Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 8–27
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

# Thai speech processing technology: A review

Chai Wutiwiwatchai [a,*], Sadaoki Furui [b]

[a] *Speech Technology Research Section, Information R&D Division, National Electronics and Computer Technology Center, 112 Pahonyothin Rd., Klong-luang, Pathumthani 12120, Thailand*

[b] *Department of Computer Science, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan*

Received 8 February 2006; received in revised form 19 September 2006; accepted 2 October 2006

**Abstract**

This paper presents a comprehensive review of Thai speech technology, from its impetus in the early 1960s to 2005. Thai is the official language of Thailand, and is spoken by over 60 million people worldwide. As with Chinese, it is a tonal language. It has a spelling system using a Thai alphabet, but has no explicit word boundaries, similar to several Asian languages, such as Japanese and Chinese. It does have explicit marks for tones, as in the languages of the neighboring countries, Laos and Vietnam. Therefore, with these unique characteristics, research and development of language and speech processing specifically for Thai is necessary and quite challenging. This paper reviews the progress of Thai speech technology in five areas of research: fundamental analyses and tools, text-to-speech synthesis (TTS), automatic speech recognition (ASR), speech applications, and language resources. At the end of the paper, the progress and focus of Thai speech research, as measured by the number of publications in each research area, is reviewed and possible directions for future research are suggested.


*Keywords:* Speech analysis; Speech processing; Speech synthesis; Speech recognition; Thai speech

## 1. Introduction

The history of Thai speech research spans over 40 years. Similar to other languages, early research was limited by computational capability. Initial research mainly involved linguistic studies, where only small sample sets were analyzed by humans regarding some specific characteristics of Thai, such as lexical tones and stress. Representative studies from this area, made during the 1960s–1970s, include that by Abramson (1962), Gandour (1979) and Luangthongkum (1977).

In the late 1980s, when computers became much more powerful and affordable, computational linguistics became a major inter-disciplinary field. Research moved from the study of the basic characteristics of Thai speech to natural language and speech processing, including the development of Thai speech synthesis systems (Luksaneeyanawin, 1989; Saravari and Imai, 1983). Such complicated systems required the parallel development of several fundamental tools for morphological and phonological analysis, which take into account the unique characteristics of the Thai language.

Research on Thai speech recognition first appeared in 1987 in a study by Pathumthan (1987). In this study, he developed a system capable of recognizing a very limited vocabulary of isolated Thai words. During the same period, several similar works on isolated digits or words were explored (Kiat-arpakul, 1996; Pensiri and Jitapunkul, 1995). Another topic widely researched was the analysis and modeling of tones (Abramson, 1962; Anivan, 1988; Gandour, 1979) and other prosody features, such as intonation and stress in Thai (Abramson and Svastikula, 1982; Potisuk et al., 1996a).

The late 1990s to the early 2000s saw a huge expansion of Thai speech research, with major developments in Thai speech recognition and synthesis. Nevertheless, no large speech corpus was available and hence most of the speech

* Corresponding author. Tel.: +662 564 6900x2230; fax: +662 564 6873.
*E-mail address:* chai@nectec.or.th (C. Wutiwiwatchai).

0167-6393/$ - see front matter 
doi:10.1016/j.specom.2006.10.004

recognition research emphasized only isolated-word recognition. In 2000, the National Electronics and Computer Technology Center (NECTEC) in Thailand initiated a project to develop the first Thai large vocabulary continuous speech recognition (LVCSR) corpus (Kasuriya et al., 2003b). This corpus was not only used for Thai continuous speech recognition research at NECTEC (Kanokphara, 2003a; Tarsaku and Kanokphara, 2002), but also for many other research at different sites. In the field of speech synthesis, systems reached a commercial level of sophistication by the end of 1990s. Most of the systems currently available on the market are based on concatenation of subword units taken from an inventory of subword units. It has also been reported that the naturalness of synthesized voices can be improved by using variable-length units selected from a large corpus of speech utterances (Sagisaka et al., 1992). NECTEC has also succeeded in improving sound quality for their TTS system using the same technique (Hansakunbuntheung et al., 2005a).

At present, research on Thai speech recognition has moved on to various applications of continuous speech recognition (Pisarn et al., 2005; Suebvisai et al., 2005; Thienlikit et al., 2004). Research areas for speech synthesis include improving intelligibility of text processors and enhancing naturalness of synthetic speech by improving prosody prediction. In addition to the basic research already described, advanced applications of speech, such as a spoken dialogue system (Suchato et al., 2005; Wutiwiwatchai and Furui, 2003b) and a speech-to-speech translation system (Schultz et al., 2004) have also been attempted. Since basic tools and resources for recognizing Thai are still limited, developing these applications remains a formidable task.

Notable reviews and surveys of progress in Thai speech research published in the past include Luksaneeyanawin (1993), Jitapunkul et al. (1998), and Thongprasert et al. (2002). All these papers mainly presented an overview of the work carried out by their own organizations. In this paper, we attempt to review all published literature on speech research in Thai, not only that conducted by particular organizations. Section 2 introduces both the writing and sound systems of the Thai language. Section 3 describes research works on fundamental analyses and tools. Sections 4 and 5 review investigations of Thai speech synthesis and recognition, respectively. Section 6 presents research on high-level speech applications. Suggestions for, and discussion on, potential research directions are given at the end of the paper.

## 2. Thai writing/sound systems

Thai is the official language of Thailand, and is spoken by over 60 million people worldwide. Thai is but one of many languages and dialects belonging to the Tai sub-family (itself a member of the Sino-Tibetan family of languages). The following subsections review the text and sound characteristics of Thai.

### 2.1. Thai writing system

Detailed information on the Thai writing system can be found in several books (Haas, 1980; Higbie and Thinsan, 2002). In this subsection, we review some important characteristics of written Thai in order to lay the foundation for explanations of research presented in the rest of this paper. Thai is a tonal language, like Chinese, and is represented in text form with the Thai alphabet. This native alphabet comprises 44 consonants, 15 basic vowels, and 4 additional tone markers. Text is written horizontally, from left to right, with no intervening spaces, to form syllables, words, and sentences. Vowels are written above, below, before, or after the consonant they modify, however the consonant is always pronounced first when the syllable is spoken. The vowel characters (and a limited number of consonants) can be combined in various ways to produce numerous compound vowels such as diphthongs.

The grammar of the Thai language is considerably simpler than the grammar of most Western languages, and for many foreigners learning Thai, this compensates for the additional difficulty of learning tones. It is a "Subject + Verb + Object" language with no definite or indefinite article, no verb conjugation, no noun declension, and no object pronouns. Most significantly, words are not modified or conjugated for tense, number, gender, or subject-verb agreement. Articles such as English "a", "an", or "the" are not used. Tenses, levels of politeness, and verb-to-noun conversion are accomplished by the simple addition of various modifying words (called "particles") to the basic subject-verb-object format. One of the major problems for Thai language processing is a lack of word boundaries and explicit sentence markers. White space can be used as sentence, phrase, and word boundaries without strict rules. An analogous example in English is the word "GODISNOWHERE", which can be perceived as "GOD IS NO WHERE", "GOD IS NOWHERE", or "GOD IS NOW HERE" depending on the context.

### 2.2. Thai sound system

A general description of Thai sound system can be found in (Tingsabadh and Abramson, 1999). Luksaneeyanawin (1992, 1993) has also published a comprehensive description of the Thai sound system, which is briefly reviewed in this subsection. Thai sound is often described in a syllable unit in the form of /$C_i$–V–$C_f^T$/ or /$C_i$–$V^T$/, where $C_i$, V, $C_f$, and T denote an initial consonant, a vowel, a final consonant, and a tonal level, respectively. The $C_i$ can be either a single or a clustered consonant, whereas the V can be either a single vowel or a diphthong. Table 1 illustrates all Thai consonants, vowels, and tones. As seen in Table 1, some of the phonemes /p, pʰ, t, tʰ k, kʰ/ can combine with each of the phonemes /r, l, w/ to form a clustered consonant. Diphthongs are double-vowels beginning with one of the vowels /i, iː, ɨ, ɨː, u, uː/ followed by /a/. Five tones in Thai can be divided into two groups: the

Table 1
Thai phonemes in IPA

| | | |
|---|---|---|
| Initial consonant, $C_i$ | Single consonant | p, $p^h$, t, $t^h$, c, $c^h$, k, $k^h$, b, d, m, n, ŋ, f, s, h, r, l, w, j |
| | Cluster consonant | 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, V | Short vowel | i, ɨ, u, e, ɜ, o, æ, a, ɔ, ia, ɨa, ua |
| | Long vowel | iː, ɨː, uː, eː, ɜː, oː, æː, aː, ɔː, iː a, ɨːa, uːa |
| Final consonant, $C_f$ | p, t, c, k, m, n, ŋ, w, j | |
| Tone, T | ā, à, â, á, ǎ | |

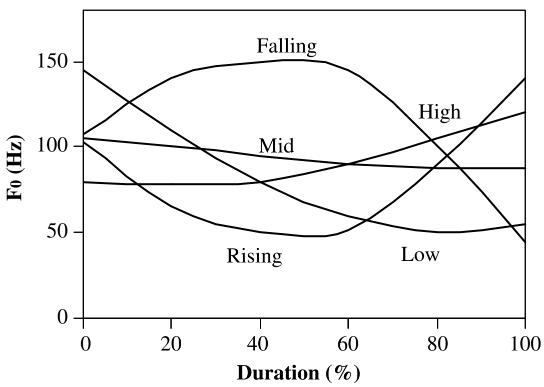


Fig. 1. F0 contours of the five Thai tones when syllables are uttered in isolation by a male speaker (Thubthong et al., 2001c).

static group consists of three tones, the high /´/, the middle /¯/, and the low /`/; the dynamic group consists of two tones, the rising /ˇ/ and the falling /ˆ/. Fig. 1 shows a graph comparing the F0 contours for the five tones that appear in Thai. Recently, some loan-words which do not conform to the rules of native Thai phonology, such as the initial consonants /br, bl, fr, fl, dr/ and the final consonants /f, s, $c^h$, l/ have begun to appear.

## 3. Fundamental analyses and tools

Certain fundamental tools are critical to research and development of advanced speech processing technology. This section summarizes existing speech, text, and pronunciation modeling tools developed mainly for Thai. One of the key issues is the analysis of the behavior of tones in Thai, which is often modulated by phrase and sentence level intonation. Other attributes of speech which have been explored for Thai include stress and pause. Work conducted on speech segmentation and basic phoneme modeling are also described. For text processing, word and sentence segmentation are reviewed. Fundamental tools for pronunciation modeling are described at the end of this section.

### *3.1. Speech analyses and tools*

#### *3.1.1. Tone analysis*

The first paper presenting an analysis of Thai tones from the age of ancient Thailand, Siam, was written by Bradley Bradley (1911). He illustrated the frequency curves of the five tones for Thai long syllables, similar to Fig. 1. Thai tones were widely investigated in the 1980s and 1990s. Thus acoustic–phonetic information for Thai tones in monosyllabic words is widely available (Abramson, 1962; Gandour et al., 1991). These earlier studies on F0 patterns found that the five Thai tones start at different F0 in general; high for the high and falling tones, low for the low and rising tones, and medium for the middle tone. Palmer (Palmer, 1969) demonstrated that the five Thai tones showed some changes in height and slope as a function of the preceding or following tone. Changes in height and slope appeared to be confined primarily to the beginning or end of the syllable (Abramson, 1979). A further study (Gandour et al., 1994) found that tonal coarticulation in Thai does not necessarily affect the whole F0 contour of adjacent syllables; carry-over effects extend forward to about 75% of the duration of the following syllables, whereas anticipatory effects extend backward to about 50% of the duration of preceding syllable. Tone studies also looked at particular conditions. For example, Burnham et al. (1996) analyzed tone perception across languages. English speakers were found to better discriminate tonal contrasts in musical contexts than in speech. However, tonal language speakers, Thai and Cantonese, perceived tonal contrasts equally well in both music and speech. Effects of speaking rate to tones were explored in (Gandour et al., 1999). Speaking rate effects on F0 contours of unstressed syllables are more extensive, both in terms of height and slope, than those of stressed syllables. Thai tones with substantial F0 movement (falling and rising) exhibited overall flatter slopes at the fast speaking rate. Modulation of syllable tones and phrasal intonation was analyzed in (Abramson and Svastikula, 1982). This study found that intonation in Thai, and perhaps other tonal languages, was not as clearly presented as intonation in non-tonal languages even in long sentences.

