# Designing Text-to-Speech Application for Learning Thai Language

NUCHAREE PREMCHAISWADI*, WICHIAN PREMCHAISWADI**
*Faculty of Information Technology, Dhurakij Pundit University
110/1-4 Prachachuen Road Laksi, Bangkok 10210, Thailand
nucharee@dpu.ac.th
**Graduate School of Information Technology, Siam University
235 Petchkasem Road, Phasi-charoen, Bangkok 10163, Thailand
wichian@siam.edu

*Abstract:* - The conversation dialogs are one of the important materials for practical used of each language. In the past, the learner could learn the correct pronunciation from tapes that come with the books. However, learners could learn to pronounce only in the limited words in the training material. When they find new words that do not exist in the book, they must guess how to pronounce it. In case of Thai language, the pronunciation of each word becomes more difficult because the Thai language is a tonal language. This means that pitches are meaningful. A word pronounced with different pitches carry different meanings. Therefore, Text-to-Speech synthesis could be coupled with computer aided learning system to provide a helpful tool to learn language. This paper presents the designing of text-to-speech application for helping foreigners in learning Thai language.

*Key-Words:* - Text-to-Speech, Thai language

## 1 Introduction

Learning a new language is very difficult without adequate support. However now a day, computer becomes cheaper and more powerful. Therefore, it is used in many applications including language teaching. The conversation dialogs are one of the important materials for practical used of each language. In the past, the learner could learn how to pronounce correctly from tapes that come with the books. However, the learners could learn to pronounce only in the limited words in the book. When they find new words that do not have in the book, they must guess how to pronounce it. Text-to-Speech synthesis can be coupled with computer aided learning system to provide a helpful tool to learn a new language. In case of Thai language, the pronunciation of each word becomes more difficult because the Thai language is a tonal language. This means that pitches are meaningful. A word pronounced with different pitches carry different meanings. There is also no spacing between words and no special mark to identify the end of a sentence. The Thai vowel forms do not follow initial consonants; some are placed before the initial consonants, some after the consonants, some above the consonants, and some underneath the consonants. The vowels that are “complex” forms (i.e. composed of more than one part) can be placed around the consonants.

Table 1. Tones (Pitches) in Thai language

| Tones (Pitchs) in Thai language | Example |
|---|---|
| 1. mid level tone | *Khaa1* (to be lodged in, here represented with the number 1) |
| 2. low level tone | *Khaa2* (Galanga, an aromatic root, (here represented with the number 2) |
| 3. falling tone | *Khaa3* (I, slave, servant, (here represented with the number 3) |
| 4. high level tone | *Khaa4* (to sell, (here represented with the number 4) |
| 5. rising tone | *Khaa5* (leg, (here represented with the number 5) |

There are five distinctive tones (pitches) in standard Thai language as shown in Table1. From the example shown in Table 1, it can be seen that it is very difficult for foreigners to pronounce a Thai word correctly. Therefore, the learners have to practices more and need some tools such as the Thai Text-to-Speech software to help them to lean how to pronounce the words correctly.

## 2 Designing Text-to-Speech Applications

Text-To-Speech or TTS provides verbal output of application related text. This involves breaking down words into phonemes; analyzing them for special handling of text such as numbers, currency amounts, inflection, and punctuation; and generating the digital audio for playback. Speech synthesis has been around for quite some time and is a relatively mature technology. The first electronic speech synthesizer was produced at AT&T's Bell Labs in 1936, with initial attempts at electronic voice recognition following in the late 1940s. Applications can use .wav or .ulaw files for prerecorded voice prompts and information delivery, but if the file is missing or the text stream is dynamically generated, prerecorded information is not sufficient. TTS technology can output dynamic data that doesn't require recording voice prompts for every possibility. Thus, sophisticated text-to-speech applications are the better alternative in situations where a prerecorded digital audio recording is inadequate or impractical.

In phonetics, an allophone is one of several similar sounds that belong to the same phoneme. A phone is a sound that has a definite shape as a sound wave, while a phoneme is a basic group of sounds that can distinguish words (i.e. changing one phoneme in a word can produce another word); native speakers of a particular language perceive a phoneme as a single distinctive sound in that language. Just like a broken piece of china that's been glued back together again doesn't look quite like the original, a word or phrase that's been assembled from phonemes often sounds a little different than the native speaker's version. The bigger the segment of sound, the more natural it sounds in the reconstruction, but using syllables or whole words as the building blocks for speech synthesis requires a very large database.

