Issues in Thai Text-to-Speech Synthesis: The NECTEC Approach [1]

Pradit Mittrapiyanuruk, Chatchawarn Hansakunbuntheung,
Virongrong Tesprasit and Virach Sornlertlamvanich
Information R&D Division,
National Electronics and Computer Technology Center (NECTEC)
Gypsum Metropolitan Building, 22nd Floor,
539/2 Sri Ayudhaya Road, Rajthevi, Bangkok 10400, Thailand
(pmittrap, chatchawarnh)@notes.nectec.or.th, (virong, virach)@nectec.or.th

ABSTRACT – This paper presents all the essential issues in developing the text-to-speech synthesis for Thai - text analysis, prosody generation and speech synthesis. In the text analysis, problems in Thai text processing can be decomposed into the models of sentence extraction, phrase boundary determination and grapheme-to-phoneme conversion. The syllable duration and F0 contour generation rules are included in the prosody generation. This is to realize the synthetic speech in the suprasegmental level. In the speech synthesis, the definition and the construction of acoustic inventory structure 'demisyllable' are presented. Furthermore, three signal-processing algorithms, amplitude normalization, the segment boundary smoothing and prosodic modification, are also presented in this topic.

KEY WORDS -- Thai text-to-speech synthesis, text analysis, prosody generation, speech synthesis, demisyllable

บทคัดย่อ -- **บทความนี้นำเสนอหัวข้อสำคัญในการวิจัยและพัฒนาระบบสังเคราะห์เสียงพูดจากข้อความภาษาไทย ประกอบด้วยการวิเคราะห์ข้อความ, การสังเคราะห์สัทสัมพันธ์และการสังเคราะห์สัญญาณเสียงพูด ในหัวข้อการวิเคราะห์ข้อความจะกล่าวถึงปัญหาที่สำคัญในการประมวลผลข้อความภาษาไทยและรายละเอียดของส่วนประกอบภายในซึ่งประกอบด้วย 3 ส่วน ได้แก่ การตัดประโยค การหาขอบเขตวลีเพื่อหยุดเว้นวรรคการอ่าน และการแปลงรูปเขียนเป็นรูปเสียงอ่าน ในหัวข้อการสังเคราะห์สัทสัมพันธ์จะกล่าวถึงกฎในการกำหนดช่วงเวลาของพยางค์และ F0 contour ซึ่งจะทำให้สามารถสังเคราะห์เสียงที่มีความสัมพันธ์ในระดับเหนือหน่วยเสียงได้ ส่วนหัวข้อการสังเคราะห์สัญญาณเสียงพูดจะกล่าวถึงโครงสร้างหน่วยเสียงแบบครึ่งพยางค์และอัลกอริทึมทางการประมวลผลสัญญาณในการปรับสัญญาณที่รอยต่อให้ต่อเนื่องและปรับสัญญาณให้มีสัทสัมพันธ์ตามที่ได้กำหนดมา**

คำสำคัญ -- การสังเคราะห์เสียงพูดจากข้อความภาษาไทย, การวิเคราะห์ข้อความ, การสังเคราะห์สัทสัมพันธ์, การสังเคราะห์สัญญาณเสียงพูด, ครึ่งพยางค์

1. Introduction

Text-to-speech synthesis is a module or system or machine that converts the input text into the acoustic speech signal that people can understand. Many kinds of applications utilized from this system are developed such as the applications for blind people e.g. screen reader, or the applications for normal people e.g. electronic mail reader using telephone interface, etc. Most of the text-to-speech synthesis systems are developed for converting the text for major languages such as English, Chinese, Japanese and the European languages. At the present, there are only few systems developed for the Thai language. Most of them lack for the continuity in their milestone and some focus on the specific point rather than the whole picture. As a result, there is no Thai text-to-speech synthesis system using in the real application. To overcome this obstacle, this work attempts to put together the jigsaw to form a complete picture. Our goal is to produce a text-to-speech synthesis that can synthesize a natural sound.

[1] This article is a reprint of the article appeared in the Proceedings of NECTEC Annual Conference 2000: ECTI Technologies for New Economics, June 2000, pp. 483-495. This paper award in category of "Best Presentation".

In retrospect, there are some research works related to Thai text-to-speech synthesis. The Luksaneeyanawin's system [1] consists of three main modules. First is the Thai text processing module, it converts a string of Thai text into a string of Thai phonological units using the syllable, word and phrase parsers. Second is the sound dictionary module. It looks up the synthesis unit for the corresponding phonological unit. Third is the synthesis by waveform concatenation module. It synthesizes the speech by using the waveform concatenation technique. Taisertavattanakul and Kanawaree [2] developed a simple but practical system. The system contains (1) the text-to-phoneme analyzer by using conversion rules and a small dictionary for exceptional words (2) the synthesizer concatenates the speech waveform from the demisyllable based acoustic inventory.

The other research works that focus on some specific points rather than the whole system are [3, 4, 5]. Kiat-arpakul, Fakcharoenphol and Keretho [3] proposes an acoustic inventory structure for Thai speech synthesis. In this work, a syllable waveform is created from the concatenation of the phoneme-based and the demisyllable-based units. Luksaneeyanawin [4] proposes a technique to transform the tonal patterns of any syllable speech units by PSOLA-based resynthesizing F0 contour. This technique takes the advantages in the reducing the number of synthesis units about 5 times. This technique stores only toneme syllabic units and synthesizes other toneme speech from these toneme units. Hansakunbuntheung [5] applies the line spectrum pair to the Thai syllabic speech synthesizer. The sound units are encoded in the form of the 20^{th} order LSP and its residues. The synthesizer can synthesize all Thai five tones and adjust the sound duration by using the pitch-synchronous overlap-add (PSOLA) technique. The details of the literature survey in the field of Thai text-to-speech synthesis can be found in Luksaneeyanawin's work [6].

In our work, the system is divided into 3 major parts: text analysis, prosody generation and speech synthesis. The main function of text analysis is to segment the input text into smaller units: sentences and phrases, and then transcribe into the phoneme description. The prosody generation then determines the prosody parameter from the information analyzed by the text analysis. The phoneme description with the prosody parameter of the text is synthesized to the speech waveform by the speech synthesis module. In this module, any synthetic speech is created by the concatenative technique based on the demisyllable units. The signal processing algorithms are involved to produce the natural synthetic speech.

In this paper, the detailed of NECTEC's Thai text-to-speech synthesis is discussed. Most parts of its are already implemented. However, they are being improved in the naturalness. Section 2 discusses the issues in text analysis. The prosody generation based on the rewriting rules is discussed in Section 3. The detail in the acoustic inventory structure and signal processing algorithm are discussed in the speech synthesis topic, Section 4.