Tone recognition in Thai has been extensively researched at Chulalongkorn University. Tungthangthum (1998) extracted pitch information from Thai monosyllables using autocorrelation and used HMM for tone classification. An evaluation on small sets of monosyllables achieved 90% accuracy. Based on the analysis by Gandour et al. (1994) described above, Thubthong et al. (2001c) introduced a novel feature called half-tone modeling, which extended an analysis frame to cover half of the neighboring frames. Application of the proposed feature in a statistical classifier based on *a posteriori* probabilities yielded approximately a 3% improvement over the baseline model without half-tone modeling. A successive work by Thubthong et al. (2001b) conducted a combination of neural networks with distinctive features for F0 slope and height for tone recognition. Outputs from three neural networks, trained by three different sets of distinctive features, were combined for the final decision. The proposed model yielded as much as a 15.6% error reduction compared to the use of single neural network. Potisuk et al. (1995) conducted an analysis–synthesis approach for tone recognition in continuous speech. The method exploited an extension of Fujisaki's model to synthesize F0 contours of possible tone sequences of a given speech signal. The resulting contours were then compared to the smoothed and normalized F0 contour of the input speech in a pattern matching module. Overall, 89% of four tone sequences were correctly classified with this method. In 1998, Kongkachandra et al. (1998) proposed a novel feature in harmonic frequency domain for tone recognition. The algorithm was to segment the frequency spectrum into overlapped frames and calculate amplitude difference values between adjacent frames. Classification of tones was achieved by applying some heuristic rules on the amplitude difference values. The model gave 90% accuracy on a monosyllable test set. All the mentioned models however, were evaluated on either isolated syllables or hand-segmented syllables from continuous speech. Further research on segmenting speech signals into sequences of syllables or phonemes is required in order to achieve a completely automatic tone recognizer for continuous speech.

As opposed to speech-only tone analysis, Mixdorff and Burnham (2005, 2004) explored another avenue. He believed that tone patterns of Thai could be partly determined using visual cues. An experiment was done by tracking movement points on speaker faces with respect to uttered words in various tones. Results, however, showed that articulatory gestures were primarily a function of the word being uttered and only secondarily a function of tones.

### 3.1.2. Intonation analysis

Intonation is another important suprasegmental feature, the analysis of which requires different approaches, depending on whether or not the language in question is tonal or not. Potisuk et al. (1999) described the difference between phrasal F0 contours in tonal and non-tonal languages. For non-tonal languages such as Japanese, the phrasal intonation is modified by F0 realization of local pitch accents, whereas for tonal languages such as Thai, lexical tones modify the overall phrasal F0 contour. Differences in the F0 movement between tonal and non-tonal languages were also extensively described by Fujisaki (2004, 2003). They carried out an analysis of F0 contours in Thai utterances based on the analysis-by-synthesis technique. It is possible to decompose the F0 contour of a sentence into its constituents, i.e., the phrase components and the accent or local tone components. The phrase F0 contour is estimated using the Fujisaki model, while the accent or local tone components are driven by local commands. Thai, like Chinese and Swedish, has been classified as a language with both positive and negative local commands as opposed to languages with only positive local commands, such as Japanese and English. The Thai middle tone has no tone command, the high tone has a positive tone command only at the latter part of the syllable, the low tone has a negative tone command, and the rising and falling tones have combinations of positive and negative tone commands. Besides the tonal characteristics, Thai has also been described as possessing a distinction between stressed and unstressed syllables, which additionally affect the shape of phrasal intonation (Hirst and Di Cristo, 1998). Mixdorff et al. (2002) also employed the Fujisaki model in Thai intonation analysis. They came to the same conclusion as that stated by Fujisaki (2004, 2003), that is, the syllabic tone can be accurately modeled by using commands of positive and/or negative polarity with the exception of the middle tone, which can be modeled by the phrase command only. Charnvivit et al. (2003) tried to classify the intonation pattern as rising, falling, or convoluted, and implemented an automatic recognition technique based on F0 contour and a neural network, which achieved over 70% accuracy. Other evidence proving the existence of intonation in Thai was presented in a study by Thubthong et al. (2002), where the performance of a tone recognizer was enhanced when the intonation feature was eliminated from test utterances.

### 3.1.3. Stress analysis

Stress is another important language-specific suprasegmental feature. For Thai monosyllabic words, all content words are strongly stressed while all grammatical words are weakly stressed unless reinforced by emphasis. Thai polysyllabic words, however, have variable syllable-stress patterns such as (LL'), (LO'), (O'O'), and (O'L') for bi-syllabic words, where L and O represent linking and non-linking syllables, and the symbol (') is a stress marker given behind a stressed syllable. The linking syllable is a syllable having a vowel /a/ and a glottal stop. Thubthong et al. (2001a) investigated detecting stressed syllables in Thai polysyllabic words. Prosodic features, related to F0, duration, and energy, were calculated for different speech units including vowel, syllable, and rhyme units. An artificial neural network was used as the classifier. The model based on rhyme units achieved the best accuracy; an 84% stress detection rate. In Thai, there exist many homographs, which can often be distinguished only by context and/or stress. This is similar to some words in English, such as "present", which is pronounced with two different stress patterns, where each carries a different meaning. An analysis and modeling of stress in Thai phrases was carried out by Potisuk et al. (1996a). Acoustic parameters such as syllable rhyme-duration, F0, and energy were used to construct a Bayesian classifier, which achieved a 96% detection accuracy. The obtained prosodic constraints were additionally incorporated in a syntactic parser based on a constraint dependency grammar (CDG) for Thai (Potisuk and Harper, 1996b).

#### 3.1.4. Pause analysis

The behavior of pauses in Thai speech has been a popular topic of studies since 1977 (Luangthongkum, 1977). Luksaneeyanawin (1985) analyzed pauses in Thai phrases by considering several influential factors such as phonetic contexts. In her work, a "pause" is defined as an perceived auditory signal which the listener is able to detect as a period of silence, a "break" is an acoustic period of silence which is instrumentally detectable in terms of a definable drop in the intensity of the speech signal, or a period where F0 cannot be detected, and a "space" refers to a blank space in written script. Pause insertion was found to be based on individual behavior. In the study, it was found that approximately 24% of pauses were not detected as breaks and the number of spaces did not correspond to the number of breaks and pauses.

#### 3.1.5. Speech segmentation

Syllable-unit based speech recognition has been pursued by some researchers since this approach makes it easier to incorporate prosodic features, such as tones, into the syllable units. Moreover, some papers have shown a high accuracy for tone recognition when applying this method to isolated syllables and hand-segmented syllables extracted from continuous speech. Automatic segmentation of continuous speech into syllable units is therefore an important issue. Jittiwarangkul et al. (2000) proposed using several prosody features including short-time energy, a zero-crossing rate, and pitch with some heuristic rules for syllable segmentation. Ratsameewichai et al. (2002) suggested the use of a dual-band energy contour for phoneme segmentation. This method decomposed input speech into a low- and a high-frequency component using wavelet transformation, computed the time-domain normalized energy of both components, and introduced some heuristic rules for selecting endpoints of syllables and phonemes based on energy contours. Although there is no comparative experiment using other typical techniques, dividing the speech signal into detailed frequency bands before applying energy-based segmentation rules seems to be an effective approach. A phoneme-segmentation experiment on 1000 Thai isolated-syllables from 10 speakers achieved an average accuracy of nearly 95%.

#### 3.1.6. Phoneme modeling

During the early 2000s, a series of research papers on Thai vowel modeling was published by Chulalongkorn University (Ahkuputra et al., 2001; Ahkuputra et al., 2000; Maneenoi, 1998; Maneenoi et al., 2000). Basic features such as formant frequencies were extracted and grouped according to vowels. This research provided the first analysis of Thai phonemes using a large real-speech corpus, not just a small set of sample utterances. There is still a lack of sufficient research on prosody analysis for Thai, however data collection is not a trivial task, and thus research on prosody modeling has been substantially delayed. Since 2000, research on basic linguistic features has decreased and the engineering of high-level speech applications such as speech recognition and synthesis has taken priority.

### 3.2. Text analyses and tools

#### 3.2.1. Morphological analysis

Morphological analysis is a process for segmenting words into sub-word units called morphemes. The morpheme is the shortest textual chunk that still has meaning. Morphological analysis not only helps in finding the pronunciation of a word, but also serves as a necessary process for reducing the number of lexical words in automatic speech recognition (ASR). For Thai, a morpheme may be as short as a syllable. Since Thai contains a large number of monosyllabic words and many polysyllabic words are concatenations of monosyllabic words, morphological analysis in Thai consists mainly in a joint process of word segmentation and syllabification. Due to lack of consensus on what generally constitutes a word in Thai, word segmentation is task-oriented. For example, in semantic or syntactic analysis, it is preferable to work with a longer compound word, whereas in phonological analysis, a syllable-like word is more suitable. In this subsection, Thai morphological analyses including both word segmentation and syllabification are reviewed.

Word segmentation in Thai is not a trivial problem. Table 2 outlines the critical problems of Thai word segmentation (Aroonmanakun, 2002; Meknavin et al., 1997a), showing the high ambiguity of Thai word segmentation, especially in the last case where a compound word can be segmented differently in different semantic contexts. Due to these Thai specific difficulties, there has been a considerable amount of work in Thai word segmentation since 1986. The accuracy, however, is still limited and more research is needed. One basic algorithm which was devised for word segmentation was the use of a dictionary with some heuristic rules (Rarunrom, 1991). Inrut et al. (2001) proposed syllabification by certain rules and then combina-

Table 2
Problems in Thai word segmentation

| Problem | Example | Possible segmentations (The line with * is correct) |
|---|---|---|
| Simple words containing word-like syllables | นาฬิกา (clock) | \|นาฬิกา (clock)\| *<br>\|นา (farm) \| ฬิ (no meaning)\| กา (crow)\| |
| Incorrect segmentation in compound words | บอกว่า (tell that) | \|บอ (mad) \| กว่า (than)\|<br>\|บอก (tell) \| ว่า (that)\| * |
| Ambiguous segmentation in compound words | ตากลม (round eye or expose wind) | \|ตา (eye) \| กลม (round)\| *<br>\|ตาก (expose) \| ลม (wind)\| * |

tion of the syllables based on a dictionary and a forward–backward search, which has enabled deciphering of some ambiguous cases. Statistical models such as part-of-speech (POS) n-gram and word n-gram have also been explored (Kawtrakul et al., 1997; Meknavin et al., 1997a). In addition to the statistical n-gram, Meknavin et al. (1997a) proposed using a machine learning algorithm to solve some ambiguous cases. Another technique utilized a weighted finite-state transducer trained by a textual corpus (Sojka and Antos, 2003). These techniques worked fairly well when an input sentence contained only words appearing in the dictionary. However, a significant problem for word segmentation in Thai is when sentences contain unknown words, such as name entities and loan words written in Thai. Techniques using a machine learning algorithm e.g. Winnow (Charoenpornsawat et al., 1998) and a decision tree (Theeramunkong and Usanavasin, 2000) have been effective in overcoming the unknown-word problem. Work by Aroonmanakun suggests that segmenting text into a sequence of syllable-like units and combining units which have high collocations can also help in these cases (Aroonmanakun, 2002; Aroonmanakun, 2005).

The models mentioned above have mostly been proposed as front-end processes for other higher-level linguistic tasks such as syntactic analysis. For Thai speech systems, the process of word segmentation is necessary before phonological analysis can be performed. Since direct word segmentation is difficult, it is more suitable to perform syllabification instead of word segmentation. Syllabification in Thai has been performed using a dictionary (Poowarawan, 1986), syllabification rules (Chanyapornpong, 1983; Khruahong et al., 2003; Mamoru and Satoshi, 2001; Sawamipakdee, 1990; Thairatananond, 1981), and a combination of rules and a statistical model (Aroonmanakun, 2002). It is commonly acknowledged that these approaches have different advantages/disadvantages, i.e. the rule-based approach requires low computational power and memory but has difficulty resolving ambiguous cases, whereas the statistical approach needs a large annotated corpus for training.

#### 3.2.2. Sentence breaking

As described in Section 2.1, Thai has no explicit sentence markers. Conventionally in Thai writing, a space is placed at the end of a sentence. However, a space does not always indicate a sentence boundary. Mittrapiyanurak and Sornlertlamvanich (2000) presented an algorithm to extract sentences from a paragraph by detecting sentence-break spaces. The algorithm considered two consecutive strings with a space in between. The strings were first segmented into word sequences with POS tagged to each word. By exploiting a POS n-gram model, it was verified whether the space was a sentence break or not. The system was trained and tested with subsets of the ORCHID corpus (Charoenporn et al., 1997), and 80% break detection and 9% false-break rates were achieved. An extension of the algorithm was proposed by Charoenpornsawat and Sornlertlamvanich (2001). Not only the POSs of surrounding words but also collocations of surrounding words and lengths of surrounding token texts were used as the features for determining whether space characters were sentence boundaries. These features were confirmed to be useful. In (Charoenpornsawat and Sornlertlamvanich, 2001), these features were extracted automatically by machine learning using the system Winnow. Winnow was also used for sentence break detection. Compared to the POS n-gram model, a 1.7% improvement of break-detection rate and a 79% reduction of false-break rate were achieved. Although these gains are substantial, the algorithms depend strongly on word segmentation and POS tagging. A larger POS tagged corpus is needed to improve all these components.

### 3.3. Pronunciation modeling tools

Pronunciation modeling is an important component for various language processing technologies such as speech recognition, text-to-speech synthesis, and spoken document retrieval. The task is to find the pronunciation of a given text by using a pronunciation dictionary and/or a phonological analyzer. This subsection describes the details of these two techniques for Thai.