The inflection in a speaker's voice is often the key to understanding the meaning of a spoken phrase. When the phonemes are connected to produce words or other sounds of speech, there are other characteristic sounds that transition certain phonemes to other phonemes. These sounds, which cannot be represented by a single symbol, are called "diphthongs". Diphthongs typically occur when pronouncing two vowel-type phonemes in succession, such as "ah" and "ee" to create the sound i [1]. Children learn inflections by imitating the speech patterns of adults until they are developed as an accent. The differences that a native speaker picks up from the tone of another's voice are difficult to describe to a non-native speaker, and even more difficult to describe to a computer. To capture these speech nuances requires adding prosody, the pitch and duration of sounds that give them additional meaning and also make them sound natural.

## 3 TTS Limitations

When designing a voice based application, it is important to understand the limitations of the TTS component and ensure you design around or generate the grammars needed to help ensure that these limitations have a minimal impact on the successful use of the application.

**Text-to-Speech Voice Quality**

Most text-to-speech engines can render individual words successfully. However, as soon as the engine speaks a sentence, it is easy to identify the voice as synthesized because it lacks human prosody -- i.e., the inflection, accent, and timing of speech. For this reason, most text-to-speech voices are difficult to listen to and require concentration to understand, especially for more than a few words at a time.

**Emotion**

Although many text-to-speech engines can parse and interpret punctuation, such as periods, commas, exclamation points, and question marks, none of the engines that are currently available can render the sound of human emotion accurately.

**Mispronunciation**

Text-to-speech engines use a set of pronunciation rules to translate text into phonemes. This is fairly easy for languages with phonetic alphabets, but it is very difficult for the English language, especially if last names are to be pronounced correctly.

(Pronunciation rules seldom fail on common words, but they almost always fail on names that are unusual or of non-English origin.). If a TTS engine mispronounces a word, the only way that the user can change the pronunciation is by entering either the phonemes, which is not an easy task, or by choosing a series of "sound-alike" words that combine to make the correct pronunciation.

**Mechanical Voice Sound**

When speaking phrases or sentences, the voice sounds mechanical and fatigues users because they have to listen intently to understand what is being said.

## 4 Designing around TTS Limitations

**Mispronunciation Issue**

Speech synthesis is driven by dictionaries, falling back for unknown words on rules for regular pronunciation. High quality speech synthesis is possible if the application designer extends the dictionary used by the application. Although this can be time consuming, it reduces misunderstanding and miscommunications for the user.

**Quality Issue**

Speech synthesis is not as good as having a trained person read the text. Content providers will want to provide prerecorded content for some parts of the application where the text is static. Prerecorded content can include music and different speakers similar to radio advertisements or news broadcasts. Use prerecorded content wherever possible so as to minimize quality issues.

**Emotion**

Text to speech dictionaries contain information on how each word is to be spoken by a speech synthesizer. This covers both phonemes and prosody (stress). The pronunciation may depend on the context in which a word occurs. As a result some limited linguistic analysis may be needed to determine which pronunciation applies.

**Mechanical Voice Sound**

Most TTS engines allow customized vocabulary pronunciations, and speech variations such as age, gender, name, pitch, speed, and volume. This permits the application designer to fine tune pronunciations. Also, multiple voices can be used in applications similar to the way color, and font size is used in GUI application.

## 5 Experimental Results

The application of the Thai text-to-speech system is designed and implement for helping foreigners in learning the Thai language. The system could generate sound from an input of Thai sentences. After the synthesis system receives input Thai text, it separates them into token according to the grammar of the Thai language [2][3]. Then, the sound of each word is generated by using the concatenation of each basic sound in Thai language. Therefore, the system could generate sound of any words in Thai Language. Although at this point of time, the system can generate sound from the variety of user input and can show how to produce sound for each basic sound. The main menu of the system is shown in Fig. 1.

