*2012 International Conference on Networks and Information (ICNI 2012)*
***IPCSIT vol. 57 (2012) © (2012) IACSIT Press, Singapore***
*DOI: 10.7763/IPCSIT.2012.V57.12*

# Thai-Speech-to-Text Transformation Using Dictionary-Based Technique

Chomtip Pornpanomchai[+], Pranrawee Ngamwongsakollert, Pornnipa Tangpitaksamer and Chavika Wonvattanakij
Faculty of ICT, Mahidol University, Bangkok, Thailand

**Abstract.** The objective of this research is to develop the computer system which can transform Thai speech into text by using a dictionary-based approach. The system is called "Thai-Speech-to-Text (TSTT)". The TSTT system is a dependent machine, which depends on a speaker. The TSTT consists of 4 modules, which are: 1) speech acquisition, 2) speech preprocessing, 3) speech recognition, and 4) result presentation. The TSTT created the system dictionary by using female sound, which consisted of 4,742 Thai words. The experiment was conducted on three scenarios, which were: 1) transform female owner sound, 2) transform female different sound, and 3) transform male different sound. The precision rates of the system are 45.15, 2.01 and 1.00 percent for female owner sound, female different sound, and male different sound, respectively.

**Keywords:** Thai-Speech-to-Text, Speech Recognition, Dictionary-Based Approach.

## 1. Introduction

The speech-to-text transformation is one of the most difficult tasks in an artificial intelligent of computer science because this task consists of many difficult problems, namely:

1) The signals of speech have no definite pattern, nevertheless, it is spoken by the same person. It will have ten patterns of speech signals if a speaker records the same word ten times.

2) A pattern of speech signal depends on speaker gender. A woman sound has a higher pitch than a man sound, as illustrated in the graph in Fig. 1. [1]

3) There are a lot of Thai words which have the same pronunciation but are written in different characters.

Even the speech-to-text application is very difficult to develop, but it is very useful for people. It can be applied to various tasks, for example: 1) doctors create patient records by using their voice, 2) lawyers can prepare their documents without typing equipment, 3) students can easily take note of a professor's lecture, and 4) the secretaries do not need a short hand note, etc. The objective of this research is to develop a computer system which can translate the Thai language speech-to-text by using a dictionary-based technique. The related works and the TSTT system will be presented in the next sections.

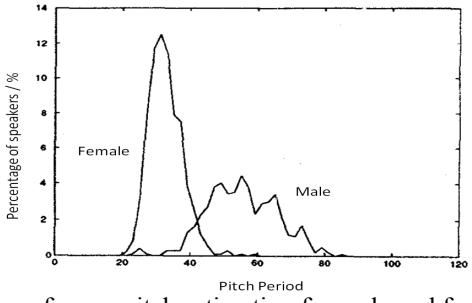



Fig. 1: Distribution of mean pitch estimation for male and female speakers [1]

[+] Corresponding author. Tel.: 662-3544333; Fax: 662-3547333.
*E-mail address*: itcpp@mahidol.ac.th.

## 2. Literature Review

Historically, a lot of researchers used many techniques to develop speech-to-text in many languages. This section reviews speech-to-text applications based on techniques and languages, as the following:

### 2.1. Neural Network-Based Approach

A neural network-based approach emulates knowledge of how a biological neural system stores and manipulates information. This artificial neural system is called "neural networks" and based on the notion that an artificial neural network can solve all problems in automatic reasoning, including a speech-to-text problem. This approach classifies sound wave patterns and generates a text format by predicTab properties of neural networks.

Lori L., et al. (2011) applied N-gram and a neural network to transform Mandarin speech into text system. Zaidi Razak, et al. (2009, 2010) used a neural network approach to transform speech into text of the Jawi script (Malay language). M.Tomalin, et al. (2010) presented Cambridge Arabic speech-to-text systems. [2-5]

### 2.2. Hidden Markov Model (HMM) Approach

The Hidden Markov Model is a statistic model, which represents an information sequence by a number of states. At each state, the system transmits to another state by probability distribution. Therefore, a sequence of changing states in HMM gives some information such as speech, word pattern, etc.