#### 3.3.1. Pronunciation dictionary

At present, there are many Thai dictionaries available both offline and online, in a monolingual format, a bilingual format such as Thai to English, or a multi-lingual format i.e., Thai to multiple languages. Dictionaries have been constructed for varying purposes, generally for translation and writing tasks. Only a few dictionaries are designed for use in natural language processing research.

Table 3
Some available Thai dictionaries

| Dictionary | Type | Size | Availability | Source |
|---|---|---|---|---|
| Royal Institute Dictionary | Monolingual with pronunciation | 33,582 | Web application | http://rirs3.royin.go.th/riThdict/ |
| NECTEC Lexitron | Bilingual (En-Th) with pronunciation | 53,000 Eng., 35,000 Thai | Publicly available | http://lexitron.nectec.or.th/ |
| NAiST Lexibase | Monolingual | 15,000 | Proprietary | http://naist.cpe.ku.ac.th/ |
| So Sethaputra Dictionary | Bilingual (En-Th) with pronunciation | 48,000 Eng., 38,000 Thai | Proprietary | http://www.thaisoftware.co.th/ |
| Longdo Dictionary | Multilingual | 6600 | Web application | http://dict.longdo.org/ |
| KDictThai | Bilingual (En-Th) | 37,018 | Publicly available | http://kdictthai.sourceforge.net/ |
| Saikam Dictionary | Bilingual (Jp-Th) | 133,524 | Web application | http://saikam.nii.ac.jp/ |
| PPA Dictionary | Monolingual with pronunciation | 60,000 | Proprietary | http://www.ppainnovation.com/ |

Table 4
Problems in Thai grapheme-to-phoneme (G2P) conversion

| Problem | Example | Pronunciation |
|---|---|---|
| Ambiguity in letter-to-sound mapping | มณฑา and มณฑป | /mōn/t$^{h-}$aː/ and /mōn/dòp/ |
| Homograph | เพลา | /p$^{h-}$eː/lāː/ or /p$^{h}$lāw/ |
| Change in vowel length | น้ำ | /nám/ must be changed to /náːm/ |
| Linking syllable | วิทยา | /wít/jāː must be changed to /wít/t$^{h}$á/jāː/ |
| Ambiguity in consonantal functionality | จักรี | /càk/rīː/ must be changed to /càk krīː/ |
| Incorrect word boundary | บอกว่า | /bɔ̄ː/kwàː/ is incorrectly pronounced as/ bɔ̀ːk/wâː/ |

Table 3 reviews Thai dictionaries available for general purposes as opposed to those for research and development. Several dictionaries with pronunciation tags as shown in Table 3 have been used for pronunciation modeling in Thai speech processing technology, for example, the use of Lexitron and PPA dictionary in Thai text-to-speech synthesis (Mittrapiyanurak et al., 2000a; Ponyanun, 2003).

### 3.3.2. Phonological analysis

Phonological analysis is one approach to finding the correct pronunciation of a given text. It is sometimes called letter-to-sound or grapheme-to-phoneme (G2P) conversion. Mapping between letters and sounds in Thai is not a straight-forward process. Table 4 illustrates several problems involved in Thai G2P (Aroonmanakun, 2005; Tarsaku et al., 2001). Using only a mapping table or a pronunciation dictionary does not provide sufficient information for solving the problems presented. A more complicated phonological analyzer is needed. The early methods proposed by Meknavin and Kijsirikul (1997b) and Narupiyakul (1999) were based on complex rules coded by linguists. To resolve ambiguous cases, statistical models such as an n-gram model (Chotimongkol and Black, 2000) and a probabilistic generalized LR parser (PGLR) (Tarsaku et al., 2001) have recently been used.

Another interesting area of research deals with the mixture of Thai and English characters, which is a somewhat common phenomenon in contemporary Thai writing. Combining native English TTS with Thai TTS often makes an unnatural synthetic sound, since speakers, accents, and other attributes such as loudness are different. One solution is to apply a G2P tool that maps English characters to Thai phonemes and then to use only the Thai TTS for synthesizing speech. Aroonmanakun et al. (2004) set up an experiment which attempted to map English text to Thai phonemes. A machine learning method was applied to automatically generate the mapping rules given an annotated corpus. The largest difficulty encountered was how to predict syllabic tones in English words. Tones (and accents) expressed in loan-words depend on several factors such as the number of syllables in each word, and this cannot be accurately predicted by simple rules.

Homograph ambiguity in Thai is another significant problem in G2P conversion. As Thai writing has no explicit word boundaries, homographs in Thai are not only limited to within words, but can additionally expand to connected words. An analogous example in English is the homograph "NOWHERE", which can be pronounced as "NO-WHERE" or "NOW-HERE", depending on its context. This is a combination problem of word segmentation, syllabification, and homograph disambiguation. Tesprasit et al. (2003a) analyzed the problem and conducted a comparative experiment using a statistical n-gram model, a decision tree, and Winnow for selecting the best phoneme sequence given Thai homographs.

## 4. Thai speech synthesis

A general structure of TTS is shown in Fig. 2; viewing this structure, research regarding TTS can be classified into four major groups. The first group focuses on the general problem of developing Thai TTS engines, while the other three groups focus on the various subcomponents of these systems, which include text processing, linguistic/prosodic processing, and waveform synthesis.

### 4.1. Development of Thai TTS

The first paper describing the development of a Thai TTS engine was published in 1983 by Saravari and Imai (1983), where a speech-unit concatenation algorithm was applied to Thai. At Chulalongkorn University in Thailand, two publications on similar techniques for TTS were presented by Luksaneeyanawin (1989) and Taisetwatkul and Kanawaree (1996). The former was a precursor to CUTalk,

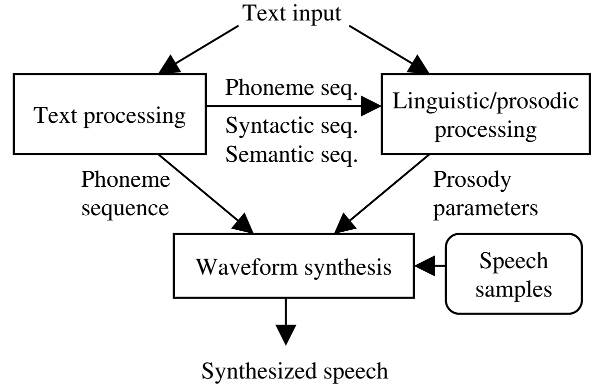


Fig. 2. An architecture of TTS.

the first commercial Thai TTS system. Two more commercial engines based on the same technique were developed a few years later. These were the Sarika (Ponyanun, 2003) and Vaja (Mittrapiyanurak et al., 2000a) systems, which used either demisyllable or diphone units for concatenation.

The synthesis approach that is recognized as giving the most human-like speech quality is based on corpus-based unit-selection. In Thai, such an approach was implemented in the latest version of Vaja (Hansakunbuntheung et al., 2005a) developed at NECTEC. Although the newest Vaja engine produces a much higher sound-quality than the former unit-concatenation based engine, the synthetic speech sometimes sounds unnatural, especially when synthesizing non-Thai words written with Thai characters. Improvement of the quality in both text processing and speech synthesis is required before state-of-the-art engines of the same level of quality as those which have been successfully developed for other languages, such as English and Japanese, can be achieved.

### 4.2. Text processing

Text processing in TTS works by finding the most probable sequence of phonemes given an input text. For Thai, there are several crucial subtasks involved as well. Thai text has no explicit sentence or word boundary markers. Space is often placed between sentences but not all the spaces are sentence boundary markers. Therefore, text is first tokenized into text chunks using some specific characters such as punctuations, digits, and spaces. The next subtask is to determine which space characters function as sentence boundary markers. Algorithms for sentence break detection were described in Section 3.2.2. In these algorithms, text chunks are segmented into word sequences, and a POS is attributed to each word. The space is determined to be a sentence break based on the POS of surrounding words. Therefore, word segmentation is considered to be one of the most crucial sub-tasks involved since it is commonly the first necessary step in the processing systems. Algorithms for Thai word segmentation have been explained in Section 3.2.1. The next subtask is called text normalization, where non-verbal words are replaced by their orthographies. Non-verbal words include numerals, symbols e.g., "$" and "=", and abbreviations. The final step in text processing is grapheme-to-phoneme conversion. This issue was discussed extensively in Section 3.3.

### 4.3. Linguistic/prosodic processing

In TTS, important prosody features include phrasal pauses, F0, duration, and energy. These basic features form higher-level suprasegmental features such as stress and intonation. In Thai, the lexical tone is another additional feature which needs to be generated. The lexical tone itself is often modified by stress and phrasal intonation. As described in Section 3.1, there are a number of research sites investigating Thai speech prosody. Some have focused exclusively on prosody analysis, whereas some have attempted to model prosody and have used the models for Thai TTS. This section focuses on only the latter.

Lexical tones in Thai are almost completely described by textual forms. Simple rules or a pronunciation dictionary are useful for examining the lexical tones. Only some specific words, mostly those which originated in the Bali or Sanskrit languages, diverge from the rules. In synthesizing continuous speech, lexical tones are modulated by phrasal or sentence intonation, which is represented by an F0 contour. Mittrapiyanurak et al. (2000b) incorporated F0 contour prediction into their TTS engine using a rule set based on the concepts of Fujisaki et al. (2003) described in Section 3.1.2.

Duration and phrasal pauses are two features important to synthesizing natural speech. Mittrapiyanurak et al. (2000b) also constructed a set of rules to predict syllable duration for their Vaja TTS engine. Klatt's method (Klatt, 1987) was applied to each syllable. The method first assigned default duration by its intrinsic property to each syllable and then the syllable duration was modified by a set of scaling factors corresponding to each successive rule. The intrinsic duration depended on phoneme identities in the syllable. Scaling factors in the rule set relied on contextual effects at the word, phrase, and sentence levels. For example, the duration of a syllable located around the end of a sentence is longer than that in the beginning portion of the sentence. Following this research, when a well-designed speech corpus for speech synthesis was developed two years later (Hansakunbuntheung et al., 2003b), a comprehensive analysis of phoneme duration in Thai continuous speech was carried out. These analysis results were used to build a duration prediction model based on multiple linear regression for the Vaja TTS (Hansakunbuntheung et al., 2003a). The model computed syllable duration by considering several factors such as phoneme identities, syllable tone, word POS, and positions in the word and the phrase. An objective test showed a high correlation of 0.8 between synthesized speech and human speech. Mixdorff et al. (2003) analyzed syllabic durations and tones of highly ambiguous Thai phrases. A listening test on synthetic utterances, in which syllabic durations were controlled by a regression model and syllabic tones and intonation were fixed, was conducted. They found a significant improvement in disambiguation for human listeners when the duration model was incorporated.

Phrasal pauses are necessary for TTS to generate natural sound. Phrase break detection has been part of ongoing investigations not only for Thai, but also for other languages (Black and Taylor, 1997; Navas et al., 2002). Research on phrasal pause prediction has been extensively investigated at NECTEC. The first implementation (Mittrapiyanurak et al., 2000b) was a rule-based method. The rules simply considered some punctuation marks and grammatical words as the ends of phrases. Later works (Hansakunbuntheung et al., 2005b; Tesprasit et al., 2003b) were based on the same idea, but used machine learning methods to predict pauses given potential features extracted from an input sentence. Tesprasit et al. (2003b) applied collocation of words and the number of syllables from the previous phrase break to the learning machine. Both C4.5 and RIPPER have been shown to outperform the simple POS n-gram model. Hansakunbuntheung et al. (2005b) extended the experiment by adding other potential features including POS context and the number of syllables and words from previous phrase and sentence breaks. Other types of machine learning including a neural network (NN) and CART were also compared. The best results were an 80% break correction rate and a 2.4% false-break rate given by the CART engine.

Additionally, after some basic research on stress analysis (Luangthongkum, 1977; Luksaneeyanawin, 1983), stressed/unstressed Thai syllables were analyzed and modeled for Thai TTS (Pansombat et al., 2002). Firstly, each syllable in a word was marked as either stressed or unstressed using a linguistic rule set. The rule set (Luksaneeyanawin, 1983; Surinpaiboon, 1985) indicated which syllable in Thai polysyllabic words was stressed by considering syllabic tone, the number of syllables and position of the syllable in the word. The stressed syllable was acoustically characterized by pitch, duration, and energy. Synthesized pitch was taken from a table containing average values of pitch for each tonal syllable, both stressed and unstressed. The average values were analyzed from a training set. Syllable duration was approximated by a set of rules. In synthesizing, default demisyllable units were concatenated. The synthesized speech was then adjusted according to the given pitch and duration parameters using TD-PSOLA. A human listening test indicated that the stress-modified speech achieved 5% more satisfaction than the normal synthesized speech.

