Sirindhorn International Institute of Technology
Thammasat University

Thesis ICTES-MS-2009-04

SPEECH SYNTHESIZER FOR THAI TEXT-TO-SPEECH (TTS) SYSTEM ON EMBEDDED DEVICE

Konlakorn Wongpatikaseree

SPEECH SYNTHESIZER FOR THAI TEXT-TO-SPEECH (TTS) SYSTEM ON EMBEDDED DEVICE

A Thesis Presented

by

Konlakorn Wongpatikaseree

Master of Engineering

Information and Communication Technology for Embedded Systems

(ICTES) Program

Sirinhorn International Institute of Technology

Thammasat University

May 2010

SPEECH SYNTHESIZER FOR
THAI TEXT-TO-SPEECH (TTS) SYSTEM ON EMBEDDED DEVICE

A Thesis Presented

by

Konlakorn Wongpatikaseree

Submitted to

Sirindhorn International Institute of Technology

Thammasat University

In partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING

Approved as to style and content by the Thesis Committee:

Advisor and Chairperson of Thesis Committee _____

(Asst. Prof. Cholwich Nattee, Ph.D)

Committee Member and
Chairperson of Examination Committee _____

(Ananlada Chotimongkol, Ph.D)

Committee Member _____

(Assoc. Prof. Thanaruk Theeramunkong, Ph.D)

Committee Member _____

(Prof. Takao Kobayashi, Ph.D)

External Examiner : Prof. Manabu Okumura

May 2010

Acknowledgement

I would like to thank my advisor, Asst. Prof. Cholwich Nattee, and co-advisor, Dr. Ananlada Chotimongkol, Assoc. Prof. Thanaruk Theeramunkong, Prof. Takao Kobayashi, for their personal encouragement as well as their academic advice.

I was so lucky to have Cholwich Nattee as my advisor. He encourage and advice the idea from the beginning to the end of my dissertation. When I had an unclear, immature idea, he helped me to grow it into a full-fledged research idea. He also gave me valuable feedback on my English written both in publication and this thesis. Without his help, I would not have been able to finish my study.

I also thank Dr. Ananlada Chotimongkol. It was always a great pleasure to work with Ananlada Chotimongkol. She was always helped me to find the answers to problems on my own by asking inductive questions of me. She helped me to realize how fun it is to do scientific research on linguistics. She also gave a good suggestion and comment on my work and English written both in publication and this thesis. Without her help, I would not have been able to finish this dissertation.

I am grateful to Assoc. Prof. Thanaruk Theeramunkong, for his suggestion on my work. He advises me to solve the problem not only in my work but also in my education. He gave the great opportunity to send me to do the pre-doctor at Japan Advance Institute of Science and Technology (JAIST) around one month. He was also suggesting me to study the PhD at JAIST.

I would also like to thank Prof. Takao Kobayashi. Although he was not stay at Thailand and I cannot discuss with him as face-to-face, he also gave the grate suggestion and comment on my work when I have a progress presentation with him.

I was very thanking to Dr. Denduang Pradubsuwan who was a senior project advisor in Bachelor degree. He gave a chance to me to get a scholarship in Master degree. He was always helped me from the first until now. He also advises me to meet with Cholwich Nattee.

I would like to thank the member of Human Language Technology (HLT) Laboratory, especially Mr. Ausdang Thangthai who helps me a lot in the first experiment. Since I do not have experience in linguistics, they help me a lot to teach me how to do research on linguistics fields. Working with them helped me to discover the fun and joy of learning in this field.

This research is financially supported by Thailand Advanced Institute of Science and Technology – Tokyo Institute of Technology (TAIST-Tokyo Tech), National Science and Technology Development Agency (NSTDA) Tokyo Institute of Technology (Tokyo Tech) and Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU).

It was fortunate for me to have excellent friends in TAIST-Tokyo Tech. There have a good time during taken course works one year in National Electronics and Computer Technology Center (NECTEC). They spent time for doing a sub-project together and playing a game together. They made one year is very short for me.

I would like to thank my family for their help. They encourage me always when I have a problem from the work. Although they do not know how to solve the problem in technical term, they can make me feel good with their love.

Last but not least, I would like to thank my special person, Ms Arunee Ratikan. She was a partner with me since the Bachelor degree until Master degree. She was a lot of helped me in second experiment. She was a speaker in our new speech corpus. She also encourages and supports me not only in research but also in my life. She made me was not stand alone when have problems.

Abstract

Speech Synthesizer for
Thai Text-to-speech (TTS) system on embedded device

May 2010

by

Konlakorn Wongpatikaseree

B.Eng, Sirindhorn International Institute of Technology
Thammasat University, 2007

This thesis presents the development of a speech synthesizer for Thai Text-To-Speech (TTS) system on embedded devices. Although the Thai TTS systems have been implemented on personal computers and the quality of most systems are acceptable, not so many systems have been developed on embedded or mobile devices, such as mobile phone. In this research, we focus on the practical aspects of the Thai TTS systems on a mobile device including application size and processing time. We aim at developing a Thai speech synthesizer that produce an output speech in real-time on the mobile device. The proposed synthesizer, Flite_Thai, is based on a unit concatenative synthesis technique, and the Flite software which is a speech synthesis engine designed for the resource-limited devices. To evaluate the proposed system, we conduct an experiment to compare the computational time, and the naturalness of the output speech. The experimental result shows that the system worked in a real-time manner. However, the quality of the generated speech is still lower than the standard level.

Hence, we propose an approach to improve the quality of the generated speech by designing a new speech corpus using a new speech unit. Non-sense carrier sentence technique is used in this design of the new speech corpus. Co-articulation affect is concerned in carrier sentence technique. Each of the carrier sentences contains a sequence of speech units without concerning the meaning. In addition, we propose a new speech unit called hybrid diphone, which is extracted from the new speech corpus. This new speech unit combines the advantage of diphone and demi-syllable. It aims to reduce the storage usage size and improve the quality. From the experiment, we found that the Mean Opinion Score (MOS) values of all the speech units that came from the non-sense carrier sentences are better than the MOS values that came from the natural sentences in the existing speech corpus, TSynC. However, among the three unit types in the news corpus, demi-syllable obtained the highest score. Although, the proposed hybrid diphone obtained higher MOS than the existing system and the diphone, it still suffers from a similar problem which is unsmooth joints between units. Some smoothing techniques are required in the future work to improve the quality of the speech generated using the hybrid diphone unit.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of Thai Text-To-Speech System

Text-To-Speech (TTS) is a software component that transforms natural language text into human speech. It is used in many ways, for example, a TTS application can be used to read messages or text on a screen while user is driving. The TTS application is also helpful for users with visual disabilities. TTS is also considered a part of assistive technology. TTS system is developed in several languages, and in this thesis we will focus on Thai TTS system.

Normally, Thai Text-To-Speech system consists of three main parts: text analysis, speech synthesis, and prosodic analysis as shown in Figure 1.1.

1. **Text analysis**

   Text analysis analyze the input text or sequence of words in order to separate and convert text into a sequence of pronunciation representative. Thai text analysis, however, is a complicated task because Thai language is different from other languages. There are tones that make a Thai language hard to understand. Moreover, Thai language does not have space between words. Thus, we have two modules to solve those problems.

   - **Word segmentation**

     Word segmentation is added in text analysis in order to separate the input text as paragraph or sentence into a sequence of words. Nevertheless, the challenge of this part is how to know which one is word. Thai language is different from western languages. For example, a sentence "ฉันกินข้าว" in Thai, which means "I eat rice", composes three words i.e. "ฉัน" (I), "กิน" (eat), and "ข้าว" (rice). It can be seen that there is no explicit marker to discriminate there words. This is similar to writing "Ieatrice" in English. So, we need to have algorithms to solve problem word segmentation such as longest matching [1], machine learning [2], and rules-based [3].

- **Grapheme-to-Phoneme conversion**

  Grapheme-to-Phoneme (G2P) [4, 5] is also another main component in this part. It is used to find the corresponding pronunciation in each word. Generally, the pronunciation is represented in a sequence of phonemes and tonal markers, for instance the pronunciation of "ฉันกินข้าว" is "ch-a-n^-4 | k-i-n^-0 | kh-aa-w^-0". Instance, the number represents the tone in Thai language.

2. **Speech synthesis**

This module is used to generate the human sound. There are several techniques to create sound. The well-known techniques for synthesizing sound are concatenative and HMM-based techniques. More details are explained in Section 2.3. For example, output sound in concatenative technique is succeeded by selected the appropriate units from the speech database. Thus, the quality of output sound is depending on speech corpus. Recording prompt in speech corpus is another point that makes a speech sound clearly. So, developer needs to design recording prompt carefully.

3. Prosodic analysis

Prosodic analysis [6] is an alternative way to make sound more natural. In this part is used to determine the prosodic features [7] such as rhythm, pitch, amplitude, in order to improve the quality of the generated speech. There are two main tasks in this part; syllabic duration analysis and Fundamental Frequency (F0). It is used to calculate duration time of each syllable and the ending between phrase or sentence, and is also used to estimate a frequency of tone and try to force a sound smoothly.

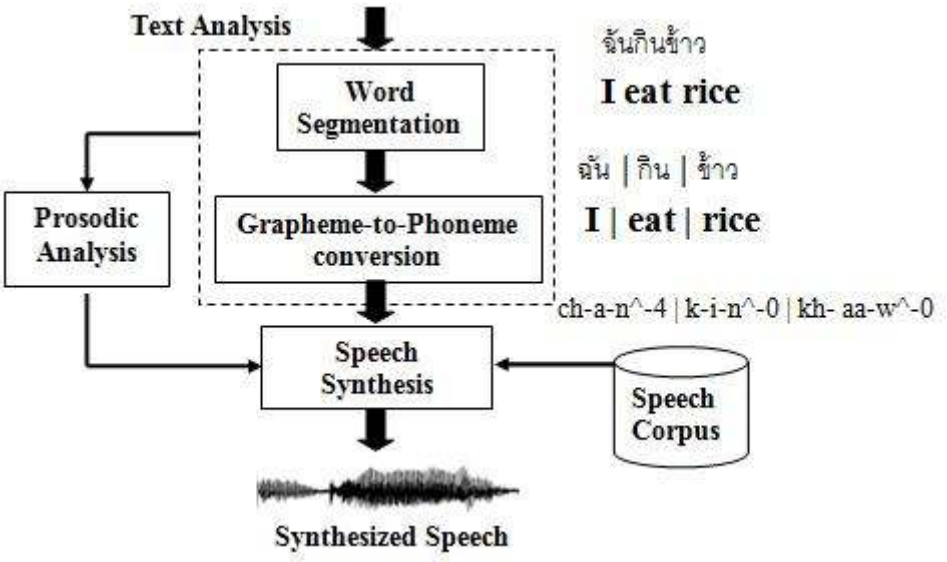

Figure 1.1 Overall process of Thai Text-To-Speech system in concatenative synthesis technique

1.2 The Problem of Thai Text-To-Speech System on Mobile Devices

Until now, text-to-speech applications have been developed on many platforms, such as personal computer, electronic dictionary devices, and mobile devices. However, a few of them work with Thai language. Thai text-to-speech systems have been developed and work maturely on the PC platform for many years. The computational time of that is quite real time and quality of sound is acceptable. However, it is still under development on the resource-limited or embedded devices. The problem of porting TTS system to mobile devices is from the limited resources in both storage and processing power. Furthermore, Thai language is a tonal language which is different from others languages. Five tones in Thai pronunciation generate more variety of sound. The number of basic sounds in Thai language is also doubled when it is compared to English. These make the size of sound database become larger than others languages. Moreover, Thai language is complicated not only in the number of basic sounds but also in phonetic of language. Since Thai language is written with no space or punctuation as English. It will affect to speech synthesis part if we cannot separate word correctly.

Speech synthesis technique is another problem of TTS on mobile devices. Because there are many techniques to synthesize sound while each technique has its own advantage and disadvantage in different aspects, for example concatenative technique is a speech synthesis technique that spends less computational time but quality of sound is not natural, whereas quality of sound in HMM-based technique is natural but spends more computational time. So to choose the appropriate technique is also a task of this research.

1.3 Objective of Thesis

Most Thai text-to-speech systems on personal computers can synthesize sound in real time with acceptable quality. However, when porting the Thai TTS systems to limited-resource systems such as mobile devices, computational time has to be reduced and quality of sound is worse than before. So the goals of dissertation are

- To choose the appropriate speech synthesis technique for mobile devices.
- To develop a Thai TTS system on mobile devices that can run on real-time system.
- To improve the quality of sound this designs a new speech corpus and speech units.
- To choose and develop an appropriate speech corpus and speech units for a Thai TTS system on mobile device.

1.4 Overview of Thesis

Since there are many tasks to do in Thai TTS system as we explain in above. However, this thesis we focus only on synthesis part of embedded device. First we analyze the strength and weakness of each speech synthesis technique because it is very important to select the appropriate technique in order to synthesize sound on mobile devices. Although some techniques are good in Thai TTS on personal computer, it does not mean that it works in the environment with limited resources. We, then, choose a concatenative technique in order to reduce the computational time. Our synthesizer, called Flite_Thai, is based on Flite [30], an open source synthesis library developed by Carnegie Mellon University because Flite is suitable for a resource-limited device as it is both small and fast. However to make it work, we have to implement it on Festival first. Because Flite is derived from Festival, it has some parameters that generate from Festival. We first create and convert attributes and parameters in order to match a format of Festival application, described in Section 3.2.2. Next, we transform TTS based on Festival to Flite software which converts file into C code in order to reduce size and to make it easy to access. Then, we port Flite_Thai to mobile device based on Windows Mobile. Finally, we evaluate and compare software among Flite_Thai, unit concatenative synthesis, and pTalk [26], HMM synthesis, with two criteria. First, we evaluate the processing speed time by counting the synthesis time manually. Second, we use MOS, Mean Opinion Score, technique to measure the quality of sound.