2. Issues in Text Analysis

The text analysis is the first part to accept the input text into the system. In practice, the input text to the system may be one or more text paragraphs. Each paragraph consists of sentences. The text may include Thai words, foreign texts (e.g. English), and other special expressions such as numeral texts, abbreviations, punctuation marks, etc. Because the aim of this work is to synthesize the speech of Thai text, the foreign text that appears in the input text will be ignored. It is impractical to process the whole input text all at once due to the limitation in memory resource and processing time. Therefore the text analysis segment the input text into smaller units for processing in other modules. In this work, the text analysis will segment the input text into a sequence of sentences. In the same time, it also determines the phrase boundary, the acceptable position to pause when reading, for synthesizing a natural sound. In addition to determination the sentence and phrase boundary, a module called 'graphme-to-phoneme' in the text analysis is also included. It converts the text into the phonological representation.

2.1 Sentence Extraction

Unlike the English or other European languages, there is no explicit sentence marker in the Thai language. It is convention to insert the space at the end of a sentence in Thai writing. But not all spaces in a paragraph are the sentence marker. They also can be used as other purposes [7,8] such as, using between phrases or cluases within a sentence, between sentences in a cohesive group of sentences, before and after numerals, etc. Mittrapiyanuruk's and Sornlertlamvanich's work [9] extended the algorithm for POS tagging in probabilistic n-gram model to discriminate the sentence break spaces from other purpose spaces. The task can be view as the classification problem. We define the space by its function into 2 different types: sentence break and non-sentence-break space and apply the statistical part-of-speech (POS) tagging as the classifier.

The block diagram of sentence extraction algorithm is shown in Figure 1. The tokenization and word segmentation stage extract a set of tokens with at least one space in between. The spaces in the set of tokens are classified by POS tagging. A token is a sequence of consecutive characters enclosed by the spaces. The special expressions e.g. numerals, abbreviations, punctuation marks, etc. are specially considered. For example, '10600' can be pronounced either in the form of digit-by-digit reading as in the phrase of 'กทม. 10600', or in the form of quantity number as in the phrase of '10600 บาท'. The normalization is needed in this process. Unlike the approaches used for English text, we do not expand this special expression into a normal text. It is embedded in the module of tokenization and grapheme-to-phoneme conversion. The tokenization module draws out this kind of text as a single token. Then the grapheme-to-phoneme conversion assigns a different phonological representation according to the text feature. Moreover, the string of characters is segmented into words because there is no explicit use of word delimiter in general Thai text [10].

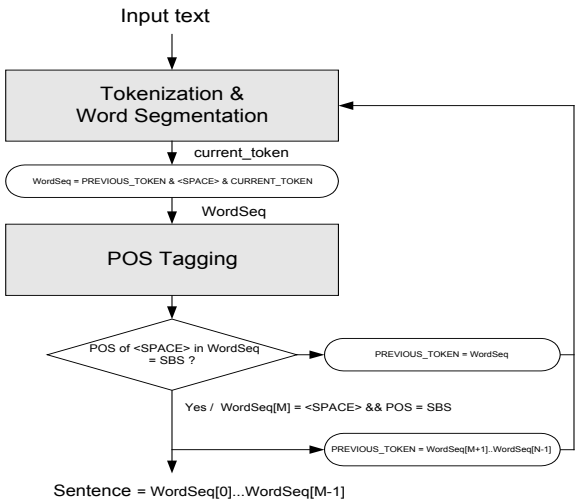

Figure 1. Block diagram of sentence extraction

The two adjacent tokens are reconstructed to the word sequence with a space in between. Any spaces in this word sequence are classified to be one of two possible classes, sentence break or non-sentence-break space. We define this classification problem in terms of statistical POS tagging. The most probable sequence of POSs and individual word-level POS assignments determines the most probable POS assignment of any word sequences. Therefore the classification task is to determine whether the POS of any spaces in the most probable sequence of POS is a sentence-break or not. We use the part-of-speech trigram model to compute the POS sequence probabilities and introduce the viterbi algorithm for computing the most probable sequence of POSs. Because it is possible that the space in the word sequence that used to be the non-sentence-break spaces are incorrectly assigned. Therefore we must scan the space between the current and previous token as well as the spaces within the previous token. If there is no sentence-break space then all of this word sequence will be used as the previous token in the next iteration. But if a sentence-break space is found then the output sentence is the whole sequence from the first word to the word just before the sentence-break space. The rest word sequence after the space is used as the first token in the next iteration. The algorithm will extract the sequence of tokens and detect the type of in-between space until the end of paragraph. It is obvious that this algorithm can solve the problem of memory resource and processing time limitations because it processes token by token rather the whole paragraph at once.

2.2 Phrase boundary determination

Normally when speakers utter a long sentence, they tend to break the sentence up into several phrases. Luksaneeyanawin [1] reports that from the study of charactersics and function of pause, the position of pause occurs averagely at each 8 syllables (S.D.=4). These phrase boundaries capture several prosodic characteristics, such as pausing at the end of phrase, lenthening in duration of phrase final syllable and the downtrend effect on F0 contour at the end of phrase, etc. If the phrase boundary is determined accurately then the prosody generation will undoubtedly play an important role in obtaining more natural speech. To accomplish the phrase boundary determination, the rule-based algorithm is developed using the output from the sentence extraction. The algorithm first determines the tentative phrase break positions in a sentence and then merges the positions using the syllable number constraint.

The algorithm places a tentative phrase break at the position of space and punctuation mark. The algorithm detects the phrase break spaces by using distinctive pattern rules derived from formal Thai writing pattern [11]. The other tentative phrase break positions are determined by scanning each pair of words from left to right in the sentence. The simple rule based on the content word/function word [12] is used to place the preliminary break position before every function word that follows a content word. In this work, the function word is a word of conjunction, preposition and relative pronoun. The content word is the rest that is not match the function word. Furthermore, the rule derived from Luksaneeyanawin's work [1] assigns the break position before/after some specific grammatical word.

At the merging step, the tentative phrase break positions in a sentence are combined to a single phrase if they do not end in the phrase-break space or punctuation and contain 10 or fewer syllables. They are combined to the following phrase until a punctuation mark or phrase-break space is found or until the number of syllables is greater than 10. This scheme of phrase combining is same as Karn's [13].

2.3 Grapheme-to-phoneme conversion

Outputs from the sentence extraction and phrase boundary determination are one sentence with the phrase break positions and the word boundaries. In this step, the phonological representation or phoneme of each word is assigned. First each word is looked up for the phoneme string from the pronunciation dictionary, about 25,000 entries. The letter-to-sound rule is developed to handle the unregistered words. The rule consists of 2 stages. First the grapheme of a word is divided into a syllable sequence and second the syllable sequence is converted to a phoneme string by simple orthographical-to-phonological mapping. The tone of each syllable is assigned by considering the phonological composition (initial consonant, vowel, final consonant) and its orthographical tone marker. Details in the tone assigment rule can be founded in Thavaranon's work [8].

