration prediction. Moreover, a single-level method with extended context was performed by merging phone- and syllable-level contexts. The purpose of utilizing extended context is to confirm that the multi-level one has more impact than a single-level one with extended context.

In summary, this section shows comparisons of *single-level*, *extended context*, and *multi-level* methods for HMM-, DNN-, GPR-based frameworks. The linguistic information used as contextual factors of all methods is summarized in Table 3.3.

3.3.1 Experimental conditions

A set of phonetically balanced sentences from T-Sync-1, the same database as the previous chapter, was used for training and evaluation. The training had 250 to 950 utterances, and the test set had 50 utterances. The phone durations of training data were obtained by forced alignment and rechecked by linguists.

The HMM-based method used context-dependent triphone hidden semi-Markov models (HSMMs) having five-state, left-to-right, no-skip model topology [34]. Decision-tree-based context clustering was performed with the minimum description length (MDL) criterion [35]. In the multi-level HMM-based method, one hundred utterances were used as a development set to find the optimal α and β in Eq. (3.5). The development set was not included in training and test set. A grid search was employed to find the optimal values of α and β that generated the lowest error in phone duration. The optimization of each set of training data was conducted separately. Figure 3.4 shows the optimal values of α and β for each training utterances set. Figure 3.5 shows the results of the grid search for 950 training utterances where the optimal values were 1.4 and 1.0 for α and β, respectively.

The DNN-based duration prediction was the framework that proposed by [7]. The network architectures were 3 and 6 hidden-layers with 1024 nodes

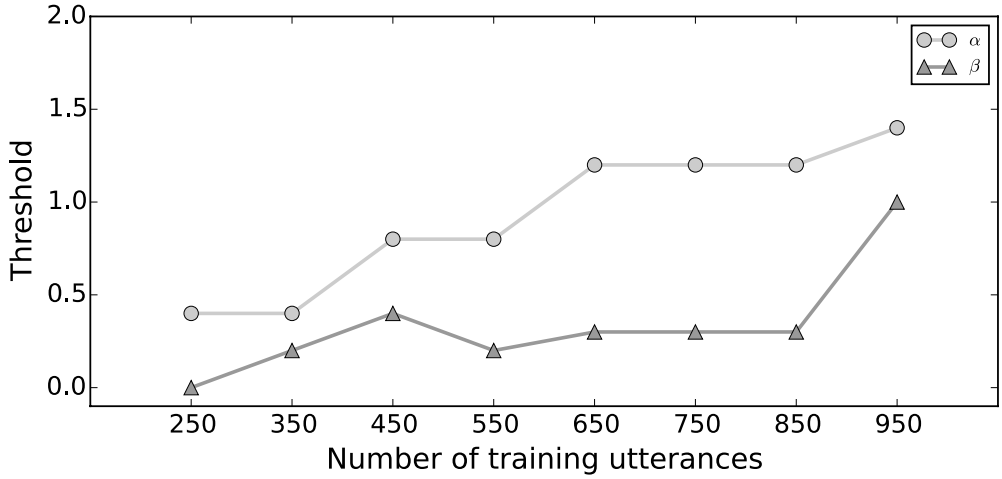


Figure 3.4: Optimal α and β values for each number of training utterances.

Table 3.3: Comparison of linguistic information as contextual factors used in the experiments.

Method	Model	Layers of Linguistic information			
		Phoneme	Syllable		Word
		Phonetic features	Tone	Phonetic features of phonemes within a syllable	Part of speech
Single-level model		✓	✓		✓
Single-level model with extended context		✓	✓	✓	✓
Multi-level model	Phone duration model	✓	✓		✓
	Syllable duration model		✓	✓	✓

of each layer. The activation function was the *tanh* function. The input and output features used in network training were normalized to zero-mean and unit variance.

GPR-based modeling was conducted by using the PIC approximation [15] and optimized the kernel function parameters using the EM-based method [14].

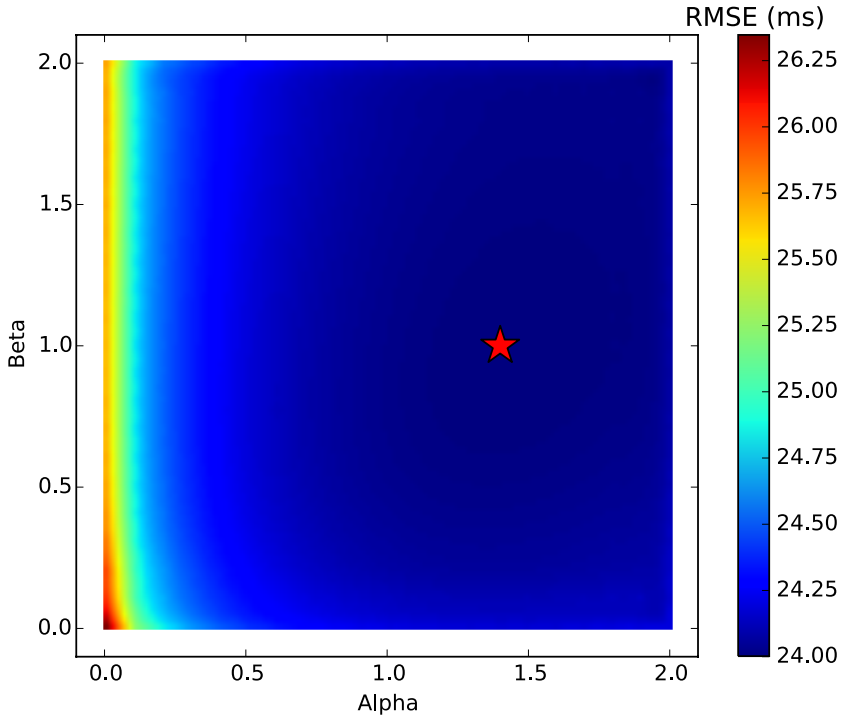
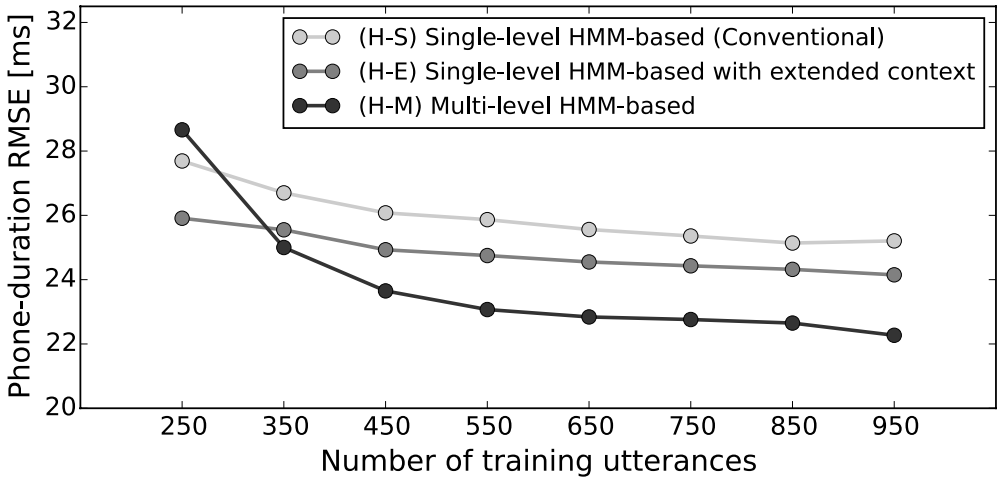


Figure 3.5: Full grid search result of optimal α and β values. Lowest distortion is marked with star.

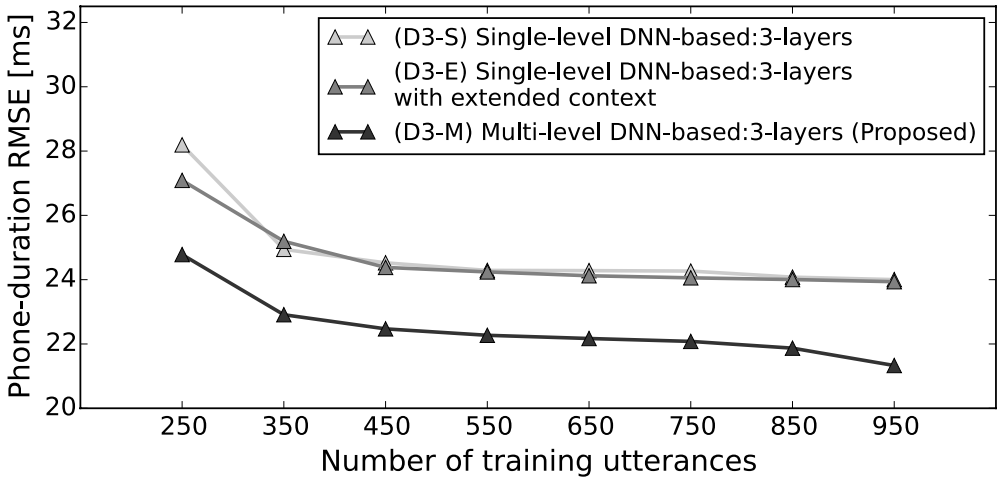
3.3.2 Objective evaluation

Figure 3.6 shows comparisons of single-level, extended context, and multi-level methods for HMM-, DNN-, and GPR-based SPSS in phone-duration distortion. The extended context method had lower phone-duration distortion than the single-level model. However, the multi-level method could achieve lower distortion than the extended context in all methods. This result could confirm that the multi-level method had more impact on reducing distortion than the extended context one.

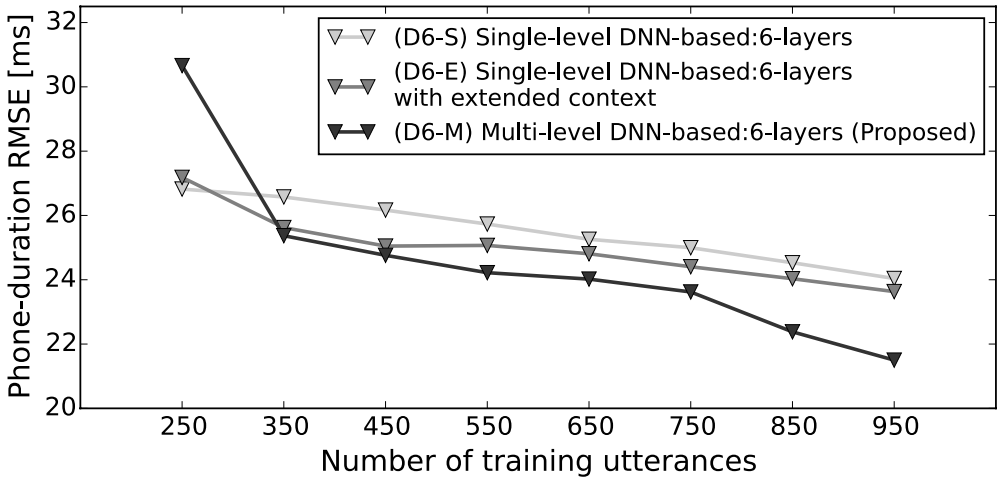
Figure 3.7 shows a comparison of multi-level methods. The results showed that the multi-level DNN-based method with 3 hidden-layers had the lowest distortion in all case. The GPR-based method had lower distortion than the HMM-based method, especially when training data was over 650 utterances. Moreover, the difference between the GPR- and DNN-based method with



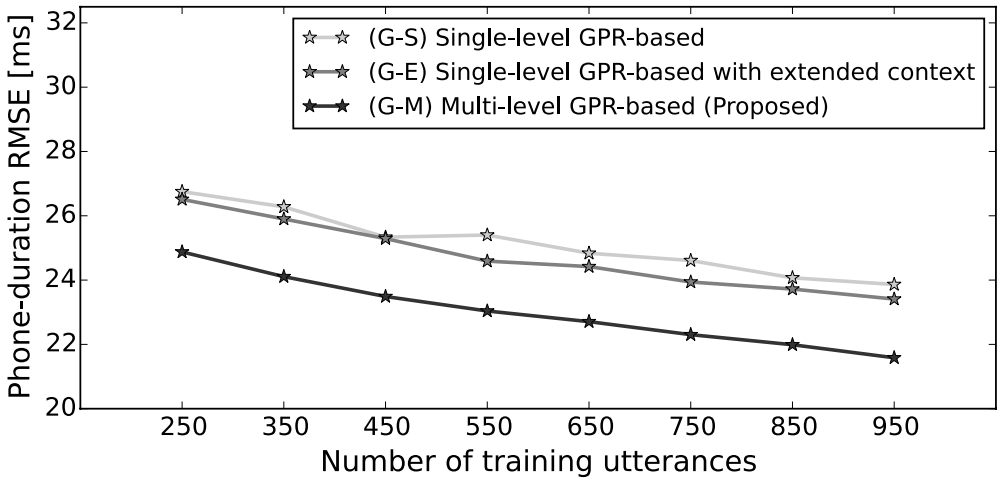
(a) HMM-based model



(b) DNN-based model with 3 hidden layers



(c) DNN-based model with 6 hidden layers



(d) GPR-based model

Figure 3.6: Comparison of phone-duration distortions among single-level model, extended context, and multi-level model for HMM-, DNN-, and GPR-based SPSS frameworks.

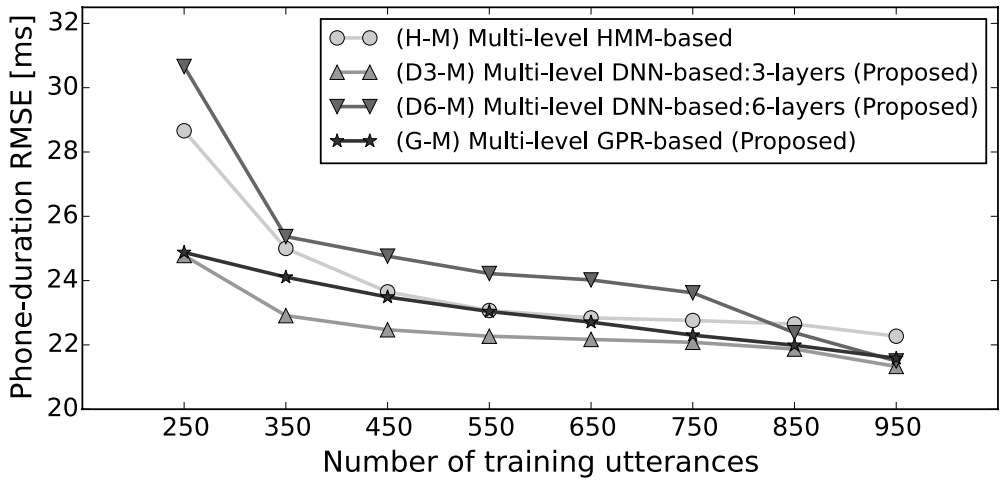


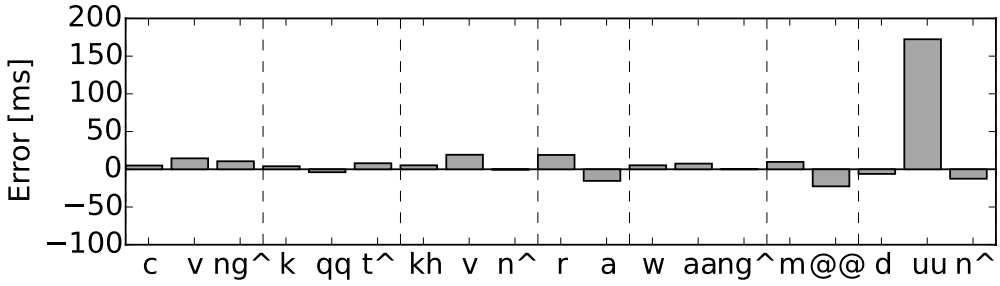
Figure 3.7: Comparison of multi-level model methods in phone-duration distortions.