Tim Ng, et al. (2008) described the BBN 2007 Mandarin speech-to-text system development by using HMM. Gyorgy Szaszak, et al. (2011) presented Hungarian speech-to-text by using an HMM technique. Ryuichi Nisimura, et al. (2008) presented Japanese speech-to-text for a web system using Java script, Julius with HMM acoustic model. [6 - 8]

### 2.3. Dictionary-Based Approach

A dictionary-based approach is a basic approach for a speech-to-text system. The pronunciation sounds are looked up in a sound wave dictionary. This method proposes a solution to reduce the system complexity. But this method will fail if a word looked for cannot be found in the dictionary.

Colin Brooks (2000) presented English speech-to-text by a dictionary-based technique for deaf, deafened and hard-of-hearing people. F. Diehl, et al. (2008) demonstrated Arabic speech-to-text by a dictionary-based approach. [9, 10]

Based on the previous works, this research has applied a dictionary-based approach with the HMM technique to recognize the Thai speech. The system analysis and design will be presented in the next section.

## 3. Methodology

This section presents the system analysis and design. First, the system conceptual diagram is presented. After that, the mapping between the system conceptual diagram and the system structure chart is illustrated.

### 3.1. Conceptual Diagram

The system conceptual diagram is shown in Fig. 2. The system starts with a user speaking Thai words to the TSTT system. After that, the TSTT converts the Thai voice into the sound wave. The TSTT processes and matches the sound wave with all the sound waves in the system dictionary. Finally, the TSTT displays the recognition result in the text format on the system graphic user interface (GUI).

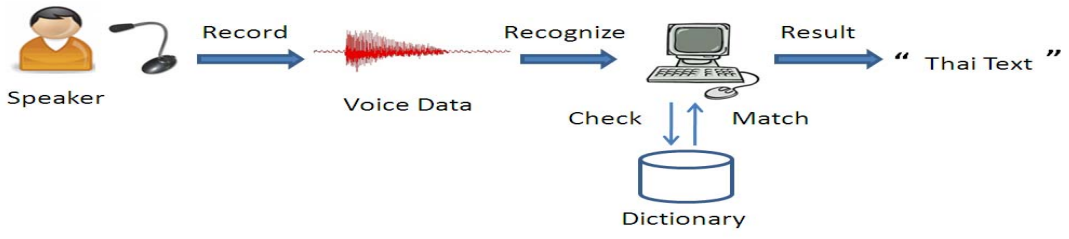



Fig. 2: The TSTT system conceptual diagram

## 3.2. System Structure Chart

The system structure chart is created by mapping the system conceptual diagram in Fig. 2. The system structure chart has four processing modules, which are 1) speech acquisition module, 2) speech preprocessing module, 3) speech recognition module, and 4) result presentation module (as shown in Fig. 3). Each module has the following details.

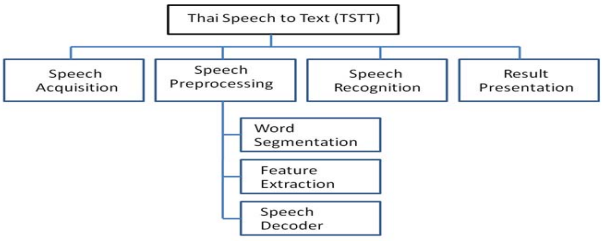



Fig. 3: The TSTT system structure chart

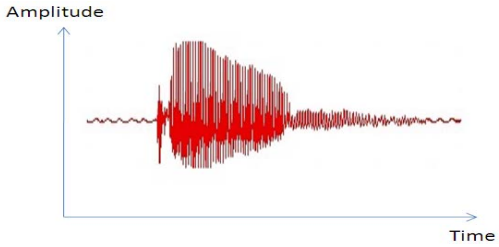



Fig. 4: The sound wave of Thai word "Kom"

**3.2.1 Speech Acquisition**– The first module is the speech acquisition routine. The TSTT system gets the speech with the sound card from a speaker and transforms the speech into the data voice format (.WAV). The sample of the Thai word "Kom" data voice format is shown in Fig. 4.