In addition, we are also improve the quality of sound by selecting the appropriate speech units for Flite_Thai in order to improve its intelligibility. We design and propose new speech units, called hybrid diphone, that combines the advantage of diphone speech unit and demi-syllable speech units. Moreover, we also design new speech corpus which uses the concept of carrier sentence technique to make a speech sound clearly. Our carrier sentence contains a speech unit or a set of similar speech units per sentence without concerning the meaning. After that, we set up the recording condition of new corpus. After we obtained record wave files, we extract three speech units: diphone, demi-syllable and hybrid diphone from new speech corpus, and evaluate the quality of sound with the existing system.

1.5 Thesis Organization

This thesis starts from a background of Thai Text-to-speech system. Then, we discuss the problem of Thai TTS and show the objective and overview of this thesis. Chapter 2 describes the phonetic of Thai language and gives an account of type of speech synthesis technique. Then we give some examples of Text-to-speech applications both in Thai and English language. Chapter 3 reveals the experiment setup and process of this thesis. Chapter 4 shows the result of my experiment. We conclude the thesis and explain the future direction in Chapter5.

Chapter 2

Background and Literature Review

In this chapter, we discuss the background and literature review. First, we explain the speech production mechanism in order to show how human produces sound. Second, we illustrate the phonology of Thai language. We discuss on consonants, vowels and tones in Thai language. Third, we talk about speech synthesis techniques. There are two speech techniques i.e. concatenative synthesis technique and HTS technique. Fourth, we compare between those techniques. Then, we present the linguistic units of speech in Thai language. There are many type of speech units such as phoneme, diphone, demi-syllable. After that, we discuss about the technique to evaluate TTS software. Finally, we show some examples of TTS systems on mobile devices.

2.1 Speech Production Mechanism

Generally, speech communication is largely used for transmission of information from a person to a person. Speech convey linguistic information, the speaker's tone and the speaker's emotion. To make the system for generating the speech effectively, it is important to understand the basic facts of how humans use speech to communicate with each other.

Before developing speech synthesis system, the task is made much easier if we know and understands how humans generate speech. Figure 2.1 shows the speech production mechanism. The speech production process involves three sub processes: source generation, articulation, and radiation. The main organs of the human body responding for producing speech are the lungs, larynx, pharynx, nose and various parts of the mouth. While the concept of producing the sound is when the diaphragm up, air is pushed up and out from the lungs, with the airflow passing through the trachea and glottis into the larynx. Then sound occurs when vocal cords are vibrated by airflow. After that, physiological structures are used to adjust that sound to each speech that human want to say.

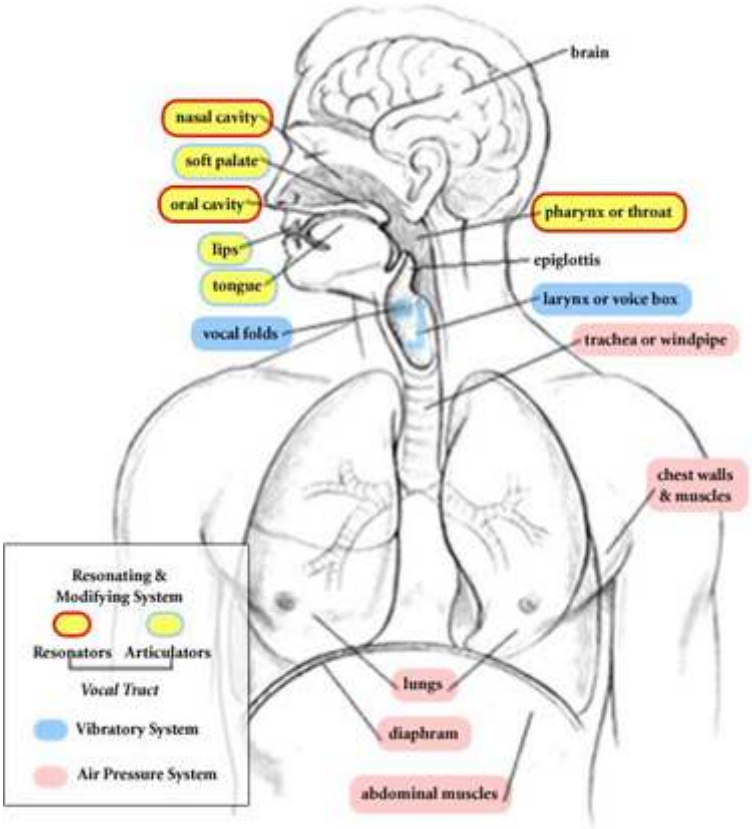

Figure 2.1 Speech production mechanisms [8]

Phonology in Thai Language

In Thai language, we derive the characteristics of the Thai language as follow:

2.1.1 Consonant

A. Consonant pronunciation system [9]

Consonants is occurred by vocal cord vibration and adjusted by structure of the mouth. However, we can classify the consonant along three aspects.

I. Voice or Voiceless

Voice consonants are occurred when they are accompanied by vocal cord vibration, and those which are not accompanied by the vibration are called unvoiced consonants.

II. Manner of Articulation

Manner of articulation is used to explain the control of airflow because path of airflow can make the different sounds. There are various types of manner of articulation such as nasal sound, fricative sound, or lateral sound.

- Nasals

Nasal sounds are produced by a complete closure in the oral cavity, and the air passes instead through the nose, for example are the initial consonant in the words ‘นาย’, ‘มี’, and ‘เงิน’.

- Fricatives

Fricative sounds are results of turbulence and noise airflow, as in the initial consonants of words ‘ฟ้า’, ‘สาม’, and ‘ห้าม’.

- Trills

Trill sounds are produced by vibrations between the articulator and the place of articulation. In Thai, we have only one consonant, ‘ร’ that produced a trill sound. For instant are the initial consonant in the words ‘รัก’, ‘รวย’, and ‘รีบ’.

- Laterals

Lateral sounds are produced by raising the tip of the tongue against the roof of the mouth so that the airstream flows past one or both sides of the tongue such as the consonants in the words ‘ลิง’, ‘ล่าม’, and ‘ลูก’.

- Approximants

Approximants sounds are a sound that is produced by bringing one articulator in the vocal tract close to another without, illustrated by the initial consonants of ‘วัน’, ‘ยาม’, and ‘ยก’.

III. Place of Articulation

Source of sound in the mouth is also can classify the consonant into each group. Place of articulation is the point of major closure in the vocal tract during its articulation.

- Labial

Labial consonants are produced by one or both lips. In some case, the lower lip approaches or closes against the upper teeth, such as in the initial consonants of words are ‘ปลา’, ‘พาน’, and ‘บ้าน’.

- Alveolar

Alveolar consonants are articulated by tongue tip approaches or closes against with upper teeth, as in consonants in the words ‘ตัว’, ‘เด็ก’, and ‘น้ำ’.

- Palatal

Palatal consonants are articulated by tongue tip or closes against with hard palate, for example in the initial consonants of words ‘จาน’, ‘ช้าง’, and ‘ยุง’.

- Velar

Velar consonant are produced by the back of the tongue closes against with the soft palate or velum as the initial consonants of words ‘ไก่’, ‘งาน’, and ‘เขียน’.

- Glottal

Glottal consonant are articulated with the glottis. For example are in the initial consonants of words ‘อ่าน’, ‘หัว’, ‘ห้อง’.

B. Type of consonant

Consonant in Thai language is also separated into three parts by position in syllable.

I. Initial Consonant

Position of initial consonant is placed in front of vowel while the numbers of those are 21 units as show in table 2.1.

II. Cluster Consonant

Position of cluster consonant is same as the initial consonant, but it have two initial consonant that pronounced cluster. Table 2.2 shows the cluster consonants that consist of 12 units.

III. Final Consonant

Position of final consonant is placed behind the vowel. Actually, final consonant is replaced by initial consonant or cluster consonant. However, it classify by sound not by characteristic, for example, word "กัด" (bite) and "รัฐ" (stage) is a different final consonant. Although each of word is replaced by "ด" and "ฐ", in Thai there use a sound "t" replaced both of them. In Thai, there are eight groups to classify the final consonant as show in Table 2.3.

Table 2.1 Consonant sound and characteristic for Thai language [9].

Manner of Articulation		**Place of Articulation**				
		Labial	**Alveolar**	**Palatal**	**Velar**	**Glottal**
Stop	**Voiceless Unaspriated**	p ป	t ฏ ต	c จ	k ก	z อ
	Voiceless aspriated	ph พ ภ ผ	th ฐ ฑ ฒ ถ ท ธ	ch ฉ ช ฌ	kh ข ฃ ค ฅ ฆ	
	Voiced	b บ	d ฎ ด			
Non-Stop	**Nasal**	m ม	n ณ น		ng ง	
	Fricative	f ฟ ฝ	s ซ ศ ษ ส			h ห ฮ
	Trill		r ร			
	Lateral		l ล ฬ			
	Approximant	w ว		j ญ ย		

Table 2.2 Cluster consonant sound and characteristic for Thai language [9].

Unaspirated Stop Set	pr ปร	tr ตร	kr กร
	pl ปล		kl กล
			kw กว
Aspirated Stop Set	phr พร ผร	thr ทร	khr คร ขร
	phl พล ผล		khl คล ขล
			khw คว ขว

Table 2.3 Final consonant sound and characteristic for Thai language [9].

Manner of Articulation		**Place of Articulation**			
		Labial	**Alveolar**	**Palatal**	**Velar**
Stop		p บ ป พ ภ ฟ	t ด ฎ ต ฏ ท ธ ฉ ฑ ฒ ฐ ฐ จ ช ซ ศ ษ ส		k ก ต ฏ ข
Non-Stop	**Nasal**	m ม	n น ณ ร ล ฬ ญ		ng ง
	Approximant	w ว		j ย	

2.1.2 Vowel

Although, both vowels and consonants are represented by phoneme, there is not a very close correspondence in Thai language. The original sound is located at the larynx. Then sound is adjusted by the structure in mouth and tongue. Normally, amplitude of vowel sound is stronger than consonant sound. Vowel in Thai language is defined in 24 units as follow:

1. **Single vowel**

Concept of single vowel is position of tongue is stable, not change until the end of vowel. While the number of single vowel are consists of 18 units as show in Table2.4.

- Short vowel

Sound of this group is short pronounced which consists of nine units as /i/, /v/, /u/, /e/, /q/, /o/, /x/, /a/, and /@/

- Long vowel

Sound of this group is long pronounced which also consists of nine units as /ii/, /vv/, /uu/, /ee/, /qq/, /oo/, /xx/, /aa/ and /@@/

2. **Diphthong**

Table 2.5 shows six vowel phonemes that are formed by making a transition from one vowel to another in sequence. So the tongue advancement and tongue height are fairly changed more than single vowel. It however classified diphthong into two groups in the similar manner to the single vowel.

- Short diphthong

Short diphthongs are combined by two short vowels. The numbers of short diphthong are three units as /ia/, /ua/, and /va/.

- Long diphthong

Long diphthongs are combined between single short and long vowel. The numbers of long diphthong are three units as /iia/, /uua/, and /vva/

Table 2.4 Single vowel with short vowel and long vowel [9].

Type	Short vowel		Long vowel	
	Grapheme	**Phoneme**	**Grapheme**	**Phoneme**
Monophthong	–ะ	|a|	–า	|aa|
	–ิ	|i|	–ี	|ii|
	–ึ	|v|	–ื	|vv|
	–ุ	|u|	–ู	|uu|
	เ–ะ	|e|	เ–	|ee|
	แ–ะ	|x|	แ–	|xx|
	โ–ะ	|o|	โ–	|oo|
	เ–าะ	|@|	ออ	|@@|
	เ–อะ	|q|	เ–อ	|qq|

Table 2.5 Diphthong vowel with short vowel and long vowel [9].

Type	Short vowel		Long vowel	
	Grapheme	**Phoneme**	**Grapheme**	**Phoneme**
Diphthong	เ–ียะ	|ia|	เ–ีย	|iia|
	เ–ือะ	|va|	เ–ือ	|vva|
	–ัวะ	|ua|	–ัว	|uua|

2.1.3 Tone

Thai is a tonal language [10] with five levels of tones i.e. middle, low, falling, high, and rising as shown in Table 2.6. It is very different from Western language. Although, a tone in Thai is making an intonation sound as accent in English, in term of meaning it is quite different. Meaning of word changes depending on the tone, even with the same initial consonant, vowel, and final consonant. For example, word 'มา', using middle tone, means 'come' in English, but word 'ม้า', using falling tone, means 'horse' in English. Tone cannot occur individual. Actually, both consonant and vowel are affected by tone, but amplitude of vowel sound is more stranger than consonant sound. Figure 2.2 shows the frequency of tone in woman sound. So we can ignore the tone in consonant sound. However, we can derive the tone in two parts based on the frequency as

- **Stable frequency**

Frequency of this case is stable in each syllable. There are three tones in this case: middle, low, and high

- **Varied frequency**

Frequency of this case is varied in one syllable. There are two tones in this case: falling and rising.

Table 2.6 Five levels of tones in Thai language [11].