3 hidden-layers became smaller at training data over 650 utterances. The DNN-based method with 6 hidden-layers had the largest distortion in all cases, except at 950 utterances which the DNN-based method with 6 hidden-layers had a comparable result with the GPR-based one.

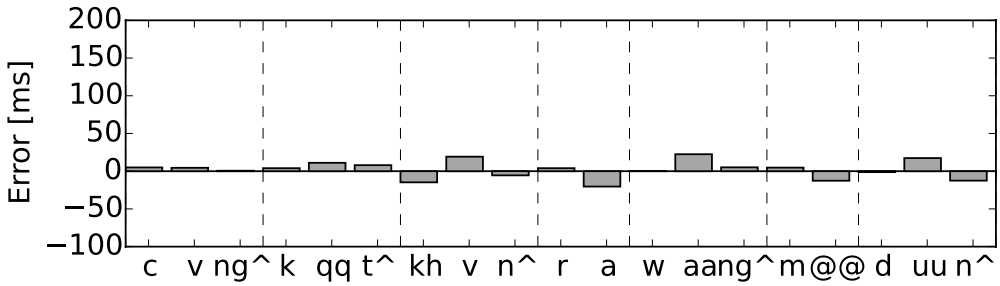
Figure 3.8, 3.9, and 3.10 illustrates an example of the duration errors obtained by HMM-, DNN-, and GPR-based methods, respectively. Each bar represents the difference between the generated phone duration and the original one. The duration sequences were predicted using 950 utterances of training data. Sentence is "... then it occurs between modules ..." in English. Vertical dashed lines are syllable boundaries. Unit of error is shown in milliseconds. In the comparison of HMM-based methods, the extend context method had smaller errors than the single- and multi-level ones, especially at the last syllable.

Moreover, the multi-level DNN- and GPR-based methods had smaller distortions than the HMM-based one, especially at the end of the sentence.

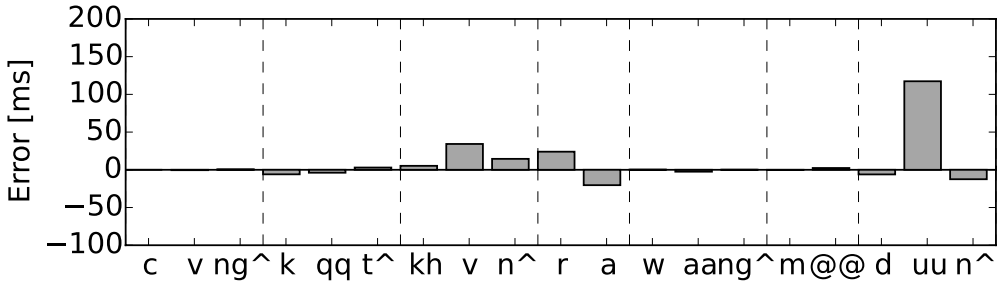
For the DNN-based method, the result of the 3-hidden-layer model is selected to show since it had less distortion than the 6-hidden-layer one. The duration distortions decreased when applied the extended context and multi-level methods.



(a) Single-level HMM-based model (H-S)

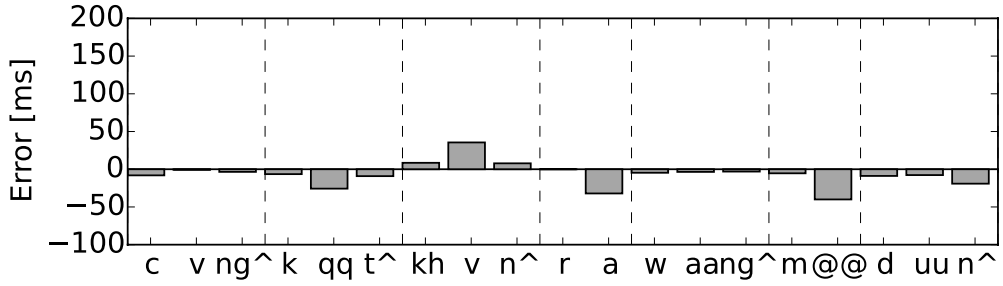


(b) Single-level HMM-based model with extended context (H-E)

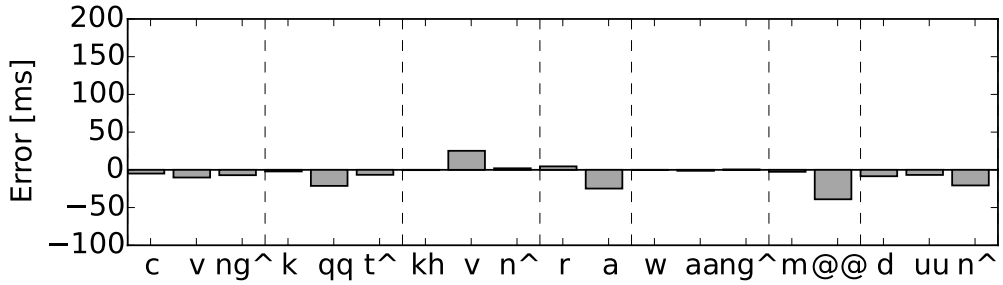


(c) Multi-level HMM-based model (H-M)

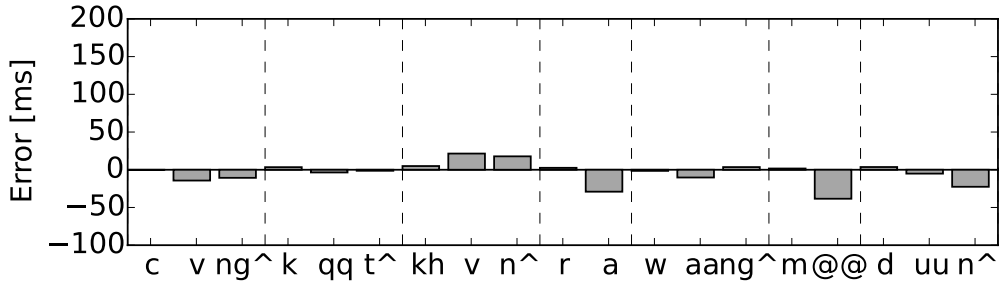
Figure 3.8: Errors of HMM-based duration prediction in terms of phone unit.



(a) Single-level DNN-based model (D3-S)



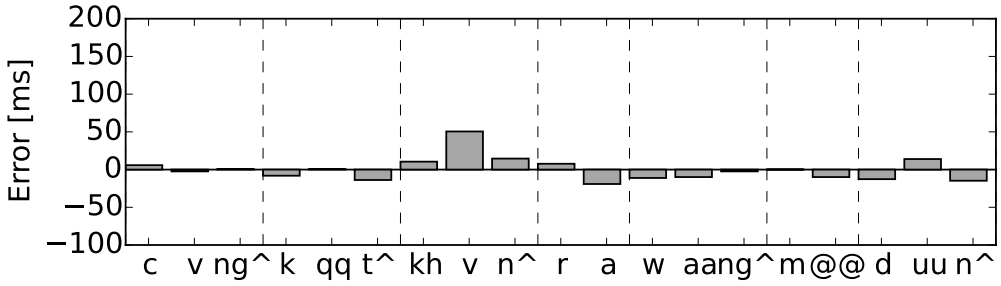
(b) Single-level DNN-based model with extended context (D3-E)



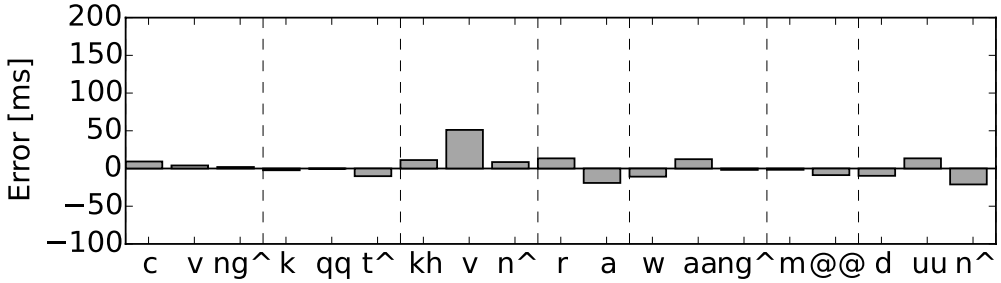
(c) Multi-level DNN-based model (D3-M)

Figure 3.9: Errors of DNN-based duration prediction with 3 hidden layers in terms of phone unit.

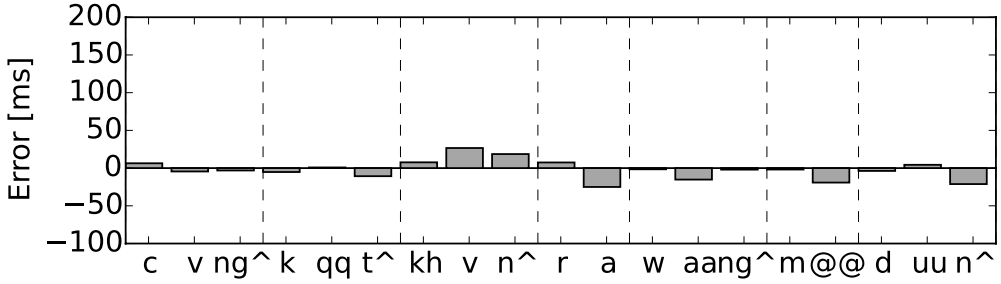
The GPR-based methods also showed a similar result as the HMM- and DNN-based ones that the multi-level method decreased the distortion from the single-level and extended context one.



(a) Single-level GPR-based model (G-S)



(b) Single-level GPR-based model with extended context (G-E)



(c) Multi-level GPR-based model (G-M)

Figure 3.10: Errors of GPR-based duration prediction in terms of phone unit.

3.3.3 Subjective evaluation

The subjective evaluation involved mean opinion score (MOS) and forced-choice preference tests in assessing the perceptual quality of predicted duration. The subjective evaluation consisted of four listening tests in which evaluated by different groups of ten Thai native speakers. The participants could repeat playback as many times as they required.

In the MOS test, the participants evaluated each sample on a five-point scale from 1 to 5 according to their satisfaction regarding naturalness. The definition of the rating was 1: bad, 2: poor, 3: fair, 4: good, and 5: excellent. Ten speech samples were randomly selected from the objective evaluation.

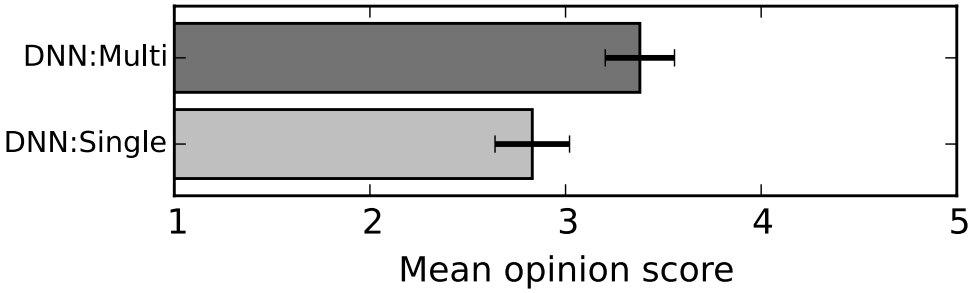


Figure 3.11: MOSs of naturalness between DNN-based duration prediction with single- and multi-level models

The first listening test evaluated the performance of single- and multi-level models for the DNN-based duration prediction. Figure 3.11 shows the MOS of DNN-based methods compared between single- and multi-level models. In this comparison, the spectral features, aperiodicity, and F0 were generated by the DNN-based method with 6 hidden-layers as described in the previous chapter. The difference between *DNN:Single* and *DNN:Multi* is the use of single- and multi-level models for duration prediction, respectively. The DNN-based method with 3 hidden-layers was used in the subjective evaluation since it showed better performance than that of 6 hidden-layers in the objective evaluation. The result showed that the multi-level DNN-based method received a statistically higher score than the single-level one (p-value=0.00002).

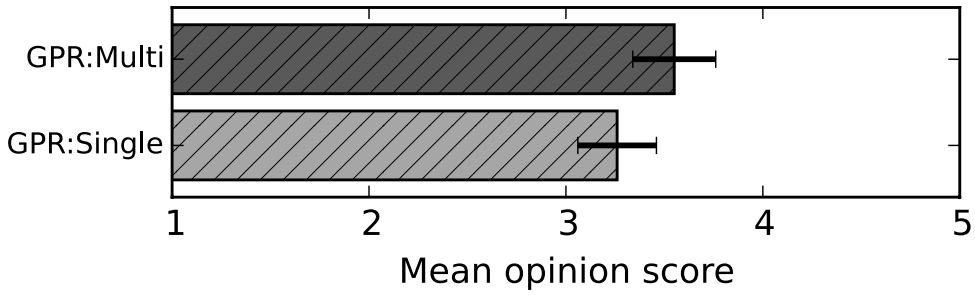


Figure 3.12: MOSs of naturalness between GPR-based duration prediction with single- and multi-level models

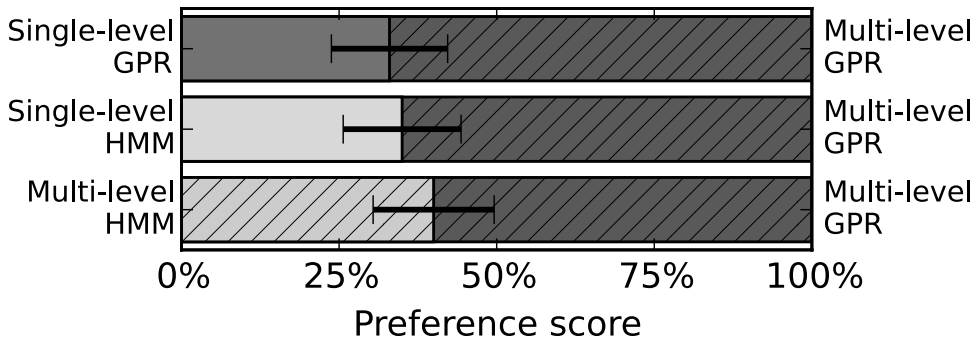


Figure 3.13: Preference scores of duration prediction with the single- and multi-level models of HMM- and GPR-based methods.

The second listening test evaluated the performance of single- and multi-level models for the GPR-based duration prediction. Figure 3.12 shows the MOS of GPR-based methods compared between single- and multi-level models. The spectral features, aperiodicity, and F0 were generated by the GPR-based framework. *GPR:Single* and *GPR:Multi* denotes the use of single- and multi-level models for duration prediction, respectively. The MOS showed that the multi-level model had a higher score than the single-level one.

In the forced-choice preference test, the participants were asked to choose the most natural one for each pair of speech samples. Ten speech samples were randomly selected from the test set for the listening tests.

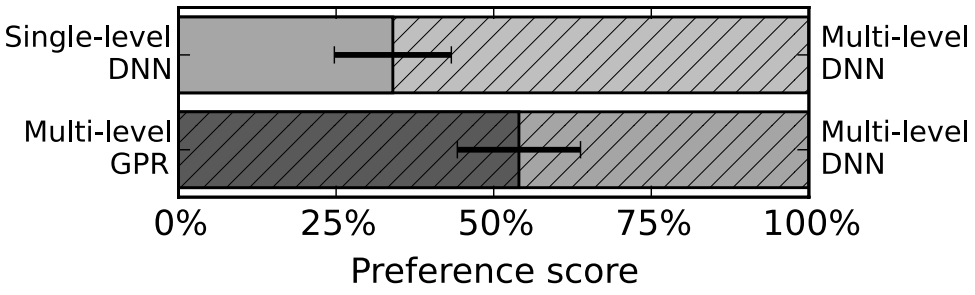


Figure 3.14: Comparison of forced-choice preference scores of naturalness between DNN- and GPR-based methods with single and multi-level duration predictions.