The first step typically called syllabification is developed using the regular expression. Rather than hard coding the rule for each syllable pattern, this work rewrites the rule in the regular expression format. All possible orthographical syllablic structure are listed in the regular expression format and compiled to be deterministic finite state automata by the lexical analyzer or 'LEX'. When it matches a syllabic pattern then the orthographical syllablic composition: initial consonant, vowel final consonant and tone marker, is returned for assigning the phonological representation. The advantages of this scheme are the flexibility in rule modification and the speed of processing time.

3. Issues in Prosody Generation

The prosody means the properties of the acoustical speech such as pitch variation, loudness and syllable length. The effects of prosody are referred to as suprasegmental phenomena [14], since it occurs in higher level than segmental level such as syllable or phoneme. It is acknowledgable by most researchers in this field that the naturalness of synthetic speech is considerably affected by the prosody. Therefore this work essentially includes the prosody generation. Many prosody parameters are generated by determining the pause position and duration and the pitch movement of utterance which represented by F0 contour. The pausing is executed by the phrase boundary determination in text analysis part.

There are two major approaches in the research of prosody generation: the rule-based method and the corpus-based method. In the rule-based approach, linguistic experts derive the factors that affect the prosody event by observing various phenomena in the natural speech, then write the rules that interplay among these factors for synthesizing the more natural speech. On the contrary, the corpus-based approach derives the prosody model from the prosodic annotated speech corpus by using machine learning algorithms such as decision tree, artificial neural network, etc. The prosodic parameters of unseen text are determined by infering from the training corpus. Lacking of the prosodic-labeled speech corpus, our prosody generation is a rule-based approach.

3.1 Syllable duration assigment rule

The first consideration when devise the durational rule is the choice of speech unit that will be affected by the rule. The contexual influences that affected the duration of different speech units are varied. Campbell and Isard [15] argue that the syllable is a suitable unit that reflects the rhythm of any utterances. This approach first predicts the syllable duration and then the smaller segment duration e.g. phone is determined from its syllable duration. Because the speech synthesizer that we use in this work is the demisyllable-based concatenative system, the speech waveform is formed by the sequence of syllabic sounds. Each syllablic waveform is created by concatenating two demisyllable units: initial and final unit. Then the timing of synthetic speech appears in the syllabic time frame. There are many linguistic works [16] conclude that Thai is a syllable-time rhythm in which the syllable is an intuitively recognizable unit for primitive people. Therefore we select the syllable as the speech unit for modeling the duration.

In timing aspect, the naturalness of any utterance occurs when the duration of every syllable in the phrase is relatively suitable. In any slow and fast utterance, the duration of syllable differs only in the absolute value but the relative value is almost the same. To accomplish this task we tailor the most favorite scheme [17] to the prosodic generation module in the syllabic framework. This scheme first assigns the base syllable duration from its intrinsic property. Then the rules are used to multiply the base duration by a specific factor. These factors are devised by investigating the natural speech in word, phrase and sentence level. In this scheme we can adjust the speaking rate by multiplying the factor to the intrisic duration without changing the rule.

For finding the intrinsic duration of each syllable, it is laborious to acquire the intrinsic duration of all Thai syllablic sounds because of its plentitude of units, which is about 27,000 [6]. To realize this process, we classify the consonants by the manner of articulation into 8 types and the vowels by the tongue advancement/short/long attribute into 12 types and use mid tone (tone 0). By the assumption that the syllables in the same group have the same intrinsic duration, we use the duration of each group representative as the duration of every syllable in the group. This method reduces the number of syllable duration patterns to 384 patterns. The duration of each unit is taken from the carrier syllable in the medial position of pronunciation. The intrinsic interval of all syllables is extrapolated using the value of the representative that has the same kind of initial consonant, final consonant and vowel. The duration of falling (tone 2) and rising (tone 4) tonal syllable is scaled-up by factor 4/3 to compensate the tonal-durational interactive effect. Moreover we also measure intrinsic pause duration which is divided into 3 types: pre-plosive pause, glottal closure pause and end of phrase pause in the same way.

For the details of rule, they are derived from Klatt's work [17]. The rule in phrase and sentence level are the same which is lengthening the phrase-final syllable duration by the factor of 1.2 and inserting the pause at the end of phrase with the intrinsic phrase pause duration. At the end of sentence, the pause duration is longer than the intrinsic phrase pause duration by the factor of 1.2. For the rule in word level, the syllables in any words that are not the non word-final position are shortened by the factor of 0.9. Other syllable, any syllables in a polysyllabic word are shortened by the factor of 0.9. The last rule considers the effect of postvocalic consonant context. It shortens the duration of syllables followed by the voiceless consonant by the factor of 0.9. Noted that these rules apply sequentially by cumulatively multiplying the initial duration with each specfied rule's factor to obtain the final duration.

3.2 F0 contour Generation rule

In natural speech, the speech is continuously uttered as strings of breaths. Each string, called phrase, consists of many sound units. The types of sound unit can be words, syllables or phonemes, etc. depending on the design purpose. Considering a particular type of the sound units, since the sound units in a phrase are produced in the same utterance, they must share some common characteristics. A characteristic, called intonation, is in the suprasegmental level of speech, relates to the tonal phenomena that affect on F0 contour of the continuous speech. In addition, there is another effect on syllabic level. There is a tone pattern when syllables are connected.

3.2.1 Intonation Rules

In the suprasegmental level, two groups of rules are defined. The first group is the downdrift phenomenon that defines how the F0 contour decreases relatively with the preceding

time. Another group concerns the pitch range of F0 contour that limits the boundary of F0 contour.

3.2.1.1 Downdrift

A downdrift, can be observed in the F0 contour across a phrase [11,18-24]. [25] shows that this downdrift also happens in Thai speech as shown in Figure 2. This phenomenon can be observed by plotting the F0 contour of a phrase containing only the mid tone (tone 0). The plot shows the downdrift effect on the F0 contour which looks like the steps of similar patterns. The reference line connects all the beginning points of F0 contours of the syllables.