### 4.4. Waveform synthesis

Generally, the state-of-the-art approach to waveform synthesis in TTS is unit concatenation. Speech units are selected from either a set of certain units, such as diphones and demisyllables, or directly from a large speech corpus. Development of Thai TTS using unit concatenation was described in Section 4.1. Some academic research has tried to explore other kinds of TTS algorithms. Kiat-arpakul et al. (1995a) proposed an acoustic inventory that contained both phoneme and demisyllable units for unit concatenation. This method certainly helps when no demisyllable unit is found in the inventory set. An articulatory synthesis method was proposed for synthesizing Thai vowels and tones (Khaorapapong et al., 2004), in which an articulatory model was constructed based on characteristics of articulated vocal-tract and sound properties such as F0. Diphthongs were synthesized by sliding from one vocal-tract vowel to another, using logarithmic and exponential functions. Results of the first and second formants of synthesized speech showed only 10% and 7.4% differences to the standard formants given in (Abramson, 1962). A Thai TTS system based on formant synthesis was proposed by Saiyot et al. (2005). The first step of the algorithm generated a spectrum using a set of rules indicating formant frequencies and magnitudes. The spectra were then modified by prosodic features including phrasal pauses, durations, intonation, syllable tones, and speaker gender. Finally, a speech waveform in time domain was generated from the modified spectrum. The main issue was the use of linear or quadratic equations for predicting F0 at each time stamp. An equation, developed from a training set, represented each of the five syllable tones in Thai. A subjective test showed that synthesizing non-tonal syllables and vowel tones achieved 83% and 97.5% accuracy, respectively. An advantage of the formant synthesizer is that it requires only a small amount of memory.

## 5. Thai speech recognition

Fig. 3 illustrates a diagrammatic representation of the major components of general ASR. For other languages, such as English, there have been major advancements in each of these areas. For Thai ASR, we cluster research issues into five groups: general Thai ASR applications, acoustic modeling, language modeling, noise robustness, and prosody.

### 5.1. Development of Thai ASR

The first discussion of Thai isolated syllable recognition appeared in the thesis of Pathumthan (1987), where distinctive features were extracted from speech utterances and used for recognition. Since 1995, many academic institutions have presented their inventions for Thai ASR. Development started with isolated word recognition, including numeral speech recognition (Pensiri and Jitapunkul, 1995; Pornsukchandra et al., 1997) and polysyllabic word recognition (Ahkuputra et al., 1997; Kiat-arpakul et al., 1995b). Most works have focused on evaluations of several recognition algorithms, such as artificial neural networks (ANN), dynamic time warping (DTW), and hidden Markov models (HMM). Results have conclusively shown that these algorithms are also efficient for Thai speech.

Some successive works have been conducted on connected word recognition using subword models such

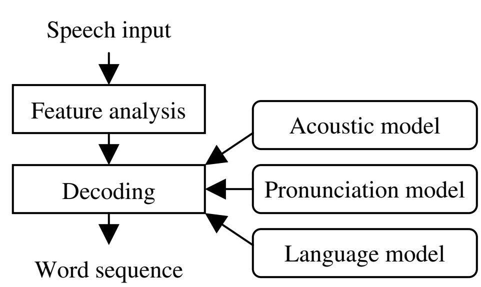


Fig. 3. An architecture of ASR.

as basic Thai phonemes modeled by HMM (Deemagarn and Kawtrakul, 2004; Karnjanadecha et al., 2001). Other research sites have focused on the use of syllable units (Thubthong and Kijsirikul, 1999b). Using syllable units is generally believed to be suitable for Thai since most Thai words are either single syllables or compounds of the syllables. Moreover, the incorporation of other dominant features such as tones and stress is easily performed with the syllable-based system. Although the argument for using the syllable unit is compelling, the problem encountered is that the number of units becomes roughly a few thousand, making context-dependent modeling impractical.

A few years ago, research was conducted on applications of continuous speech recognition (CSR) with limited vocabularies (less than 1000 words), such as an email access system (Sattayapanich and Kawtrakul, 2003), a telephone banking system (Kittipiyakul et al., 2004), and a spelling recognition system (Pisarn and Theeramunkong, 2004a; Pisarn and Theeramunkong, 2004b; Pisarn et al., 2005). All of these works applied state-of-the-art speech recognition algorithms. There were several papers describing the development of large vocabulary continuous speech recognition (LVCSR) for Thai (Kanokphara et al., 2003b; Suebvisai et al., 2005; Tarsaku and Kanokphara, 2002; Thienlikit et al., 2004). In (Thienlikit et al., 2004), an automatic approach to lexicon building and language modeling for Thai LVCSR in the domain of newspaper reading was proposed. The system covered 5000 of the most frequent words found in the newspaper over a 1-year period. With an approximate perplexity of 140, a nearly 10% word-error rate was obtained. In (Suebvisai et al., 2005), a large set of Thai utterances from GlobalPhone (Schultz, 2002) and NECTEC-ATR corpus (Kasuriya et al., 2003a) was used to build a general acoustic model, and a statistical language model in a medical domain was built from a 350,000-word training set provided by the Defense Language Institute (DLI). After optimizing the phone inventory and pronunciation dictionary, the best system achieved approximately a 20% word-error rate. In this work, introduction of either a tone-explicit acoustic model or tonal features did not improve the overall system performance. To improve LVCSR performance, Kanokphara et al. (2003b) and Tarsaku and Kanokphara (2002) applied the same concept for modeling pronunciation variation. Explicit models of pronunciation variation prepared by linguists were incorporated in the speech decoding network. The modified decoding network provided more accurate phoneme labeling.

In addition to the state-of-the-art HMM modeling, there were several trials using other recognition engines (speech decoders) for Thai speech recognition, including an ANN (Maneenoi et al., 1997), a combination of fuzzy logic and ANN (Wutiwiwatchai et al., 1998a; Wutiwiwatchai et al., 1998b), and support vector machines (SVM) (Thubthong and Kijsirikul, 2000b). These papers showed that these algorithms were also applicable to Thai isolated word recognition.

### 5.2. Acoustic modeling

As mentioned above, CSR has been extensively explored since 2000. One of the fundamental issues in Thai CSR is to optimize recognition units for use in acoustic modeling. Besides the common phoneme units, Kanokphara (2003a) evaluated the use of larger units, in which the basic phoneme units were combined considering syllable structures. As described in Section 2.1, an initial consonant of a Thai syllable can be either a single consonant or a cluster, and the vowel part can be either a single vowel or a diphthong. Since some consecutive phonemes, such as clustered consonants and diphthongs, are difficult to acoustically segment, it is more advantageous to define them separately from the single-phoneme units. Furthermore, in contrast to English, in Thai, consonants are pronounced differently dependent on their location in the syllable, either in an initial or final position. The final consonants are therefore defined separately from the initial consonants. Kanokphara (2003a) carried out an experiment, where a basic set of 35 single phonemes was compared with a proposed set of 76 units based on the above idea. Results showed much improved efficiency for the larger set over the basic set. In another similar work, Maneenoi et al. (2002) conducted a comparative experiment between a set of onset-rhyme units (a rhyme consists of a V and a $C_f$) and the basic phoneme unit, and found that the onset-rhyme scheme was superior to the phoneme unit scheme. Another advantage of the onset-rhyme inventory is that syllable tones are mostly expressed in the rhyme portion and thus only the rhyme units are considered when incorporating tone features. It is noted that combining several phonemes is actually a concept native to context-dependent modeling, which is commonly acknowledged to be superior to context-independent models.

Development of an accurate ASR system requires a large speech corpus. During the first stages of Thai CSR research, when no such large corpus was available, attempts were made to construct a Thai acoustic model by mapping from acoustic models of other languages, specifically Japanese (Kasuriya et al., 2002). Suebvisai et al. (2005) also attempted to deal with the lack of available data by using an acoustic model of some other language as an initial model for Thai. Three acoustic models were compared in their experiment. The first acoustic model was trained on only a small set of Thai utterances taken from the NECTEC-ATR speech corpus (Kasuriya et al., 2003a). The second model was a multilingual seed model retrained by the same Thai utterance set. The last one was an English-mapped model. In the experiment, the monolingual Thai acoustic model clearly showed the highest performance, and the multilingual seed model outperformed the English one.

Optimizing acoustic models under conditions where training data is sparse has also been studied. For example, to construct a recognition dictionary, some have proposed an automatic process using a speech corpus. The dictionaries built from the speech corpus thus cover pronunciation variation of words. Kanokphara et al. (2003b) introduced some heuristic rules to an automatic dictionary generation process, due to the fact that only a small Thai speech corpus was available. Some dominant pronunciation variations of Thai, such as the ambiguity of the two phonemes /r/ and /l/ and the shortening of some long vowels, could be captured from the small corpus using the heuristic rules. Another optimization task proposed by Kanokphara and Carson-Berndsen (2004) was an automatic approach for building a decision tree used in tied-state acoustic modeling. The approach required an initial small set of linguistic rules for clustering phonemes. The initial set was then expanded to a full decision tree based on a given speech corpus. It is noted that this novel approach is a compromise between an entirely handcrafted decision tree and a completely data-driven approach.

In the area of development of speech features used in acoustic modeling, Sricharoenchai et al. (2002) proposed an auditory model, which was evaluated for Thai vowel recognition. The model applied a bark scale corresponding to human perception of spectral features and conducted a simple distance measurement for recognition. Karnjanadecha and Kimsawad (2002) evaluated a feature set based on discrete cosine transformation (DCT). The proposed feature set, called DCTC, captured the global spectral shape of each speech frame using DCT basis vectors modified by a bilinear frequency warping function, which simulated human hearing. When applied to Thai digit string tasks with different restrictions on string length, an approximately 7% improvement was obtained by DCTC when compared to the normal Mel-frequency cepstral coefficients (MFCC). It should be mentioned as well that the development of speech features is not a Thai-specific research issue.

### 5.3. Language modeling

State-of-the-art language modeling is currently based on either a finite-state grammar (regular grammar) or an n-gram model. The finite-state grammar is often used in tasks where a restricted grammar is applicable, such as in form filling and voice dialing. The n-gram model is useful when unrestricted grammar is needed, such as in dictation. Kasemsiri et al. (2000) realized the problem of an overly restricted grammar based on the finite-state model coded by hand, and proposed using an abduction algorithm to generate possible word sequence hypotheses given language constraints. The finite-state grammar was automatically modified to cope with new word sequence hypotheses.

To date, Thai has turned out to be one of the most ambiguous languages encountered in automatic language processing research. As mentioned previously, Thai writing has neither explicit word nor sentence boundaries. Thai is also different from ideographic languages like Chinese and Japanese, as it has a spelling system where connected characters can be pronounced differently in different contexts. Moreover, the definition of words is not clear as there are many ways to construct compound words. Therefore, developing an adequate method for segmenting Thai text into a sequence of words still remains an unresolved problem. Without a precise automatic tool for word segmentation, developing a large speech and text corpus for CSR is a costly and time-consuming process. Thienlikit et al. (2004) tackled this problem by developing a completely automatic approach to word segmentation, G2P, dictionary generation, and text-corpus construction. Instead of using a word segmentation tool based on a handcrafted dictionary, he proposed using another segmentation tool developed by Aroonmanakun (2002). This tool segmented Thai text into a sequence of syllable-like units, namely pseudo-morphemes (PMs), using PM-structure rules and a PM n-gram model. The segmented PMs were then combined to form new words according to mutual information values. The automatic approach achieved a performance level almost equal to the basic semi-automatic approach. However, one disadvantage of the proposed method was the large number of lexical words generated in the automatically-constructed dictionary. Furthermore, as a data-driven process, words generated by the approach were likely to be specific only to the given training text.

### 5.4. Noise robustness

Papers regarding noise robustness have not been Thai-specific but have been implemented for Thai as case studies. Similar to outcomes for other languages, although isolated word recognition can achieve high accuracy in laboratories, it fails to be effective in real situations, where input speech is often corrupted by environmental and channel noises. Work which has tried to improve system robustness has been based on the development of robust speech features, such as a trajectory feature (Muenpinij, 2004) and a wavelet-based feature (Thatphithakkul and Kruatrachue, 2004).

One related paper explored which of two proposed phoneme sets – a 35 unit set comprising all the basic Thai phonemes, or a larger syllable-structure based set composed of 76 units – was most suitable for ASR under noisy conditions (Boonpiam et al., 2005). Results showed no effect of noisy environment on the phoneme inventory set. This indicates that the 76-phone set used in several works (Boonpiam et al., 2005; Kanokphara, 2003a) is optimal for Thai regardless of noise conditions.

### 5.5. Prosody in ASR

In Thai, as different tones result in different lexical words, analysis of tones is a critical aspect for Thai speech recognition. In Section 3.1.1, we reviewed relevant papers describing tone analysis, modeling, and classification. In this section, studies incorporating the knowledge of tones or the tone model into ASR are reviewed.