The third listening test compared the multi-level models of HMM- and GPR-based duration prediction. Figure 3.13 showed the result of forced-choice preference test. The result showed comparisons between the multi-level GPR-based method and single- and multi-level HMM-based ones. Spectral features, aperiodicity, and F0 of all methods were trained and generated in the GPR-based framework because it showed the lowest distortions in the objective evaluation of the previous chapter. The difference was predicted phone duration in which *Single-level HMM*, *Multi-level HMM*, *Single-level GPR*, and *Multi-level GPR* denote the duration-prediction methods. The result showed that the multi-level GPR-based method had statistically higher scores than the single-level GPR-, single-level HMM-, and multi-level HMM-based ones with p-values of 0.0003, 0.002, and 0.04, respectively.

The last listening test compared the multi-level models of GPR- and DNN-based duration prediction. Figure 3.14 showed the result of forced-choice preference test of the multi-level DNN-based compared with the single-level DNN- and multi-level GPR-based method for duration prediction. Spectral features, aperiodicity, and F0 of all methods were trained and generated in the GPR-based framework. Thus, the difference was only in the phone-duration prediction. *Single-level DNN*, *Multi-level DNN*, and *Multi-level GPR* denote the duration-prediction methods. The multi-level DNN-based method outperformed the single-level DNN-based one (p-value=0.0007) and

was comparable with the multi-level GPR-based one (p-value=0.4).

3.4 Conclusion

This chapter describes a multi-level model method, called two-stage duration prediction, for GPR-based SPSS using phone- and syllable-duration models. In this technique, a syllable-duration model is used to predict syllable duration; then the predicted syllable duration is incorporated as an additional context for the phone-duration model. A syllable-level context set was designed to train the syllable-duration model. The syllable-level context consists of phonetic features of phonemes in a syllable, linguistic information of a syllable and longer unit, and relative positioning information. This technique is not only limited to the GPR-based SPSS but also can apply to the DNN-based one. The experimental results showed that the multi-level method outperformed the single-level one in both GPR- and DNN-based SPSS. Moreover, the multi-level GPR-based method showed a better result than the multi-level HMM-based one.

Chapter 4

Duration Prediction Using Multiple Gaussian Process Experts for GPR-Based Speech Synthesis

This chapter describes an alternative method to incorporate multi-level models for GPR-based duration prediction. This technique uses two-level models, phone- and syllable-level models. The method combines multiple models by product of experts. First, the phone- and syllable-duration models are individually trained. The predictive distribution of syllable-duration model is reformulated in term of phone duration. Then, the predictive distributions of syllable- and phone-duration models can combine by product of Gaussians. The means of combined predictive distributions are used as predicted durations for synthetic speech. The experimental result showed that the proposed method outperformed the conventional single-level GPR-based duration prediction.

4.1 Introduction

The conventional duration modeling uses a phone-duration model for SPSS [8, 36, 41]. Duration is an important factor in the perception of naturalness.

For Thai, the duration is the most dominant factor of various linguistic functions such as stress and co-articulation. However, a single phone-duration model is ineffective to predict duration sequence, since many factors that affect duration belong to longer units.

Various techniques have been proposed to incorporate multi-level models for improving prosody modeling. A joint maximization was used to combine state-level model and longer units for prosody generation [39]. A speaking rate-dependent hierarchical prosodic model (SP-HPM) [55] was proposed to utilize a hierarchical structure including prosodic-acoustic features, linguistic information, and prosody structure for speaking rate modeling. A product of expert framework was incorporated to train multiple acoustic models for speech synthesis [56]. Our previous chapter had shown the effectiveness in using syllable-duration model for prediction. However, the syllable-duration model was not explicitly used for prediction.

This chapter describes a technique to use phone- and syllable-duration models explicitly for duration prediction. First, phone- and syllable-level models are trained individually. Then, the predictive distributions were combined by product of Gaussians. The predicted duration can be obtained by calculating model parameters of the product. The experiments were conducted to evaluate the performance of the proposed method.

4.2 Duration prediction by multiple GP-experts

The phone- and syllable-duration models are trained individually by using the same contextual factors as the two-stage duration prediction. The difference is that this technique does not includes the syllable duration as a context in phone-duration model. When synthesizing, the predictive distributions of phone- and syllable-duration models are combined in a similar way as described in [57]. The predictive distributions of syllable duration and phone duration are expressed by $p(\mathbf{d}_T^s|\mathbf{d}^s, \mathbf{X}^s, \mathbf{X}_T^s)$ and $p(\mathbf{d}_T^p|\mathbf{d}^p, \mathbf{X}^p, \mathbf{X}_T^p)$, respectively. $\mathbf{X}^s$ and $\mathbf{X}^p$ are input variables of syllable and phone duration models, respectively. Matrix forms of syllable durations $\mathbf{d}^s$ and phone durations $\mathbf{d}^p$

are output variables of syllable and phone duration models, respectively. Finally, the product of predictive distributions is given by

$$p(\mathbf{d}_T^p|\mathbf{d}^s,\mathbf{d}^p,\mathbf{X},\mathbf{X}_T) \quad = \quad \frac{1}{Z}p(\mathbf{d}_T^s|\mathbf{d}^s,\mathbf{X}^s,\mathbf{X}_T^s) \;\cdot\; p(\mathbf{d}_T^p|\mathbf{d}^p,\mathbf{X}^p,\mathbf{X}_T^p) \quad (4.1)$$

$$\mathbf{d}_T^s = [d_1^s, d_2^s, ..., d_n^s]^\intercal \quad (4.2)$$

$$\mathbf{d}_T^p = [d_{1,1}^p, d_{1,2}^p, d_{2,1}^p, ..., d_{n,m(n)}^s]^\intercal \quad (4.3)$$

where $\mathbf{d}_T^s$ and $\mathbf{d}_T^p$ are matrix forms of syllable and phone durations of test data, respectively, and Z is a normalization term. To combine the predictive distributions, the relationship between output variables of distributions is defined as syllable duration d_i^s is determined by the sum of phone durations $d_{i,j}^p$ within the syllable as follows:

$$d_i^s = \sum_{j=1}^{m(i)} d_{i,j}^p \quad (4.4)$$

where $m(i)$ is the number of phones in *i-th* syllable, whose value is 2 or 3 (for Thai). It can be written in a matrix form using a transformation matrix $\mathbf{W}$ as follows:

$$\mathbf{d}_T^s = \mathbf{W}\mathbf{d}_T^p. \quad (4.5)$$

For example, a sentence has 3 syllables and respective syllables have 3, 2, and 3 phones. Then, the matrix form is expressed as

$$\begin{matrix} \mathbf{d}_T^s \\ \\ \begin{bmatrix} d_1^s \\ d_2^s \\ d_3^s \end{bmatrix} \end{matrix} \quad = \quad \begin{matrix} \mathbf{W} \\ \\ \left[\begin{array}{cccccccc} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right] \end{matrix} \quad \begin{matrix} \mathbf{d}_T^p \\ \\ \left[\begin{array}{c} d_{1,1}^p \\ d_{1,2}^p \\ d_{1,3}^p \\ \hline d_{2,1}^p \\ d_{2,2}^p \\ \hline d_{3,1}^p \\ d_{3,2}^p \\ d_{3,3}^p \end{array}\right] \end{matrix} . \quad (4.6)$$

Since the predictive distribution of syllable duration is Gaussian, it can be reformulated in terms of phone duration in the same way as the formulation

of trajectory HMM framework [58] as follows:

$$p(\mathbf{d}_T^s|\mathbf{d}^s, \mathbf{X}^s, \mathbf{X}_T^s) = \mathcal{N}(\mathbf{W}\mathbf{d}_T^p; \boldsymbol{\mu}_T^s, \boldsymbol{\Sigma}_T^s) \tag{4.7}$$

$$p(\mathbf{d}_T^p|\mathbf{d}^s, \mathbf{X}^s, \mathbf{X}_T^s) = \mathcal{N}(\mathbf{d}_T^p; \mathbf{P}\mathbf{r}, \mathbf{P}) \tag{4.8}$$

$$\mathbf{P} = (\mathbf{W}^\intercal {\boldsymbol{\Sigma}_T^s}^{-1}\mathbf{W})^{-1} \tag{4.9}$$

$$\mathbf{r} = \mathbf{W}^\intercal {\boldsymbol{\Sigma}_T^s}^{-1}\boldsymbol{\mu}_T^s. \tag{4.10}$$

Since both predictive distributions are Gaussian, Eq. (4.1) can be rewritten by Gaussian as follows:

$$\begin{aligned} p(\mathbf{d}_T^p|\mathbf{d}^s, \mathbf{d}^p, \mathbf{X}, \mathbf{X}_T) &= \frac{1}{Z'}\mathcal{N}(\mathbf{d}_T^p; \mathbf{P}\mathbf{r}, \mathbf{P}) \cdot \mathcal{N}^p(\mathbf{d}_T^p; \boldsymbol{\mu}_T^p, \boldsymbol{\Sigma}_T^p) \\ &= \mathcal{N}(\mathbf{d}_T^p; \boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D). \end{aligned} \tag{4.11}$$

where the mean and covariance of the predictive distribution are given by

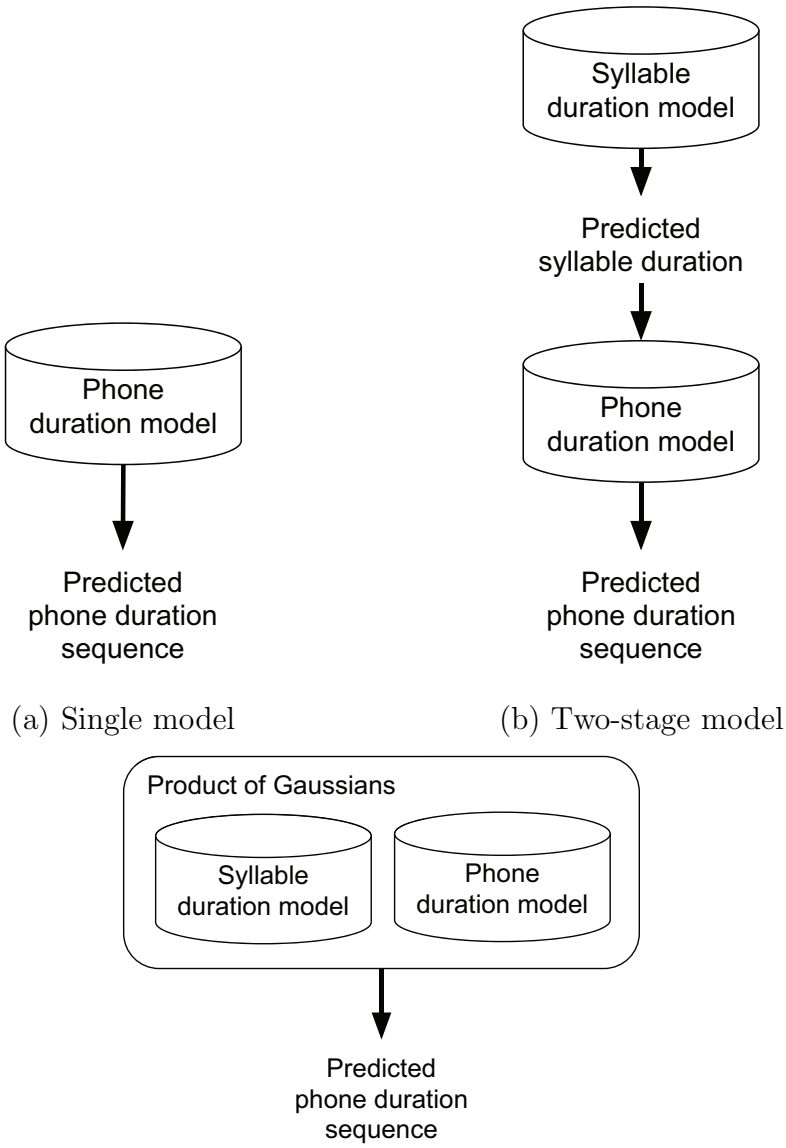
$$\boldsymbol{\mu}_D = \boldsymbol{\Sigma}_D(\mathbf{r} + {\boldsymbol{\Sigma}_T^p}^{-1}\boldsymbol{\mu}_T^p) \tag{4.12}$$

$$\boldsymbol{\Sigma}_D^{-1} = \mathbf{P}^{-1} + {\boldsymbol{\Sigma}_T^p}^{-1}. \tag{4.13}$$

The mean $\boldsymbol{\mu}_D$ is used as the predicted phone-duration sequence.

4.3 Experiments

The experiments were conducted to evaluate the performance of the proposed technique, product of Gaussian process (GP) experts. In the experiments, three techniques, *single model*, multi-level model by *two-stage* prediction, and *product of Gaussian process (GP) expert* model were compared in objective and subjective evaluations. Figure 4.1 summarizes these prediction approaches. The *single model* was the conventional GPR-based duration prediction [13] that uses a single-phone model for duration prediction. The *two-stage model* is described in Chapter 3. The two-stage model firstly predicts syllable duration, and then the phone-duration model is trained by using predicted syllable duration as an additional context. The proposed technique, product of GP-expert model, combines phone and syllable duration models for phone duration prediction as described in Section 4.2. In



(a) Single model

(b) Two-stage model

(c) Product of Gaussian process experts (Proposed technique)

Figure 4.1: Comparison of prediction models.

the product of GP-experts model, both phone- and syllable-duration models are trained separately and combine the predictive distributions for phone-duration prediction.

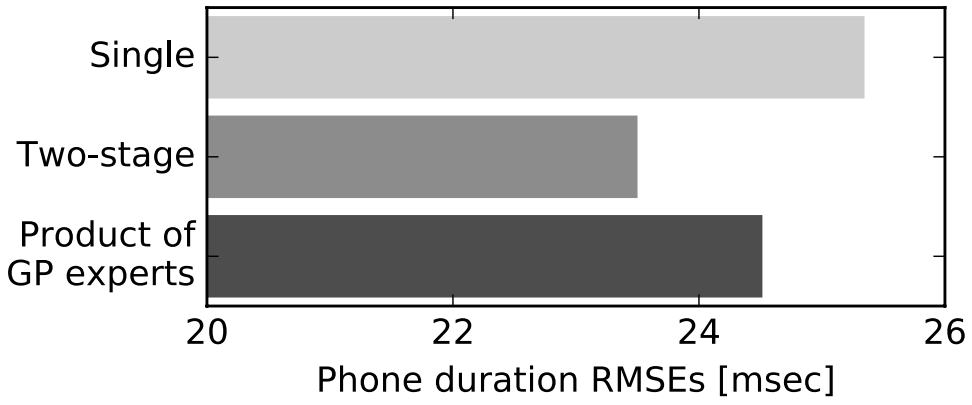


Figure 4.2: Phone duration distortion

4.3.1 Experimental condition

The speech database, T-Sync-1, was used in this experiment. The training and test sets were 450 utterances and 50 utterances, respectively. The test set was not included in the training data. The contextual variables of phone- and syllable-level models are the same as shown in 2.5 and 3.2, respectively. The kernel function was the same as used in Chapter 2 and 3 for phone- and syllable-level models, respectively.

Spectral features, aperiodicity, and F0 models were trained to generate speech parameters for the subjective evaluation. Speech signals sampled at a rate of 16kHz. Acoustic features were extracted by STRAIGHT [33] with 5-ms frame shift. The acoustic feature vector consisted of the 0-39th mel-cepstral coefficients, 5-band aperiodicity, log F0, and their delta and delta-delta coefficients. GP model training was employed partially independent conditional (PIC) approximation [15], and the kernel function parameters were optimized by EM-based method [14].