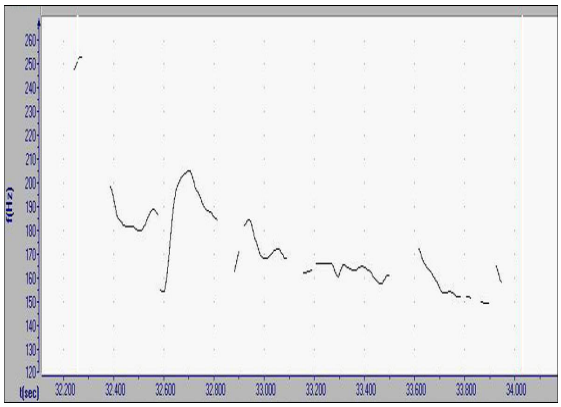

Figure 2. An example of downdrift on F0 contour of Thai speech

To simplify the effect, the downdrift is estimated by a linear declining slope of the F0 contour. Since one system is for a female speech, the slope was computed for a female speech prototype. As a result of the experiment, the declination is 30 hertz per second. This declination can be represented by a linear line, called a reference line, as shown in Figure 3. This line is used as the reference of the starting or ending point of F0 contour of each tone.

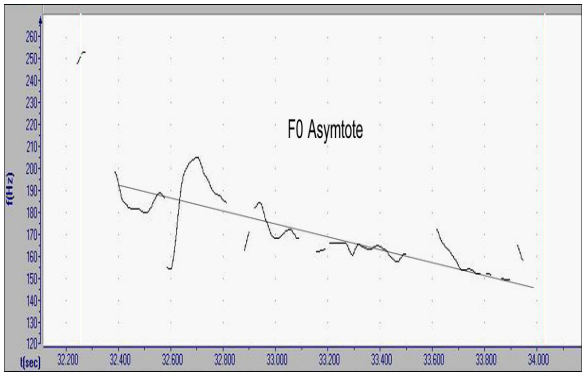

Figure 3 A reference line that expresses a simiplified version of downdrift

3.2.1.2 Pitch Range

An obvious difference between male and female speech is the pitch range. Generally, a female speech is more perceptible higher than a male speech. The pitch range specifies how high and how low the pitch level can reach.

To determine a pitch range, an observation on a female's speech prototype is done by measuring the maxima and minima of F0 levels. The boundaries of pitch range are applied to the system to limit the level of synthetic F0 contour with downdrift effect. If F0 value of syllabic F0 contour exceed these limits, the level of the syllabic F0 contour will be reset to the starting F0 of the same phrase.

3.2.2 Tone Rules

After processing on the suprasegmental level, here, a syllabic level will be discussed. In this level, there are two parts. The first part explains where the tone contours should be located and another part explaining the effect of adjacent syllables on a tone contour.

3.2.2.1 Tonal contour location

This part explains how to locate the tone contours. When tone contours are concatenated, the locations of tone contours are different depending on the situation. In a reading speech, the Mid, Low and Rise tone start at the reference line while the others end at the reference line. However, there are some special cases that the tone contours do not conform to this rule. If there is a stress syllable in a phrase, its tone contour level will be shifted up. Since only the reading speech synthesizing is the goal of this system, all tone contours are based on this rule.

3.2.2.2 Coarticulation Effect

When a syllabic sound is voiced, the following one is effected and vice versa. This effect on connected speech has been reported by Gandour, Potisuk and Dechongkit [26]. The study on tonal height reports that the anticipatory effect extends forward to about 75% of the duration of the following syllable and, similarly, the carry-over effect extends backward to about 50% of the duration of preceding syllable. This work uses the above duration during smoothing the F0 contour at the syllabic junction.

4. Issues in Speech Synthesis

After all necessary parameters for the synthesis are determined. This part will use these parameters to determine which sound units should be selected and how these units should be processed to synthesize a high natural synthetic speech. The parameters can be classified into two groups. First group is generated from the text analysis consisting of phonemic lists of a phrase. These lists are used to select the relevant units. Another one is generated from the prosody generation consisting of the duration and the F0 contours. All are used in signal processing to improve the naturalness of synthetic speech. Before describing the detail of the synthesis techniques, the synthesis unit structure will be detailed.

4.1 Synthesis Scheme

In the synthesis work, there are various types of synthesis units used in concatenative speech synthesis such as words, syllables, demisyllables, phonemes, diphones, triphones etc. Each type has different advantages depending on the purpose of each system. In this work, demisyllable is selected because it has a reasonable number of sound units and acceptable quality. Although, its sound quality at syllable boundary is not quite natural as real speech but this problem can probably

be improved by signal processing as being presented in this system.

4.1.1 Demisyllable

Demisyllable is the unit being the initial and final halves of a syllable. On the idea that a speech waveform is constructed by splicing the syllabic segment. A syllabic waveform is created from the proper initial and final demisyllable unit. Both units are segmented from a syllable at the stable vowel part. In general, Thai syllable has a structure of "C(C)VC" [1], therefore, the syllable is segmented into two portions, "C(C)V" and "VC". The initial unit contains a single consonant or double consonants linking with a partial vowel. The final unit contains a partial vowel linking with a final consonant and, also, tonal characteristic. Figure 4 shows an example of a demisyllable unit.

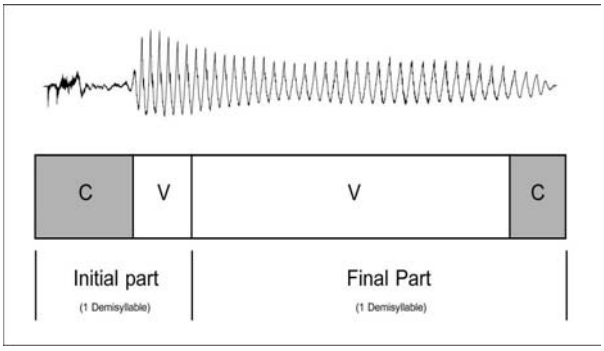

Figure 4. Demisyllable unit structure

4.1.2 Demisyllable inventory structure

In the previous section, we have already explained what the demisyllable-based concatenative synthesis is and how it works. Therefore, in this section we will discuss about how to list all units in the inventory and how many units are necessary in this system to construct the entire Thai syllables.

A Thai syllable sound can generally be characterized by four elements: initial consonant, vowel, final consonant and tone. Considering the Thai syllable structure, there are 4 patterns; CV (ปา [pa;0], รี [ri;1]), CCV (ปลา [pla;0], ครู [khru;0]), CVC (ปาด [pa;t1], กาก [ka;k1]), CCVC (ปราบ [pra;t1], กวาด [kwa;t1]). The first letter "C" represents an initial consonant while "CC" denotes a consonant cluster. The vowel of a syllable is represented by the letter "V" and the final consonant represented by the last letter "C". The Arabic numerals (0-4) represents the tone.