**3.2.2 Speech Preprocessing**– The second module is the preprocessing routine. This module is the processes which prepare the sound wave clear enough to extract all the features for the speech recognizing. The preprocessing module consists of the sub-modules namely, 1) word segmentation, 2) feature extraction, and 3) speech decoder. Each sub-module has the following details.

a) **Word Segmentation** – This sub-module separates a silence voice out of the Thai word sound by using the amplitude of the sound wave. The TSTT system detects the silence of the sound wave by using the signal energy method (as shown in Equation 1) and spectral centroid method (as shown in Equation 2).The amplitude of silence voice is very close to zero.

$$E(i) = 1/N \sum_{i=1}^{N} | X_i(n) |^2 \quad (1)$$

Where
E (i) is the energy value
$X_i$ (n) is the voice samples of the $i^{th}$ frame
N is the frame length

$$C_i = [ \sum_{i=1}^{N} (k+1) \bar{X}_i(k) ] / [ \sum_{i=1}^{N} Xi(k) ] \quad (2)$$

Where
$C_i$ is the spectral centroid
$X_i(k)$ is the discrete Fourier transform (DFT) coefficients of frame i
N is the frame length

b) **Feature Extraction** –This sub-module extracts the voice features by using Mel-frequency cepstral coefficient (MFCC), as shown in Equation 3 and vector quantization (VQ) processes. The MFCC is setting up hamming windows in 30 milliseconds and moving every 10 milliseconds. Each window has got signal for Fourier transformation.

$$f_{mel} = 2595 * \log 10 ( 1 + f/700) \quad (3)$$

Where

$f_{mel}$ is Mel scale
$f$ is frequency in Hertz (Hz.)

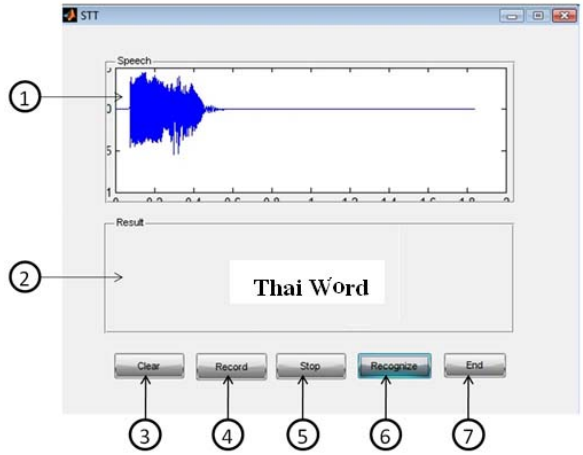
The vector quantization process is setting up a codebook, which is a Tab to keep voice signal of Thai words representative. The TSTT system uses a codebook to recognize Thai speech.

c) **Speech Decoder** – The TSTT system applies a speech decoder method to calculate the probability of data sequence in the codebook. The probability of data sequence is very useful for the Hidden Markov Model (HMM), which is used for matching voice signal of Thai words.

**3.2.3 Speech Recognition –** The TSTT system employs the HMM technique to map unknown speech with the speech in the dictionary. The HMM computes all the ways and uses maximum likelihood decision to get a recognition result.

**3.2.4 Result Presentation –** The final module is result presentation. Based on the graphic user interface in Fig. 5, the user screen consists of two image boxes and five command buttons. The two image boxes are: 1) the Thai speech signal image box (label number 1) and 2) the text recognition results image box (label number 2). Moreover, the five command buttons have the following details: 1) the <Clear> button for clearing all graphic user interface screen (label number 3), 2) the <Record> button for recording the Thai speech for recognition (label number 4), 3) the <Stop> button for stopping recording user sound (label number 5), 4) the <Recognize> button for recognizing the user speech (label number 6), and 5) the <End> button for exiting the TSTT system (label number 7).

Tab. 1 The tstt system precision rate




| Type of user | Precision |
|---|---|
| Female owner sound | 45.15 % |
| Female different sound | 2.01 % |
| Male different sound | 1.00 % |

Fig. 5: The TSTT system graphic user interface

## 4. Experimental Results

This part presents an experimental result of the Thai speech-to-text system, which is developed and based on the concepts and design mentioned in the previous part. In this system, the experiment's results are focused on the usability and the effectiveness of the system.

### 4.1. Usability Proof

In this section, we will analyze the usability of the Thai speech-to-text system. The input data is Thai speech loaded into the system. Then, the TSTT system uses computer software to match Thai words with a sound wave file. Finally, the system displays the Thai speech in the text format. The usability concept is proved by observing whether the user can record the voice and the system can transform it into a text format, as a student takes note of a professor's lecture.

### 4.2. Effectiveness Proof

The effectiveness test is done in the same way as that in the previous section, but we focus on the correctness of the reading result. The TSTT system data dictionary contains 4,742 Thai words. The experiment was conducted on three scenarios, which are the female owner sound with the system dictionary, the female different sound, and the male different sound. Each user tested 100 Thai sentences. The experimental results are shown in Tab 1. The precision of the TSTT system are 45.15, 2.01, and 1.00 percent for the female owner sound, the female different sound, and the male different sound, respectively.