4.3.2 Objective evaluation

The objective evaluation used phone- and syllable-duration distortions between original and predicted durations. Figure 4.2 and 4.3 shows root mean square (RMS) errors of phone and syllable durations, respectively. The pro-

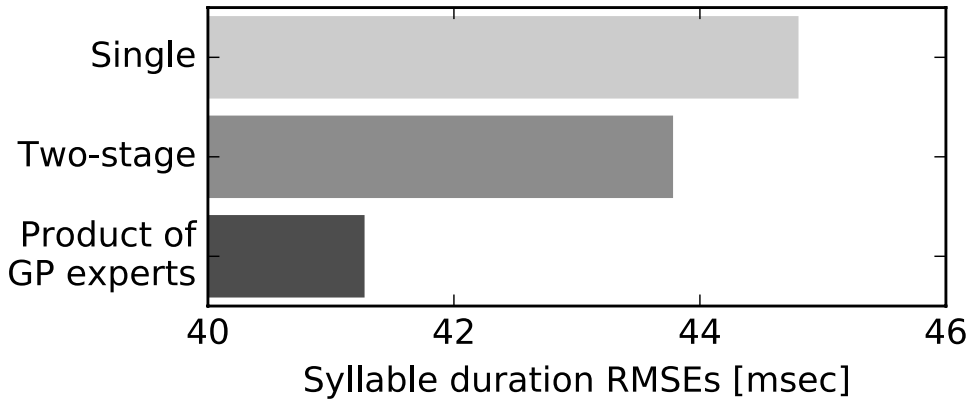


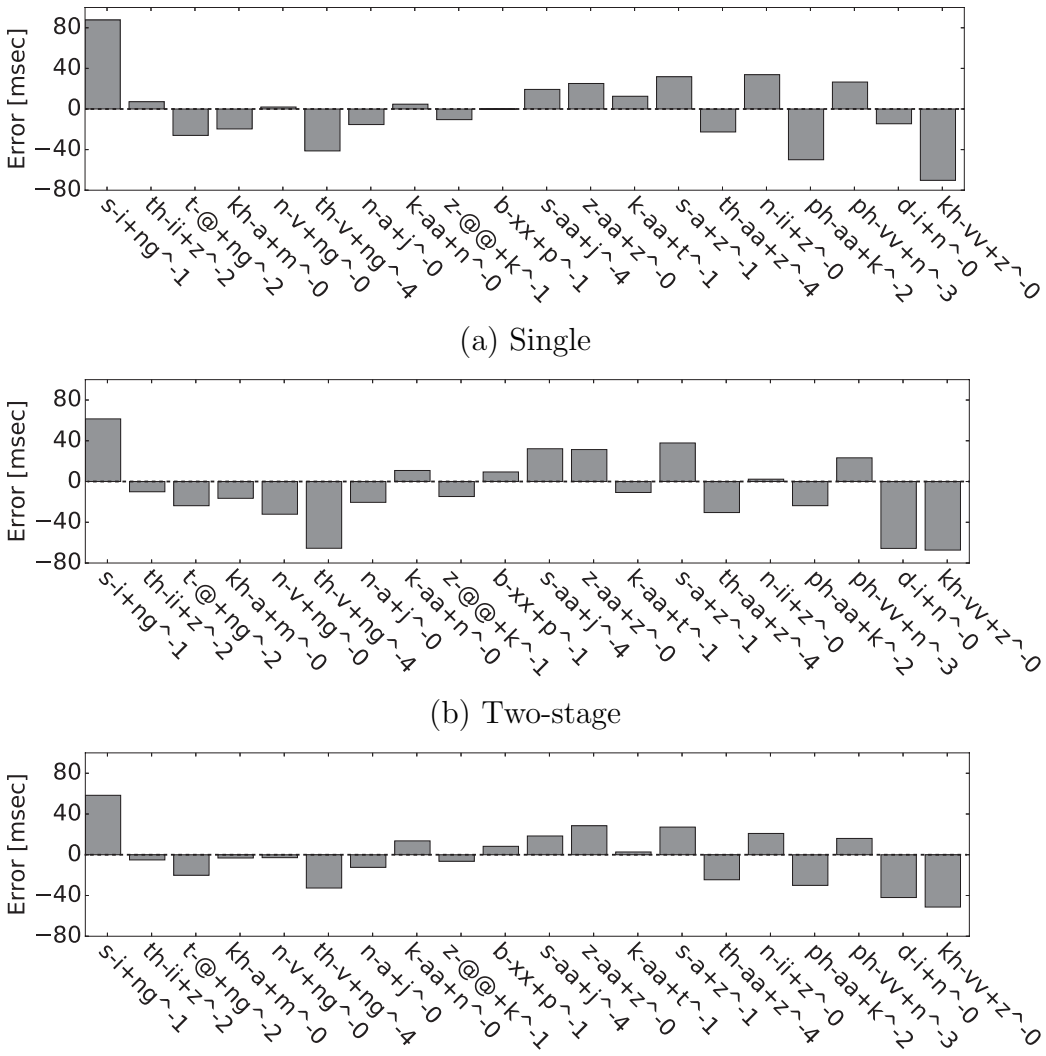
Figure 4.3: Syllable duration distortion

posed method, product of GP-experts, had lower RMS errors than the single-level model in both phone- and syllable-duration distortions. The two-stage method had lower RMS error than the proposed method in phone-duration distortion. However, the proposed method had lower RMS error than the two-stage method in syllable-duration distortion.

Figure 4.4 shows an example of syllable-duration errors in a sentence. Each bar represents the difference between the predicted syllable duration and the original. In the figure, the proposed method showed smaller errors than the other method in almost syllables.

4.4 Subjective evaluation

The subjective evaluation involved with mean opinion score (MOS) and the forced-choice preference test. Participants were ten Thai native speakers. Each person evaluated ten speech-samples that were randomly selected from the test set. In the MOS test, the participants evaluated each sample on a five-point scale from 1 to 5 according to their satisfaction in the naturalness of syllable and phone duration. The definition of the rating was 1: bad, 2: poor, 3: fair, 4: good, and 5: excellent. Participants could repeat playback as many times as they required for evaluation. Figure 4.5 shows the MOS scores with 95% confidence intervals. The result showed that the product of



(a) Single

(b) Two-stage

(c) Product of GP-expert technique

Figure 4.4: Comparison of duration prediction errors in syllable unit. The sentence is "... the points of concern in design of antenna at ground station is ..." in English.

GP-expert method received a higher score than the single-model one with statistically significant (p-value=0.019). The two-stage method had a higher score than the product of GP-expert one, but the difference was insignificant (p-value=0.2).

In the forced-choice preference test, the participants were asked to choose

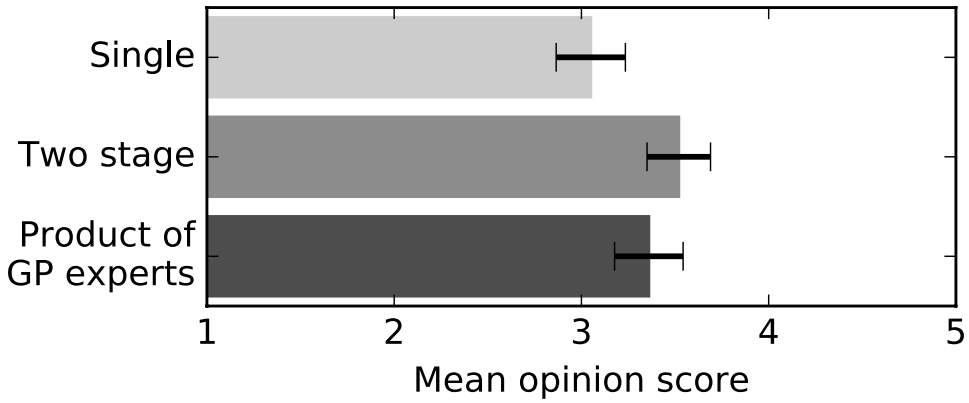


Figure 4.5: Result of MOS test in subjective evaluation of naturalness.

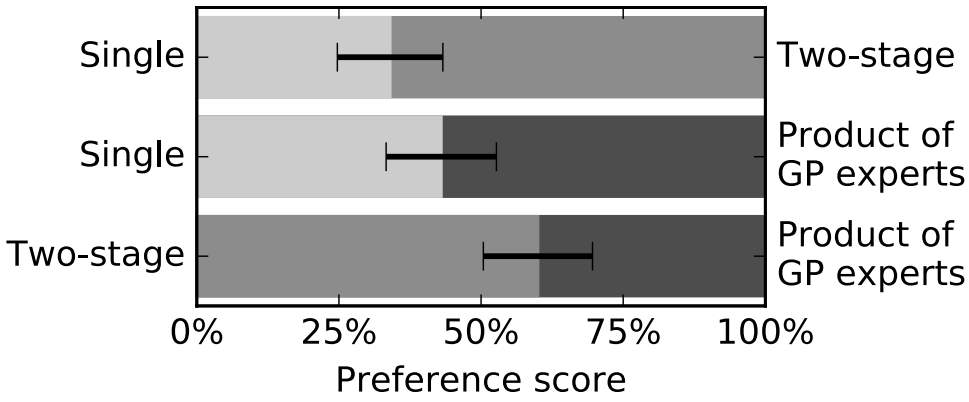


Figure 4.6: Result of forced choice preference test in subjective evaluation of naturalness.

more natural one regarding phone and syllable durations for each pair of speech samples. The participants could repeat playback as many times as they required in the same way as the MOS test. Figure 5.3 shows the resultant score of preference test. The result showed that the product of GP-expert method received higher preference than the single-model one. The two-stage method had a higher score than the product of GP-expert one. The reason might be that the perception of stress intensity is highly dependent on the duration of vowel and final-consonant than that of an entire

syllable. This means that even though the duration of initial consonant is very long, the participants may not perceive it as stressed syllable if vowel and final-consonant durations are short. Therefore, the accuracy of phone durations is more significant in the perception of naturalness than syllable durations.

4.5 Conclusion

This chapter describes an alternative method for multi-level model GPR-based duration prediction. This technique firstly trains phone- and syllable-duration models independently. The relationship between syllable and phone duration was expressed as syllable duration equals to the sum of phone durations. In prediction, the predictive distributions of syllable- and phone-duration models are combined by product of Gaussian process experts. Then, mean of the product is used as the predicted phone duration. The objective evaluation showed that the product of GP expert method outperformed the single-model one in both objective and subjective evaluations. The product of GP expert method had lower RMS error than the two-stage one in syllable-duration distortion. The results of the subjective evaluation showed that the product of GP expert method was comparable with the two-stage one. In future work, experiments will be conducted with a larger number of syllables since Thai syllables are quite complex and the current amount of syllable data might be insufficient.

Chapter 5

Enhanced F0 Generation for GPR-Based Speech Synthesis Considering Syllable-Based Prosodic Features

This chapter describes a technique to consider a syllable-level model to enhanced F0 generation for GPR-based SPSS. This technique uses discrete cosine transform (DCT) coefficients as output variables of the syllable-level model. The DCT coefficients are extracted from log F0 contour in syllable unit. F0 contours are generated by jointly maximizing predictive distributions of frame- and syllable-level models. Objective and subjective evaluations showed improvement in F0 generation when using a small amount of training data, approximately 30 minutes.

5.1 Introduction

Fundamental frequency (F0) contour is a prosodic feature which contributes to linguistic functions such as stress, accent, emphasizing, and tone. Various techniques have been proposed to F0 modeling for improving the naturalness of synthetic speech. However, the conventional F0 modeling technique is a single-level modeling. For example, HMM-based SPSS uses state-level F0

modeling, and DNN- and GPR-based SPSS uses frame-level one. In HMM-based SPSS, various attempts have been made to improve F0 generation performance by incorporating multiple models. In [59], discrete cosine transform (DCT) was used to parameterize log F0 contour for syllable-level modeling, and F0 contours were generated by maximizing the log-likelihood of phone- and syllable-level models. The product of experts was utilized to combine multiple experts for prosody generation [56]. In the DNN-based SPSS, the syllable-level network was trained by using suprasegmental features, and the output of syllable-level network was integrated as inputs of frame-level network [60]. Hierarchical and additive architectures of DNN were proposed to model suprasegmental features for prosody generation [61].

In Chapter 2, the use of GPR-based F0 generation has shown better performance than the HMM- and DNN-based F0 generation. However, a single frame-level model is insufficient to model F0 at suprasegmental units, especially when the amount of training data is small. This chapter examines a multi-level model approach to improve F0 generation performance in GPR-based speech synthesis by considering a syllable-level model.

In GPR-based method, speech parameters are generated by the parameter of predictive distribution for given contexts. The use of predictive distribution enables various statistical techniques to combine multiple models. This chapter uses a joint maximization to utilize multi-level models for F0 generation. First, the frame- and syllable-level models are separately trained. The input and output variable of syllable-level are syllable-level context and DCT coefficients extracted from F0 contour of a syllable, respectively. In the synthesizing, the combined predictive distributions of frame- and syllable-level models are maximized. The objective and subjective evaluations were conducted to measure improvement of the proposed method.

5.2 Multiple GP models for F0 generation

In the proposed technique, the model training of frame- and syllable-level model are performed individually. The frame-level model is the same setting as the conventional GPR-based SPSS [13]. For a syllable-level model, the input variables are the syllable-level context that contains linguistic and po-

sitional information of syllable, word, and utterance units as shown in Table 3.2. The output variables are K DCT-coefficients, the 0-th to $(K-1)$-th DCT coefficients, extracted from log F0 contour in syllable unit. Unvoiced frames were interpolated before DCT coefficient extraction. Then, the syllable-level model is trained in the same manner as the conventional GPR-based SPSS.

In synthesizing phase, let $\mathbf{S}_N$ and $\mathbf{X}^s$ be the matrix forms of DCT coefficients and syllable-level contexts of training data, respectively. Let $\mathbf{S}_T$ and $\mathbf{X}^s_T$ be the matrix forms of test data. $\mathbf{Y}_N$ and $\mathbf{Y}_T$ are the feature vector of log F0 sequence of training and test data, respectively. Then, the predictive distribution of frame-level model is given by

$$p(\mathbf{Y}_T|\mathbf{Y}_N, \mathbf{X}, \mathbf{X}_T) = \mathcal{MN}(\mathbf{Y}_T|\mathbf{M}, \mathbf{\Sigma}, \mathbf{V}) \tag{5.1}$$

and that of syllable-level model is given by

$$p(\mathbf{S}_T|\mathbf{S}_N, \mathbf{X}^s, \mathbf{X}^s_T) = \mathcal{MN}(\mathbf{S}_T|\mathbf{M}_s, \mathbf{\Sigma}_s, \mathbf{V}_s) \tag{5.2}$$

where $\mathbf{M} = [\mathbf{M}_0 \quad \mathbf{M}_1 \quad \mathbf{M}_2]$, $\mathbf{\Sigma}$, $\mathbf{M}_s = [\mathbf{M}_{s,0} \quad \cdots \quad \mathbf{M}_{s,K-1}]$, and $\mathbf{\Sigma}_s$ are mean and covariance matrices of the frame- and syllable-level predictive distributions, respectively. $\mathbf{V} = \mathrm{diag}[\sigma_0^2, \sigma_1^2, \sigma_2^2]$ and $\mathbf{V}_s = \mathrm{diag}[\sigma_{s,0}^2, \cdots, \sigma_{s,K-1}^2]$ are variances of each dimension of training data for corresponding models.

In the GPR-based method, the frame-level output variables $\mathbf{Y}_T$ contain log F0 sequence including its delta and delta-delta as follows:

$$\mathbf{Y}_T = \begin{bmatrix}\mathbf{C} & \Delta\mathbf{C} & \Delta^2\mathbf{C}\end{bmatrix} \tag{5.3}$$

$$= \begin{bmatrix}\mathbf{W}_0\mathbf{C} & \mathbf{W}_1\mathbf{C} & \mathbf{W}_2\mathbf{C}\end{bmatrix} \tag{5.4}$$

where $\mathbf{C}$ is static feature vector of log F0 sequence and $\mathbf{W}_n$ is a window matrix for calculating the n-th dynamic features of log F0 sequence as described in [58].