Over the years, there have been many attempts to utilize tones in ASR for Thai as well as for other tonal languages,

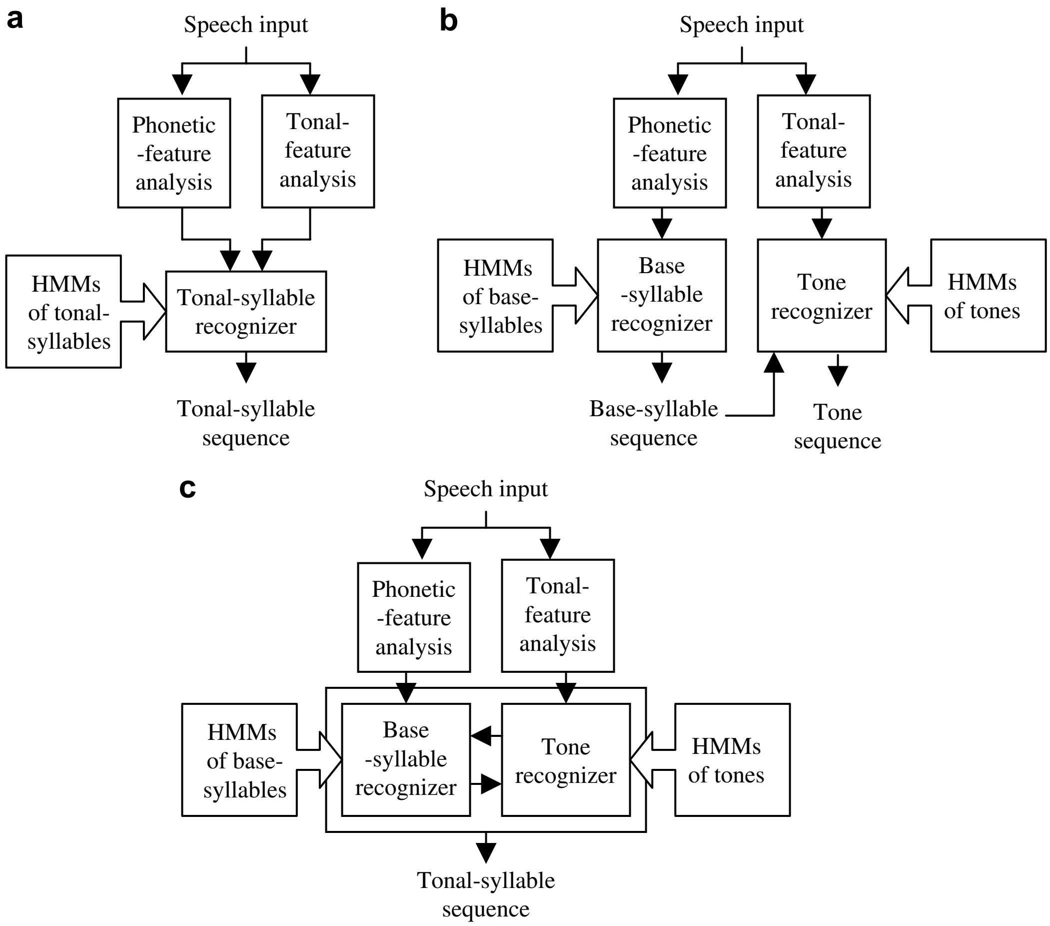


Fig. 4. Three ways to incorporate tones in ASR: (a) joint detection, (b) sequential detection, and (c) linked detection (Demeechai and Makelainen, 2001).

such as Chinese. Here, we classify the published papers into two groups. The first group has attempted to incorporate tone features into isolated word recognition (Thubthong and Kijsirikul, 1999a; Thubthong and Kijsirikul, 2000a). In addition, Thubthong and Kijsirikul (2000a) conducted studies using other prosody features including stress and duration for improving the ASR accuracy.

The second group of works investigated adding tones into continuous speech recognition (CSR). A paper by Demeechai and Makelainen (2001) concluded that using tones in the state-of-the-art CSR could be performed in three ways as shown in Fig. 4. The first way (Fig. 4a), called a joint detection model, involved adding a set of tone features to speech feature vectors. Phoneme/syllable HMMs could be designed as either tone or non-tone explicit models. The second way (Fig. 4b) was termed a sequential detection model, for which a two-pass processing method was developed. The first pass involved a general speech recognition system which yielded phoneme/syllable outputs with time-alignments in the form of either a phoneme/syllable lattice or N-best hypotheses of phoneme/syllable sequences. The second pass used five HMMs representing five Thai tones to rescore the output of the first pass. The last way (Fig. 4c), a linked detection model, involved a one-pass method that computed base-phoneme/syllable (regardless of tones) probabilities and tone probabilities simultaneously, using the Viterbi decoding process. A comparative experiment showed that the best performance was given by the last method. Since the most difficult problem in using tones is the high variation of tone features, a successive work (Demeechai and Makelainen, 2000) tried to reduce the variation by incorporating heuristic rules to modify the tone-feature values.

Speech recognition for clinical applications is another area of investigation pursued by many research sites around the world. In Thailand, Thubthong and Kayasith (2004a,b) developed a specific ASR application for Dysarthric persons. Tone features were also taken into account to improve the recognition accuracy.

## 6. Thai speech applications

Since basic research of Thai speech, as well as the development of large Thai resources, is still limited and, in certain areas, still in the initial stages, research into high-level speech applications presents quite a challenge. According to past literature reviews, three applications including spoken dialogue systems, speech translation, and speaker recognition have been explored. It is noted that the speaker recognition issue is not particularly language-dependent.

However, it is worth mentioning here for a complete review of speech technology evaluated with Thai speech data.

### 6.1. Spoken dialogue systems

Wutiwiwatchai and Furui (2003b) initiated the first complete Thai spoken dialogue system in the domain of hotel reservation. Their subsequent works investigated a novel spoken language understanding (SLU) component (Wutiwiwatchai and Furui, 2006; Wutiwiwatchai and Furui, 2003a; Wutiwiwatchai and Furui, 2004a), which was suitable for languages with weak grammars such as Thai. The SLU model consisted of three parts: concept extraction based on weighted finite-state automata; goal identification using a pattern classifier; and, concept-value extraction based on simple rules. The model was claimed to be able to handle sentences with highly flexible grammar. Moreover, the SLU model can be trained by a partially annotated corpus and hence is expected to be applicable to other languages and dialogue domains as well. A more efficient method of concept extraction introduced in their subsequent work was a hybrid statistical and structural semantic model, which was implemented on the basis of weighted finite-state automata (Wutiwiwatchai and Furui, 2004b).

In 2005, Suchato et al. (2005) set up another Thai spoken dialogue system in the domain of call routing, which helped connect a desired person to a customer. They evaluated two dialogue flows, a one-step request where the customer input both the contact name and his/her department, and a two-step request where the name of the department needed to be input prior to the name of the person within that department. Although the system was simple, it showed a higher potential for the use of this speech recognition engine over a telephone network.

### 6.2. Speech translation

An important pioneering work proposed by Schultz et al. (2004) was an English–Thai speech-to-speech translation system in the medical diagnosis domain. The paper described a two-way speech-to-speech translation system between Thai and English for dialogues in the limited medical domain, where the English speaker was a doctor and the Thai speaker was a patient. The system consisted of three major parts, speech recognition, language translation, and speech synthesis. A Thai speech recognizer was built using a seed multi-lingual acoustic model and retrained by the hotel reservation speech utterances taken from the NECTEC-ATR speech corpus, disregarding tone information. The lexicon used for ASR contained 734 words in the medical diagnosis domain. The recognizer, which had been evaluated with speech utterances from the same domain, achieved 85% word accuracy. For language translation, text in the source language was parsed and converted to an interlingua via machine translation using the Interchange Format (IF), and target-language sentences were generated from the parsed result. The IF was customized to cope with Thai syntax. Three Thai-specific characteristics were addressed in the IF; the use of a special term to indicate the gender of a person, affirmation expressions having meanings other than simply "yes", and the separation of words from the main verb to indicate feasibility and other modalities. To generate Thai sentences, manually written semantic/syntactic grammars and lexicons were employed. In the Thai speech synthesizer, a limited-domain Festival system covering 522 words was created. It contained 235 sentences selected from the hotel reservation part of the NECTEC-ATR corpus (Black and Lenzo, 2000). A year later, a particular investigation to improve the ASR part of the system was made (Suebvisai et al., 2005). An optimized ASR module was acoustically trained with a joint set which comprised the Thai GlobalPhone corpus and Thai Babylon Medical-domain corpus. Rather than a single pronunciation dictionary, a multi-variant dictionary was created using simple rules such as omitting the difference between /l/ and /r/. Incorporating tone features into ASR gave no clear improvement. The best ASR module achieved 18.2% word error rate.

### 6.3. Speaker recognition

Research works on text-dependent speaker identification (Kasuriya et al., 2001; Wutiwiwatchai et al., 1999) and speaker verification (Wutiwiwatchai et al., 2001) using Thai utterances have also been carried out. However, the algorithms used have been mostly language-independent. Several pattern matching algorithms such as DTW, HMM, and Gaussian mixture model (GMM) have been applied with some well-known speech features, such as MFCC and linear prediction coefficients (LPC). Only one paper concerning Thai thus far, by Tanprasert et al. (1999), has been presented. They conducted a speaker-identification experiment using text prompts with different tones. Six text prompts, one for each of the five Thai tones and one for mixed tones, were read by speakers. This experiment clearly showed that a text with mixed tones was the most effective for speaker recognition.

## 7. Language resources

Kawtrakul et al. (2002) summarized Thai language resources which are used mainly for text processing. This section extends that summary by reviewing both Thai text and speech resources. Dictionaries which are a fundamental resource for language processing are also mentioned at the end of this section.

### 7.1. Text resources

Table 5 gives details for some Thai text corpora mainly used for language processing research. The NECTEC ORCHID corpus (Charoenporn et al., 1997) is a medium size text resource with several annotations including

Table 5
Thai text resources

| Corpus | Organization | Size | Details |
|---|---|---|---|
| ORCHID | NECTEC | 568,316 words | – Thai junior encyclopedias and NECTEC technical papers<br>– Sentence/word segmentation<br>– Part-of-speech tagged |
| NAiST corpus | Kasetsart University | 60,511,974 words | – Magazines<br>– Sentence/word segmentation |

part-of-speech (POS), word and sentence boundaries, and pronunciation. It contains approximately 43,000 sentences covering 570,000 words from Thai junior encyclopedias and NECTEC technical papers. ORCHID has been widely exploited in many areas of Thai language processing and also has been used as the original text source for several speech corpora such as a phonetically-balanced set (Wutiwiwatchai et al., 2002) and a prosody-annotated corpus for speech synthesis, named TSynC (Hansakunbuntheung et al., 2003b).

The NAiST text corpus (Kawtrakul, 1995) was created with the primary aim of collecting magazine documents for training and evaluating a writing assistance system. The system functioned to assist writers in proofing their documents. It is a very large text database with word and sentence boundary tags. It has also been employed in other research such as Thai noun phrase analysis (Pengphon et al., 2002) and named entity recognition (Chanlekha et al., 2002).

### 7.2. Speech resources

There are only a few research sites that have consistently contributed to speech research and published papers describing the development and assessment of Thai speech corpora. Most research sites have developed their own in-house corpora which have not been publicly distributed. Table 6 summarizes the current available Thai speech resources. The first corpus, namely the ThaiARC, has provided a set of digitized audio in various styles of Thai speech such as royal speeches, academic lectures, and oral literatures. The purpose of the corpus, however, is linguistic learning, not speech processing.

The speech technology section of NECTEC is one major organization that develops large Thai speech resources (Shuichi et al., 2000; Sornlertlamvanich and Thongprasirt, 2001). A progress report of the project was given in 2002 (Thongprasert et al., 2002). In the same year, Wutiwiwatchai et al. (2002) presented a procedure to select a phonetically-balanced sentence set, creating a subset extracted from several continuous speech recognition corpora developed in NECTEC. Tarsaku and Kanokphara (2002) constructed a semi-automatic tool for phoneme-boundary annotation. Regarding corpora for speech recognition research, two corpora of NECTEC have been described in two separately published papers. The first one, namely the NECTEC-ATR Thai speech corpus (Kasuriya et al., 2003a), was made under the collaboration of NECTEC and ATR in Japan. It contained three sets: a set of 5000 frequently used words, a set of phonetically balanced sentences, and a set of hotel reservation dialogues. The other

Table 6
Thai speech resources

| Corpus | Organization | Purpose | Details |
|---|---|---|---|
| ThaiARC (http://thaiarc.tu.ac.th) | Thammasat University | An archive of Thai digitized audio/speech for learning purpose | – Samples of Thai dialects<br>– Samples of various Thai speech styles<br>– Samples of Thai regional folktales<br>– Samples of Thai poetry |
| NECTEC-ATR | NECTEC | Various Thai speech utterances for ASR research | – Financial support by ATR, Japan<br>– A set of 5000 frequently-used words<br>– A set of phonetically-balanced sentences<br>– A set of hotel reservation dialogues<br>– 54 h from 48 speakers (24 males/24 females), reading style in a clean environment |
| LOTUS (http://www.nectec.or.th/rdi/lotus) | NECTEC | Well-designed speech utterances for 5000-word dictaion systems | – A set of phonetically-distributed sentences<br>– Three 5000-word covered sets for training, development testing, and evaluation testing<br>– 70 h from 48 speakers (24 males/24 females), reading style in clean and office environments |
| TSynC-1 | NECTEC | Corpus-based unit-selection Thai speech synthesis | – Triphone and tritone-coverage speech utterances<br>– 13 h from a fluent-speaking female speaker<br>– Prosody-tags prepared for corpus-based unit-selection speech synthesis |
| GlobalPhone (http://www.cs.cmu.edu/~tanja/GlobalPhone) | Canegie Mellon University | Multilingual speech corpus for LVCSR systems | – Newspaper reading<br>– Over 300 h from 1500 native speakers of more than 15 languages<br>– For Thai, 20 h from 90 native speakers covering 14,039 sentences, 260,000 words, and 7400 distinctive words |

corpus, LOTUS, was the first Thai large vocabulary continuous speech recognition (LVCSR) corpus, which contained several sets of read speech covering 5000 words from a domain of magazine and feature articles. NECTEC also developed a speech corpus, called TSynC-1, containing 13 h of fluent speech of one speaker. It was used for corpus-based unit-selection speech synthesis. Prosody tags as well as comprehensive phoneme boundary markers were given in the corpus.