Tone	Tone marker	Symbol in TTS system	Example	Example meaning in English
middle	-	0	นา	A paddy
low	-่	1	หน่า	(a nickname)
falling	้	2	หน้า	Face
high	-๊	3	น้า	Aunt/uncle (younger than one's parents)
rising	-๋	4	หนา	Thick

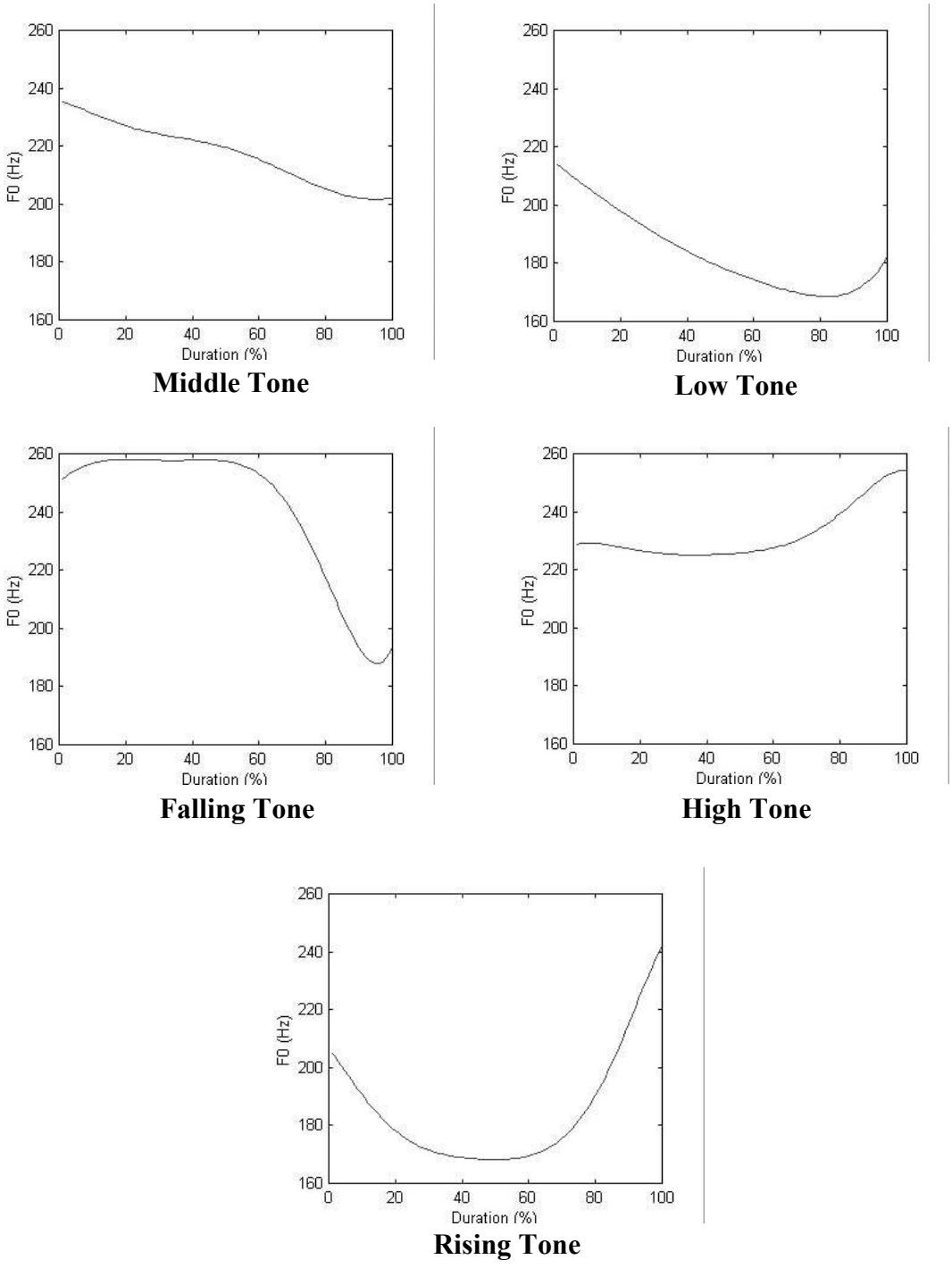

Figure 2.2 Frequency of tone in woman sound

2.2 Speech Synthesis Techniques

Speech synthesis is a transformation of the linguistic message represented in form of the phonetic symbols to an acoustic signal. In the part, speech synthesis technology was developed mainly in research. However, today speech synthesis has come to use widely. Speech synthesis is used in many ways depending on the task. For example, reading machine for blind people is an application used to generate a speech sound form message or text. Nevertheless, in order to get the good speech synthesis system, two well-known techniques have been proposed for synthesizing output speech that is similar to human speech. The first technique is concatenative synthesis, and the second is HMM-based synthesis.

2.2.1 Concatenative Synthesis

The concept of concatenative synthesis technique [12] is to connect several segments of pre-recorded speech at the required time to compose the messages. The sound of pre-recorded speech is converted from analog signal to digital signal in order to reduce size before collecting units in the database. The number of units and the size of the speech segments in a database largely affect the quality of the synthesized speech, and another problem of this technique comes from the joint positions between speech units. However, there are many ways to solve the quality of sound in this problem such as Line Spectrum Pair – LSP [13] and Time Domain Pitch Synchronous Overlap Add – TD-PSOLA [7]. Moreover, the computational time of concatenative synthesis is suitable for real-time speech generation. There are two main techniques of concatenative synthesis. Figure 2.3 shows the concept of concatenative synthesis.

2.2.1.1 Unit Concatenation

Unit concatenation is a technique that concatenates small fixed-size speech units together. The main consideration for this technique is the size of speech units. There are many types of the speech units which range from the level of phone to syllable. For example, individual phones can be split into two half-phones. This is called a diphone or they can be merged into multi-phone units as syllable. The advantage and disadvantage of each type is not same, for instant, although syllable speech units can generate a natural sound, the number and size of units in speech corpus is quite large. It does not work for a small device. On the other hand, even though the diphone speech unit produces a sound worst than sound of syllable speech unit, the size of speech corpus is quite smaller than that of syllable. So the consumption of each unit is different depending on task. Another point, this technique overcomes the limitation of word problem, because we can make a new word from the sub-word which without making a new recording. To make a new word with used a sub-word, we try to keep the transition region in each sub-word. The quality of sound is better if each unit can represent the transition between phones. However, quality of speech in concatenation of sub-word is a problem because there are many joints between speech units. Mostly, speech units that are commonly used are diphone and demi-syllable because for both of them the number of units required to cover one language is not too large. Moreover, the

quality of synthesized sound is acceptable. More detail of speech unit is discussed in section 2.5.

2.2.1.2 Corpus Unit Selection

Among the concatenative approaches, corpus unit selection or unit selection is a technique that can generate the most natural output speech. This technique uses a large speech database of pre-recorded utterances. For each utterance, the boundaries of various types of units, e.g. sentence, word, syllable, diphone and phone, are labeled. During the run time, a weighted decision tree is used to select the best chain of units from the database. This technique can be divided into two categories: general domain and limited domain. For the first category, the database is quite large since it has to cover phonetic and prosodic variations in a language. A larger database usually yields more natural synthesized speech. Moreover, the possible size of the database could grow up to gigabytes. The problem of this type is hard to add even a single new word without making a new recording when the message component is missing in the database. Limited domain synthesis is suitable for the case where possible sentences that have to be synthesized are limited. This could occur when a synthesizer is used in a specific domain such as weather forecast and talking clock. It is require a small corpus in order to cover all necessary patterns in the domain.

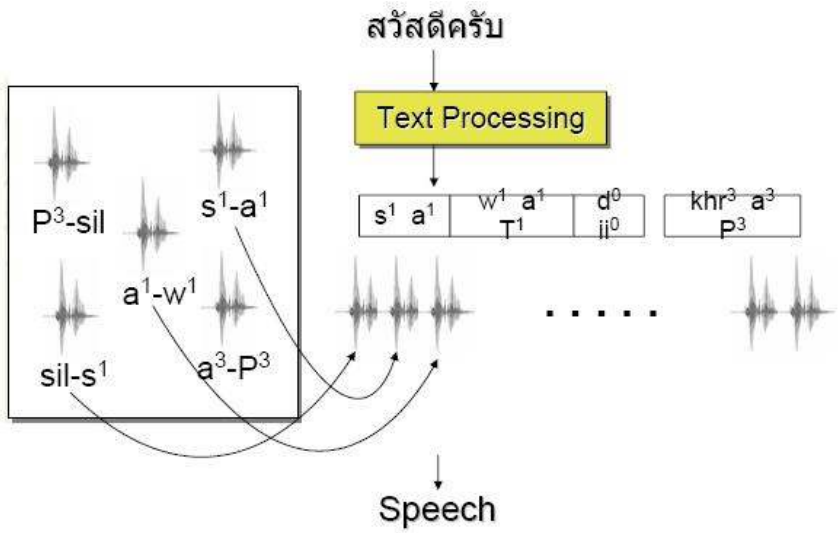

Figure 2.3 Concept of concatenative synthesis

2.2.2 HMM-based synthesis (HTS)[14]

Since, the quality of sound of corpus-based speech synthesis can generate natural-sounding high quality speech. However, it is very hard to improve the quality of sound in term of duration, intonation or speaking rate. For constructing human like talking machines, speech synthesis systems are required to have an ability of speech such as

various speaking styles, emotional expression etc. This technique consists of two parts: training part and synthesis part, as shown in Figure 2.4. In the training part, the excitation parameters and spectrum parameters are extracted from a speech database. Then, these parameters are modeled by a context-dependent HMM. In the synthesis part, the context-dependent HMMs which match the pronunciation of the input text are concatenated. Excitation and spectrum parameters generated by the HMMs are passed through the Mel Log Spectrum Approximation (MLSA) filter to generate a waveform. The quality of the speech synthesized by this technique is quite natural. It, however, requires more computational time than other techniques.

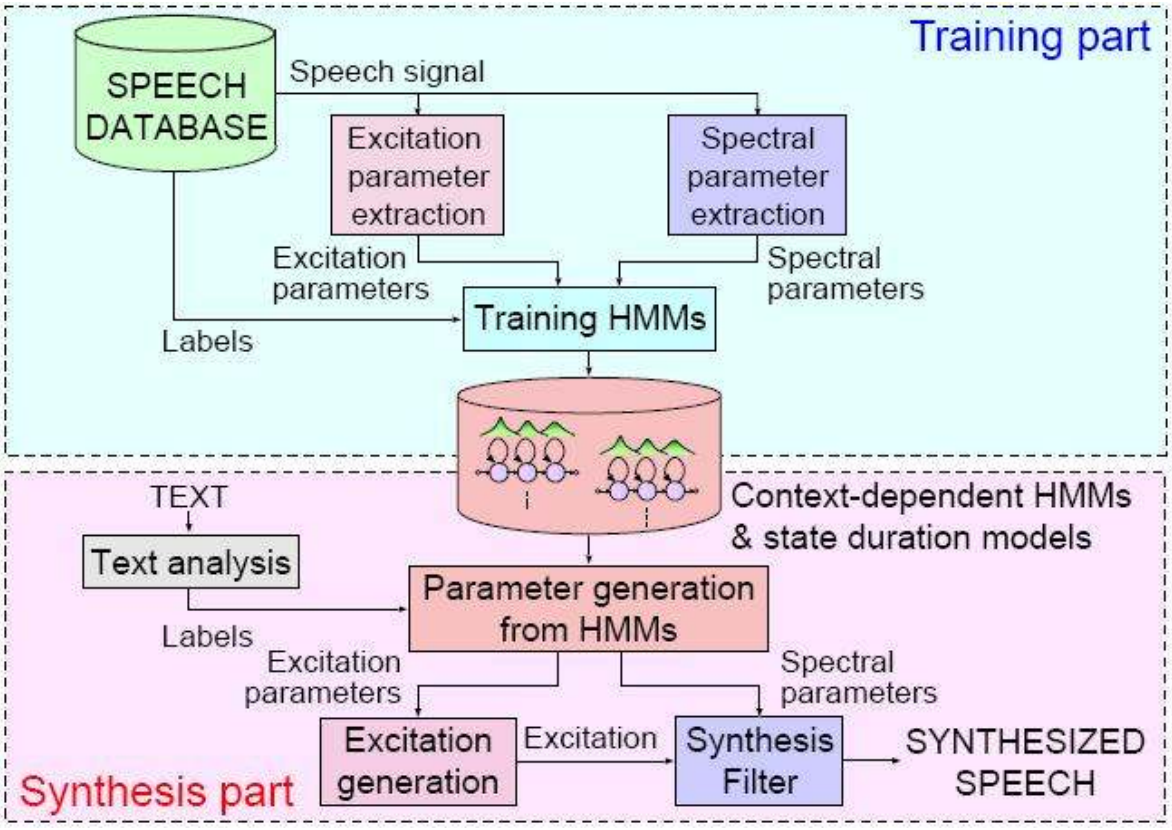

Figure 2.4 Process of HMM-based speech synthesis technique with training part and synthesis part. [15]

2.3 Comparison of Speech Synthesis Technique

Comparing between the two speech synthesis methods, a concatenative speech synthesizer, especially a unit selection synthesizer, usually has a bigger model than HTS. However, it requires less processing time than the HTS as the latter has to re-synthesize speech given a set of parameters. In terms of quality, both approaches produce output speech with comparable quality. Post-processing techniques can be applied to both approaches to make the synthesized speech sounds more natural and smooth.

Since unit concatenation, corpus unit selection and HTS have both advantage and disadvantage, we have to study each technique which one is suitable to use in this thesis.

First steps, we compare between the main speech synthesis technique, concatenative synthesis and HTS. Since, the thesis aims at the synthesizer on embedded devices. We select the concatenative synthesis technique. There are two reasons for this. The first reason is that size of speech corpus in HTS is a small size, whereas size of it in the concatenative synthesis is varied, fewer to larger. It depends on type of speech units in synthesizer. So size of speech corpus is not a main point because both of them can generate sound by a small database. The other reason is that computational time is spent a lot in HTS whereas Computational time in Concatenative synthesis is quite real-times system.