- Initial consonant: There are 44 consonantal letters in Thai. These letter represent 21 phonemes, grouped by traditional Thai grammarians into 3 classes; the high class, the middle class and the low class. These classes are very important in determining the tone of a syllable [6]. All of them can occur in the initial position such as กาด [ka;t], บน [bon], โยน [jo;n], etc. Some phonemes can be clustered to produce two different types of sound together such as ปล- /pl/, คว- /khw/, ตร- /tr/, etc. Our system also includes some phonemes for producing some loan words namely, consonant clusters and final consonants such as ดร- /dr/, ฟร- /fr/, บร- /br/, -ฟ /f/, -ล /l/, -ส /s/, etc. The list of Thai consonants is shown in Table 1 and 2.
- Vowel: There are 2 types of vowel; monophthong and diphthong. Monophthong can be classified into 2 groups; 9 long vowel sounds such as –า /a;/, ู /-u;/, แ- /x;/ and 9 short vowel sounds such as –ะ /a/, ุ /u/, แ-ะ /x;/, etc. Diphthong consists of 6 vowel sounds; 3 long vowel sounds such as เ-ีย /i;a/, เ-ือ /v;a/, -ัว /u;a/ and 3 short vowel sounds such as เ-ียะ /ia/, เ-ือะ /va/, -ัวะ /ua/ [27]. The list of Thai vowels is shown in Table 3.
- Final consonant: The final consonant consists of 9 phonemes, including open syllable such as ราด [ra;t1], กับ [kap1], จะ [ca?1], ยาว [ja;w0], etc. and 4 phonemes come from foreign language such as ball [bol], half [haf]. The reason of adding some borrowed phonemes in our inventory is that pronouncing a loan word close to the native pronunciation is preferable.
- Tone: There are 5 levels in Thai phonology represented in our inventory with the Arabic numerals (0-4); low-level [◌̀ , 1], mid-level [no tone marker, 0], high-level [◌̋ , 3], falling [◌̂ 2] and rising [◌̌ , 4]. The syllable structure and the initial consonantal phoneme are very important to assign a tone for each syllable.

In this work, the demisyllable-based inventory for the initial part is constructed by creating all combination of 38 initial consonantal phonemes and 9 monophthongs (only short vowels sounds), resulting 342 units. The reason why only the short monophthongs are selected is the characteristics of short and long vowels are the same at the beginning. For final part, there are 1,163 units divided into 2 sets; 804 Thai phonemes and 359 phonemes of loan words. As a result, the total number of our inventory is 1,505 units. The combination is shown in table 4.

4.1.3 Construction phase

After a sound inventory has been designed, the next step is to select the speaker. A female speaker is selected based on the result of the listening test. All test sentences are read by 5 female speakers at a natural speaking speed and recorded directly to the computer. The speaker whose voice is natural and correctly pronounced in Thai standard is finally selected.

In our early work, the speaker pronounced a set of meaningless syllable sounds or logatoms and the recording was done in the office room environment, using a high quality microphone. The result was unsatisfactory because the synthesized speech was over stressed. Moreover, audible discontinuities occurred at the concatenated boundary. In the current work, we improve our synthesized speech by recording the target syllable with a frame sentence rather than syllable unit to solve the over stress problem and recording in

[1] This article is a reprint of the article appeared in the Proceedings of NECTEC Annual Conference 2000: ECTI Technologies for New Economics, June 2000, pp. 483-495. This paper award in category of "Best Presentation".

the studio to get the clear voice. The female speaker reads each sentence 3 times for recording an initial part and a final part. From the recorded files, select the best one for manual segmentation. To segment a syllable into the initial and the final parts we have to find the segmenting point by listening and using the spectrogram to cut at a zero crossing point.

This process is an important step that has an effect on the quality of synthetic speech. However, such defects can be corrected by digital signal processing technique.

4.2 Speech Signal Processing

Applying signal-processing techniques to the synthesis part is a way to improve the naturalness. We apply the signal processing techniques (1) to normalize the amplitude of sound units to the same scale, (2) to capture the prosody parameter from the prosody generator into the synthetic speech waveform (this function makes change to pitch contour (F0 contour) and duration of speech signal.), (3) to smooth the discontinuity at the vicinity of concatenation. The considered discontinuities are the pitch variation, spectral mismatch and amplitude abruption.

4.2.1 Amplitude normalization

In a concatenation of sound units, the problem about abrupt change of amplitude at the concatenating point between the units. This problem occurs because each unit comes different syllable. For this reason, each unit has different amplitude. This defect makes synthetic speech sound fluctuated. To decrease this defect, this system normalizes the amplitude of all demisyllable-base inventory units to be the same standard.

First, the standard amplitude of each vowel is calculated. The frame sentence "เธอบอกให้พูด…ไปเรื่อยๆ" (/thq:?0 b@:k1 haj2 phu:t2 …. paj0 rv:aj2 rv:aj2/) with different target vowels in between is selected to pronounce and record for measuring the reference amplitude of the corresponding vowel. Then, each sentence is multiplied by a ratio that makes the mean of amplitude of the syllable that preceding the reference vowel, in this case is "พูด" (/phu:t2/), to be the same. Now, all reference vowel sounds present their own amplitude characteristic. These adjusted reference vowel sounds are used in calculating the amplitude reference of each vowel. Finally, these references are used in multiplying all demisyllable-based inventory units to have the same amplitude at the junction.

4.2.2 Prosodic modification

To modify prosody, this system uses Time-Domain Pitch Synchronous Overlap-Add (TD-PSOLA) [28] technique, a widespread technique which modifies pitch and duration of synthetic speech. This technique is based on dividing speech into subframes that partially overlapped on each other and each subframe is synchronized using pitch. To alter prosody, these subframes are slid to the preferred positions to modify prosody.

This technique is divided into three steps. First step, each unit in the entire inventory is marked at pitches in voiced part and, for unvoiced part, at virtual pitches every 4 millisecond as shown in Figure 5. All pitch markers are also labeled with its type whether voiced or unvoiced. Second step, the new pitch contour (F0 contour) is calculated. Actually, this new pitch contour was prepared from prosodic generating part. Next, the new pitch contour will be mapped with the old one using linear interpolation as shown in figure 6. This interpolation will specify which old pitch should be mapped to the new one so there are some old pitch missing or duplicating depending on this modification be to decrease or to increase duration. Final step, window function will process at each marked pitch on the original inventory units as shown in figure 7. Each window function is two-pitchwidth hanning window. By changing the displacement between marked pitch and overlapping some part of window at the edge, the prosody-modified synthetic speech is generated.