Multilingual GlobalPhone corpus, under the GlobalPhone project (Schultz, 2002), is a very large collection of read speech, mainly in the newspaper domain. More than 300 h of transcribed speech were spoken by 1500 natives from over 15 languages including Thai. The Thai data set consists of 20 h from 90 native speakers in Bangkok, Thailand; each read 160 sentences on average. In total, 14,039 sentences covering 260,000 words and 7400 distinctive words were recorded. The corpus was mainly designed for language-independent acoustic modeling and fast-bootstrapping of ASR systems for new languages (Schultz and Waibel, 1997; Suebvisai et al., 2005).

### 7.3. Dictionaries

Dictionaries are one of the most necessary components for many language processing technologies. Thai dictionaries have been partly described in Section 3.3.1 and Table 3. They are available in one of the following three ways. "Publicly available": fully-downloadable for research and development purposes, "web application": interactively searchable over the internet, or "proprietary": usable with a license agreement. In Thai, dictionaries have been widely employed in several language processing areas such as sentence breaking (Charoenpornsawat and Sornlertlamvanich, 2001), word segmentation for machine translation (Sornlertlamvanich, 1993), grapheme-to-phoneme conversion for text-to-speech synthesis (Mittrapiyanurak et al., 2000a) and language modeling (Kawtrakul et al., 2002).

## 8. Conclusion and future directions

This paper has reviewed the research and development of Thai speech processing technology, covering a gap of more than ten years since the last complete review paper was published by Luksaneeyanawin in 1993 (Luksaneeyanawin, 1993). In recent years, research has expanded considerably, as shown in Fig. 5, which plots the number of publications, which were accessible to us, in their respective speech areas. The graph shows that, with few exceptions, the number of publications has been steadily increasing year by year. Distributions of publications in each specific area are illustrated in Fig. 6. From the chart, we can see that research on fundamental speech analysis has decreased, except in the area of tone analysis. Tone analysis and research on TTS continues to be a focus, while there are steadily increasing efforts directed toward ASR and high-level speech applications.

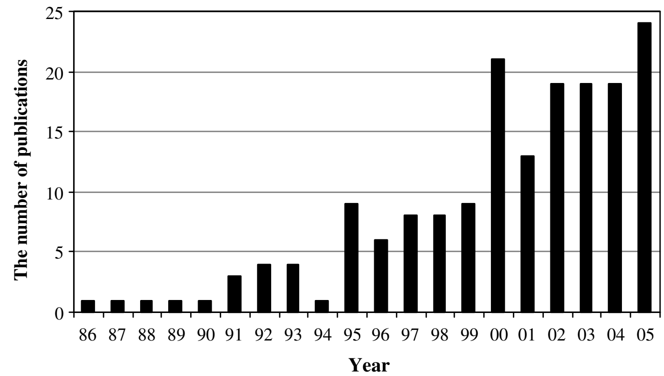


Fig. 5. Distribution of publications in Thai speech processing technology.

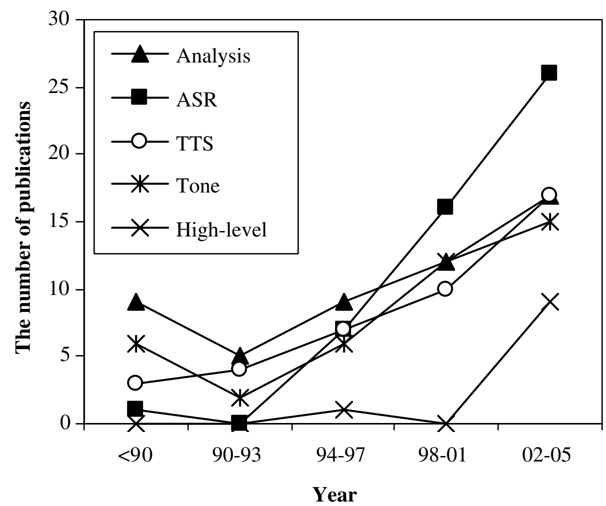


Fig. 6. The number of publications in Thai speech processing technology distributed by topic.

Although Thai ASR has been widely investigated, there is still a lack of research at state-of-the-art level; likewise, there has been no benchmark evaluation of LVCSR. One of the most important obstacles delaying the development of a Thai LVCSR is the lack of a large annotated text corpus for language modeling. Constructing such a corpus requires an accurate word segmentation tool, which could shorten the hand-correction process. Current segmentation tools produce more than 90% accuracy but the accuracy is significantly lower when applied to text in new domains (Aroonmanakun, 2002). Due to the difficulty currently encountered in constructing LVCSR engines, researchers have tended to focus on other Thai-specific issues, such as tones. Tone analysis and modeling itself is useful for TTS, but provides little improvement for ASR as shown in many publications on Chinese. The current state-of-the-art TTS, based on variable-length unit-selection, has been carried out at only one research site, NECTEC. Hence, the advancement of Thai TTS remains slow. Considering the aforementioned limitations, the following are significant issues the authors would suggest Thai researchers to turn their attention to:

(1) Fundamental research
   – Improving Thai word segmentation and G2P conversion with systems which can be easily adapted to new domains.
   – Analyzing and predicting prosodies using a large, real (not only collected in laboratories) speech corpus.
   – Developing large standard speech/text corpora in various domains, environments, and speaking styles.
(2) TTS
   – Optimizing unit-selection algorithms and corpus creation so that the size of speech corpus is reduced and hence adding new speakers is made feasible.
   – Analyzing the most prominent suprasegmental features that can enhance speech naturalness. For example, in Thai, tone and intonation were found to be minor problems compared to duration, phrase-break, and stress.
(3) ASR
   – Developing LVCSR algorithms that comply with incomplete word segmentation and G2P conversion.
   – Evaluating with real noisy or spontaneous speech instead of read speech.
(4) High-level applications
   – With limitations presented in the existing fundamental tools, research and development on high-level applications such as spoken dialogue systems, speech-to-speech translation, speech summarization, etc. remain challenging. Several medium-level modules such as speech understanding, syntactic/semantic parsers, and text generation are needed. Extensive collaboration among research sites in order to create standard resources for research and development is a key to achieving future goals.

## Acknowledgements

The authors would like to thank all researchers who have contributed to Thai speech processing and apologize for any paper unintentionally overlooked in our literature review. Opinions and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of any particular research organization. The authors would also like to express thanks to anonymous reviewers for their constructive suggestions and comments on the first version of the manuscript.

## References

Abramson, A.S., 1962. The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments, Vol. 20. Indiana U. Research Center in Anthropology, Folklore, and Linguistics, Bloomington.

Abramson, A.S., 1979. The coarticulation of tones: an acoustic study of Thai. In: Panupong, V., Kullavanijaya, P., Tingsabadh, K., Luangthongkum, T. (Eds.), Studies of Tai and Mon-Khmer phonetics. Chulalongkorn University.

Abramson, A.S., Svastikula, K., 1982. Intersections of tone and intonation in Thai. In: International Congress of Linguistics.

Ahkuputra, V., Jitapunkul, S., Luksaneeyanawin, S., 2001. Modeling vowel coarticulations from acoustic signal: a case study using Thai language with nine spreading monophthongs. International Journal of Phonetics.

Ahkuputra, V., Jitapunkul, S., Luksaneeyanawin, S., Maneenoi, E., 2000. Direct classification of Thai monophthongs on two-dimenstional acoustic-phonetic feature spaces in linear, Mel, bark, and bark difference frequency scales, Acoustics Society of America (ASA) Meeting.

Ahkuputra, V., Jitapunkul, S., Ponrsukchandra, W., Luksaneeyanawin, S., 1997. A speaker-independent Thai polysyllabic word recognition using Hidden Markov model. In: IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing (PACRIM), pp. 593–599.

Anivan, S., 1988. Evaluation of Bangkok tones, International Symposium on Language and Linguistics.

Aroonmanakun, W., 2002. Collocation and Thai word segmentation. In: Joint International Conference of SNLP-Oriental COCOSDA, pp. 68–75.

Aroonmanakun, W., 2005. A unified model of Thai word segmentation and romanization. In: Pacific Asia Conference on Language, Information and Computation (PACLIC), pp. 205–214.

Aroonmanakun, W., Thubthong, N., Wattuya, P., Kasisopa, B., Luksaneeyanawin, S., 2004. Automatic Thai transcriptions of English words, Southeast Asian Linguistics Society Conference 14 (SEALS 14).

Black, A., Lenzo, K., 2000. Limited domain synthesis. In: International Conference on Spoken Language Processing (ICSLP), Beijing, China.

Black, A., Taylor, P., 1997. Assigning phrase breaks from part-of-speech sequences. In: European Conference on Speech Communication and Technology (EUROSPEECH), pp. 995–998.

Boonpiam, V., Thatphithakkul, N., Wutiwiwatchai, C., 2005. A study of acoustic modeling for Thai noisy speech recognition. In: Electrical Engineering Conference (EECON), Thailand, pp. 1053–1056 (in Thai).

Bradley, C.B., 1911. Graphic analysis of the tone-accents of Siamese language. Journal of American Oriental Society.

Burnham, D., Francis, E., Webster, D., Luksaneeyanawin, S., 1996. Perception of lexical tone across languages: evidence of a linguistic mode of processing. In: International Conference on Spoken Language Processing (ICSLP), pp. 2514–2517.

Chanlekha, H., Kawtrakul, A., Varasrai, P., Mulasas, I., 2002. Statistical and heuristic rule based model for Thai named entity recognition. In: Joint International Conference of SNLP-Oriental COCOSDA, pp. 326–329.

Chanyapornpong, S., 1983. A Thai Syllable Separation Algorithm. Asian Institute of Technology, Pathumthani, Thailand.

Charnvivit, P., Thubthong, N., Maneenoi, E., Luksaneeyanawin, S., Jitapunkul, S., 2003. Recognition of intonation patterns in Thai utterances. In: European Conference on Speech Communication and Technology (EUROSPEECH), pp. 137–140.

Charoenporn, T., Sornlertlamvanich, V., Isahara, H., 1997. Building a large Thai text corpus – part-of-speech tagged corpus: ORCHID, Natural Language Processing Pacific Rim Symposium (NLPRS).

Charoenpornsawat, P., Kijsirikul, B., Meknavin, S., 1998. Feature-based Thai unknown word boundary identification using Winnow. In: IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), pp. 647–650.

Charoenpornsawat, P., Sornlertlamvanich, V., 2001. Automatic sentence break disambiguation for Thai. In: International Conference on Computer Processing of Oriental Languages (ICCPOL), pp. 231–235.

Chotimongkol, A., Black, A., 2000. Statistically trained orthographic-to-sound models for Thai. In: International Conference on Spoken Language Processing (ICSLP), pp. 533–554.

Deemagarn, A., Kawtrakul, A., 2004. Thai connected digit speech recognition using Hidden Markov models. In: International Conference on Speech and Computer (SPECOM), pp. 731–735.

Demeechai, T., Makelainen, K., 2000. New method for mitigation of harmonic pitch errors in speech recognition of tone languages. In: IEEE Nordic Signal Processing Symposium (NORSIG), pp. 303–306.

Demeechai, T., Makelainen, K., 2001. Recognition of syllables in a tone language. Speech Communication 33 (3), 241–254.

Fujisaki, H., 2004. Information, prosody, and modeling – with emphasis on tonal features of speech. In: International Conference on Speech Prosody, pp. 1–10.

Fujisaki, H., Ohno, S., Luksaneeyanawin, S., 2003. Analysis and synthesis of F0 contours of Thai utterances based on the command-response model. In: International Congress of Phonetic Sciences.

Gandour, J., 1979. Perceptual dimensions of tone: Thai. In: Liem, N.D. (Ed.), Southeast Asian Studies. Australian National University.

Gandour, J., Potisuk, S., Dechongkit, S., 1994. Tonal coarticulation in Thai. International Journal of Phonetics.

Gandour, J., Potisuk, S., Ponglorpisit, S., Dechongkit, S., 1991. Inter- and intraspeaker variability in fundamental frequency of Thai tones. Speech Communication 10 (4), 355–372.