After we choose the Concatenative synthesis, we then compare the unit concatenation and Corpus unit selection technique. Unit concatenation technique is selected in order to make a speech synthesizer in this thesis. Because of, size of speech corpus, size of unit concatenation is smaller than corpus-based. Since unit concatenation is used a fix-sized speech unit that a small size units, diphone or demi-syllables inasmuch as corpus unit selection is use a large speech unit as sentence or word. So copus-based is not suitable for synthesizer on embedded system. Moreover, the Corpus-based is impossible to cover all of word in Thai language and it cannot make an unknown word when missing in the database but in Unit concatenation is not. It can make an unknown word by concat between sub-words.

From the recent work, Tokuda et al. [14] compare the advantage and disadvantage of unit selection technique and HTS technique as show in Figure 2.5. They conclude that, the advantage of unit selection is High quality at waveform because they concatenate from original wave files, on the other hand output in HTS is generated by a parameter so it had buzzy in some point. However, in the disadvantage of unit selection is a joint position between units so concatenative point is not smooth when concatenating. In addition, word in each language is varied, so it is very hard if store all unit to cover all word in corpus. Thus hit or miss it depends on database, but in HTS the output file it will smooth than unit selection because they train and adjust the wave form so, it don't have a gap between unit.

Unit selection	HTS
Advantage: • High quality at waveform level Disadvantage: • Discontinuity • Hit or miss	Disadvantage: • Vocoded speech (buzzy) Advantage: • Smooth • Stable
• Large run-time data	• Small run-time data
• Fixed voice	• Various voices

Figure 2.5 The comparison between Unit selection and HTS [14].

2.4 Linguistic Units of Speech in Thai language

Since the choice of speech unit greatly affects the quality of the concatenative synthesis and the size of the required database, the speech unit is one of the main considerations especially for the unit concatenation technique. There are many types of speech units; each type has different properties. Table 2.7 summarizes the number of distinct units for each type of speech unit in Thai.

Table 2.7 The number of distinct unit for each type of speech unit in Thai.

Unit type	Number of Units
Phoneme (Thai phoneme only)	65 [16]
Diphone (include tone)	9,090 [17]
Demi-Syllable (include tone)	1,505 [18]
Syllable (include tone)	26,928 [17]

For a small unit such as phoneme, only a small number of distinct units are required to cover the whole language. Thus, the corresponding database is small. However, at run time, we have to concatenate many small units together in order to generate output speech. A large number of concatenating points make the synthesized speech less smooth. A large speech unit such as syllable, on the other hand, requires a large number of distinct units in order to cover one language. Thus, the corresponding database is large, but only a small number of units are required to generate a new utterance. Thus, the synthesized speech sounds more natural. Next, we are going to explain in each unit.

2.4.1 Phoneme

Phonemes [19] are the smallest units of sound for generating speech. The number of phonemes is depending on a language. In Thai, there are 65 phonemes: 32 initial consonants, 24 vowels, and 9 final consonants, for example, word "ครับ" (yes) is a word in Thai that consists of three phonemes as /khr/, /a/, /p/.

2.4.2 Diphone

Diphone [20] is an adjacent pair of phones. In concatenative speech synthesis, a is generally diphone unit defined from the middle of one phone to the middle of its adjacent phone as shown in Figure 2.6. But in Thai language, this is not enough because Thai is a tonal language. Thus, to represent tones in Thai diphone, each phone is attached with one of the five Thai tones. There are up to five tonal phones, shown in Table 2.6, for each Thai phone. In TsynC, Thai speech corpus designed by NECTEC, there are 89 phones, included Thai phoneme and foreign phoneme, and 326 tonal phones; hence, the number of tonal diphones in Thai becomes very large, namely 9,090 units, when compared with other languages such as English which has 1,936 units [20].

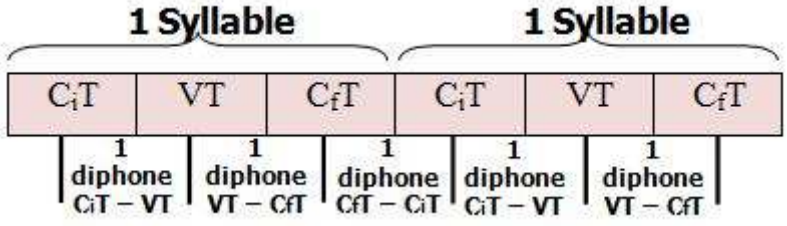

Figure 2.6 The structure of a diphone speech unit.

In Figure 2.6, "Ci" represents an initial consonant and "Cf" denotes a final consonant while vowel and tone are represented by the letter "V" and "T" respectively. A diphone unit is designed to capture the co-articulation between phones, so it keeps the boundary between adjacent phones. However, one drawback of this technique is discontinuity within a consonant. Since the speech signal of a consonant is non-periodic unlike that of a vowel, the conjugation of two halves of a consonant is not so smooth. Another drawback is a large size database since we have to add tones to every phone.

2.4.3 Demi-syllable

Demi-syllable [18, 21] is a sub-syllable unit, which consists of two parts: initial part and final part. Figure 2.7 shows the structure of a demi-syllable. A syllable is divided into two parts at the place in a vowel where the speech signal becomes a steady period. The initial part contains a single consonant or double consonants and a small part of a vowel. The final part contains the rest of the vowel and a final consonant.

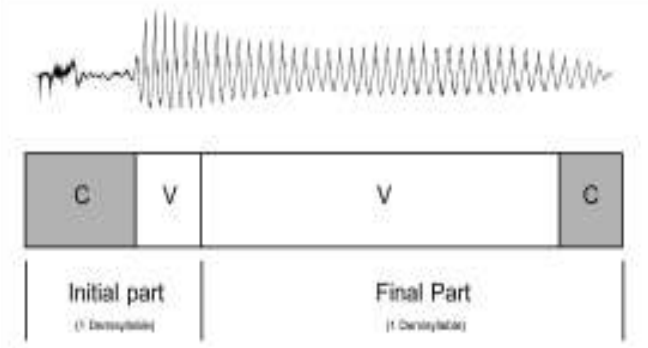

Figure 2.7 The structure of a demi-syllable speech unit.

The advantage of the demi-syllable unit is a small database. When we analyze where a tone actually occurs in a syllable using phonetic properties of phones, a tonal characteristic basically occurs only in vowels. Thus, we do not have to include a tone in a consonant. In the structure of the demi-syllable, a large part of a vowel is found in the final part; hence, we only have to include a tone in the final part not the initial part. As a result, the size of a speech database is decreased. In Thai, there are 1,505 demi-syllables. Furthermore, the quality of concatenated sound within a syllable is better than that of concatenated diphones because there is only one concatenating point and it occurs at the steady part of a vowel. A drawback of the demi-syllable unit is, however, discontinuity between syllables. Demi-syllable only stores co-articulation within a syllable. It does not keep co-articulation between syllables.

2.4.4 Syllable

Syllable is a sound that human pronounces in one time even it has meaning or not. So sometimes syllable, it is look like a word if that syllable has a meaning. Structure of Thai syllable, has at least three sounds associated: consonant, vowel, and tone followed as:

$$S = C_1(C_2)V^T(C_3)(C_4)$$

where
C_1 is represented a initial consonant
C_1C_2 is represented a cluster consonant
V is denoted a vowel
C_3 is denoted a final consonant
C_3C_4 is denoted a final cluster consonant
T is denoted a tone
() is an optional symbol.

2.5 Technique to evaluate TTS system

There are two techniques that used to evaluate quality of sound in Text-To-Speech system.

2.5.1 Mean Opinion Score (MOS)

Mean Opinion Score [22] is a technique used to evaluate the quality of sound by tester. Before evaluating the sound, developer should tell the testers what are mention, such as the goal of system is intelligibility, testers are asked as “Do you know what the sound talks about?”, “Do you understand what the sound says.” Then, the testers have to evaluate the sound and mark the score. Score of evaluating is between 1 to5, as shown in Table 2.8. After that score from testers is averaged and becomes a MOS value.

Table 2.8 Score of evaluating in MOS technique.

Value	Quality
1	Bad
2	Poor
3	Fair
4	Good
5	Very good

2.5.2 Comparison Category Rating (CCR)

Comparison Category Rating [22] is a technique to compare a sound two system which one is better. The method is similar with a MOS, but meaning of score is different. Because this technique is used to compare 2 systems, score is separated in 2 sides. Positive value is mean, quality of sound in system 1 is better than system 2, on the other hand negative value is mean, quality of sound in system 2 is better than system 1 as shown in Table 2.9.

Table 2.9 Score of evaluating in CCR technique.

Value	Fondness
3	Much better
2	Better
1	Slightly better
0	Same
-1	Slightly worse
-2	Worse
-3	Much worse

2.6 Existing TTS systems on mobile devices

Early works on developing a TTS system on a mobile device focused mainly on migrating an existing TTS system from a resourceful platform to a resource-limited platform [23]. Most of the efforts were spent on code optimization and database compression. Since the space was quite limited, only a small diphone database could be utilized which reduced the quality of the synthesized speech. To improve output speech quality, some researcher attempted to apply a unit selection technique on resource limited devices. Tsiakoulis, et al. [24], used a statistical selection method to make a unit selection database small enough for an embedded device without much reduction in speech quality. Pucher, and Frohlich [25], on the other hand, used a large unit selection database but synthesized an output speech on a server and then transferred the wave form to a mobile device over a network.

A few research studies have been carried out to develop a Thai speech synthesizer on a mobile device. Torruangwathana [26] proposed a Thai user interface on mobile phone device for blinds, called pTalk, that can read messages and menus on mobile devices. In pTalk, a simple text analysis technique was utilized. When an input text is entered into the text analysis part, pTalk will do text normalization that deletes symbols from a sentence such as space, $, # etc. and then use longest matching as word segmentation. The output word will look up in dictionary. Finally, synthesizer based on Hidden Markov Model (HMM) is applied since it requires the size database less than the concatenation synthesis.

In 2007, Chinathimatmongkhon [27] proposed experiment between speech quality and complexity of the system for implements TTS on hand-held devices (PDA). For the speech synthesizer, HMM-based algorithm is selected instead of concatenation method. The computational time decreases from length of impulse response parameter and dimension of feature vector in experiment. The experiment on 564 Thai words from news website give accuracy 68% segment correct. 90% of time processing is spent for TTS process in step of synthesis filter. Average result of MOS score is range 3.5 to 4. So the computation of system can be reduced while the quality of sound is not decay. Another embedded Thai synthesizer was developed by Nokia [28]. This synthesizer is a concatenative synthesizer and is available for specific models of Nokia mobile phones such as E50, E65, and 6120 classic**.**

Chapter 3

Thai Synthesizer on Flite Text-To-Speech Engine

This chapter presents the design methods and the process of work in this thesis. First of all we discuss on the design methods. Since in last chapter we compare between the speech syntheses techniques completely. We are definitely that we use the unit concatenation technique to make a speech synthesizer on mobile devices then we explain the speech engine used in this thesis, Festival and Flite software. Section 3.2, we show the step of building Thai Text-To-Speech system on mobile devices. Moreover, we reveal a technique to improve the quality of sound. A new speech corpus is created by the carrier sentence technique. We design a new speech unit, called hybrid-diphone, which combine the diphone and demi-syllable speech units.

3.1 Design Methods

The goal of this thesis is to develop a Thai TTS system that can operate in real-time, so we compare each speech synthesis technique in order to find the appropriate technique to use as a synthesizer on embedded device. As the result, we found that the technique to reach the goal of real-time system is unit concatenation technique, discussed on Section 2.4.

We develop a Thai TTS based on Flite software, because it suitable for implementing a fast and small TTS application. Moreover, it supports developers who would like to develop TTS systems in other languages. Since, Flite or Festival-lite, is a small speech engine that derived from Festival software. So we have to learn both of speech engines in detail.

3.1.1 Festival Software

Festival [29] is an open-source software that is designed as a Text-To-Speech system run on personal computer or a server base on English language. It is a stable, run-time speech synthesis engine. Moreover, it supports to port to multilingual. Festival was developed by University of Edinburgh. The system was primarily developed under UNIX. Basically, there are two main parts as the general TTS system i.e. Text analysis and Speech synthesis.

In text analysis, the degree of difficulty of this part depends on the text type and language. In English, the text or sentence has space between words so the task is easier in term of word segmentation than Thai language that there is no space or delimiter between words. However, the task is also hard to select the correct pronunciation in dictionary. In Festival, we use the dictionary based technique to find the pronunciation of each word in dictionary. Nevertheless, the drawback of this technique is limited number of words in dictionary. It is impossible to add all possible words into the dictionary. Thus, we need a method to solve this problem i.e. letter-to-sound rules (LTS). LTS rules is technique that creates a pronunciation of out of vocabulary word in dictionary. Festival provides a method for building rule sets automatically or using hand-writing depend on language.

In Speech synthesis, there are two speech synthesis techniquess applied in this system: unit selection, and diphone synthesis. Although, the size of corpus is large, the quality of sound and computational time is the reason to choose this technique used in this part. The concept of unit selection is differing from diphone synthesis. There is more than one example of the unit and some mechanism is used to select between them at run-time. Size of corpus is varied depending on concerning of the target of system.

Diphone synthesis makes an absolutely fixed choice about which units exist. In general, a diphone is described by phone-phone transition. These make sure for effecting never go over more than two phones. The size of corpus is smaller than unit selection, English is not too large much. There are 1,396 units in English diphone; acceptable when compare with a big size corpus in another language such as China or Thai. However, the quality of sound is worse than unit selection because there are a lot of concatenation points in diphone synthesis. It is not smoothly. Nevertheless, Festival has method to solve this problem.