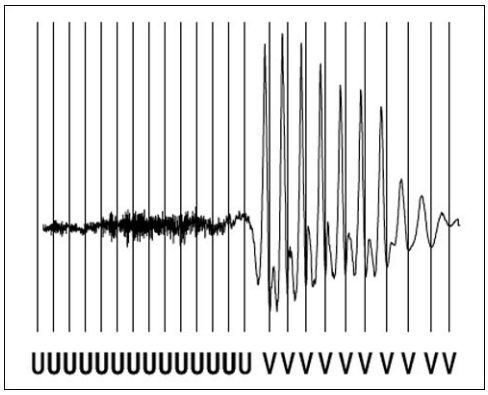

Figure 5. Pitch marking and labeling

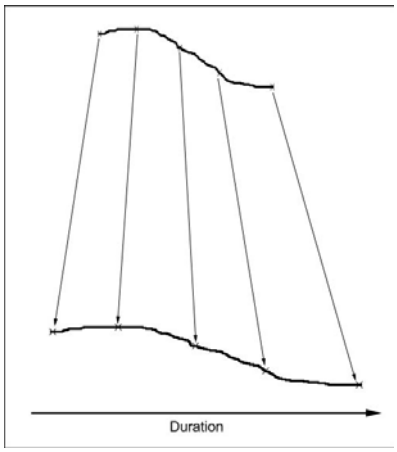

Figure 6 Duration Scaling.

[1] This article is a reprint of the article appeared in the Proceedings of NECTEC Annual Conference 2000: ECTI Technologies for New Economics, June 2000, pp. 483-495. This paper award in category of "Best Presentation".

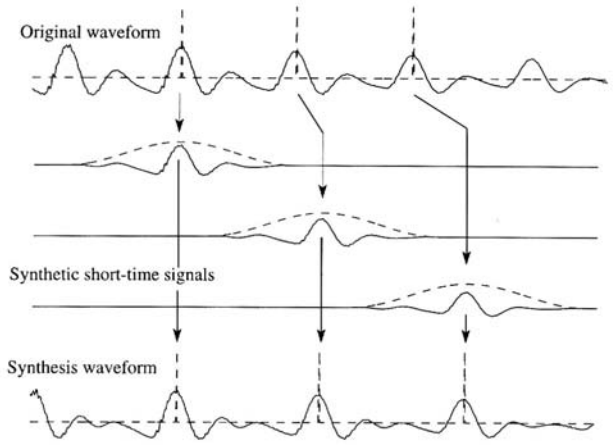

Figure 7. TD-PSOLA

4.2.3 Concatenated boundary smoothing

Due to the inventory unit used in this system is demisyllabic unit, so there appears some quality problems that happen at the intrasyllabic and intersyllabic concatenated point. These problems are a discontinuity of pitch, amplitude and, especially, spectrum. For the pitch discontinuity, it was solved in the prosodic modification part and, also, the amplitude discontinuity was already solved in the amplitude normalization part. The discontinuity of pitch and amplitude can be solved in time domain while the spectrum discontinuity must be solved in frequency domain. To solve the spectrum discontinuity, the speech signal is transformed to some representation in frequency domain. In this system, the Line Spectrum Pairs (LSP) [29] representation derived from LPC coefficients is selected.

The advantages of this representation are; (1)its parameters correspond to speech fomants, work like formant coding, (2) this representation is stable on interpolation. Figure 8 shown an example of a relation between formants computed by using LPC and their corresponding LSP parameters.

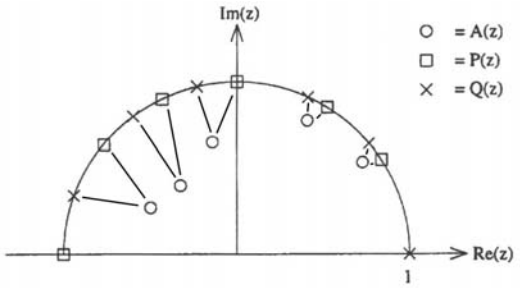

Figure 8. Example of relation between formants (circles) and its LSP parameters (squares and crosses)

This smoothing method is applied to this system to smooth the synthetic speech at the connecting points. To implement this method, first, several subframes at the junction are removed from the preceding demisyllable and the following one. Then the LSP parameter of subframes at the edges are computed as the reference parameters. The linear interpolation between these reference parameters is calculated to replace the removed subframes. This method is shown in figure 9.

4.2.4 Cross-syllable coarticulation modeling

When more than one syllable are connected together, there are some cross-syllable coarticualtion between adjacent syllables. These phenomena make the natural speech distinct from the synthetic speech. The effects of these phenomena are classified into two types. The first type is a prosody alteration, which is effected by the adjacent syllables. This effect was computed in the prosodic modification part. Another effect is a waveform interaction with neighbor syllables. Since the natural speech is the continuous speech but not the syllabic speech. Some syllabic speech signal is continuously transformed to the adjacent one. To improve the quality of the synthetic speech, this system includes these effects in synthesis part.

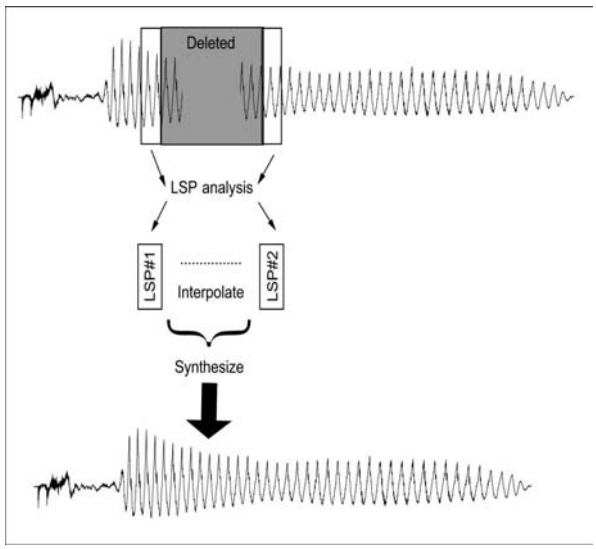

Figure 9. LSP Smoothing method

There are two connection types derived by investigating the natural speech. The first type is a simple touching. This occurs when the initial consonant of preceding syllable or final consonant of next one is unvoiced. Another one is an assimilated connection. This occurs when both the initial consonant of preceding syllable and final consonant of next one are voiced. In the implementation, this system uses the digital signal processing technique, LSP smoothing, as described above to simulate these connecting. The example is shown in figure 10. Figure (10a) shows the simple touching type and the assimilated connecting type is shown in figure (10b).

[1] This article is a reprint of the article appeared in the Proceedings of NECTEC Annual Conference 2000: ECTI Technologies for New Economics, June 2000, pp. 483-495. This paper award in category of "Best Presentation".