Gandour, J., Tumtavitikul, A., Satthamnuwong, N., 1999. Effects of speaking rate on Thai tones. Phonetica 56, 123–134.

Haas, M.R., 1980. The Thai System of Writing. Program in Oriental Languages. Spoken Language Services, Inc., New York.

Hansakunbuntheung, C., Rugchatjaroen, A., Wutiwiwatchai, C., 2005a. Space reduction of speech corpus based on quality perception for unit selection speech synthesis. In: International Symposium on Natural Language Processing (SNLP), pp. 127–132.

Hansakunbuntheung, C., Thangthai, A., Wutiwiwatchai, C., Siricharoenchai, R., 2005b. Learning methods and features for corpus-based phrase break prediction on Thai. In: European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1969–1972.

Hansakunbuntheung, C., Tesprasit, V., Siricharoenchai, R., Sagisaka, Y., 2003a. Analysis and modeling of syllable duration for Thai speech synthesis. In: European Conference on Speech Communication and Technology (EUROSPEECH), pp. 93–96.

Hansakunbuntheung, C., Tesprasit, V., Sornlertlamvanich, V., 2003b. Thai tagged speech corpus for speech synthesis. In: International Conference on Speech Databases and Assessments (Oriental-COCOSDA).

Higbie, J., Thinsan, S., 2002. Thai Reference Grammar: The Structure of Spoken Thai. Orchid Press, Bangkok.

Hirst, D., Di Cristo, A., 1998. A survey of intonation systems. In: Hirst, D., Di Cristo, A. (Eds.), Intonation Systems. Cambridge University Press, Cambridge, pp. 1–45.

Inrut, J., Yuanghirun, P., Paludkong, S., Nitsuwat, S., Limmaneepraserth, P., 2001. Thai word segmentation using combination of forward and backward longest matching techniques. In: International Symposium on Communications and Information Technology (ISCIT), pp. 37–40.

Jitapunkul, S., Luksaneeyanawin, S., Ahkuputra, V., Wutiwiwatchai, C., 1998. Recent advances of Thai speech recognition. In: IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), pp. 173–176.

Jittiwarangkul, N., Luksaneeyanawin, S., Jitapunkul, S., 2000. Syllable segmentation algorithm for Thai connected speech. In: International Symposium on Natural Language Processing (SNLP), pp. 243–251.

Kanokphara, S., 2003a. Syllable structure based phonetic units for context-dependent continuous Thai speech recognition. In: European Conference on Speech Communication and Technology (EUROSPEECH), pp. 797–800.

Kanokphara, S., Tesprasit, V., Thongprasirt, R., 2003b. Pronunciation variation speech recognition without dictionary modification on sparse database. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 764–767.

Kanokphara, S., Carson-Berndsen, J., 2004. Automatic question generation for HMM state tying using a feature table. In: Australian International Conference on Speech Science and Technology (SST), pp. 322–327.

Karnjanadecha, M., Kimsawad, P., 2002. A comparison of front-end analyses for Thai speech recognition. In: International Conference on Spoken Language Processing (ICSLP), pp. 2141–2144.

Karnjanadecha, M., Kimsawad, P., Tanthanakit, P., 2001. HMM based speech recognition of continuous Thai digits. In: International Symposium on Communications and Information Technology (ISCIT), pp. 271–274.

Kasemsiri, W., Kimpan, C., Kongkachandra, R., 2000. Refined language modeling for Thai continuous speech recognition. In: Symposium on Natural Language Processing (SNLP), pp. 252–262.

Kasuriya, S., Jitsuhiro, T., Kikui, G., Sagisaka, Y., 2002. Thai speech recognition by acoustic models mapped from Japanese. In: Joint International Conference of SNLP-Oriental COCOSDA, pp. 211–216.

Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Jitsuhiro, T., Kikui, G., Sagisaka, Y., 2003a. NECTEC-ATR Thai speech corpus. In: International Conference on Speech Databases and Assessments (Oriental-COCOSDA).

Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Kanokphara, S., Thatphithakkul, N., 2003b. Thai speech corpus for speech recognition. In: International Conference on Speech Databases and Assessments (Oriental-COCOSDA).

Kasuriya, S., Wutiwiwatchai, C., Achariyakulporn, V., Tanprasert, C., 2001. Text-dependent speaker identification via telephone based on DTW and MLP. In: IASTED International Conference on Modeling, Identification, and Control (MIC), pp. 190–194.

Kawtrakul, A., 1995. A Lexibase model for writing production assistant system. In: International Symposium on Natural Language Processing (SNLP).

Kawtrakul, A., Suktarachan, M., Varasai, P., Chanlekha, H., 2002. A state of the art of Thai language resources and Thai language behavior analysis and modeling. In: COLING 2002 post-conference workshops: the 3rd Workshop on Asian Language Resources and International Standardization, Taiwan.

Kawtrakul, A., Thumkanon, C., Poowarawan, Y., Varasrai, P., Suktarachan, M., 1997. Automatic Thai unknown word recognition. In: Natural Language Processing Pacific Rim Symposium (NLPRS), pp. 341–348.

Khaorapapong, T., Karnjanadecha, M., Inthavisas, K., 2004. Synthesis of vowels and tones in Thai language by articulatory modeling. In: International Conference on Spoken Language Processing (INTERSPEECH), pp. 1909–1912.

Khruahong, S., Nitsuwat, S., Limmaneepraserth, P., 2003. Thai syllable segmentation for text-to-speech synthesis by using suited-syllable-structure mapping. In: International Symposium on Communications and Information Technology (ISCIT).

Kiat-arpakul, R., 1996. Speech Recognition: Small-Vocabulary Isolated-Word Phone-Based System. Kasetsart University.

Kiat-arpakul, R., Fakcharoenphol, J., Keretho, S., 1995a. A combined phoneme-based and demisyllable-based approach for Thai speech synthesis. In: International Symposium on Natural Language Processing (SNLP), pp. 361–369.

Kiat-arpakul, R., Hemunnopjit, P., Keretho, S., 1995b. A phoneme-based isolated-word approach towards Thai numeral speech recognition. In: International Symposium on Natural Language Processing (SNLP), pp. 354–360.

Kittipiyakul, S., Pornsukchandra, W., Triamlumlerd, S., 2004. Telephone banking by speech recognition. In: ECTI International Conference (ECTI-CON).

Klatt, D.H., 1987. Review of text to speech synthesis conversion for English. Journal of the Acoustic Society of America 59, 1208–1221.

Kongkachandra, R., Pansang, S., Sripramong, T., Kimpan, C., 1998. Thai intonation analysis in harmonic-frequency domain. In: IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), pp. 165–168.

Luangthongkum, T., 1977. Rhythm in Standard Thai. University of Edinburgh, Edinburgh.

Luksaneeyanawin, S., 1983. Intonation in Thai. University of Edinburgh.
Luksaneeyanawin, S., 1985. The phonetic correlates and the functions of speech pause in Thai. In: International Conference on Sino-Tibetan Language and Linguistics.
Luksaneeyanawin, S., 1989. A Thai text-to-speech system. In: Regional Workshop on Computer Processing of Asian Languages (CPAL), pp. 305–315.
Luksaneeyanawin, S., 1992. Three-dimensional phonology: a historical implication. In: International Symposium on Language and Linguistics, pp. 75–90.
Luksaneeyanawin, S., 1993. Speech computing and speech technology in Thailand. In: International Symposium on Natural Language Processing (SNLP), pp. 276–321.
Mamoru, S., Satoshi, H., 2001. Thai morphological analyses based on the syllable formation rules. Journal of Information Processing, 15.
Maneenoi, E., 1998. Thai Vowel Phoneme Recognition Using Artificial Neural Networks. Chulalongkorn University, Bangkok.
Maneenoi, E., Ahkuputra, V., Luksaneeyanawin, S., Jitapunkul, S., 2002. Acoustic modeling of onset-rhyme for Thai continuous speech recognition. In: Australian International Conference on Speech Science and Technology (SST), pp. 462–467.
Maneenoi, E., Jitapunkul, S., Luksaneeyanawin, S., Ahkuputra, V., 2000. Thai monophthongs classification using CDHMM. The Journal of the Acoustic Society of America 108 (5), 2575.
Maneenoi, E.J., S., Ahkuputra, V., Wutiwiwatchai, C., 1997. Modification of BP algorithm for Thai speech recognition. In: Natural Language Processing Pacific Rim Symposium (NLPRS), pp. 343–348.
Meknavin, S., Charoenpornsawat, P., Kijsirikul, B., 1997a. Feature-based Thai word segmentation. In: Natural Language Processing Pacific Rim Symposium (NLPRS), pp. 41–48.
Meknavin, S., Kijsirikul, B., 1997b. Thai grapheme-to-phoneme conversion. In: Burnham, D. (Ed.), Interdisciplinary Approaches to Language Processing: The International Conference on Human and Machine Processing of Language and Speech, NECTEC.
Mittrapiyanurak, P., Sornlertlamvanich, V., 2000. The automatic Thai sentence extraction. In: International Symposium on Natural Language Processing (SNLP), pp. 23–28.
Mittrapiyanurak, P., Hansakunbuntheung, C., Tesprasit, V., Sornlertlamvanich, V., 2000a. Issues in Thai text-to-speech synthesis: the NECTEC approach. In: NECTEC Annual Conference, Bangkok, pp. 483–495.
Mittrapiyanurak, P., Hansakunbuntheung, C., Tesprasit, V., Sornlertlamvanich, V., 2000b. Improving naturalness of Thai text-to-speech synthesis by prosodic rule. In: International Conference on Spoken Language Processing (ICSLP), pp. 334–337.
Mixdorff, H., Burnham, D., 2005. Auditory–visual perception of syllabic tones in Thai. In: Auditory–Visual Speech Processing (AVSP).
Mixdorff, H., Charnvivit, P., 2004. Visual cues in Thai tone recognition. In: International Symposium on Tonal Aspects of Languages (TAL), pp. 143–146.
Mixdorff, H., Luksaneeyanawin, S., Charnvivit, P., Thubthong, N., 2003. Analysis and modeling of syllable duration for Thai speech synthesis. In: International Congress of Phonetic Sciences.
Mixdorff, H., Luksaneeyanawin, S., Fujisaki, H., Charnvivit, P., 2002. Perception of tone and vowel quality in Thai. In: International Conference on Spoken Language Processing (ICSLP), pp. 753–756.
Muenpinij, B., 2004. Automatic Thai-speech Recognition System by Using Normalization Along Trajectory. King Monkut University of Technology Thonburi, Bangkok (in Thai).
Narupiyakul, L., 1999. The phonetic transcription of Thai words. In: International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS).
Navas, E., Hernáez, I., Ezeiza, N., 2002. Assigning phrase breaks using CARTs for Basque TTS. In: International Conference on Speech Prosody, pp. 527–531.
Palmer, A., 1969. Thai tone variants and the language teacher. Language Learning 19, 287–299.
Pansombat, N., Leelarasamee, E., Luksaneeyanawin, S., 2002. Synthesis of stressed and unstressed syllables in Thai. In: Electrical Engineering Conference (EECON), pp. EL54–EL58 (in Thai).
Pathumthan, T., 1987. Thai Speech Recognition Using Syllable Units. Chulalongkorn University, Bangkok (in Thai).
Pengphon, N., Kawtrakul, A., Suktarachan, M., 2002. Word formation approach to noun phrase analysis for Thai. In: Joint International Conference of SNLP-Oriental COCOSDA, pp. 277–282.
Pensiri, R., Jitapunkul, S., 1995. Speaker-independent Thai numeral voice recognition by using dynamic time warping. In: Electrical Engineering Conference (EECON), Pattaya, pp. 977–981.
Pisarn, C., Theeramunkong, T., 2004a. Speed compensation for improving Thai spelling recognition with a continuous speech corpus. Lecture Notes in Artificial Intelligence, 100–111.
Pisarn, C., Theeramunkong, T., 2004b. Thai spelling recognition using a continuous speech corpus. In: International Conference on Computational Linguistics (COLING), pp. 529–534.
Pisarn, C., Theeramunkong, T., Cercone, N., 2005. Spelling recognition for two commonly-used Thai spelling methods. In: IASTED Conference on Artificial Intelligence and Applications (IASTED-AIA), pp. 453–312.
Ponyanun, P., 2003. PPA Salika (Thai text-to-speech synthesis system). In: International Symposium on Research-and-Development Innovation Design.
Poowarawan, Y., 1986. Dictionary-based Thai syllable separation. In: Electrical Engineering Conference (EECON) (in Thai).
Pornsukchandra, W., Jitapunkul, S., Ahkuputra, V., 1997. Speaker-independent Thai numeral speech recognition using LPC and the back propagation neural network. In: Natural Language Processing Pacific Rim Symposium (NLPRS), pp. 585–588.
Potisuk, S., Gandour, J., Harper, M.P., 1996a. Acoustic correlates of stress in Thai. Phonetica 53 (4), 200–220.
Potisuk, S., Harper, M.P., 1996b. CDG: an alternative formalism for parsing written and spoken Thai. In: International Symposium on Language and Linguistics: Pan-Asiatic Linguistics, Salaya, Thailand, pp. 1177–1196.
Potisuk, S., Harper, M.P., Gandour, J., 1995. Speaker-independent automatic classification of Thai tones in connected speech by analysis–synthesis method. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 632–635.
Potisuk, S., Harper, M.P., Gandour, J., 1999. Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method. IEEE Transactions on Speech and Audio Processing 7 (1), 91–102.
Rarunrom, S., 1991. Dictionary-based Thai Word Separation. Chulalongkorn University, Bangkok.
Ratsameewichai, S., Theera-umpon, N., Vilasdechanon, J., Uatrongjit, S., Likit-anurucks, K., 2002. Thai phoneme segmentation using dual-band energy contour. In: International Technical Conference on Circuits/Systems Computers and Communications (ITC-CSCC), pp. 110–112.
Sagisaka, Y., Kaiki, N., Iwahashi, N., Mimura, K., 1992. ATR n-talk speech synthesis system. In: International Conference on Spoken Language Processing (ICSLP), Canada, pp. 483–486.
Saiyot, S., Thipjaksurat, S., Mitatha, S., 2005. Thai speech rules for text-to-speech by using formant synthesis approach. In: National Computer Science and Engineering Conference (NCSEC), Bangkok, Thailand, pp. 521–529.
Saravari, C., Imai, S., 1983. A demisyllable approach to speech synthesis of Thai, a tonal language. Journal of Acoustic Society of Japan (ASJ) 4 (2), 97–106.
Sattayapanich, T., Kawtrakul, A., 2003. Speaker-dependent continuous Thai speech recognition system for e-mail accessing. In: National Computer Science and Engineering Conference (NCSEC).
Sawamipakdee, D., 1990. Developing Thai Analysis Structure Software Under Unix. Thammasat University, Bangkok, Thailand (in Thai).
Schultz, T., 2002. GlobalPhone: a multilingual speech and text databased developed at Karlruhe University. In: International Conference on Spoken Language Processing (ICSLP), pp. 345–348.