Prosody analysis part is used to make output sound become smoothly. Duration and intonation is also making sound more natural. To perform the process, we have to extract pitchmark and build linear predictive coding (LPC) coefficients first. Then, we use these parameters to adjust the quality of sound. The task to improve of this software is still on-going. Festival is consists of a set of C++ objects and some parts in Scheme language. Although, it is suitable for doing synthesis task, it cannot be ported to a small device.

3.1.2 Flite Software

Flite [30] is a small speech synthesis engine with fast run-time suitable for a device that has limited resources in terms of computational power and memory, such as an embedded system and a mobile device. Flite was developed by Carnegie Mellon University (CMU) and was derived from Festival. Festival was designed to run on a personal computer or a server. Therefore, it needs rich resources. Festival is quite slow and big since it uses C++ code, so it cannot be ported to a mobile device. In addition, Festival loads the data into an internal structure; therefore, it is time consuming.

Flite was designed to solve these problems, so that it runs on a device with limited resources but still has equivalent sound quality as Festival. To reduce the size of Festival, Flite converted the code from C++ to C. C is much more portable than C++ as well as it offers much lower level control of the size of the objects and data structure. Furthermore, it uses a compact representation of a finite state machine to represent a lexicon. In the part of

a speech database, LPC coefficients and encoded residual are chosen since it is smaller than the full pulse coded modulated signal (PCM). Moreover, there is no loading or copying of data into internal structure because Flite declared them explicitly as read-only variables and can be put into ROM. The voices in the speech database also get compiled into static structures.

After, we have background knowledge of Festival and Flite software. In the next step, we will build the Thai text-to-speech system on embedded devices.

3.2 Building Thai Text-To-Speech System on Mobile Devices.

Generally, Flite, a TTS system on mobile devices, has been implemented to generate speech in several languages. It works well in English and German. In this thesis, we use Flite as the basic TTS system to develop Thai TTS system on mobile devices. However, modifying Flite to work with Thai language is a complicated task because Thai language is a tonal language. It makes a speech corpus quite large so the speech application requires a large source database. In part of synthesis, it is impossible to collect all pre-recording sound into small device in order to make a speech synthesizer. Nowadays, Flite can solve this problem which converts the pre-recording sound to the C code. It reduces the size of corpus.

To use Flite as a text-to-speech engine for Thai, many components have to be adapted. First, the text analysis module has to be modified to be able to take Thai text as input and determine its pronunciation. For this, a word segmentation component is added to Flite together with a Thai pronunciation dictionary. Since in Thai language there do not have between words, so we cannot know what words appear in the sentence. Word segmentation module is used to separate the sentence into words, while Thai pronunciation dictionary is utilized in order to find a word pronunciation. For the speech synthesis module, we use a word pronunciation from text analysis part as a input in synthesis part. In this thesis, a Thai speech database is based on TSynC corpus. The other modules of synthesis are discussed below.

3.2.1 Thai Speech Database and Pronunciation Dictionary

Since our synthesizer uses a Unit concatenation technique. We need a Thai speech corpus in order to label speech units. Our speech database is based on TSynC [17], a Thai speech synthesis corpus developed by NECTEC. TSynC was designed as a database for a unit selection synthesizer. This corpus consists of 5,200 sentential utterances read by one female speaker which is equivalent to approximately 14 hours of speech. These utterances were selected to cover all Thai and foreign lent phones. Thai phones, 65 phones in total, are presented in Table 3.1 in the International Phonetic Alphabet (IPA), a standardized system of written phonetic notation intended to be able to represent the sounds of all world languages. In addition, Thai is a tonal language with five levels of tones and four tonal markers. Although TSynC covers all Thai phones and tones, it does not cover all diphones (i.e. all pairs of phones) as some diphones do not occur or hardly occur in Thai. We use LEXITRON, a Thai-English dictionary [31], as our pronunciation dictionary. LEXITRON contains 19,317 entries of both Thai words and English loan words such as อิเล็กทรอนิกส์

(electronics), เช็ค (check), เซนติเมตร (centimeter) etc. Each entry consists of words and word pronunciations.

Table 3.1 The number of Thai phones (Ph) and characters (Ch) in The International Phonetic Alphabet (IPA).

Type		List	Ph	Ch
Initial Consonant	**Single**	p,p^{h},t,t^{h},c,c^{h},k,k^{h},b,d,m,n,ŋ ,f,s,h,r,l,w,j	32	44
	Cluster	pr,pl,tr,kr,kl,kw,p^{h}r, p^{h}l,t^{h}r,k^{h}r,k^{h}l,k^{h}w		
Vowel	**Short**	i,ɨ,u,e,ɜ,o,æ,a,ə,ia, ɨa,ua	24	16
	Long	iː,ɨː,uː,eː,ɜː,oː,æː,aː, əː,iːa, ɨːa,uːa		
Final Consonant		p,t,c,k,m,n,,w,j	9	37

3.2.2 Synthesis Part

After, we already have prepared the database. The next step is to make a TTS system on Festival software. Figure 3.1 shows the process of Festival_Thai.

We convert a Thai speech synthesis database into the format that follows he guidelines of Festival. To do so, we define a Thai phone set and create a speech database that iscompatible with Festival. In TSynC, our speech database resource, a syllable structure is defined as follow:

$$| Ci – V – Cf - T |$$

where Ci is an initial consonant, V is a vowel, Cf is a final consonant, and T is a tone.

To facilitate a tonal language such as Thai in Festival, one simple solution is to integrate tones into phones. Hence, the structure of a Thai syllable in Festival becomes:

$$| CiT – VT – CfT |$$

To define phoneset, since some of these phone and tone combinations do not occur in Thai, there are only 326 combinations of phone and tone in TSynC. Thus, a set of Thai phones in Festival, a tonal phone list, contains only 326 phones instead of 445 phones which is the number of Thai phones with foreign phones (89) multiplied by the number of tones (5).

Since TsynC is available phone boundary in each sentence of sound. Next step, we can generate diphone by applying the concept of mid of phone to another mid of phone as show in Figure 3.2.

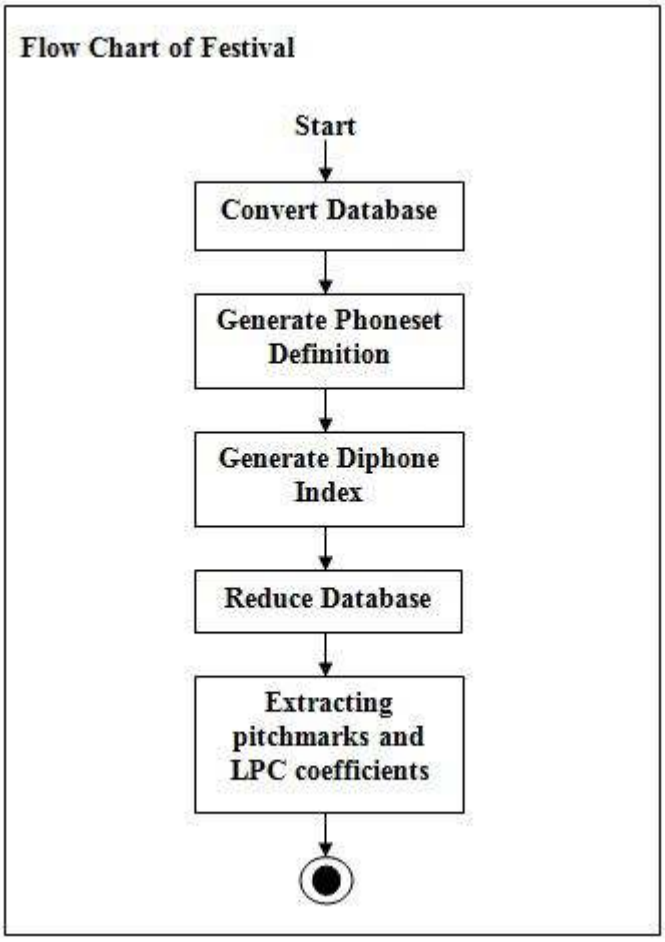

Figure 3.1 Process of Festival_Thai.

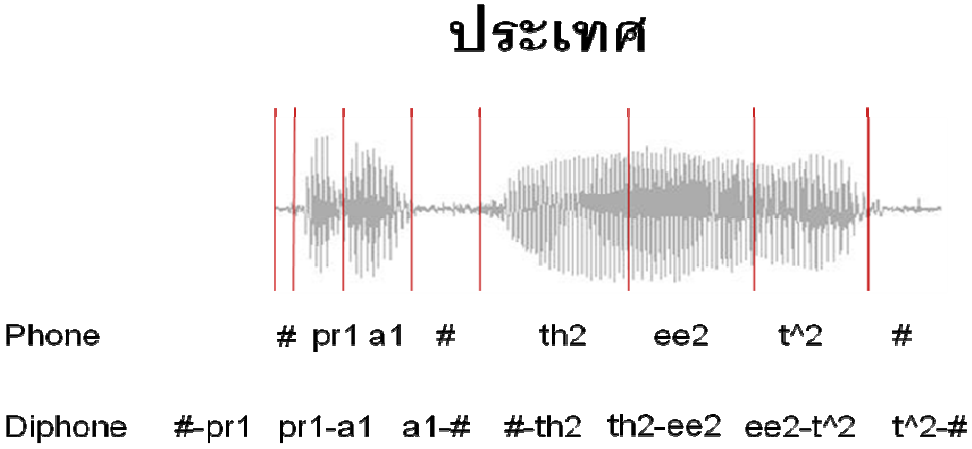

Figure 3.2 Label word "ประเทศ" or "nation" into diphone.

There are 9,901 unique diphones in TSynC. However, many diphones occur many times in the database. So, we use a First In First Out (FIFO) algorithm to choose only the first occurrence of each diphone in the database. Next, we generate a diphone index that points to a position in a wave file that contains each diphone. We then reduce the size of the database by keeping only the indexed parts of the wave files plus. We also keep 0.5 second of speech before and after each indexed part for a concatenating purpose. Nevertheless, a speech database of 9,901 diphones is still too large to be compiled. For comparison, the number of diphones in English is 1,619. Hence, we further reduce the size of the database by removing some diphones that hardly occurs such as some foreign phone and tone combinations. At this stage, the size of our speech database reduces to 150 MB and contains 9,090 diphones.

Then, we extract pitchmark wave sound and do the post-processing step that moves the predicted pichmark to the nearest waveform peak. Then, we find the mean and average power for each vowel overall files by power normalization technique. Finally, we extract LPC parameters and LPC residual files for each file in the diphone database.

At this stage, we got the Thai TTS standard version with Festival on PC. After that, we transfer all information of Festival to Flite software. Figure 3.3 shows the process of constructing Flite_Thai.

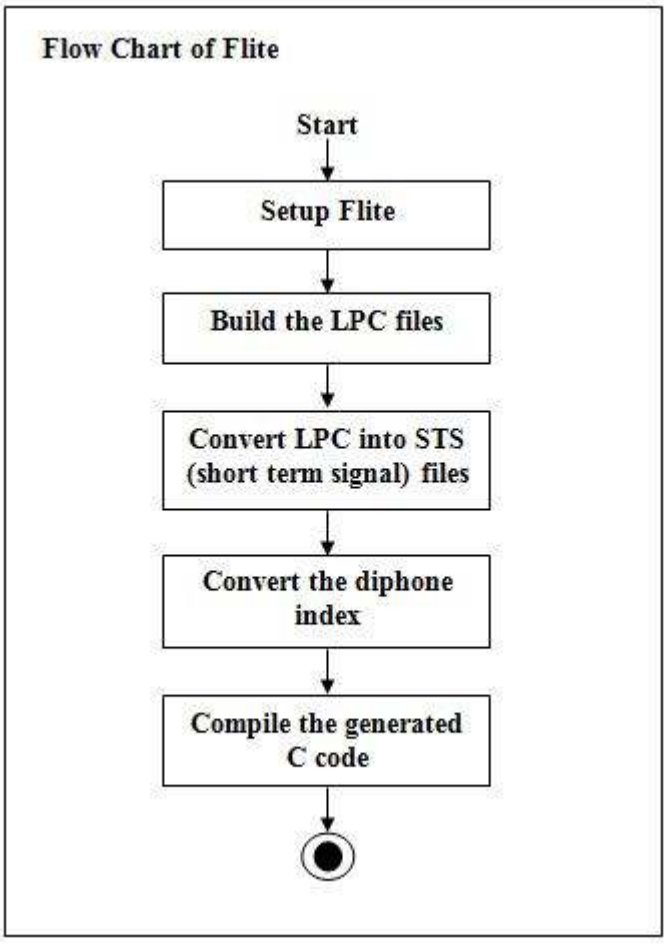

Figure 3.3 Process of Flite_Thai.

First of all, we have to setup Flite in order to transfer some parameters from Festival to Flite. Then, we convert the speech database into C code which reduces the size of the database to 44 MB. This is done by extracting pitchmarks and LPC coefficients. These features are then converted into short term signals (STS) and then C code. The size of the speech database in C code is 44 MB. Finally, after compiling this C code together with the C code of the pronunciation dictionary and the source code of Flite engine, the size of the excitable Thai version of Flite (Flite_Thai) is only 10 MB. Table 3.2 summarizes the size at each step of the compression process.

Table 3.2 Summarizes the size at each step of the compression process.