In syllable-level model, the transformation between DCT coefficients $\mathbf{S}_T$ and log F0 contours is defined in a similar manner as the frame level as follows:

$$\mathbf{S}_T = \begin{bmatrix}\mathbf{S}_0 & \ldots & \mathbf{S}_{K-1}\end{bmatrix} \tag{5.5}$$

$$= \begin{bmatrix}\mathbf{E}_0\mathbf{C} & \ldots & \mathbf{E}_{K-1}\mathbf{C}\end{bmatrix} \tag{5.6}$$

where $\mathbf{E}_k$ is a window matrix for calculating the k-th DCT coefficient of log F0 sequence in syllable unit in a similar manner as described in [39].

$\mathbf{S}_k$ is the sequence of the k-th coefficient. An example of $\mathbf{E}_k$ is given as follows:

$$\begin{matrix} \mathbf{S}_k \\ \begin{bmatrix} s_{k,0} \\ \vdots \\ s_{k,J} \end{bmatrix} \end{matrix} = \begin{matrix} \mathbf{E}_k \\ \begin{bmatrix} \mathbf{E}_{k,0} & & \\ & \ddots & \\ & & \mathbf{E}_{k,J} \end{bmatrix} \end{matrix} \begin{matrix} \mathbf{C} \\ \begin{bmatrix} \mathbf{C}_0 \\ \vdots \\ \mathbf{C}_J \end{bmatrix} \end{matrix} \tag{5.7}$$

where $s_{k,j}$, $\mathbf{E}_{k,j}$, and $\mathbf{C}_j$ correspond to the DCT coefficient, the DCT matrix, and the static feature of log F0 sequence for the k-th order and j-th syllable. More specifically, the j-th syllable has T_j frames, then $\mathbf{E}_{k,j}$ is given by

$$\mathbf{E}_{k,j} = \frac{2}{T_j}\left[\cos\left[\frac{\pi}{T_j}k(0+\frac{1}{2})\right] \;\cdots\; \cos\left[\frac{\pi}{T_j}k(T_j-1+\frac{1}{2})\right]\right]. \tag{5.8}$$

The predictive distribution of the k-th DCT coefficient is defined by $p(\mathbf{S}_k|\mathbf{S}_N, \mathbf{X}^s, \mathbf{X}^s_T) = \mathcal{MN}(\mathbf{S}_k|\mathbf{M}_{s,k}, \mathbf{\Sigma}_s, \sigma^2_{s,k})$. Then, the log joint predictive distribution is given by

$$\begin{aligned} \log p(\mathbf{C}) \propto{}& \log p(\mathbf{Y}_T|\mathbf{Y}_N, \mathbf{X}, \mathbf{X}_T) \\ &+ \sum_{k=0}^{K-1} \alpha_k \log p(\mathbf{S}_k|\mathbf{S}_N, \mathbf{X}^s, \mathbf{X}^s_T) \qquad (5.9) \\ \propto{}& -\frac{1}{2}\mathrm{Tr}\big[\mathbf{V}^{-1}(\mathbf{Y}_T - \mathbf{M})^\top \mathbf{\Sigma}^{-1}(\mathbf{Y}_T - \mathbf{M})\big] \\ &- \frac{1}{2}\sum_{k=0}^{K-1} \alpha_k \mathrm{Tr}\big[\sigma^{-2}_{s,k}(\mathbf{S}_k - \mathbf{M}_{s,k})^\top \\ &\qquad \mathbf{\Sigma}_s^{-1}(\mathbf{S}_k - \mathbf{M}_{s,k})\big]. \qquad (5.10) \end{aligned}$$

By applying the definitions in Eqs. (5.3) and (5.5), the predictive distribution can be rewritten in terms of $\mathbf{C}$ as follows:

$$\begin{aligned} \log p(\mathbf{C}) \propto{}& -\frac{1}{2}\sum_{i=0}^{2} \mathrm{Tr}\big[\sigma_i^{-2}(\mathbf{W}_i\mathbf{C} - \mathbf{M}_i)^\top \mathbf{\Sigma}^{-1}(\mathbf{W}_i\mathbf{C} - \mathbf{M}_i)\big] \\ &- \frac{1}{2}\sum_{k=0}^{K-1} \alpha_k \mathrm{Tr}\big[\sigma^{-2}_{s,k}(\mathbf{E}_k\mathbf{C} - \mathbf{M}_{s,k})^\top \\ &\qquad \mathbf{\Sigma}_s^{-1}(\mathbf{E}_k\mathbf{C} - \mathbf{M}_{s,k})\big] \qquad (5.11) \end{aligned}$$

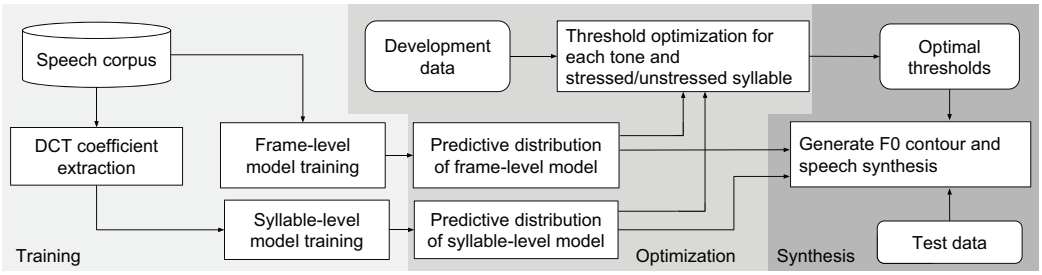


Figure 5.1: Block diagram of the proposed method using multiple models for GPR-based F0 generation.

where α_k is a threshold for the k-th coefficient. The generated F0 contour can be obtained by maximizing $\log p(\mathbf{C})$.

In the experiment, the optimization of the α_0 are separated from the other thresholds $\alpha_*(= \alpha_1 = \alpha_2 = \cdots = \alpha_{K-1})$. In conclusion, the log likelihood of predictive distribution is expressed as

$$\begin{aligned}\log p(\mathbf{C}) \propto & -\frac{1}{2}\sum_{i=0}^{2}\mathrm{Tr}\left[\sigma_i^{-2}(\mathbf{W}_i\mathbf{C}-\mathbf{M}_i)^\top\mathbf{\Sigma}^{-1}(\mathbf{W}_i\mathbf{C}-\mathbf{M}_i)\right] \\ & -\frac{1}{2}\alpha_0\mathrm{Tr}\left[\sigma_{s,0}^{-2}(\mathbf{E}_0\mathbf{C}-\mathbf{M}_{s,0})^\top\mathbf{\Sigma}_s^{-1}(\mathbf{E}_0\mathbf{C}-\mathbf{M}_{s,0})\right] \\ & -\frac{1}{2}\alpha_*\sum_{k=1}^{K-1}\mathrm{Tr}\left[\sigma_{s,k}^{-2}(\mathbf{E}_k\mathbf{C}-\mathbf{M}_{s,k})^\top\right. \\ & \qquad\left.\mathbf{\Sigma}_s^{-1}(\mathbf{E}_k\mathbf{C}-\mathbf{M}_{s,k})\right]. \end{aligned} \tag{5.12}$$

An overview of training and synthesizing processes is shown in Figure 5.1.

5.3 Experiments

The experiments were conducted to evaluate the performance of *multiple* models compare to *single* model for F0 generation in the GPR-based speech synthesis framework. The single model method was the conventional GPR-based F0 generation used in [62]. The multiple model method was the proposed one as described in Section 5.2. A set of phonetically balanced sentences of Thai speech corpus, T-Sync-1 developed by NECTEC [32] was used in the experiment. The training data contained 250 and 450 utterances ap-

Table 5.1: Distortions between original and generated log F0 contours of stressed syllable.

Number of utterances for training	Method	RMS erros of log F0 [cent] Stressed syllable				
		Tone 0	Tone 1	Tone 2	Tone 3	Tone 4
250	Single	82.68	84.97	101.7	104.66	104.64
	Multiple	79.37	79.86	96.37	103.44	100.43
450	Single	77.65	75.82	96.7	99.39	91.57
	Multiple	76.49	75.82	96.6	99.39	85.6

proximately 27 and 48 minutes in duration, respectively. The development set for finding optimal threshold values consisted of 50 utterances. The test set consisted of 50 utterances that were not included in the training data nor the development set. In the multiple model method, a grid search was performed to find optimal thresholds α_0 and α_*. The grid search was conducted separately under a condition of tone-type and stressed/unstressed.

Speech signals were sampled at a rate of 16 kHz. Spectral features, aperiodicity, and log F0 were extracted by STRAIGHT [33] with a 5-ms frame shift. For subjective evaluation, mel-ceptral, aperiodicity, and phone duration were trained with the conventional GPR-based SPSS as described in Chapter 2. The acoustic feature vector consisted of the 0th to 39th mel-cepstral coefficients, five-band aperiodicity, log F0, and their delta and delta-delta coefficients. The output variables of syllable-level model were zeroth to second DCT coefficients of log F0 contour in syllable unit, since it is sufficient to represent F0 pattern in syllable unit. The training was performed with partially independent conditional (PIC) approximation [15]. The EM-based method was used for optimizing the parameter of the kernel function.

5.3.1 Objective evaluation

Root-mean-square (RMS) errors between original and generated F0 contour were used for the objective evaluation. The RMS errors of stressed, unstressed, and all syllables are shown in Figure 5.1, 5.2, and 5.3. The multiple

Table 5.2: Distortions between original and generated log F0 contours of unstressed syllable.

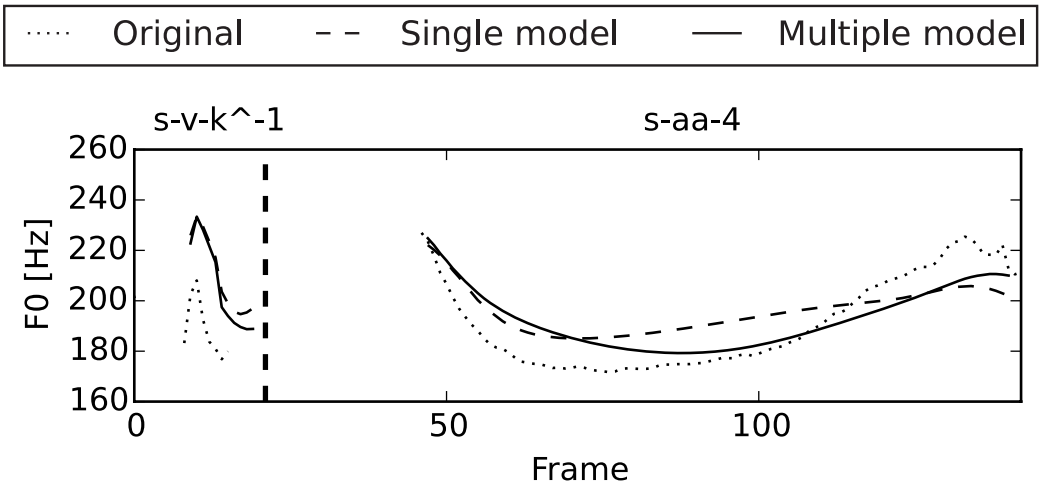
Number of utterances for training	Method	RMS erros of log F0 [cent]				
		Stressed syllable				
		Tone 0	Tone 1	Tone 2	Tone 3	Tone 4
250	Single	106.75	128.73	117.30	105.87	108.21
	Multiple	101.63	128.73	111.29	105.87	93.84
450	Single	101.07	117.66	110.07	104.7	92.11
	Multiple	99.15	117.66	109.9	104.7	91.3

Table 5.3: Distortions between original and generated log F0 contours of all syllable.

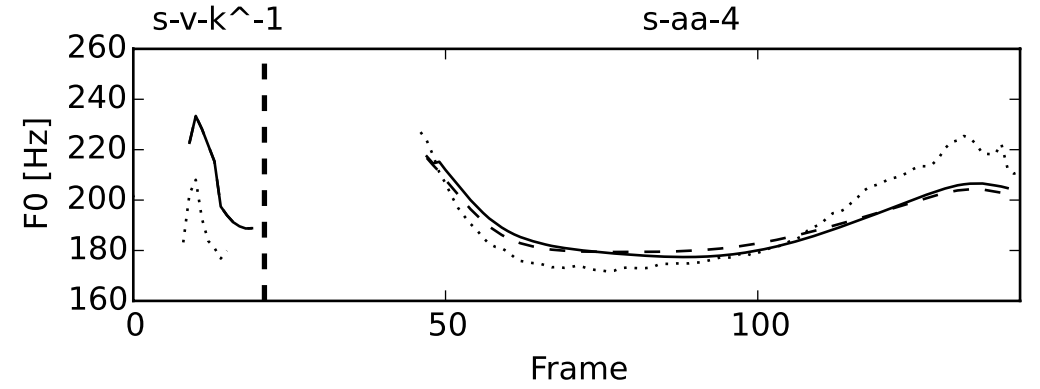
Number of utterances for training	Method	RMS erros of log F0 [cent]
		Stressed syllable
		All
250	Single	108.07
	Multiple	101.79
450	Single	100.64
	Multiple	99.71

model method showed significant improvement when using 250 utterances for training data. The multiple model method could reduce RMS errors in all tones of stressed syllables and tones 0, 2, and 4 of unstressed syllables. When the amount training data was 450 utterances, it had a notable improvement in tone 4 of stressed syllables and a small improvement in tones 0 and 2 of stressed and unstressed syllables. In contrast, there was no improvement in tones 1 and 3. A reason might be that tone 1 and tone 3 of unstressed syllables are usually dominated by unvoiced frames which causes errors in the extraction of DCT coefficients. In all syllable, 250-utterance case showed significant reduction of log F0 distortion, while 450-utterance case also showed an improvement.

Figure 5.2 shows an example of F0 contours generated by single and



(a) F0 contour generated using 250 utterances



(b) F0 contour generated using 450 utterances

Figure 5.2: Example of generated F0 contours of single and multiple model methods. The example word is "s-v-kˆ-1, s-aa-4" meaning "education". The digits indicate a tone-type of syllables.

multiple model methods. Figure 5.2(a) shows an F0 contour generated by 250 utterances of training data. It can be seen that the multiple model method could generate F0 contour closer to the original one than the single model one, especially at the last syllable "s-aa-4" (rising tone). 5.2(b) shows an F0 contour generated by 450 utterances of training data. There was no significant difference between the single and multiple model methods.

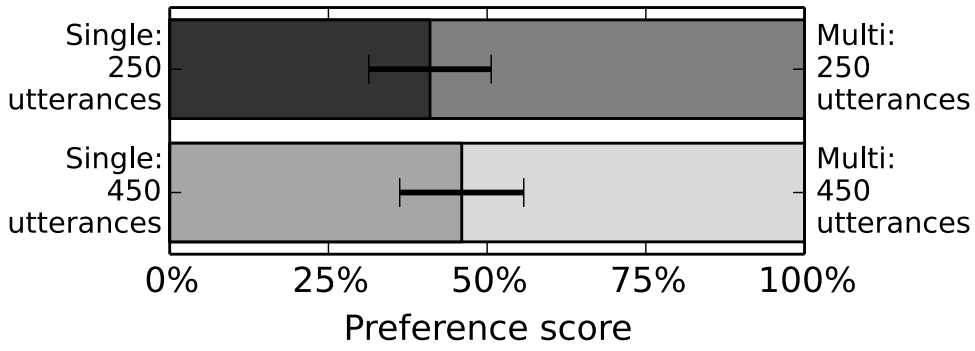


Figure 5.3: Result of forced choice preference test in subjective evaluation of naturalness.

However, it was seen that the multiple model method generated a slightly wider range of F0 change. From this result, it can be seen that the multiple model method could enhance F0 generation, especially when the training data is small.