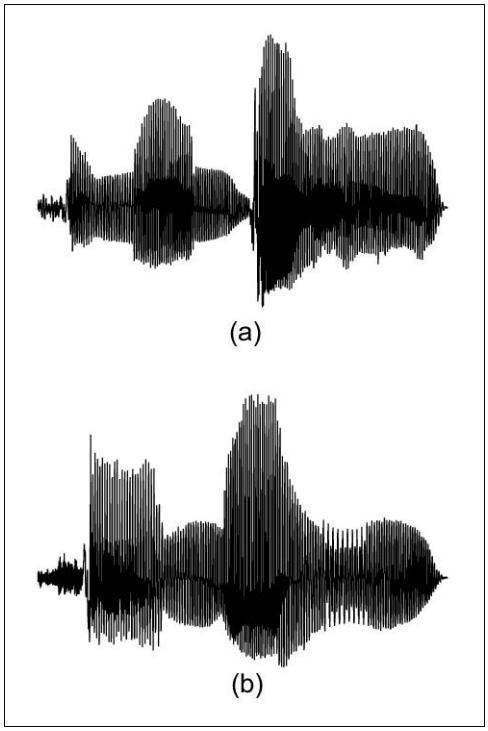

Figure 10. Syllable connections:
(a) a simple touching
(b) an assimilated connection

5. Future work

After implementing the system with the above approach and evaluating the synthetic speech, we found that the quality was acceptable. However, the improvement of the naturalness of the speech is suggested.

In the aspect of text analysis, there are two major points: the input text segmentation and the grapheme-to-phoneme conversion that need improvement. The improvement in grapheme-to-phoneme conversion will boost the system in the way to synthesize the correctly pronounced speech but not the naturalness. The most important problem is how to handle the homograph disambiguation. The homograph is that a word has more than one possible phonological representation such as the word 'เพลา' can be transcribed into '/phe:0-la:0/' or '/phlaw3/' depending on its context. The input text segmentaion includes the tasks of sentence extraction and phrase break determination. This task has an affect on the naturalness of synthetic speech because its result implies the determination of the pause position that has the major role in the prosody generation. The break positions are so ambiguous that even the native Thai people also can not judge decisively whether they are actual break. The rule-based approach has been introduced for the task. An advantage of rule based approach is that it is easy to developed. But it also has a drawback in handling the problem that has several interacting factors and high degree in ambiguity like the prosody parameter prediction task. The statistical or corpus based method is another alternative. In the future, we plan to apply the corpus-based approach to both the text analysis and prosody generation.

In the aspect of speech synthesis, other acoustic inventory stucture such as diphone, triphone, syllable and the non-uniform unit, etc. is in our consideration. Also the automatic unit selection algorithm which works well in other language systems is studied to replace the manual speech segmentation in inventory construction phase. Furthermore, other advance topics such as the voice transformation and more sophisticated synthesizer are also our interesting topics.

To pursue these future works, it is evident that the large-scale prosody-labeled speech corpus be indispensible for us. Our next step is to design and to develop the speech corpus which be labeled with the complete information.

Acknowledgement

We would like to thank Ms.Tanakorn Wiboon, an intern student from Department of Computer Engineering, Kasertsart University, for her contribution in the implementation of grapheme-to-phoneme conversion module.

References

[1] Luksaneeyanawin, S. et., al., 1992. A Thai text-to-speech system, Proceeding of 4th NECTEC Conference, pp.65-78 (in Thai).

[2] Taisertavattanakul, S. and Kanawaree, W., 1995, Thai speech synthesizer. Unpublished senior project report, Department of Computer Engineering, Chulalongkorn University (in Thai).

[3] Kiat-arpakul, R., Fakcharoenphol, J. and Keretho, S., 1995, A combined phoneme-based and demisyllable approach for Thai speech synthesis, Proceedings of the 2nd Symposium on Natural Language Processing SNLP'95, pp 361-369.

[4] Luksaneeyanawin, S., 1995, Tone transformation, Proceedings of the 2nd Symposium on Natural Language Processing SNLP'95, pp. 345-360.

[5] Hansakunbuntheung, C., Leelarasmee, E., 1999, Thai syllabic speech synthesis based on line spectrum pair, 22nd Thailand Electrical Engineering Conference, pp.521-524 (in Thai).

[6] Luksaneeyanawin, S., 1993. Speech computing and speech technology in Thailand, Proceedings of the Symposium on Natural Language Processing in Thailand, pp.276-321.

[7] Danvivathana, N., 1987, The Thai writing system, Forum Phoneticum 39, Helmut Buske Verlag Hamburg.

[8] Thavaranon, K., 1978, Spacing in Thai Writing, M.A.Thesis Department of Thai Chulalongkorn University (in Thai).

[9] Mittrapiyanuruk, P. and Sornlertlamvanich, V., 2000, The automatic Thai sentence extraction, Proceeding of 4th Symposium on Natural Language Processing (SNLP'2000).

[1] This article is a reprint of the article appeared in the Proceedings of NECTEC Annual Conference 2000: ECTI Technologies for New Economics, June 2000, pp. 483-495. This paper award in category of "Best Presentation".

[10] Sornlertlamvanich, V., 1993, Word segmentation for Thai in machine translation system, Machine Translation, NECTEC pp 556-561 (in Thai).
[11] Thavaranon, K., 1978, Spacing in Thai writing, Master Thesis, Department of Thai, Chulalongkorn University (in Thai).
[12] Sileverman, K., 1987, The Structure and Processing of Fundamental Frequency Contours, Ph.D. Thesis, University of Cambridge.
[13] Karn, H., 1996, Design and evaluation of a phonological phrase parser for Spanish text-to-speech, Proceedings of the Fourth International Conference on Spoken Language Processing, Vol. 3, pp. 1696-1699.
[14] Dutoit, T., 1997, Introduction to text-to-speech synthesis, Kluwer Academic Publishers.
[15] Campbell, W.N. and Isard, S.D., 1991, Segment durations in a syllable frame, Journal of Phonetics, Vol. 19, pp37-47.
[16] Luangthongkum, T., 1977, Rhythm in standard Thai, Unpublished Ph.D. Thesis, Univerity of Edinburge.
[17] Klatt, D.H., 1987, Review of text to speech synthesis conversion for English, Journal of Acoustic Society America, Vol 82, pp.737-793.
[18] t'Hart, J., and Cohen, A., 1973, Intonation by rule: a perceptual quest., Journal of Phonetics, 1:309:327.
[19] t'Hart, J., and Collier, R., 1975, Integrating different levels of intonation analysis, Journal of Phonetics, 3:235-255.
[20] Pierrehumbert, J. B., 1980, The phonology and phonetics of English intonation., PhD Thesis, Published by University of Edinburgh.
[21] Cooper, W.E. and Sorensen,J.M., 1981, Fundamental Frequency in Sentence Production., Springer-Verlag., 1981.
[22] Liberman, M. and Pierrehumbert, J., 1984, Intonational invariance under changes in pitch range and length., In Aronoff, M. and Oehrle, R T., editors, Language Sound Structure., MIT Press.
[23] Fujisaki, H. and Kawai, H., 1988, Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese, In International Conference on Speech and Signal Processing. IEEE.
[24] Taylor, P.A., 1992, A Phonetic Model of English Intonation, PhD Thesis, Edinburgh.
[25] Luksaneeyanawin, S., 1983, Intanation in Thai, Unpublished PhD Thesis, University of Edinburgh.
[26] Gandour, J. T., Potisuk, S., and Dechongkit, S., 1994, Tonal Coarticulation in Thai, Journal of Phonetics, vol 22, pp.477-492.
[27] Khanitthanan, W., 1990, Phasa lae Phasasart, Thammasat University Press. (in Thai)
[28] Charpentier, F. and Moulines, E., 1989, Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones, European Conference on Speech Communication and Technology, vol. I, pp. 013-019.
[29] Klejin W.B. and Paliwal, K.K., 1995, Speech coding and synthesis, Elsevier Science.