Database & Application	Size
TsynC	768 MB
Diphone database	150 MB
Diphone database in C	44 MB
Flite_Thai	10 MB

3.3 Improving the Quality of Sound

Although, Flite_Thai can run in real-time on a mobile phone, having a system that runs on real-time is not enough; the quality of the sound should also be good. A synthesis technique is the one that affects the quality of the sound. For example, the quality of the sound from a concatenative technique depends on a speech corpus and speech units because it utilizes only a set of pre-recorded units. On the other hand, the quality of the sound from a HMM-based synthesis (HTS) techniqueis quite natural because the output sound is re-created from speech parameters generated from a statistical model trained from a speech corpus; hence there is no discontinuity from speech concatenation.

Since, the result in term of quality of sound in Flite_Thai is not natural, discussed in detail in section 4.2.2. So in this point, we focus on improve the existing system, Flite_Thai, in terms of the quality of synthesized speech. In Flite_Thai, a speech corpus, TsynC was used. TsynC was designed for a unit selection synthesizing technique, so a large set of natural sentences were recorded. The unit selection technique, then, chooses the portions of speech in the corpus that best match the input text; the sizes of these portions varied. The unit concatenation employed in Flite_Thai, on the other hand, uses a fixed size unit, and thus, requires more clearly articulated speech. So, the TsynC corpus is not appropriate for Flite_Thai. We design a new speech corpus and a new speech unit to solve the problem of output sound quality. The new speech unit, a hybrid diphone, combines the advantages of both a diphone and a demi-syllable. The new speech corpus consists of three types of speech units: diphone, demi-syllable and hybrid diphone recorded using non-sense carrier sentences.

3.4 Hybrid diphone

Since, we need to improve the quality of diphone speech units. We design a new speech unit to improve the quality of concatenated speech and reduce the size of the database. This unit, a *hybrid diphone* [32], is designed based on a diphone unit but incorporates some characteristics of a demi-syllable unit to benefit from its advantage i.e. a smaller database. It is solve the problem of size in diphone and problem of discontinuity at the boundary between syllables in demi-syllable.

Generally, a syllable structure in Thai can be denoted as "Ci-V-Cf". With this structure, there are three types of boundaries between phones i.e., "Ci-V", V-Cf", and "Cf-Ci". The first and two types, Ci-V and V-Cf, are used concepts of demi-syllable. Because we need to reduce number of Ci-V (do not include tones). However, we change a little bit in consonant boundary in demi-syllable. We move the boundary of Ci-V and V-Cf to mid consonant and add diphone Cf-Ci to keep the co-articulation between syllables. Figure 3.4 shows the structure of the hybrid diphone.

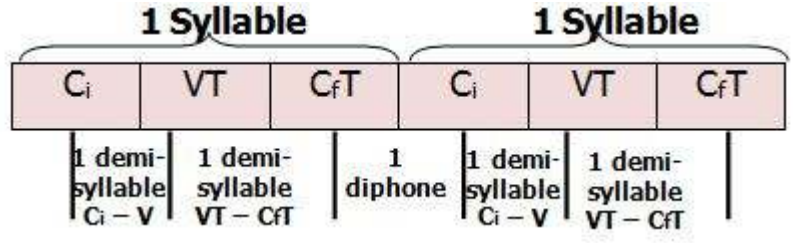

Figure 3.4 Structure of Hybird diphone.

Even though, the hybrid diphone combines the advantages of both the diphone and demi-syllable, it still has a problem of discontinuity within a consonant because the consonant is cut into halves in the diphone unit. Table 3.3 illustrates the advantages and disadvantages of diphone, demi-syllable and hybrid diphone.

Since the hybrid diphone uses all of the demi-syllable units together with only "Cf-Ci" diphone units, the total number of the hybrid diphone units is quite small.

Table 3.3 The advantages and disadvantages of diphone, demi-syllable and hybrid diphone

	Advantage	Disadvantage
Diphone	• Preserving the transition between phones	• Large Speech database • Discontinuity within a consonant after concatenating
Demi-syllable	• Small speech database • Good speech quality within a syllable	• Discontinuity at the boundary between syllables
Hybrid diphone	• Medium size speech database • Good speech quality within a syllable • Preserving the transition between phones	• Discontinuity within a consonant after concatenating

3.5 Building a New Thai Speech Corpus

Flite_Thai uses a diphone concatenation technique with the diphones extracted from the TsynC database which was designed for corpus unit selection synthesis. We believe that using a speech database that is designed especially for diphone concatenation should improve the quality of the synthesized sound. In this section, we first describe the technique to reduce the number of the hybrid diphones required for a Thai speech database by carefully looking at co-articulation characteristics between phones. Next, we describe the design of carrier sentences which will be used as recording prompts when we construct a new speech corpus. Finally, we explain how to identify and extract different types of speech units from the corpus.

3.5.1 Reduce the Size of a Speech Corpus without Affecting the Quality of the Synthesized Speech

A diphone speech corpus, such as TsynC, is usually quite large and requires a lot of space which could be a problem on resource-limited devices. To solve this problem, Wutiwiwatchai, et al. [33] proposed a technique that can reduce the number of diphones in the database without affecting the quality of the synthesized sound. In Thai there are some pairs of phones that have less or no co-articulation between them as illustrated in Table 3.4.

Table 3.4 Phoneme pair with weak co-articulation.

Group of phoneme pair with weak co-articulation	
Stop	Stop
Stop	Nasal
Nasal	Stop
Nasal	Fricative
Fricative	Fricative
Fricative	Stop
Fricative	Nasal
Lateral	Stop
Lateral	Fricative

So, we can ignore some diphones that contain weak co-articulation or short pauses for example word "เทศ" and "ไทย", final consonant (t^_3) of first word is a stop sound and initial consonant (th_0) of second word is a stop sound thus short pause is occur between phoneme as show in Figure 3.5 Whereas, co-articulation between phonemes is shows in Figure 3.6, word "ไทย" and "มี" are pairs of approximant (j^_0) and nasal (d_3) sound.

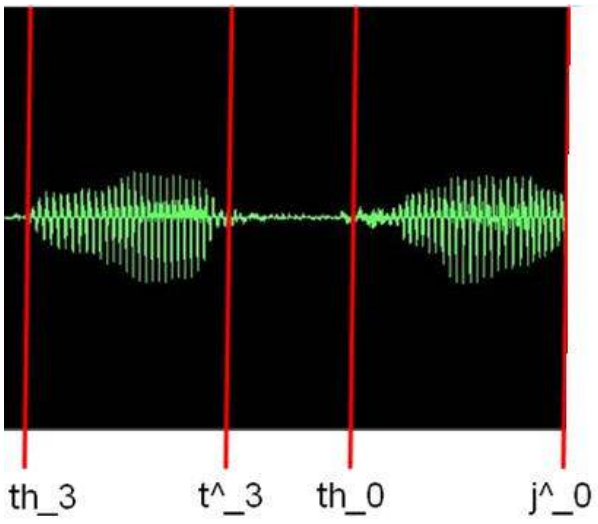

Figure 3.5 Phoneme pair (t^_3 – th_0) with weak co-articulation.

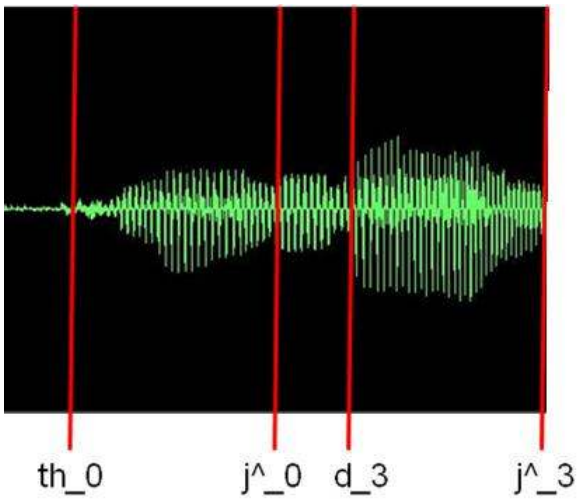

Figure 3.6 Co-articulation of Phoneme pair (j^_0 – d_3).

In addition, some syllabic tones affect only voiced phones. So, we replace all tone variations of a phone with a mid tone (represented by 0) if that phone is unvoiced sound. Table 3.5 shows the some weak pairs in case of unvoiced sound.

Table 3.5 Weak pair in case of unvoiced sound.

Weak pairs in case of unvoiced sound	
Unvoiced Cf	Voiced Ci
Unvoiced Cf	Unvoiced Ci
Voiced Cf	Unvoiced Ci
Vowel	Unvoiced Ci

For example, a diphone “n4-th3” is a voiced-unvoiced diphone. Therefore, we can replace “th3” with “th0”. In the same way, “th1”, “th2” and “th4” can be replaced by “th0”.

Moreover, we can use phonetic rules to reduce the diphones that hardly occur [19]. Some cluster consonant cannot followed by some vowel [34] as shown in Table 3.6, for instant, although cluster consonant “kw” or “กว” in Thai can concatenate with vowel “ee” or “–ี” in Thai to word “กวี”, it cannot pronounce as one syllable. Word is pronounced as “/kr0-a0-z^0/w0-ee0-z^0/”. In the same way, we have some cases that vowel and final consonant is not occurring in Thai writing rules [34] as shown in Table 3.7.

Table 3.6 Some cluster consonant cannot followed by some vowel.

Cluster consonant		Vowel					
		v, vv, va, vva	**u, uu, ua, uua**	**e, ee**	**o, oo**	**@, @@**	**q, qq**
kw	กว	X	X	X	X	X	X
kr	กร						
khw	คว	X	X			X	X
tr	ตร						X
pl	ปล						X
phr	พร, ผร						X
phl	พล, ผล	X					

Table 3.7 Some vowel cannot followed by some final consonant.

Vowel		Final consonant							
		p^	t^	k^	m^	n^	ng^	j^	w^
i	–ิ							X	
ii	–ี						X	X	X
v	–ึ							X	X
vv	–ื			X			X	X	X
u	–ุ								X
uu	–ู							X	X
e	เ–ะ							X	
ee	เ–							X	
x	แ–ะ			X				X	
xx	แ–							X	
o	โ–ะ							X	X
oo	โ–								X
@	เ–าะ	X	X	X					X
@@	ออ								X
q	เ–อะ	X	X	X	X			X	X
qq	เ–อ								X
iia	เ–ียะ							X	
vva	เ–ือ								X
uua	–ัว								X

To reduce the units, we can decrease the number of units in case of tone. Since tone is a main point that make a large size of corpus. So, we reduce tone by modulation of three classed consonant rule. We can classify consonant into three classes, high consonant, mid consonant and low consonant. However tone is not only modulating base on three classed consonant but also property of syllable. Live syllable and dead syllable are a property of syllable that makes tone changing [35].

Live syllable is an open syllable with a long vowel, or a closed syllable with a live consonant ending, -m^, -n^, -ng^, -j^, -w^.

Dead syllable is an open syllable with a short vowel, or a closed syllable with a dead consonant ending, -p^, -t^, -k^. In the latter case, the tone rules require further distinction between long dead syllable and short dead syllable. We can modulation tone base on three classed consonant as Table 3.8.

Table 3.8 Modulation tone base on three classed consonant.

Property of syllable	Three class consonant		Tone				
			Middle	Low	Falling	High	Rising
Live syllable	Mid consonant		กา	ก่า	ก้า	ก๊า	ก๋า
	High consonant			ข่า	ข้า		ขา
	Low consonant		คา		ค่า	ค้า	
Dead syllable	Mid consonant			กะ	ก้ะ	ก๊ะ	ก๋ะ
	Short vowel	High consonant		ขะ	ข้ะ		
		Low consonant			ค่ะ	คะ	ค๋ะ
	Long vowel	High consonant		ขาก	ข้าก		
		Low consonant			คาก	ค้าก	ค๋าก

After applying all technique to the speech database, the number of diphones in the TsynC database reduces from 9,090 units to 5,820 units, which is about 36% reduction. Table 3.9 shows the number of units for diphone, demi-syllable, and hybrid diphone.

Table 3.9 The number of diphone, demi-syllable and hybrid diphone.

	Diphone	Demi-syllable	Hybrid diphone
Number of units	5,820	1,505	2,830

3.5.2 Carrier Sentences

To collect a speech corpus, we use a nonsense carrier sentence [20] to record a speech corpus because the pronunciation of a phone is clearer in a carrier sentence than in a natural sentence as the phone in the natural sentence is affected by context. We design a set of carrier sentences based on diphone units because a set of diphone speech units covers both demi-syllable and hybrid diphone units. To support this, the first types of boundaries between phones "Ci-V" in demi-syllable and hybrid diphone do not have tone, while in diphone there is all tone in "Ci-V". In this paper, we design two types of carrier sentences, one for a diphone within a syllable and another one for a diphone across syllables.

a) Diphone within a Syllable

Some tones in some syllables are very hard to pronounce individually. Nevertheless, Thai people are acquainted with pronouncing all five tones in one sentence. For each diphone, we then design a carrier sentence which contains five tonal diphones (with five levels of tones) of that diphone. Moreover, we add a syllable "ทา" ("t0-aa0") at the beginning and at the end of the sentence as fillers in order to avoid the articulation affects of the beginning and the end of the sentence. The sound "t" and "aa" are chosen since they do not affect other phones. For example, a carrier sentence for a diphone "k-aa" is "SIL-t0-aa0-k0-aa0-k1-aa1-k2-aa2-k3-aa3-k4-aa4-t0-aa0-SIL" or "ทากาก่าก้าก๊าก๋า" in Thai. The number (i.e. 0 to 4) in the sentence represents a tone in each phone. This pattern is used to collect Ci-V, V-Cf, silence-Ci, V-silence and Cf-silence.

b) Diphone across Syllables

Another set of carrier sentences is used to capture the transition between syllables. Some diphones in this case may have different tones, for example, a diphone Cf-Ci that consists of a final-consonant of one syllable and an initial-consonant of the next syllable. Phonemes "z" and "aa" are added to the beginning of the carrier sentence so that when combining with the final consonant in a diphone it can be pronounced as one syllable. We use "z" because it is a general character which can produce all five levels of tones. Furthermore, the phoneme "aa" is also added at the end of the sentence to make an initial consonant of a diphone pronounceable. For example, for a diphone "uu1-r3", we use a sentence "SIL-t0-aa-z1-aa1-uu1-r3-aa3-t0-aa0-SIL" or "ทาอ่ายร้าทา" in Thai as a carrier.