5.3.2 Subjective evaluation

The subjective test was conducted by a forced-choice preference test to evaluate perceptual in naturalness of generated F0 contours. Spectral features, aperiodicity, and duration of all speech samples were generated by the conventional single-level GPR-based speech synthesis [13, 14]. The difference between respective methods was log F0 contours that were generated by single and multiple models with 250 and 450 utterances of training data. Ten Thai native speakers participated in the subjective evaluation, and each person evaluated ten samples randomly selected from 50 test samples.

In the forced choice preference test, the participants were asked to choose more natural-sounding one regarding tone and intonation for each pair of samples. The participants could repeat playback as many times as they required. The result of forced choice preference test is shown in Figure 5.3. It can be seen that the multiple model method received higher preference score than the single model method in both 250 and 450 utterances of training data.

5.4 Conclusion

This chapter has proposed a multiple model method to enhance F0 generation for GPR-based speech synthesis. In the proposed technique, frame- and syllable-level GP models were trained. In syllable-level GP model, DCT coefficients extracted from log F0 contour in syllable unit were used as output variables. When synthesizing phase, the predictive distributions of frame- and syllable-level models were jointly maximized for generating log F0 contours. A grid search method was used to find optimal thresholds for each tone and stressed/unstressed conditions. The objective evaluation results showed that the multiple model method had less distortion than the single model method, especially when the training data size was relatively small. The subjective evaluation results also showed that the multiple model method could generate more natural-sounding F0 contours than the single model one. From this point, the proposed technique enables the multiple GP model to be combined and showed an improvement in the experiment. This paper examined the multiple model method using speech data of a professional speaker which has clear articulation. In future work, the multiple model method will be examined using non-professional speakers' utterances which have imperfect articulation and cause unnatural-sounding generated F0 contours in synthetic speech.

Chapter 6

Unsupervised Stress Information Labeling Using Gaussian Process Latent Variable Model for Statistical Speech Synthesis

In Thai language, stress is an essential factor that affects naturalness and has a crucial role in meaning. It is seen that a speech synthesis model that is trained with lack of stress causes incorrect tones and ambiguity in the meaning of synthetic speech. The previous studies have shown that manually annotated stress information improves the naturalness of synthetic speech [63, 64]. However, high time consumption is a drawback of the manual annotation. In this chapter, an unsupervised learning technique, called Bayesian Gaussian process latent variable model (Bayesian GP-LVM), was used to put stress annotation on the given training data automatically. Stress-related features were projected onto a latent space in which syllables was easier classified into stressed/unstressed classes. The stressed/unstressed information was used as an additional context in GPR-based speech synthesis. Experimental results showed that the proposed technique improves the naturalness of synthetic speech as well as the accuracy of stressed/unstressed classification.

6.1 Introduction

In speech synthesis, the main goal is to generate speech which is natural-sounding and has the intended meaning. Prosody is an important factor that has a significant influence on naturalness and meaning of speech. Various techniques have been used to model prosodic features for generating natural-sounding speech. In tonal languages, the tone is a significant factor used for distinguishing the lexical or grammatical meaning of speech. Since Thai is a tonal language and tone is very sensitive in perception, a tone-separated tree structure was proposed to remove tone-dependency on the context in tree-based context clustering for HMM-based SPSS [65]. Moreover, pitch contour varies diversely in continuous speech, and thus only tone-type context is not sufficient in F0 modeling. To model diversity of F0 contour in each tone, tone geometrical features that represent the shape of F0 contour were proposed in F0 modeling for speech synthesis [19]. Another technique for modeling prosodic feature in Thai is a modified version of Tilt model called T-Tilt [20, 21] which was successfully used for representing prosody in accentual languages. Since co-articulation affects F0 contour shape but tone nuclei are less affected by adjacent syllables, a tone nucleus model was used in F0 modeling and generation [66]. Furthermore, due to the fact that vowel part of a syllable receives small effect from neighboring syllables and contains the main prosodic feature of a syllable, an F0 modeling using only vowel part instead of entire syllable was proposed to reduce complexity and improve accuracy in tone recognition [67].

In addition to tone, stress is another important factor in Thai language which affects naturalness and meaning of a sentence. The use of stress information can improve the accuracy of tone recognition [68]. The previous study showed that manually annotated stress information could reduce F0 and duration distortions in the HMM-based speech synthesis [64]. To alleviate the problem of high cost of manual labeling, [63] proposed an unsupervised labeling technique for classifying syllables into stress-related classes based on F0 movement and syllable duration. However, problems remain; some tones have low F0 movement in both stressed and unstressed cases, and error in F0 extraction may cause a high F0 variance in a syllable.

In this paper, to overcome the problems, this chapter proposes a new unsupervised labeling technique for stress annotation. This technique utilizes a dimensionality reduction technique, called Bayesian Gaussian process latent variable model (Bayesian GP-LVM), to project prosodic features onto latent space in which similarity of prosodic features can be easily measured by using the distance between latent variables. In [69], the latent variables of Bayesian GP-LVM were directly used as additional context. In contrast, the proposed technique clusters the latent variables into simple stressed/unstressed classes and uses the obtained class information as an additional context. This method enables intended stress position to be given to label sequence. Experiments examine stressed/unstressed classification performance to evaluate the effectiveness of the use of latent variables. Then Gaussian process regression (GPR)-based speech synthesis [12] was used as speech synthesis framework for listening tests, which could generate more natural-sounding speech than the HMM-based one [16]. Lastly, this chapter assessed the performance of newly added context though objective and subjective tests.

6.2 Unsupervised stress information labeling

6.2.1 Stress in Thai

Stress is a major factor in which contributes to diversity of prosodic features [26]. It affects naturalness and meaning of speech. As described in [28], the position of a stressed syllable influences meaning of phrase level. The position of stressed syllable is unknown and cannot be obtained from a text because Thai is a non-lexical stress language. However, various studies of stress agree that stressed syllables are usually isolated syllable, a syllable at the end of a phrase, and emphasized syllable or word [25, 26, 28, 70]. Regarding acoustic characteristics, it is known that stressed syllables have F0 contours similar to typical F0 contours and long durations, whereas unstressed syllables are otherwise [25]. Additionally, durations of stressed syllables also depend on final consonant [31].

6.2.2 Bayesian Gaussian process latent variable model

In this chapter, stress annotation was conducted in an unsupervised way. For this purpose, a dimensionality reduction technique, Bayesian GP-LVM [71], was performed to reduce the complexity of stress-related features, specifically, F0 contour and duration. Bayesian GP-LVM is robust to overfitting and unhelpful data points. Moreover, it can determine most dominant dimensions of the nonlinear latent space. Next, unsupervised clustering was employed on the latent variables obtained from Bayesian GP-LVM training.

In Bayesian GP-LVM, the output variables Y are observed, and the input variables Z are fully unobserved and treated as latent variables. In the model training, the stress-related features were used as the observed variables Y. In Bayesian GP-LVM training, the marginal likelihood of data is given by

$$p(Y) = \int p(Y|Z)p(Z)dZ. \tag{6.1}$$

Then a variational distribution $q(Z)$ is introduced to approximate the posterior of latent variables $p(Z|Y)$ as follows:

$$q(Z) = \prod_{i=1}^{N} \mathcal{N}(z_i|\mu_i, S_i) \tag{6.2}$$

where μ_i, and S_i are mean and covariance. A variational lower bound $\mathcal{F}$ is derived as

$$\mathcal{F} \leq \log p(Y) \tag{6.3}$$

$$\mathcal{F} = \langle \log p(Y|Z) \rangle_{q(Z)} - KL(q(Z)|p(Z)) \tag{6.4}$$

where $\langle \cdot \rangle_{q(Z)}$ is the expectation with respect to $q(Z)$. The variational parameters μ_i, and S_i are obtained by maximizing the lower bound.

In stressed/unstressed clustering, the means μ_i of variational distribution were used as features in unsupervised learning. The stressed/unstressed classes obtained from the clustering are used as an additional context in GPR-based SPSS. The stress context is represented by binary values: 1 for stressed syllable and 0 for unstressed syllable. The similarity of stress context is calculated by a linear kernel.

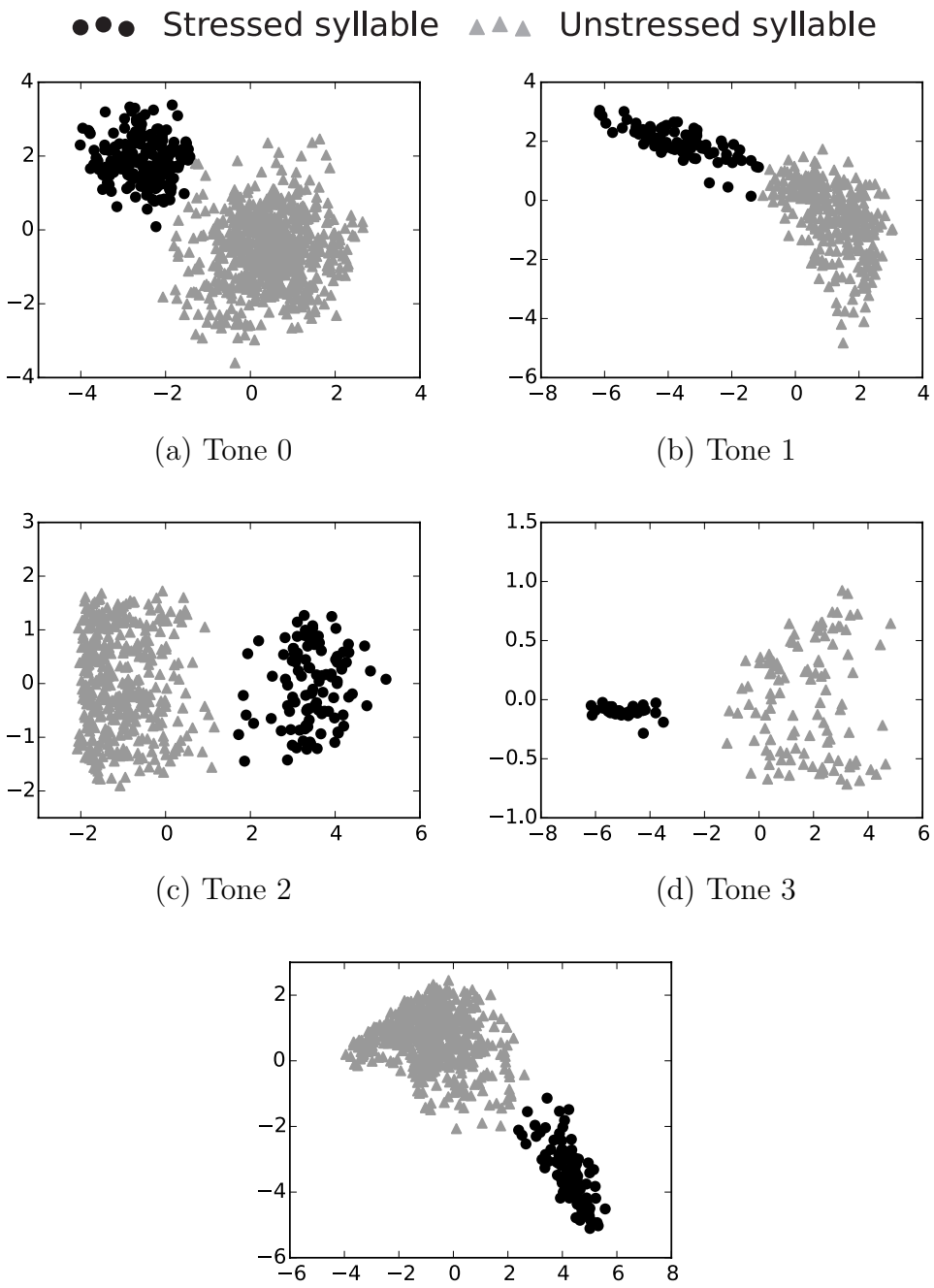
6.3 Experiments

First, unsupervised learning was performed to annotate stress information to context set of speech synthesis. Then acoustic-feature generation was performed to measure the improvement by considering stress information in the GPR-based speech synthesis. A set of phonetically balanced sentences of Thai speech database T-Sync-1 from NECTEC [32] was used for training and evaluation. Speech signals were sampled at a rate of 16kHz. STRAIGHT [33] was used to extract spectral features, aperiodicity, and F0 with 5-ms frame shift.

6.3.1 Stressed/unstressed annotation

The training set contained 329 utterances, 10741 syllables in total. The numbers of stressed and unstressed syllables were 1491 and 9250, respectively. Bayesian GP-LVM training was performed by using stress-related features, log F0 contour and duration in syllable-unit, as observed variables. In this experiment, prosodic features in initial consonant part were omitted because they did not have significant differences between stressed and unstressed syllables. Log F0 contour in the unvoiced region was interpolated by using a third-order polynomial. Since durations of respective syllables are not equal, the log F0 contour was normalized into 50 samples, and its delta and delta-delta were also included in the observed variables. As a result, the observed variables had 150 dimensions of log F0 contour information and 1 dimension of syllable duration. The dimensionality of latent variables was 10 dimensions. Bayesian GP-LVM was trained by using squared exponential kernel as a covariance function and 100 inducing points. The model optimization was conducted by using scaled conjugate gradient method. Since the characteristics of stress depend on tone-type and final consonant type of syllable, Bayesian GP-LVMs were separately trained based on tone-type and final consonant of syllables. Syllables were grouped based on their final consonant into three types: non final consonant syllable, nasal final consonant syllable, and non-nasal final consonant syllable.

The accuracy of stressed/unstressed classification was measured to evaluate the effectiveness of the latent variable model. Density-based clustering



(a) Tone 0

(b) Tone 1

(c) Tone 2

(d) Tone 3

(e) Tone 4

Figure 6.1: Visualization of stress-related features in latent space by keeping most two dominant dimensions from the projections of syllables with nasal final consonant.

using DBSCAN algorithm was performed [72] to cluster the latent variables. Then each cluster was identified to be stressed or unstressed class by observing distribution of a small set of labeled training data in the latent space. Stressed/unstressed classification performance was compared with that obtained by using observed variables. The accuracy was calculated by evaluating consistency with the manual stress annotation. The results are shown in Tables 6.1 and 6.2. It can be seen that the use of the latent variables provides higher F1-scores than the observed variables. The accuracy of tone 2 and 4 is higher than other tones because these tones are dynamic ones whose characteristics of stressed syllable are much different from the unstressed ones. The differences between stressed and unstressed static tones, i.e., tones 0, 1 and 3, are not so large as the dynamic ones. Figure 6.1 shows the visualization of the observed variables in latent space by projecting syllables that have nasal final consonant. The stress annotation in the figure was obtained by the unsupervised learning.

Table 6.1: Accuracy of stressed/unstressed syllable classification with observed variables. Values represent F1-scores.

Positive class	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4	All
Unstressed	0.975	0.934	0.987	0.943	0.936	0.96
Stressed	0.808	0.542	0.93	0.401	0.641	0.708

Table 6.2: Accuracy of stressed/unstressed syllable classification with latent variables. Values represent F1-scores.

Positive class	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4	All
Unstressed	0.983	0.972	0.99	0.98	0.989	0.983
Stressed	0.87	0.792	0.982	0.818	0.94	0.88

6.3.2 Experimental conditions for objective/subjective evaluation

Synthetic speech with and without using stress information context was generated to clarify the effectiveness of the stress context. The stress information

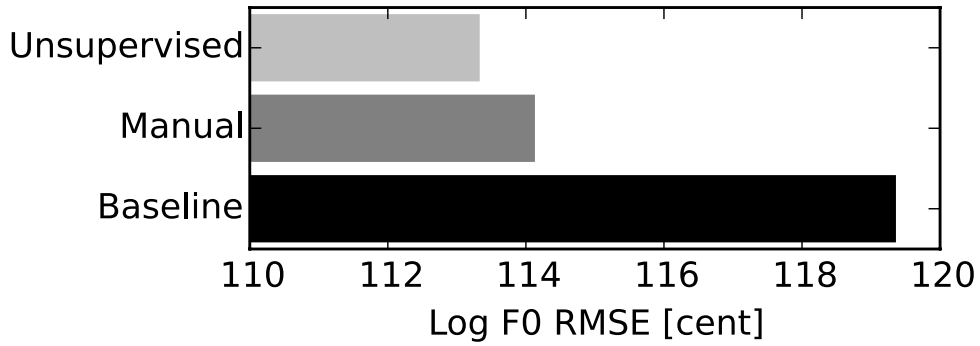


Figure 6.2: Log F0 distortions between original and synthetic speech.

is obtained from the manual labeling and the unsupervised labeling described in Section 6.3.1.