Virach Sornlertlamvanich is the acting director of Information Research and Development Division of the National Electronics and Computer Technology(NECTEC) of Thailand since 1992. He received the B.Eng. and M.Eng. degrees from Kyoto University, in 1984 and 1986, respectively. From 1988 to 1992, he joined NEC Corporation and involved in the Multi-lingual Machine Translation Project supported by MITI. He received the D.Eng. degree from Tokyo Institute of Technology in 1998. His research interests are natural language processing, lexical acquisition and information retrieval.

Pradit Mittrapiyanuruk received bachelor degree in electrical engineering from King Mongkut's University of Technology Thonburi (KMUTT) in 1994 and master degree in electrical engineering Chulalongkorn University in 1996. Then he joined NECTEC in August 1996. He had involved in the projects i.e. Integrated Receiver&Decoder (IRD), Text Retrieval Database, and speech synthesis. Currently, he mainly works for the NECTEC's Thai text-to-speech synthesis project. His research interests are speech synthesis, speech recognition and multimedia signal processing.

Chatchawarn Hansakunbuntheung received his Bachelor degree and Master degree in electrical engineering from Chulalongkorn University in 1998 and 2000 respectively. He has joined the NECTEC at the Software and Language Engineering Laboratory (SLL) since 2000. He started his work in the research and development group of the Thai Text-to-Speech project since. At the present, he is involved in Thai Text-to-speech project and Thai speech corpus project. His research interests are speech technology and Natural Language processing.

Virongrong Tesprasit has joined NECTEC in April 1996 after receiving her BA.(Linguistics) degree from Thammasat University. She had joined both Royal Institute Dictionary Development Network Project and Development of Thai Corpus Base Project. At present, she researches on Thai Text-to-Speech synthesis Project. Her research interests are Phonetics and Speech Technology.

[1] This article is a reprint of the article appeared in the Proceedings of NECTEC Annual Conference 2000: ECTI Technologies for New Economics, June 2000, pp. 483-495. This paper award in category of "Best Presentation".

Table 1. Phonetic symbol of Thai consonant

Thai Letter	Phonetic Symbol	
	Initial	Final (including open syllable)
ก	/k/	/k/
ข, ฃ, ค, ฅ, ฆ	/kh/	
ง	/ng/	/ng/
จ	/c/	
ฉ, ช, ฌ	/ch/	
ซ, ศ, ษ, ส	/s/	
ญ, ย	/j/	/j/
ฎ, ด	/d/	/t/
ฏ, ต	/t/	
ฐ, ฑ, ฒ, ถ, ท, ธ	/th/	
ณ, น	/n/	/n/
บ	/b/	/p/
ป	/p/	
พ, ภ, ผ	/ph/	
ฟ, ฝ	/f/	
ม	/m/	/m/
ร	/r/	
ล, ฬ	/l/	
ว	/w/	/w/
ห, ฮ	/h/	-
อ	/?/	-

Table 2. Consonant Cluster

Thai Letter	Phonetic Symbol of Consonant Cluster		English Letter	Phonetic Symbol of Consonant Cluster	
	Initial	Final		Initial	Final
ปร-	/pr/	-	br-	/br/	-
ปล-	/pl/	-	bl-	/bl/	-
ตร-	/tr/	-	fr-	/fr/	-
กร-	/kr/	-	fl-	/fl/	-
กล-	/kl/	-	dr-	/dr/	-
กว-	/kw/	-	f-	-	/f/
พร-, ผร-	/phr/	-	l-	-	/l/
พล-, ผล-	/phl/	-	s-	-	/s/
ทร-	/thr/	-	ch-	-	/ch/
คร-, ขร-	/khr/	-	-	-	-
คล-, ขล-	/khl/	-	-	-	-
คว-	/khw/	-	-	-	-

[1] This article is a reprint of the article appeared in the Proceedings of NECTEC Annual Conference 2000: ECTI Technologies for New Economics, June 2000, pp. 483-495. This paper award in category of "Best Presentation".

Table 3. Phonetic Symbol of Thai Vowel

Monophthong				Diphthong				Vowel Letter			
Short Vowel		Long Vowel		Short Vowel		Long Vowel		Short Vowel		Long Vowel	
◌ะ	/a/	◌า	/a;/	เ◌ียะ	/ia/	เ◌ีย	/i;a/	◌ำ	/am/	-	-
◌ิ	/i/	◌ี	/i;/	เ◌ือะ	/va/	เ◌ือ	/v;a/	ไ◌, ใ◌	/aj/	-	-
◌ึ	/v/	◌ื	/v;/	◌ัวะ	/ua/	◌ัว	/u;a/	เ◌า	/aw/	-	-
◌ุ	/u/	◌ู	/u;/								
เ◌ะ	/e/	เ◌	/e;/								
แ◌ะ	/x/	แ◌	/x;/								
โ◌ะ	/o/	โ◌	/o;/								
เ◌าะ	/@/	◌อ	/@;/								
เ◌อะ	/#/	เ◌อ	/#;/								

Table 4. Combination of Demisyllable based Inventory for Final Part

Vowel	Final Consonant		
	Dead Syllable (3)	Live Syllable (5)	Open Syllable (1)
Monophthong			
Short vowels (9)	Mid, Low, Falling, High	Mid, Low, Falling, High	Mid, Low, Falling, High
Long vowels (9)	Mid, Low, Falling, High	Mid, Low, Falling, High, Rising	Mid, Low, Falling, High, Rising
Diphthong			
Short vowels (3)	-	-	Low, Falling, High
Long vowels (3)	Low, Falling, High	Mid, Low, Falling, High, Rising	Mid, Low, Falling, High, Rising

[1] This article is a reprint of the article appeared in the Proceedings of NECTEC Annual Conference 2000: ECTI Technologies for New Economics, June 2000, pp. 483-495. This paper award in category of "Best Presentation".