The difference of carrier sentence and natural sentence is a co-articulation in sentence. Figure 3.7 shows units in the natural sentence that recorded by TSnyC corpus. It has some co-articulation affect between units. So it is not clear when collect that unit. However, Figure 3.8 shows units in the carrier sentence. In each syllable, is very clear because it does not have any affect from the previous syllable.

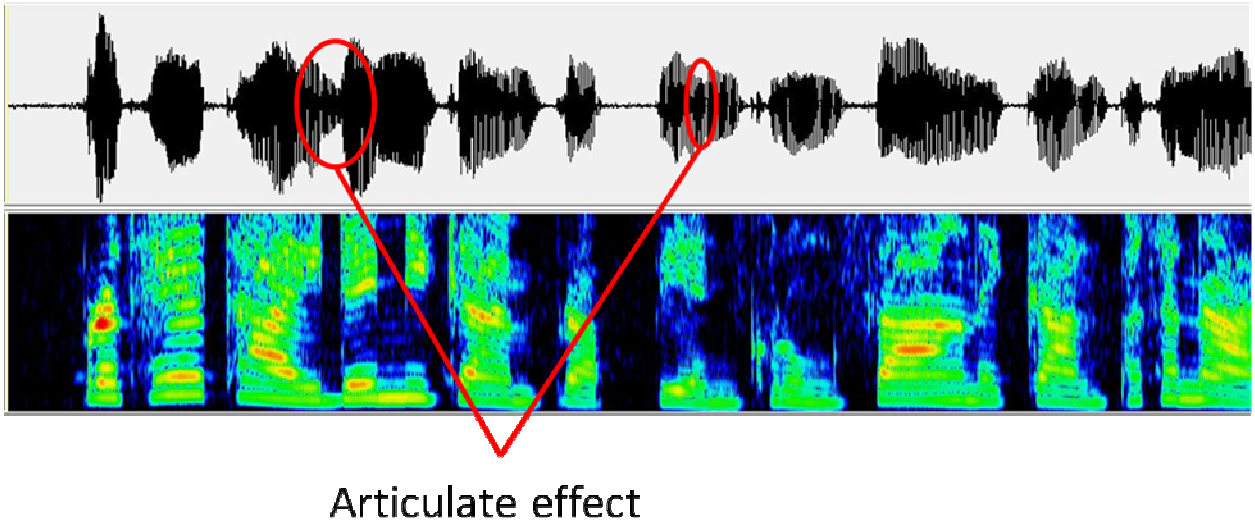

Figure 3.7 Natural sentence that recorded by TSynC.

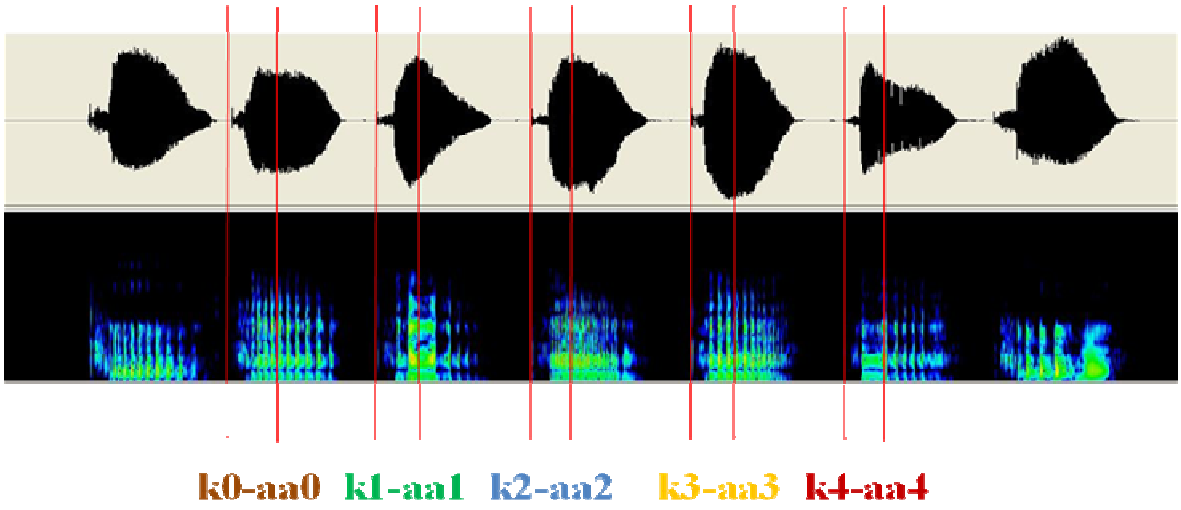

Figure 3.8 Carrier sentence as “ทากาก่าก้าก๊าก๋า”.

3.5.3 Recording Condition

To create a new speech database, we asked a Thai female to read carrier sentences designed in the previous section in a clear articulation manner. The recording was done in a silent room with a digital audio tape recorder (DAT Recorder). The sentences were recorded as 16-bit/sample raw data at a sampling rate of 44.1 kHz.

3.5.4 Label Wave File

After recording all of carrier sentences, the last step of corpus preparation is to label wave files. In order to extract all three types of speech units (i.e. diphone, demi-syllable, and hybrid diphone) from carrier sentences, three kinds of boundaries in the carrier sentences are needed to be marked: (1) a phone boundary, (2) the beginning of the steady part of a vowel, and (3) the middle point of a phone as shows in figure 3.9.

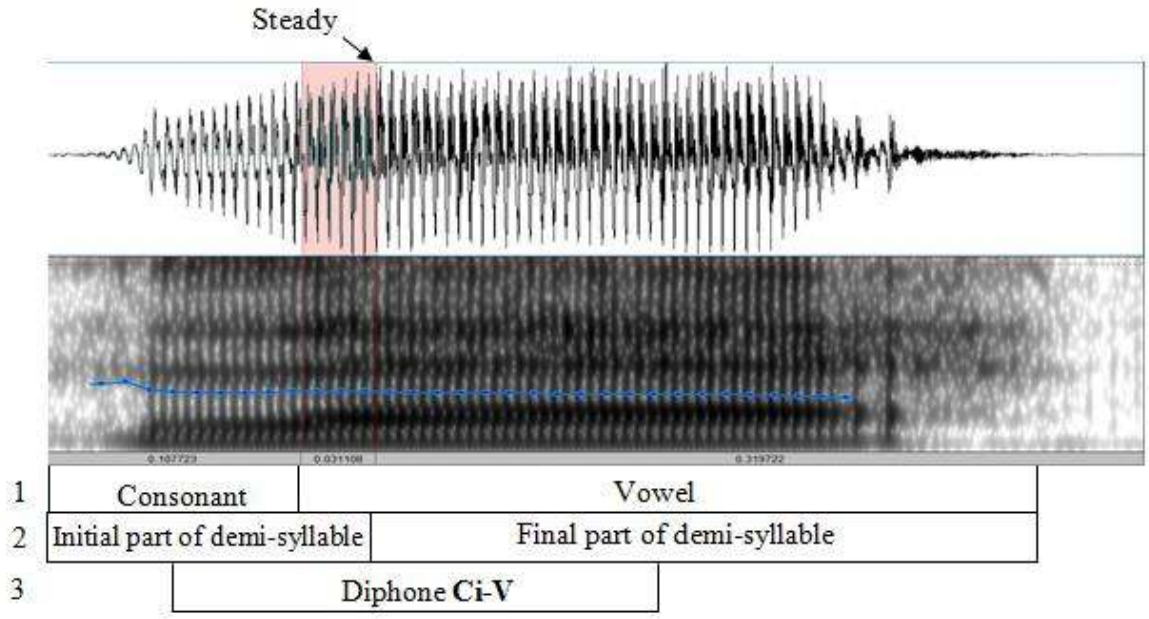

Figure 3.9 Steady point in demi-syllable.

The first two kinds of the boundaries are manually annotated while the last one can be identified automatically. The beginning of the steady part of a vowel necessary for marking a demi-syllable unit can be obtained by using a spectrogram. We mark the boundary of a diphone unit by using the middle point between two consecutive phone boundaries. This can be computed easily when we marked all the phone boundaries. The boundaries of a hybrid diphone unit can be identified using the boundaries of demi-syllables and diphones. All kinds of boundaries are marked at the zero-crossing point in order to get smooth sound when two units are concatenated.

Chapter 4

Experiments and Results

4.1 Experiment Design

The goal of our research is to develop a Thai speech synthesizer that can produce an output speech in real-time on a mobile device, and qualities of sound should be acceptable. We evaluate this research using two experiments.

- The first experiment is evaluating of Flite_Thai (first version)
- The second experiment is measuring of a new speech corpus and speech unit.

4.2 Evaluation of Flite_Thai

We conducted the experiments to compare Flite_Thai and pTalk software. Flite_Thai is a concatenative speech synthesizer that we develop based on Flite. pTalk is an HMM-based speech synthesizer as described in section 2.7. Both systems use the TSynC database and LEXITRON dictionary. The experiment was run on a HTC touch 3450. The detailed specification of this mobile phone is Windows Mobile 6.1 Professional using TI's OMAPTM850, 201 MHz processor with 128 MB RAM. Since Flite was developed on Linux, we also have to port the source code of Flite_Thai to a Window Mobile platform. The size of Flite_Thai on Window Mobile is about 10 MB while the size of pTalk is about 12 MB. In our experiments, we compared both synthesizers based on two criteria: processing speed and sound quality.

4.2.1 Processing Speed Test

Measurement of processing speed is done by counting the synthesis time manually. For both Flite_Thai and pTalk systems, we started measuring the time when a "play" button is pressed until the first speech sound is pronounced. In this experiment, we defined four test sets. Each test set contains two sentences with the same number of words and syllables. Different test sets have different number of words as shown in Table 4.1. We also report the number of syllables in each test set as it also affects the processing speed.

Table 4.1 The result of processing speed both Flite_Thai and pTalk.

Test Set	1	2	3	4
#words	10	20	30	40
#syllables	21	44	70	90
Processing time of Flite_Thai (second)	0.83	1.42	1.73	2.26
Processing time of pTalk (minute)	> 2.0	-	-	-

The processing time that Flite_Thai uses to generate speech is shown in the Table 4.1. The result of each test set is presented by the average value. Unfortunately, we are not able to report the average processing time of pTalk on each test set since it takes too long to synthesis some sentences. In general, pTalk takes longer than 2 minutes to synthesize one sentence. Flite_Thai, On the other hand, takes only about 2 second to synthesize a 40-words sentence. Hence, Flite_Thai is at least 60 times faster than pTalk.

4.2.2 Speech Quality Evaluation

In this section, Mean Opinion Score (MOS) is chosen to measure the synthesizers' performance in term of sound quality. Scores in the experiment are represented between 1 and 5 (least natural to most natural) as explained Section 2.6.1. The synthesized speech outputs from both Flite_Thai and pTalk were evaluated by 10 subjects who are between 25 – 35 years old (30 years in average). Each subject was asked to listen 3 sets of 10 wave files. The first set is a human natural speech taken from the TSynC database. The second set is a synthesized speech generated by Flite_Thai while the third set is a synthesized speech generated by pTalk. The results, the average MOS values, are shown in table 4.2.

Table 4.2 The MOS values of original sound (TSynC), Flite_Thai and pTalk.

Source	MOS
Original Sound	4.5
Flite_Thai Sound	1.47
pTalk Sound	2.93

From the experimental results, pTalk obtained higher score than Flite_Thai, but the quality of its output speech is still not closed to the natural speech. The difference in terms of synthesized speech quality between Flite_Thai and pTalk comes from two main points,

training data and prosody analysis. pTalk, which is an HMM-based synthesizer, was trained from the entire TSynC database while Flite_Thai, which is a diphone concatenation synthesizer, utilizes only a set of selected diphones that is a subset of the TSynC database. When the same amount of data is utilized such as the case of a unit selection synthesizer both an HMM-based synthesizer and a concatenative synthesizer produce output speech with comparable overall quality. Furthermore, an HTS system usually produces a smoother speech than a concatenative system, especially a diphone concatenation which has many concatenating parts. So in the next experiment, we improve the quality of sound as discuss in section 3.3

4.3 Evaluation of Speech Units and Corpus

We compare the sound qualities of the synthesizer outputs when different speech corpora and speech units are used. The Mean Opinion Score (MOS) is used for evaluating the intelligibility of the synthesized speech. The output three wave files were evaluated by ten evaluators who are between 22 to 26 years old. Each evaluator had to listen to the synthesized speech from all four conditions. The results are shown in Table 4.3. The result in the first row comes from the existing Flite_Thai system which uses diphones extracted from natural sentences in the TsynC corpus. All other results are the cases when the new speech database designed is used.

Table 4.3 The MOS values of diphone, demi-syllable and hybrid diphone which different recording prompt, natural sentence and carrier sentence.