The training set contained 329 utterances, approximately 50 minutes in total, and 40 utterances were used for evaluation which were not included in the training set. The acoustic-feature vector consisted of the 0-39th mel-cepstral coefficients, 5-band aperiodicity, log F0, and their delta and delta-delta coefficients. The acoustic models were trained by using PIC approximation [15] and EM-based optimization [14]. The context set of a GPR-based model described in [62] was used for the baseline context set. In the proposed technique, stress information was incorporated as an additional context in the context set. The manual stress labeling was the same set as used in [64]. In the test set, stress information was manually annotated.

6.3.3 Objective evaluation

Objective tests were evaluated by log F0 and duration distortions between synthetic and original speech samples. The result of log F0 distortion is shown in Figure 6.2. In the figure, “Baseline” represents the result without using stress information, “Manual” and “Unsupervised” represent the results with stress information by manual labeling and automatic labeling using the proposed technique, respectively. It is seen that the stress context gives smaller log F0 distortion than the baseline. It is noted that there is a small difference between manual and unsupervised labeling cases. Figure

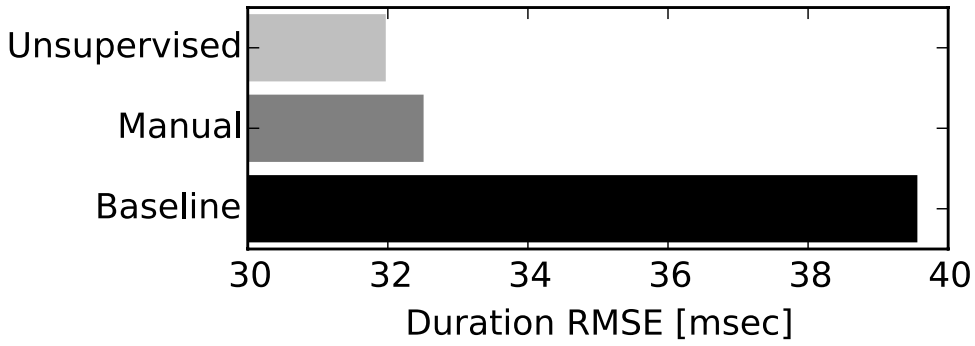


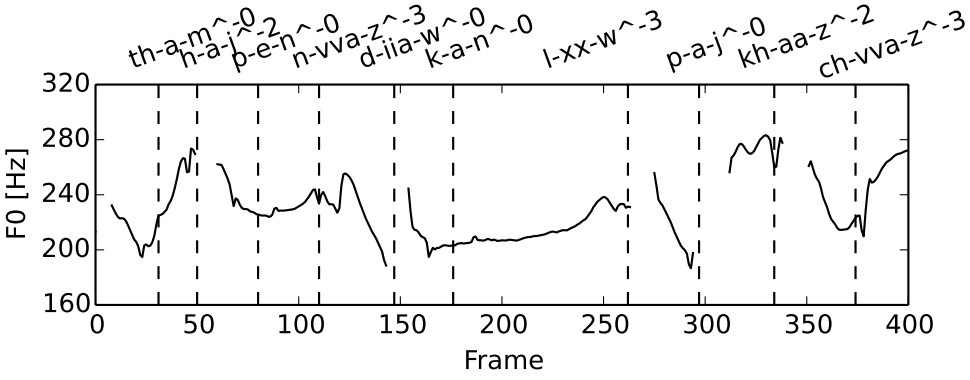
Figure 6.3: Duration distortions between original and synthetic speech.

6.3 shows the duration distortions. It shows the similar result to that of log F0 distortions that the stress context can reduce distortion, and there is no significant difference between the manual and unsupervised labeling.

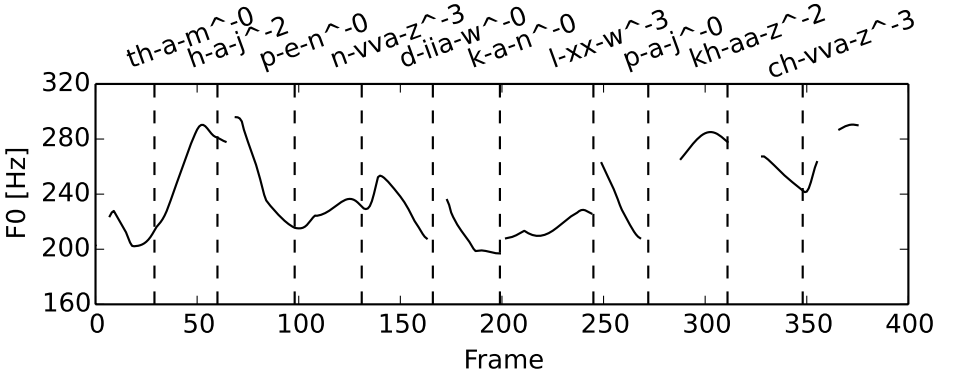
Figure 6.4 shows an example of F0 contours and syllable durations of original, baseline, manual and unsupervised labeling. It can be seen that the results for manual and unsupervised labeling are closer to the original than the baseline, especially at the seventh syllable (l-xx-w^3). In this example, the baseline produced ambiguous meaning because the incorrectness of the seventh syllable affects the meaning of the sentence. By giving the stress context onto the seventh syllable, the synthesized speech can have a unique meaning.

6.3.4 Subjective evaluation

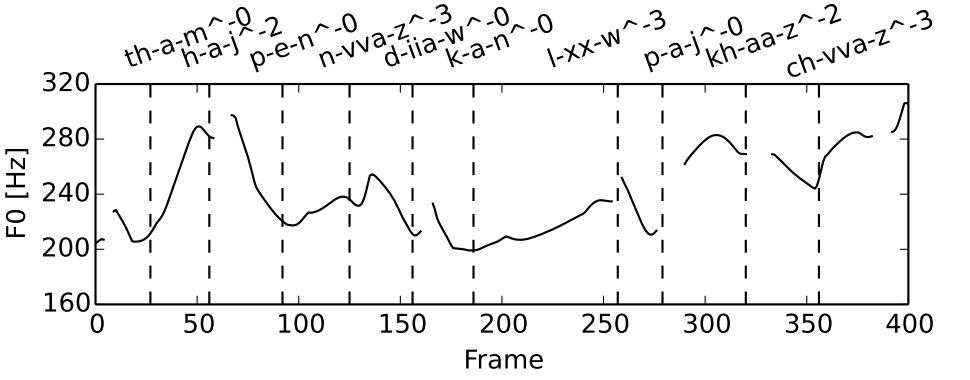
The subjective evaluation was performed by mean opinion score (MOS) and forced-choice preference tests. Ten Thai-native speakers participated the evaluation. Ten synthetic speech samples were randomly selected from the test set of the objective evaluation. Speech samples were evaluated on a five-point scale corresponding to perception in the naturalness of synthetic speech. The definitions of scores were 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent. The participants could listen to the sample as many time as they required for ensuring in quality. Figure 6.5 shows the result of MOS test with 95% confidence interval. Synthetic speech with stress context received



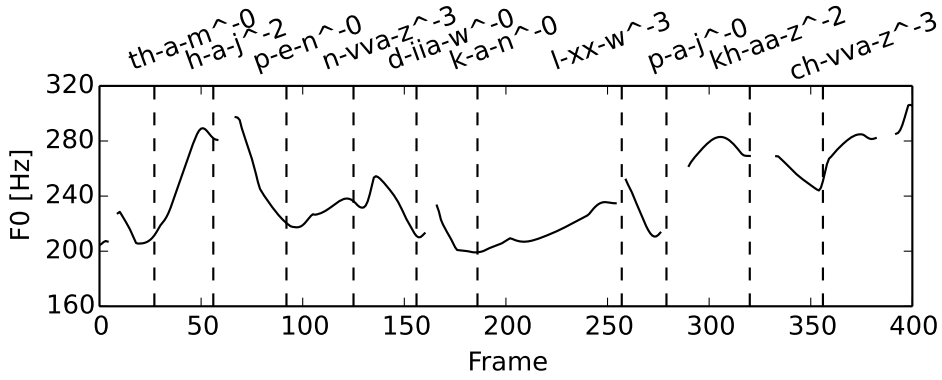
(a) Original



(b) Baseline



(c) Manual



(d) Unsupervised

Figure 6.4: Example of F0 contours and syllable duration compared with original. The sentence means "... bring mixed milk into sterilization ... ". The number suffixed to each syllable indicates its tone type.

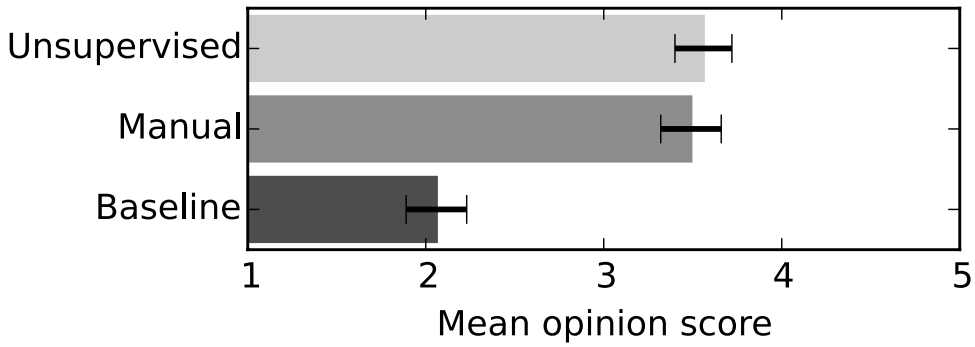


Figure 6.5: Result of mean opinion score test.

a higher score than the baseline. Moreover, there was no significant difference between the manual and unsupervised labeling.

In the forced-choice preference test, the participants were asked to choose more natural-sounding and clear meaning of speech from each pair of samples. The participants could repeat playback in the same way as MOS test. Figure 6.6 shows the result of the preference test. It can be seen that the proposed technique could achieve a higher score than the baseline. Additionally, there was no significant difference between manual and unsupervised labeling.

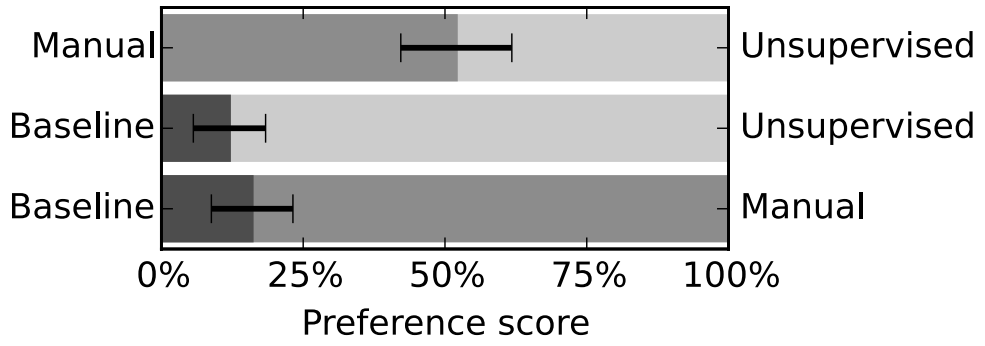


Figure 6.6: Result of forced choice preference test.

6.4 Conclusion

This chapter describes an unsupervised labeling technique for giving stress information to context set for speech synthesis. Stress is an essential factor which affects naturalness and meaning of an utterance. An unsupervised learning technique, called Bayesian Gaussian process latent variable model, was employed to obtain stress information automatically. The prosodic features of stress, log F0 contour and syllable duration, were used as observed variables of Bayesian GP-LVM training. Then a latent variable model training was performed to project the observed variable into a latent space in which can easily classify syllables into stressed and unstressed ones. The objective and subjective results showed that the proposed technique was comparable to the manual labeling and also outperformed the baseline. Future work will utilize the latent variable model into longer speech unit than syllable unit in which the prosodic feature is affected by various factors.

Chapter 7

Conclusions

7.1 Summary of the Thesis

This thesis described approaches to improve Thai prosody modeling.

Chapter 1 described the background of speech synthesis including the related topics and its application. This chapter described the problems of the conventional HMM-based SPSS and the GPR-based SPSS which was proposed to overcome the limitation of the HMM-based SPSS. Although the GPR-based SPSS had successfully alleviated the problems of HMM-based method, the prosody modeling was still imperfect. Therefore, the scope of study focused on improving the prosody modeling for the GPR-based SPSS. Chapter 2 explained the GPR-based SPSS and the implementation for Thai. This chapter showed a comparison of GPR-, HMM-, and DNN-based SPSSs.

Chapter 3 describes a multi-level model method to incorporate multiple GP models for duration prediction, called a two-stage method. In this technique, phone- and syllable- duration model are trained. The syllable duration is predicted, then used it as an additional context in training and prediction of phone-duration model. The two-stage method was applied into GPR- and DNN-based SPSSs. The result showed that the multi-level model outperformed the single-level model in GPR- and DNN-based SPSSs. Moreover, the multi-level GPR-based SPSS showed a better result than the multi-level HMM-based one.

Chapter 4 described an alternative technique for multi-level model GPR-

based duration prediction, called the product of Gaussian process experts. This technique individually trains phone- and syllable-duration models. When synthesizing the predictive distributions of the phone- and syllable-duration models are combined by the product of experts. Then, the mean of the product of Gaussian process experts is used as predicted phone duration. The product of GP expert method outperformed the single-level model in the objective and subjective evaluation, while had more accuracy than the two-stage method in syllable-duration distortion.

Chapter 5 described a multi-level model technique for GPR-based F0 generation. This technique separately trains the frame- and syllable-level models. The output variable of syllable-level GP is DCT-coefficients extracted from F0 contour in a syllable. The predictive distributions of the frame- and syllable-level models are jointly maximized for F0 generation. The multi-level model method could enhance the accuracy and naturalness of generated F0 contour, especially when the amount of training data was small.

Chapter 6 described an automatic stress annotation. This technique uses an unsupervised learning, called Bayesian Gaussian process latent variable model, to analyze the acoustic features of stress in latent space. By representing acoustic features in latent space, the similarity between syllables can be easily observed. Therefore, stressed/unstressed clustering was performed by using latent variables as the input. The experiments showed that the use of latent variable yielded higher accuracy than the observed variable in stressed/unstressed classification. Moreover, the use of stress information could improve performance in prosody generation.

7.2 Future Work

Future work will focus on applying the proposed techniques onto different speakers, and languages. Moreover, additional experiments will be conducted such as using multi-level prosody generation in both F0 and duration for subjective evaluation, and compares the result with multi-level DNN-based SPSS.

The conventional GPR-based SPSS uses on a single GP model in express-

ing the relation between contextual factor and acoustic features. Recently, various GP-based techniques have been proposed to improve the performance of GP-based modeling such as deep Gaussian processes [73], recurrent Gaussian processes [74], and chained Gaussian processes [75]. The future work will examine these variation of GP-based models for speech synthesis.

This thesis showed improvement of naturalness when stress information was incorporated into speech synthesis. This means that stress information should be considered in the future Thai-speech corpus. However, stress annotation is a preliminary method to alleviate the problem because the real problem is lack of syntactic structure in T-sync-1. In fact, Thai stress relies on syntactic structure of sentence, and T-sync-1 does not provide such information. For example, there is no difference between noun phrase and word in the corpus. This is the main reason of incorrect stress in the synthetic speech. In the future, annotation of syntactic structure must be conducted in order to improve the quality of prosody modeling.