Schultz, T., Alexander, D., Black, A., Petersen, K., Suebvisai, S., Waibel, A., 2004. A Thai speech translation system for medical dialogs. In: Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL).

Schultz, T., Waibel, A., 1997. Fast bootstrapping of LVCSR systems with multilingual phoneme sets. In: European Conference on Speech Communication and Technology (EUROSPEECH), pp. 371–373.

Shuichi, I., Hisao, K., Idomuso, D., Yong-Ju, L., Sornlertlumvanich, V., Hsiao-Chuan, W., Ren-Hua, W., 2000. Overview of recent activities in East Asia: speech corpora and assessment. In: International Conference on Speech Databases and Assessments (Oriental-COCOSDA), pp. 8.

Sojka, P., Antos, D., 2003. Context sensitive pattern based segmentation: a Thai challenge. In: Hall, P., Rao, D. (Eds.), Proc. of EACL Workshop on Computational Linguistics for South Asian Languages – Expanding Synergies with Europe, Budapest.

Sornlertlamvanich, V., 1993. Word segmentation for Thai in a machine translation system, Machine Translation. National Electronics and Computer Technology Center, Bangkok, pp. 50–56.

Sornlertlamvanich, V., Thongprasirt, R., 2001. Speech technology and corpus development in Thailand. In: International Conference on Speech Databases and Assessments (Oriental-COCOSDA).

Sricharoenchai, A., Kawtrakul, A., Ongroungrueng, S., 2002. Thai vowel phoneme recognition based on auditory model. In: Joint International Conference of SNLP-Oriental COCOSDA, pp. 334–337.

Suchato, A., Punyabukkana, P., Mana-aporn, P., Pumprao, T., Kawewong, A., 2005. EngCC: a Thai language speech-automated contact center. In: International Symposium on Natural Language Processing (SNLP), pp. 247–252.

Suebvisai, S., Charoenpornsawat, P., Black, A., Woszczyna, M., Schultz, T., 2005. Thai automatic speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 857–860.

Surinpaiboon, S., 1985. A Stressed/unstressed Syllable System in Thai Polysyllabic Words. Chulalongkorn University, Bangkok, Thailand (in Thai).

Taisetwatkul, S., Kanawaree, W., 1996. Thai Speech Synthesis. Chulalongkorn University, Bangkok (in Thai).

Tanprasert, C., Wutiwiwatchai, C., Sae-Tnag, S., 1999. Text-dependent speaker identification using neural network on distinctive tone marks, IJCNN.

Tarsaku, P., Kanokphara, S., 2002. A study of HMM-based automatic segmentation for Thai continuous speech recognition system. In: Joint International Conference of SNLP-Oriental COCOSDA, pp. 217–220.

Tarsaku, P., Sornlertlamvanich, V., Thongprasirt, R., 2001. Thai grapheme-to-phoneme using probabilistic GLR parser. In: European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1057–1060.

Tesprasit, V., Charoenpornsawat, P., Sornlertlamvanich, V., 2003a. A context-sensitive homograph disambiguation in Thai text-to-speech synthesis. In: Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL), pp. 103–105.

Tesprasit, V., Charoenpornsawat, P., Sornlertlamvanich, V., 2003b. Learning phrase break detection in Thai text-to-speech. In: European Conference on Speech Communication and Technology (EUROSPEECH), pp. 325–328.

Thairatananond, Y., 1981. Towards the Design of a Thai Text Syllable Analyzer. Asian Institute of Technology, Pathumthani, Thailand.

Thatphithakkul, N., Kruatrachue, B., 2004. Denoise speech recognition based on wavelet transform using threshold estimation. In: Electrical Engineering Conference (EECON), Thailand (in Thai).

Theeramunkong, T., Usanavasin, S., 2000. Non-dictionary-based Thai word segmentation using decision trees. In: The First International Conference on Human Language Technology Research (HLT), pp. 1–5.

Thienlikit, I., Wutiwiwatchai, C., Furui, S., 2004. Language model construction for Thai LVCSR. In: Acoustic Society of Japan (ASJ) Meeting.

Thongprasert, R., Sornlertlamvanich, V., Cotsomrong, P., Subevisai, S., Kanokphara, S., 2002. Progress report on corpus development and speech technology in Thailand, Oriental-COCOSDA.

Thubthong, N., Kayasith, P., 2004a. Incorporated tone model speech recognition for Thai dysarthria. In: The 11th Biennial International Conference of the International Society for Augmentative and Alternative Communication (ISAAC).

Thubthong, N., Kayasith, P., 2004b. Tone modeling for Thai dysarthric speech recognition, Rehabilitation Engineering and Assistive Technology Society of North America Annual Conference (RESNA).

Thubthong, N., Kijsirikul, B., 1999a. Improving isolated Thai digit speech recognition using tone modeling. In: Electrical Engineering Conference (EECON), Bangkok, pp. 163–166.

Thubthong, N., Kijsirikul, B., 1999b. A syllable-based connected Thai digit speech recognition using neural network and duration modeling. In: IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Phuket, Thailand, pp. 785–788.

Thubthong, N., Kijsirikul, B., 2000a. Improving connected Thai digit speech recognition using prosodic information. In: National Computer Science and Engineering Conference (NCSEC), Bangkok, pp. 63–68.

Thubthong, N., Kijsirikul, B., 2000b. Support vector machines for Thai phoneme recognition. In: International Conference on Intelligent Technologies (INTECH), pp. 206–213.

Thubthong, N., Kijsirikul, B., Luksaneeyanawin, S., 2001a. Stress and tone recognition of polysyllabic words in Thai speech. In: International Conference on Intelligent Technologies (INTECH), Bangkok, pp. 356–364.

Thubthong, N., Kijsirikul, B., Pusittrakul, A., 2001b. A method for isolated Thai tone recognition using combination of neural networks. Computational Intelligence 18 (3), 313–335.

Thubthong, N., Pusittrakul, A., Kijsirikul, B., 2001c. Tone recognition of continuous Thai speech under tonal assimilation and declination effects using half-tone model. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFK) 9 (6), 815–825.

Thubthong, N., Kijsirikul, B., Luksaneeyanawin, S., 2002. Tone recognition in Thai continuous speech based on coarticulation, intonation and stress effects. In: International Conference on Spoken Language Processing (ICSLP), pp. 1169–1172.

Tingsabadh, K., Abramson, A.S., 1999. Illustrations of the IPA: Thai. Handbook of the International Phonetic Association. Cambridge University Press, Cambridge.

Tungthangthum, A., 1998. Tone recognition for Thai. In: IEEE Asia Pacific Conference on Circuits and Systems (APCCAS).

Wutiwiwatchai, C., Achariyakulporn, V., Kasuriya, S., 2001. Improvement of speaker verification for Thai language. In: European Conference on Speech Communication and Technology (EUROSPEECH), pp. 775–778.

Wutiwiwatchai, C., Achariyakulporn, V., Tanprasert, C., 1999. Text-dependent speaker identification using LPC and DTW for Thai language. In: IEEE Region Ten Conference on Digital Signal Processing Applications (TENCON), pp. 674–677.

Wutiwiwatchai, C., Cotsomrong, P., Suebvisai, S., Kanokphara, S., 2002. Phonetically distributed continuous speech corpus for Thai language. In: International Conference on Language Resources and Evaluation (LREC), pp. 869–872.

Wutiwiwatchai, C., Furui, S., 2006. A multi-stage approach for Thai spoken language understanding. Speech Communication 48 (3–4), 305–320.

Wutiwiwatchai, C., Furui, S., 2003a. Confidence scoring for ANN-based spoken language understanding. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 566–571.

Wutiwiwatchai, C., Furui, S., 2003b. Pioneering a Thai language spoken dialogue system, Acoustic Society of Japan (ASJ) Meeting.

Wutiwiwatchai, C., Furui, S., 2004a. Belief-based nonlinear rescoring in Thai speech understanding. In: International Conference on Spoken Language Processing (ICSLP), pp. 2129–2132.

Wutiwiwatchai, C., Furui, S., 2004b. Hybrid statistical and structural semantic modeling for Thai multi-stage spoken language understanding, Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL), pp. 2–9.

Wutiwiwatchai, C., Jitapunkul, S., Luksaneeyanawin, S., Ahkuputra, V., 1998a. Thai polysyllabic word recognition using fuzzy-neural network. In: International Conference on Spoken Language Processing (ICSLP), pp. 767–780.

Wutiwiwatchai, C., Jitapunkul, S., Maneenoi, E., Luksaneeyanawin, S., 1998b. A new strategy of fuzzy-neural network for Thai numeral speech recognition. In: International Conference on Spoken Language Processing (ICSLP), pp. 763–766.

**Chai Wutiwiwatchai** received his B.Eng. (the first honor) and M.Eng. degrees in electrical engineering from Thammasat and Chulalongkorn University in Thailand in 1994 and 1997, respectively. In 2001, he received a scholarship from the Japanese government to pursue a Ph.D. at the Furui laboratory at Tokyo Institute of Technology in Japan and graduated in 2004. He has been conducting research in the Speech Technology Section of the National Electronics and Computer Technology Center (NECTEC) in Thailand since 1997 and now acts as the Head of Speech Technology Section. His research interests include speech and speaker recognition, natural language processing, and human–machine interaction. He is a member of the Acoustical Society of Japan.

**Sadaoki Furui** received B.S., M.S. and Ph.D. degrees in mathematical engineering and instrumentation physics from Tokyo University in Tokyo, Japan in 1968, 1970 and 1978, respectively. He is currently a Professor at the Tokyo Institute of Technology in the Department of Computer Science. He is engaged in a wide range of research encompassing speech analysis, speech recognition, speaker recognition, speech synthesis and multimodal human–computer interaction and has authored or coauthored over 700 published articles. From 1978 to 1979, he held the post as visiting researcher at the Acoustics Research Department of Bell Laboratories in Murray Hill, New Jersey, working on speaker verification. He is a Fellow of the IEEE, the Acoustical Society of America and the Institute of Electronics, Information and Communication Engineers of Japan (IEICE). He has served as President for the Acoustical Society of Japan (ASJ) from 2001 to 2003, the Permanent Council for International Conferences on Spoken Language Processing (PC-ICSLP) from 2000 to 2004, and the International Speech Communication Association (ISCA) from 2001 to 2005. He was a Board of Governor of the IEEE Signal Processing Society from 2001 to 2003. He has served on the IEEE Technical Committees on Speech and MMSP and on numerous IEEE conference organizing committees. He has been Editor-in-Chief of both the Journal of Speech Communication and the Transactions of the IEICE. He is an Editorial Board member of Speech Communication, the Journal of Computer Speech and Language, and the Journal of Digital Signal Processing. He has received both Senior Award and the Society Award from the IEEE Signal Processing Society (1989, 2006). He has received the Yonezawa Prize and the Paper Award from the IEICE (1975, 88, 93, 2003), and the Sato Paper Award from the ASJ (1985, 87). He has received Achievement Awards from the Ministry of Science and Technology and the Ministry of Education, Sports, Culture and Science and Technology, Japan (1989, 2006). He has also received the Technical Achievement Award and the Book Award from the IEICE (2003, 1990). In 1993, he served as an IEEE SPS Distinguished Lecturer.