Speech Unit	Recording Prompt	Intelligibility (MOS)
Diphone	natural sentence	2.73
Diphone	carrier sentence	2.88
Demi-syllable	carrier sentence	3.72
Hybrid diphone	carrier sentence	3.13

Since the proposed technique is also developed on Flite_Thai and the synthesize technique is also uses a unit concatenation technique. The computational time is similar to the existing system. The goal of the experiment is to find the appropriate speech corpus and speech units for Flite_Thai. From our experiment, the MOS values obtained when the speech units from the new corpus are used are higher than the MOS values obtained when the units from the existing corpus, TsynC, are used. This is because nonsense carrier sentences are more clearly articulated than natural sentences. The experimental results show that the proposed hybrid diphone unit is able to provide better intelligibility than the traditional diphone unit, but requires fewer speech units. Because speech units that come from carrier sentence are clearly in sound more than natural sound, TSynC, it affects to output sound. Since the quality of sound in unit concatenation technique is depending on pre-recording sound.

However, demi-syllable is the speech unit that generates speech with the best intelligibility and requires the fewest speech units. One possible reason that makes the output from the hybrid diphone noticeably less intelligible than the output from the demi-syllable is the number of concatenating points. The hybrid diphone basically requires more concatenating points than the demi-syllable. In this experiment, the concatenation was done without applying any smoothing technique. The evaluators easily detected the discontinuity between the hybrid diphone speech units. However, we expected that this problem can be improved by the smoothing technique.

Chapter 5

Conclusion

This thesis had two related goals. The first is to develop a Thai Text-To-Speech system on embedded devices, or mobile phones that produces an output speech in real-time manner. The second is to improve the quality of sound in term of intelligibility by proposing a new speech unit called a hybrid diphone, and nonsense carrier sentence, a technique to prepare a speech corpus. Moreover, the size of corpus is reduced by new speech corpus and speech units.

We first focused more on the synthesis speed rather than the sound quality. Our synthesizer is based on Flite, an open source synthesis library, suitable for implementing on a fast and small TTS application. To use Flite as a text-to-speech engine for Thai, many components have to be modified. We modified several components in Flite to make it support Thai. The modifications include a tonal phone set for Thai which integrates tones into Thai phonemes, and a reduced-size speech database. We transform a unit selection database into a diphone database by selecting only necessary diphones. We conducted an experiment to compare our speech synthesizer with pTalk, an HMM-based speech synthesizer, both in terms of speed and sound quality measured by a subjective listening test. For the result of the first goal, we found that our system is much faster, approximately 60 times faster, and more stable than pTalk. However the quality of our output speech may not be as good as the output from pTalk. The hypothesis of that problem is speech corpus and speech units in Flite_Thai.

To support the second goal, we aimed at designing the appropriate speech unit and creating the new speech corpus for Flite_Thai in order to improve its intelligibility. We designed a new speech corpus that consists of three different speech units: *demi-syllable*, *diphone* and our newly proposed speech unit called *hybrid diphone*. The hybrid diphone combines the advantages of diphone and demi-syllable speech units. It solves the problem of speech quality within a syllable; however, it introduces a lot of concatenating joints similar to the diphone. A TTS system on mobiles devices suffers more from the concatenating joints as it has limited resources to apply a smoothing technique.

Although, the hybrid diphone is not the best speech unit, but we found that recording prompt affects quality of sound. We use a non-sense carrier sentence technique for recording a new speech corpus since we focus more on clear articulation of each speech unit. The technique to prepare a speech corpus with carrier sentences provides a clearer pronunciation than using natural sentences and is more suitable for a unit concatenation technique used in Flite_Thai. Co-articulation affect is concerned in carrier sentence technique. Our carrier sentence contains a speech unit or a set of similar speech units per sentence without concerning the meaning. We designed a new carrier sentence to make it easier to record speech units by including five speech units from all tone variations in one sentence.

For one of the achievements of this thesis, we could say that Flite_Thai is the first noncommercial Thai speech synthesizer that can run in real-time on a mobile phone. Since Flite is an open source software, it can be widely distributed and further developed. Moreover, there are plenty of rooms for improvement a quality of sound. A more optimal way to select a list of diphones from a non-sense carrier sentence database could be investigated. A fully implemented prosodic analysis with duration and intonation modeling could also improve the naturalness of the synthesized speech. In addition, each speech unit has different pitch value, so using the average pitch value at the transition between units could improve the smoothness of the sound. Furthermore discovering the appropriate position of stable signal in vowel boundary is important to make it clearer.

Reference

[1] J.Inrut, P. Yuanghirun, S. Paludkong, S. Nitsuwat, and P. Limmaneepraserth, "Thai word segmentation using combination of forward and backward longest matching techniques", in the proceeding of ISCIT, pp 37 – 40, 2001.

[2] V. Sornlertlamvanich. "Word Segmentation for Thai in Machine Translation System". Machine Translation, pp. 50 – 66. NECTEC, 1993.

[3] S. Mekanavin, P. Charenpornsawat, and B. Kijsirikul, "Feature-based Thai Words Segmentation", in the proceedings of the Natural Language Processing Pacific Rim Symposium, pp. 41 – 48, Phuket, Thailand, 1997.

[4] P. Tarsaku, V. Sornlertlamvanich, R. Thongprasirt, "Thai Grapheme-to-Phoneme using Probabilistic GLR Parser", in the proceeding of Eurospeech, volume 2, pp. 1057 – 1060, 2001.

[5] P. Charoenpornsawat and T. Schultz, "Example-Based Grapheme-to-Phoneme Conversion for Thai", in the proceeding of Interspeech 2006, pp. 1268 - 1271, 2006.

[6] A. Rugchatijaroen, A. Thangthai, S. Saichum and C. Wutiwiwatchai, "Prosody-based Naturalness Improvement in Thai Unit-selection Speech Synthesis" in the Proceeding of ECTI-CON 2007, Chiang Rai, Thailand.

[7] F. Charpentier and E. Moulines, "Pitch synchronous waveform processing techniques of European", in the Proceeding of Speech Communication and Technology, vol. I, pp. 013 – 019, 1989.

[8] http://www.voiceproblem.org/anatomy/learning.asp

[9] S. Luksaneeyanawin, et al. "A Thai text-to-speech system" in the Proceeding of 4th NECTEC conference, pp. 65 – 78, (in Thai).

[10] J.T. Gandour, S. Potisuk, S. Dechongkit, "Tonal Coarticulation in Thai", Journal of Phonetics, vol 22, pp. 447 – 492.

[11] http://en.wikipedia.org/wiki/Thai_language

[12] A. Hunt and A. Black, "unit selection in a concatenative speech synthesis system using a large speech database" Proceedings of ICASSP 96, vol 1, pp. 373 – 376, Atlanta, Georgia.

[13] A.C. Smith and J.J.D. van Schalkwyk, "Line-spectrum pairs-a review" in the Proceeding of Southern African Conference on Communications and Signal Processing (COMSIG 88), pp. 7 – 11, 1988.

[14] K. Tokuda, H. Zenl, and A. W. Black, "An HMM-BASED SPEECH SYNTHESIS SYSTEM APPLIED TO ENGLISH", IEEE 2002 Workshop on Speech Synthesis, Santa Monica, USA, Sept. 2002.

[15] K. Tokuda and H. Zen, "Fundamentals and recent advances in HMM-based speech synthesis" Tutorial at Interspeech 2009, Brighton, U.K., Sept. 2009.

[16] K. Wongpatikaseree, A. Ratikan, A. Thangthai, A. Chotmongkol and C. Natte, "A Real-time Thai Speech Synthesizer on a Mobile Device", in the Proceeding of 8th Symposium on Natural Language Processing (SNLP-2009), 20-21 October 2009, Bangkok Thailand.

[17] C. Hansakunbuntheung, V. Tesprasit, and V. Sornlertlamvanich, "Thai Tagged Speech Corpus for Speech Synthesis", in the Proceeding of the Oriental COCOSDA, pp. 97 – 104, 2003

[18] P. Mittrapiyanuruk, C. Hansakunbuntheung, V. Tesprasit and V. Sornlertlamvanich, "Issues in Thai Text-to-Speech Synthesis: The NECTEC Approach", in the Proceeding of NECTEC Annual Conference, Bangkok, Thailand, pp. 483 – 495, 2000.
[19] T. Soontornphand, "Design and Development of Concatenative Thai Speech Synthesis using A Small Speech Corpus", Master thesis, Chulalongkorn University, 2008.
[20] Kevin A. Lenzo and Alan W Black, "Diphone Collection and Synthesis", in the Proceeding of Sixth International Conference on Spoken Language Processing, ICSLP 2000, Beijing, China, Oct 2000.
[21] Torsak S., Atiwong S., Proadpran P., "Concatenative Thai Speech Synthesis Using a Small Speech Corpus", in the Proceeding of The 12th National Computer Computer Science and Engineering Conference, NCSEC 2008, Chonburi, Thailand, Nov 2008.
[22] X. Huang, A. Acero and H. Hon, "Spoken language processing: a guide to theory, algorithm, and system development", New Jersey: Prentice Hall PTR, 2001.
[23] Hoffmann, R. et al., "A Multilingual TTS System with less than 1 MByte Footprint for Embedded Applications", in the Proceeding of ICASSP, 2003.
[24] P. Tsiakoulis, A. Chalamandaris, S. Karabetsos, and S. Raptis, "A Statistical Method for Database Reduction for Embedded Unit Selection Speech Synthesis", in the Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 – April 4, Las Vegas, Nevada, USA, 2008
[25] M. Pucher, and P. Frohlich, "A User Study on the Influence of Mobile Device Class, Synthesis Method, Data Rate and Lexicon on Speech Synthesis Quality", in the Proceeding of Interspeech 2005, Lisbon, Portugal, 2005.
[26] P. Torruangwatthana, and T. Nuchpithak, "Thai User Interface on Mobile Phone Device for Blinds", senior project report, Sirindhorn International Institute of Technology, Thammasat University, 2009.
[27] N. Chinathimatmongkhon, A. Suchato, and P. Punyabukkana, "Implementing Thai Text-to-Speech Synthesis for Hand-held Devices", in the Proceeding of ECTI-CON 2008, Krabi, Thailand, May 2008.
[28] NOKIA Connecting People, Text-to-Speech Available: http://www.nokia.co.th/support-and-software/software/text-to-speech.
[29] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system", Available: http://www.cstr.ed.ac.uk/projects/festival.html, 1998.
[30] A. W. Black and K. A. Lenzo, "Flite: a small fast run-time synthesis engine", in the Proceeding of the 4th ISCA Workshop on Speech Synthesis, 2001.
[31] LEXITRON, a Thai-English electronic dictionary, Available: http://lexitron.nectec.or.th.
[32] K. Wongpatikaseree, A. Ratikan, A. Chotimongkol, P. Chootrakool, C. Nattee, T. Theeramunkong and T. Kobayashi, "A Hybrid Diphone Speech Unit and a Speech Corpus Construction Technique for a Thai Text-to-Speech System on Mobile Devices", in the Proceeding of ECTI-CON International Conference 2010, 19-21 May 2010, Chiang Mai, Thailand.
[33] C. Wutiwiwatchai, S. Saychum, A. Rugchatijaroen, "An Intensive Design of a Thai Speech Synthesis Corpus", in the Proceeding of the SNLP 2007, Bangkok, Thailand.
[34] K. Nacaskul, "Sound System in Thai Language", Chulalongkorn University Printing House, Bangkok, Thailand, 1998, (in Thai).
[35] Pitsri Kramonlawech, "All Tone in Thai Language", Horatanachai Publishers, Bangkok, 2005, (in Thai)

List of Publications

Thesis Title: Speech Synthesizer for Thai Text-To-Speech (TTS) System on Embedded Device

Name of Student: Mr. Konlakorn Wongpatikaseree

School: School of Information and Communication Technology for Embedded Systems (ICTES)

Thesis Committee:

Advisor and Chairperson of Thesis Committee:	Asst. Prof. Cholwich, Ph.D.
Committee Member and Chairperson of : Examination Committee	Ananlada Chotimongkol, Ph.D.
Committee Member:	Assoc. Prof. Thanaruk Theeramunkong, Ph.D.
Committee Member:	Prof. Takao Kobayashi, Ph.D.

Year of Graduation: 2nd Semester / 2009

International Conference Proceedings

- Konlakorn Wongpatikaseree, Arunee Ratikan, Ausdang Thangthai, Ananlada Chotimongkol and Cholwich Nattee, "A Real-time Thai Speech Synthesizer on a Mobile Device", *in the Proceeding of 8th Symposium on Natural Language Processing (SNLP-2009)*, Bangkok Thailand, 20-21 October 2009, pp. 42 - 47.

- Konlakorn Wongpatikaseree, Arunee Ratikan, Ananlada Chotimongkol, Patcharika Chootrakool, Cholwich Nattee, Thanaruk Theeramunkong and Takao Kobayashi, "A Hybrid Diphone Speech Unit and a Speech Corpus Construction Technique for a Thai Text-to-Speech System on Mobile Devices", *in the Proceeding of ECTI-CON International Conference 2010*, Chiang Mai, Thailand, 19-21 May 2010.

- Arunee Ratikan, Konlakorn Wongpatikaseree, Ananlada Chotimongkol, Ausdang Thangthai and Cholwich Nattee, "Comparison of Text analysis for Thai Text-to-Speech (TTS) application on mobile devices", *in the Proceeding of The first International Conference on Information and Communication Technology for Embedded Systems (ICICTES-2010)*, Pathumthani, Thailand, 28-30 January 2010.

- Arunee Ratikan, Konlakorn Wongpatikaseree, Ananlada Chotimongkol, Cholwich Nattee, Thanaruk Theeramunkong, and Manabu Okumura, "Combining a Pronunciation Dictionary of Multiple-Sized Units and a Rules-Based Technique for Thai G2P Conversion", *in the Proceeding of The 7th International Joint Conference on Computer Science and Software Engineering (JCSSE 2010)*, Bangkok, Thailand, 12-14 May 2010.