## 5. Conclusion

The TSTT system fulfills the research objective, which is to develop a computer system that can transform Thai speech into text. The system created a Thai speech-to-text dictionary by recording 4,742 Thai

words in the system. The TSTT system is a dependent machine because it cannot transform Thai speech with the different recording sound. The precision of the system is quite low because of many reasons, namely:

1) The experiment was conducted on a normal microphone, which cannot capture a clear voice.
2) The TSTT was developed by using the simple dictionary-lookup method, which gives a low precision rate. But it is an easy method for implementing the system.
3) Some Thai words have similar pronunciation sound but different written ways.

This research is just the beginning stage for developing the TSTT system. We hope that the TSTT system can help people to more easily create a document in the near future. We need more man power and time to solve the TSTT difficult problems.

## 6. References


[1] David F.Marston. Gender Adapted Speech Coding. *The International conference on Acoustics, Speech, and Signal Processing 1998*, pp. 357-360, Seattle, Washington, USA, May 1998.

[2] Lori Lamel, Jean-Luc Gauvain, Viet Bac Le, Ilya Oparin and Sha Meng. Improved Models for Mandarin Speech-to-text Transcription. *The International Conference on Acoustics, Speech, and Signal Processing 2011*, pp. 4660-4663, Prague, Czech Republic 2011.

[3] Zaidi Razak, Siti Rabeah Sumali, Mohd Yamani Idris, Ismail Ahmedy and Mohd Yakub. Review of Hardware Implementation of Speech-to-text Engine for Jawi Character. *The International Conference on Science and Social Research*, pp. 565-568, Kuala Lumpur, Malaysia, December 2010.

[4] Zaini Arifah Othman, Zaidi Razak, Nor Aniza Abdullah and Mohd Yakub. Jawi Character Speech-to-Text Engine Using Linear Predictive and Neural Network for Effective Reading. *The Third Asia International Conference on Modeling & Simulation*, pp. 348-352, Bali, Indonesia, May 2009.

[5] M. Tomalin, F. Diehl, M.J.F. Gales, J. Park and P.C. Woodland. Recent Improvements to the Cambridge Arabic Speech-to-Text Systems. *The International Conference on Acoustics, Speech, and Signal Processing 2010,* pp. 4382-4385, Dallas, Texas, USA, March 2010.

[6] Tim Ng, Bing Zhang, Kham Nguyen and Long Nguyen. Progress in The BBN 2007 Mandarin Speech-to-Text System. *The International Conference on Acoustics, Speech and Signal Processing 2008*, pp.1537-1540, Las Vegas, Nevada, USA, March-April 2008.

[7] Gyorgy Szaszak, Akos Mata Tundik and Klara Vicsi. Automatic Speech-to-Text Transformation of Spontaneous Job Interviews on The HuComTech Database. *The International Conference on Data Cognitive Infocommunications*, pp. 1-4, Budapest, Hungary, July 2011.

[8] Ryuichi Nisimura, Jumpei Miyake, Hideki Kawahara and Toshio Irino. Speech-to-Text Input Method for Web System Using Java Script. *The Spoken Language Technology Workshop 2008*, pp. 209-212, Goa, India, December 2008.

[9] Colin Brooks. Speech-to-Text System for Deaf, Deafened and Hard-of-Hearing People. *The IEEE Seminar on Speech and Language Processing for Disabled and Elderly People*, pp. 5/1-5/4, Ref. No. 2000/025, April 2000.

[10] F. Diehl, M.J.F. Gales, M. Tomalin and P.C. Woodland. Phonetic Pronunciations for Arabic Speech-to-Text Systems. *The International Conference on Acoustics, Speech and Signal Processing 2008*, pp. 1573-1576, Las Vegas, Nevada, USA, March-April 2008.



**Chomtip Pornpanomchai** received his B.S. in general science from Kasetsart University, M.S. in computer science from Chulalongkorn University and Ph.D. in computer science from Asian Institute of Technology. He is currently an assistant professor in the faculty of Information and Communication Technology, Mahidol University, Bangkok, Thailand. His research interests include artificial intelligence, pattern recognition and object-oriented systems. Email: itcpp@mahidol.ac.th.