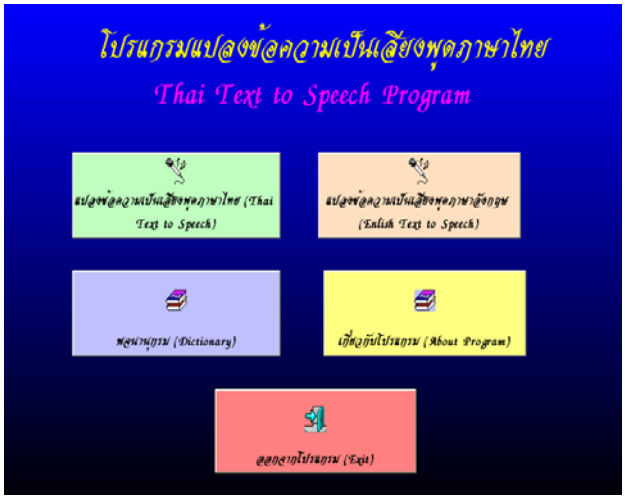


Fig.1 Main menu

In order to help the learners to know how a sentence is separated into each word. This is necessary because there is no blank space or other punctuations used for separating each word from others as in English language. The result of this process is shown in Fig. 2. This also can help the learners to known that the mispronunciation is come from the quality of the system or not. In Thai sentence, if the word boundaries can not identified correctly, it leads to the incorrect pronunciation. For example, the string "ตากลม" can be separated into two different ways with different meaning and

pronunciations. The first one is "ตา"(eye) and "กลม"(round) and its pronounced "ta klom". For the second case, it is separated into "ตาก"(expose) and "ลม"(wind) and pronounced "tak lom". All of alphabetic characters of the Thai language together with their pronunciation are also prepared for the learners as shown in Fig. 3.

In testing the system, however, there are still some words that the system could not correctly produce and need to be improved in the future in term of quality issue and emotion.

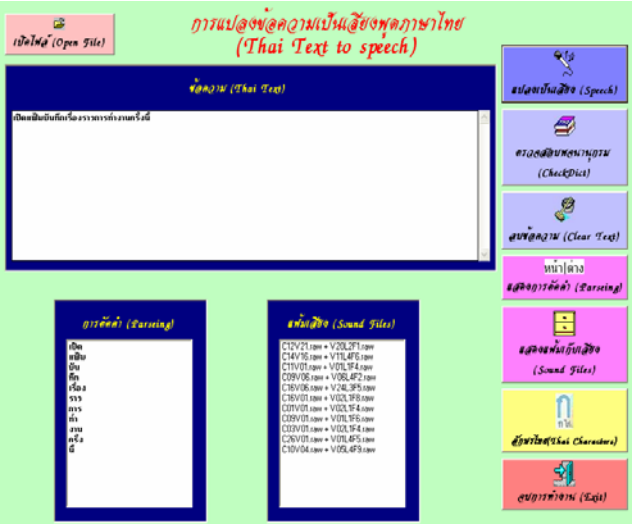


Fig.2. Example of the separated words in a sentence.

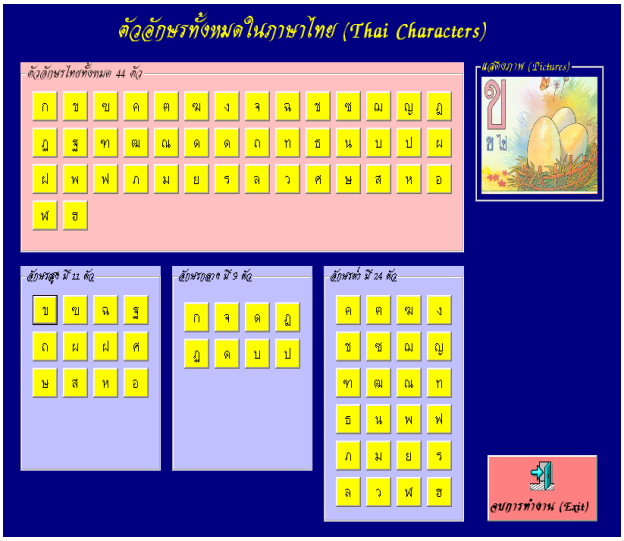


Fig. 3. Thai alphabetic characters and their pronunciation.

## 6 Conclusion

This paper presents the designing guidelines of text-to-speech application for helping foreigners in learning Thai language. The concept was implemented and tested. However, there are still some words that the system could not correctly produce and need to be improved in the future in term of both quality issue and emotion. In designing a text-to-speech application, we need to consider not only technology but also user interface and the quality issue of the sound produced from the system. The system should also provide information for the learners as much as possible.

*References:*

[1] Cater, John P., Electronically Hearing: Computer Speech Recognition, Howard W. Sams & Co., Indianapolis, IN, 1984.

[2] Pantumetha K. (1998). Thai Language Characteristics. Ramkamhaeng University.

[3] Dutoit T. (1997). An Introduction to Text-to-Speech Synthesis, Dordrecht: Kluwer.