Bibliography

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. EUROSPEECH, pp.2347–2350, 1999.

[2] J.J. Odell, "The use of context in large vocabulary speech recognition," University of Cambridge, 1995.

[3] T. Nose and T. Kobayashi, "Speaker-independent HMM-based voice conversion using adaptive quantization of the fundamental frequency," Speech Commun., vol.53, no.7, pp.973–985, 2011.

[4] Y. Ijima, T. Matsubara, T. Nose, and T. Kobayashi, "Speaking style adaptation for spontaneous speech recognition using multiple-regression HMM," Proc. INTERSPEECH, pp.552–555, 2009.

[5] T. Nose, M. Kanemoto, T. Koriyama, and T. Kobayashi, "HMM-based expressive singing voice synthesis with singing style control and robust pitch modeling," Computer Speech & Language, vol.34, no.1, pp.308–322, 2015.

[6] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," IEICE Trans. on Information and Systems, vol.90, no.9, pp.1406–1413, 2007.

[7] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," Proc. ICASSP, pp.7962–7966, 2013.

[8] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent

neural network with recurrent output layer for low-latency speech synthesis," Proc. ICASSP, pp.4470–4474, 2015.

[9] Z.H. Ling, S.Y. Kang, H. Zen, A. Senior, M. Schuster, X.J. Qian, H.M. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," IEEE Signal Processing Magazine, vol.32, no.3, pp.35–52, May 2015.

[10] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," Proc. ICASSP, pp.3844–3848, May 2014.

[11] T. Koriyama, T. Nose, and T. Kobayashi, "Frame-level acoustic modeling based on Gaussian process regression for statistical nonparametric speech synthesis," Proc. ICASSP, pp.8007–8011, 2013.

[12] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical parametric speech synthesis based on Gaussian process regression," IEEE J. Selected Topics in Signal Process., vol.8 (2), pp.173–183, 2014.

[13] T. Koriyama and T. Kobayashi, "Prosody generation using frame-based Gaussian process regression and classification for statistical parametric speech synthesis," Proc. ICASSP, pp.4929–4933, 2015.

[14] T. Koriyama, T. Nose, and T. Kobayashi, "Parametric speech synthesis based on Gaussian process regression using global variance and hyperparameter optimization," Proc. ICASSP, pp.3834–3838, 2014.

[15] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical nonparametric speech synthesis using sparse Gaussian processes," Proc. INTERSPEECH, pp.1072–1076, 2013.

[16] T. Koriyama and T. Kobayashi, "A comparison of speech synthesis systems based on GPR, HMM, and DNN with a small amount of training data," Proc. INTERSPEECH, pp.3496–3500, 2015.

[17] S. Chomphan and T. Kobayashi, "Implementation and evaluation of an HMM-based Thai speech synthesis system," Proc. INTERSPEECH, pp.2849–2852, 2007.

[18] S. Chomphan and T. Kobayashi, "Tone correctness improvement in speaker dependent HMM-based Thai speech synthesis," Speech Commun., vol.50 (5), pp.392–404, 2008.

[19] S. Chomphan and T. Kobayashi, "Tone correctness improvement in speaker-independent average-voice-based Thai speech synthesis," Speech Commun., vol.51, pp.330–343, 2009.

[20] A. Thangthai, N. Thatphithakkul, C. Wutiwiwatchai, A. Rugchatjaroen, and S. Saychum, "T-Tilt: a modified Tilt model for F0 analysis and synthesis in tonal languages," Proc. INTERSPEECH, pp.2270–2273, 2008.

[21] A. Thangthai, A. Rugchatjaroen, N. Thatphithakkul, A. Chotimongkol, and C. Wutiwiwatchai, "Optimization of T-Tilt F0 modeling," Proc. INTERSPEECH, pp.508–511, 2009.

[22] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," The Journal of the acoustical society of America, vol.107, no.3, pp.1697–1714, 2000.

[23] V. Chunwijitra, T. Nose, and T. Kobayashi, "A tone-modeling technique using a quantized F0 context to improve tone correctness in average-voice-based speech synthesis," Speech Commun., vol.54, pp.245–255, 2012.

[24] C. Wutiwiwatchai and S. Furui, "Thai speech processing technology: A review," Speech Commun., vol.49 (1), pp.8–27, 2007.

[25] S. Potisuk, J. Gandour, and M. Harper, "Acoustic correlates of stress in Thai," Phonetica, vol.53, no.4, pp.200–220, 1996.

[26] P. Peyasantiwong, "Stress in Thai," Papers from a Conference on Thai Studies in Honor of William J. Gedney. Michigan Papers on South and Southeast Asia, Center for South and Southeast Asian Studies, University of Michigan, Ann Arbor, pp.19–39, 1986.

[27] S. Hiranburana, "Changes in the pitch contours of unaccented syllables in spoken Thai," Tai Phonetics and Phonology, pp.23–27, 1972.

[28] S. Luksaneeyanawin, "Intonation in Thai," University of Edinburgh, 1983.

[29] J. Gandour, A. Tumtavitikul, and N. Satthamnuwong, "Effects of speaking rate on Thai tones," Phonetica, vol.56, pp.123–134, 1999.

[30] J. Gandour, S. Potisuk, S. Dechongkit, and S. Ponglorpisit, "Tonal coarticulation in Thai disyllabic utterances: a preliminary study," Linguistics of the Tibeto-Burman Area, vol.15, pp.93–110, 1992.

[31] S. Potisuk, J. Gandour, and M.P. Harper, "Vowel length and stress in Thai," Acta Linguistica Hafniensia, vol.30 (1), pp.39–62, 1998.

[32] C. Hansakunbuntheung, A. Rugchatjaroen, and C. Wutiwiwatchai, "Space reduction of speech corpus based on quality perception for unit selection speech synthesis," Proc. of the 6th Symposium on Natural Language Processing (SNLP), pp.127–132, 2005.

[33] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun., vol.27 (3-4), pp.187–207, 1999.

[34] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," IEICE Trans. on Information and Systems, vol.E90-D (5), pp.825–834, 2007.

[35] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," The Journal of the Acoustical Society of Japan (E), vol.21, no.2, pp.79–86, 2000.

[36] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis.," Proc. ICSLP, pp.29–32, 1998.

[37] N. Iwahashi and Y. Sagisaka, "Statistical modelling of speech segment duration by constrained tree regression," IEICE Trans. on Information and Systems, vol.83, no.7, pp.1550–1559, 2000.

[38] J. Yamagishi, H. Kawai, and T. Kobayashi, "Phone duration modeling using gradient tree boosting," Speech Commun., vol.50, no.5, pp.405–415, 2008.

[39] Y. Qian, Z. Wu, B. Gao, and F.K. Soong, "Improved prosody generation by maximizing joint probability of state and longer units," IEEE Trans. on Audio, Speech, and Language Processing, vol.19, no.6, pp.1702–1710, 2011.

[40] J.P. Teixeira and D. Freitas, "Segmental durations predicted with a neural network," Proc. EUROSPEECH/INTERSPEECH, pp.169–172, ISCA, 2003.

[41] P. Nagy and G. Németh, "DNN-based duration modeling for synthesizing short sentences," Proc. of International Conference on Speech and Computer (SPECOM), pp.254–261, Springer, 2016.

[42] G.E. Henter, S. Ronanki, O. Watts, M. Wester, Z. Wu, and S. King, "Robust TTS duration modelling using DNNs," Proc. ICASSP, pp.5130–5134, 2016.

[43] S.H. Chen, W.H. Lai, and Y.R. Wang, "A new duration modeling approach for Mandarin speech," IEEE Trans. on Speech and Audio Processing, vol.11, no.4, pp.308–320, 2003.

[44] A. Lazaridis, P.E. Honnet, and P.N. Garner, "SVR vs MLP for phone duration modelling in HMM-based speech synthesis," 2014.

[45] J.P. Van Santen, "Contextual effects on vowel duration," Speech Commun., vol.11, no.6, pp.513–546, 1992.

[46] O. Goubanova and S. King, "Bayesian networks for phone duration prediction," Speech Commun., vol.50, no.4, pp.301–311, 2008.

[47] W.N. Campbell, "Syllable-based segmental duration," in Talking Machines: Theories, Models, and Designs, Elsevier, North-Holland, Amsterdam, pp.211–224, 1992.

[48] W.N. Campbell and S.D. Isard, "Segment durations in a syllable frame," Journal of Phonetics, vol.19, no.1, pp.37–47, 1991.

[49] K.S. Rao and B. Yegnanarayana, "Modeling durations of syllables using neural networks," Computer Speech & Language, vol.21, no.2, pp.282–295, 2007.

[50] I. Sainz, D. Erro, E. Navas, and I. Hernáez, "A hybrid TTS approach for prosody and acoustic modules," Proc. INTERSPEECH, pp.333–336, 2011.

[51] A. Lazaridis, I. Mporas, T. Ganchev, G. Kokkinakis, and N. Fakotakis, "Improving phone duration modelling using support vector regression fusion," Speech Commun., vol.53 (1), no.1, pp.85–97, 2011.

[52] A. Lazaridis, T. Ganchev, I. Mporas, E. Dermatas, and N. Fakotakis, "Two-stage phone duration modelling with feature construction and feature vector extension for the needs of speech synthesis," Computer Speech & Language, vol.26, no.4, pp.274–292, 2012.

[53] Y. Wang, M. Yang, Z. Wen, and J. Tao, "Combining extreme learning machine and decision tree for duration prediction in HMM based speech synthesis," Proc. INTERSPEECH, pp.2197–2201, 2015.

[54] B. Gao, Y. Qian, Z. Wu, and F.K. Soong, "Duration refinement by jointly optimizing state and longer unit likelihood," Proc, INTERSPEECH, pp.2266–2269, 2008.

[55] S.H. Chen, C.H. Hsieh, C.Y. Chiang, H.C. Hsiao, Y.R. Wang, Y.F. Liao, and H.M. Yu, "Modeling of speaking rate influences on Mandarin speech prosody and its application to speaking rate-controlled TTS," IEEE/ACM Trans. on, Audio, Speech, and Lang. Process., vol.22, pp.1158–1171, July 2014.

[56] H. Zen, M.J. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," IEEE Trans. on Audio, Speech, and Language Processing, vol.20, no.3, pp.794–805, 2012.

[57] N.C. Pilkington, H. Zen, and M.J. Gales, "Gaussian process experts for voice conversion," Proc. INTERSPEECH, pp.2761–2764, 2011.

[58] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," Computer Speech & Language, vol.21, pp.153–173, 2007.

[59] J. Latorre and M. Akamine, "Multilevel parametric-base F0 model for speech synthesis.," Proc. INTERSPEECH, pp.2274–2277, 2008.

[60] M.S. Ribeiro, O. Watts, and J. Yamagishi, "Syllable-level representations of suprasegmental features for DNN-based text-to-speech synthesis," Proc. INTERSPEECH, pp.3186–3190, 2016.

[61] X. Yin, M. Lei, Y. Qian, F.K. Soong, L. He, Z.H. Ling, and L.R. Dai, "Modeling F0 trajectories in hierarchically structured deep neural networks," Speech Commun., vol.76, pp.82–92, 2016.

[62] D. Moungsri, T. Koriyama, and T. Kobayashi, "Duration prediction using multi-level model for GPR-based speech synthesis," Proc. INTERSPEECH, pp.1591–1595, 2015.

[63] D. Moungsri, T. Koriyama, and T. Kobayashi, "HMM-based Thai speech synthesis using unsupervised stress context labeling," Proc. APSIPA ASC, 2014. http://www.apsipa.org/proceedings.htm.

[64] D. Moungsri, T. Koriyama, T. Nose, and T. Kobayashi, "Tone modeling using stress information for HMM-based Thai speech synthesis," Proc. Speech Prosody 7, pp.1057–1061, May 2014.

[65] S. Chomphan and T. Kobayashi, "Design of tree-based context clustering for an HMM-based Thai speech synthesis system," Proc. of Sixth ISCA Workshop on Speech Synthesis (SSW6), pp.160–165, 2007.

[66] O. Krityakien, K. Hirose, and N. Minematsu, "Generation of fundamental frequency contours for Thai speech synthesis using tone nucleus model," Proc. INTERSPEECH, pp.1037–1041, 2013.

[67] J. Chaiwongsai and Y. Miyanaga, "Improved tone model for low complexity tone recognition," Proc. SICE Annual Conference (SICE), pp.1124–1129, 2014.

[68] N. Thubthong, B. Kijsirikul, and S. Luksaneeyanawin, "Stress and tone recognition of polysyllabic words in Thai speech," Proc. Int. Conf. Intelligent Technologies, pp.356–364, 2001.

[69] D. Moungsri, T. Koriyama, and T. Kobayashi, "Tone modeling using Gaussian process latent variable model for statistical speech synthesis," Proc. Speech Prosody 8, pp.1014–1018, 2016. http://www.isca-speech.org/archive/sp2016/.

[70] A.S. Abramson, "Lexical tone and sentence prosody in Thai," International Congr. Phonetics Science, pp.380–387, 1979.

[71] M.K. Titsias and N.D. Lawrence, "Bayesian Gaussian process latent variable model," Proc. International Conference on Artificial Intelligence and Statistics, pp.844–851, 2010.

[72] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," Proc. International Conference on Knowledge Discovery and Data Mining (KDD), pp.226–231, 1996.

[73] A. Damianou and N. Lawrence, "Deep Gaussian processes," Artificial Intelligence and Statistics, pp.207–215, 2013.

[74] C.L.C. Mattos, Z. Dai, A. Damianou, J. Forth, G.A. Barreto, and N.D. Lawrence, "Recurrent Gaussian processes," arXiv preprint arXiv:1511.06644, 2015.

[75] A.D. Saul, J. Hensman, A. Vehtari, and N.D. Lawrence, "Chained Gaussian processes," Artificial Intelligence and Statistics, pp.1431–1440, 2016.

List of Publications

Publications Related to This Thesis

Journal paper

1. Decha Moungsri, Tomoki Koriyama, Takao Kobayashi,
   "GPR-based Thai speech synthesis using multi-level duration prediction,"
   Speech Communication, Vol 99, pp.114-123 (2018.05).

International Conference

1. Decha Moungsri, Tomoki Koriyama, Takao Kobayashi,
   "Enhanced F0 generation for GPR-based speech synthesis considering syllable-based prosodic features,"
   Proc. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017, Paper ID:47, (2017.12).

2. Decha Moungsri, Tomoki Koriyama, Takao Kobayashi,
   "Duration prediction using multiple Gaussian process experts for GPR-based speech synthesis,"
   Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, pp.5945-5948 (2017.03).

3. Decha Moungsri, Tomoki Koriyama, Takao Kobayashi,
   "Unsupervised stress information labeling using Gaussian process latent variable model for statistical speech synthesis,"

Proc. 17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016, pp.1591-1595 (2016.09).

4. Decha Moungsri, Tomoki Koriyama, Takao Kobayashi,
"Tone modeling using Gaussian process latent variable model for statistical speech synthesis,"
Proc. 8th Speech Prosody Conference, Speech Prosody 2016, pp.1014-1018 (2016.06).

5. Decha Moungsri, Tomoki Koriyama, Takao Kobayashi,
"Duration prediction using multi-level model for GPR-based speech synthesis,"
Proc. 16th Annual Conference of the International Speech Communication Association, INTERSPEECH 2015, pp.1591-1595 (2015.09).

6. Decha Moungsri, Tomoki Koriyama, Takao Kobayashi,
"HMM-based Thai speech synthesis using unsupervised stress context labeling,"
Proc. 2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2014, Paper ID:1138, (2014.12).

7. Decha Moungsri, Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"Tone modeling using stress information for HMM-based Thai speech synthesis,"
Proc. 7th Speech Prosody Conference, Speech Prosody 2014, pp.1057–1061 (2014.05).