Available online at www.sciencedirect.com
ScienceDirect
Information Sciences 178 (2008) 122–136
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

Thai spelling analysis for automatic spelling speech recognition

Chutima Pisarn [a,1], Thanaruk Theeramunkong [b,*]

[a] Faculty of Technology and Environment, Prince of Songkla University, 80 Moo 1 Vichitsongkram Road, Kathu, Phuket 83120, Thailand
[b] Sirindhorn International Institute of Technology, Thammasat University, 131 Moo 5 Tiwanont Road, Bangkadi, Muang, Pathumthani 12000, Thailand

Received 3 April 2006; received in revised form 3 April 2007; accepted 6 June 2007

Abstract

Spelling speech recognition can be applied for several purposes including enhancement of speech recognition systems and implementation of name retrieval systems. This paper presents a Thai spelling analysis to develop a Thai spelling speech recognizer. The Thai phonetic characteristics, alphabet system and spelling methods have been analyzed. As a training resource, two alternative corpora, a small spelling speech corpus and an existing large continuous speech corpus, are used to train hidden Markov models (HMMs). Then their recognition results are compared to each other. To solve the problem of utterance speed difference between spelling utterances and continuous speech utterances, the adjustment of utterance speed has been taken into account. Two alternative language models, bigram and trigram, are used for investigating performance of spelling speech recognition. Our approach achieves up to 98.0% letter correction rate, 97.9% letter accuracy and 82.8% utterance correction rate when the language model is trained based on trigram and the acoustic model is trained from the small spelling speech corpus with eight Gaussian mixtures.

Keywords: Spelling analysis; Spelling speech recognition; Automatic speech recognition; Hidden Markov model

1. Introduction

Recently, several works on automatic speech recognition (ASR) for continuous speech have been published in the context of either systems that rely on dictionaries [2,22] or those that can recognize out-of-vocabulary circumstances [4,8,13,26,35]. In the situation of misrecognition and out-of-vocabulary words, a practical and efficient solution to assist the ASR is to equip a system with a spelling speech recognition subsystem in which users can make spelling pronunciation of a word letter by letter. Moreover, spelling speech recognition is a challenging task with high interest for directory assistance services or other applications where a large number of proper names or addresses are handled. Many spelling speech recognition systems were widely developed for several languages, including English [7,16,27], Spanish [24,25], Portuguese [23] and German [3]. Most of

* Corresponding author. Tel.: +66 (0)2 5013505x2004; fax: +66 (0)2 5013505x2001.
E-mail addresses: chutimap@phuket.psu.ac.th (C. Pisarn), thanaruk@siit.tu.ac.th (T. Theeramunkong).
[1] Tel.: +66 (0)7 6276146; fax: +66 (0)7 6276102.

0020-0255/$ - see front matter
doi:10.1016/j.ins.2007.06.034

the existing systems assume a restricted domain by concentrating on how to retrieve the correct name from a given name list (e.g., telephone directory) when a spelling utterance is put in. In [16], a tree-based lexical fast match scheme that utilizes spelling speech recognition was proposed to create a shorter list of candidate English names from very large list. Consisting of a free letter recognizer, a fast matcher, and a re-scorer, the system retrieved the correct name 97.6% of the time. There has recently been a number of spelling speech recognition applications developed for some specific purposes, such as car navigation systems [7,27] with an interactive dialog speech-driven module. In [24,25], the hypothesis-verification approach for recognizing Spanish continuously spoken spelled names over the telephone was proposed. In the hypothesis stage, a set of potential letter sequences derived from an HMM recognizer are fed into a dynamic programming (DP) alignment module. As the result, *n*-best names are retrieved from the dictionary to form a dynamic grammar incorporated with earlier HMMs in the verification stage. This method can retrieve names with a recognition rate of up to 89.6% when the dictionary is composed of approximately 10,000 proper names (mostly city names). In [23], a Portuguese subject-independent general-purpose system was proposed for recognizing an isolated letter over a telephone line using HMM. For German language [3], Bauer and Junkawitsch introduced a fallback strategy into a spelling recognition system to prevent erroneous word recognition of city names over the telephone directory task using the HMM approach.

Although several works on Thai speech recognition have been conducted during the last few years, most of them are limited to some specific tasks such as tone recognition [32], digit recognition [6,30,31] and isolated speech recognition [1,10,11]. There have been few attempts [15,20] to propose a method to recognize continuous speech with large-scale vocabulary. Moreover, putting these systems in practice is still far from solved. The hindrance in broadening research in this area is a lack of a large-scale Thai speech corpus. Recently, with the collaboration between National Electronics and Computer Technology Center (NECTEC) of Thailand and Advanced Telecommunication Research International Institute (ATR) of Japan, a corpus named the NECTEC-ATR Thai continuous speech corpus [12] has been developed. The corpus consists of 5131 isolated words, 16380 sentences and 50 dialogues of hotel reservation speech. There have been several works [21,36] using this corpus on a specific domain or under a quite limited environment. Despite some work on Thai spelling speech recognition (e.g. [19,20]), the area still requires significant investigation. As an interesting domain, spelling speech is much less complex than large-scale vocabulary continuous speech recognition since the vocabulary is limited to the size of the Thai alphabet, thus allowing us to use information of character (letter) sequences to guide the recognition process. Unlike other languages, there are several spelling styles in the Thai language. A simple style is similar to spelling in English where each character is spelled phonetically with its character pronunciation. For instance, the phonetic representation of "หมา" (dog) is */h-@@4/ /m-@@0/ /z-aa0/* in the Thai Phonetic Set (TPS) notation, corresponding to */h2//m2//a2/* in the IPA notation. This is analogous to */d-ii0/ /z-oo0/ /c-ii0/* for 'dog' IPA: /ɦdiː/ /ɦʔoː/ /ɦtɕiː/) in English. The correspondence between TPS and IPA can be found in the Appendix.

There are three more styles in Thai spelling, each introducing additional syllables to assist the listener in clearly understanding spelling utterances. Of these four styles, the style that adds the representative meaningful word after the pronunciation of a character is the most widely used. This paper presents an approach to recognize spelling utterances of this spelling style. However, at present there is no standard corpus that we can use as a precedent for training a spelling speech recognition system. Based on the above background, there are three objectives of this work. The first one is to examine the possibility of applying an existing Thai continuous speech corpus to spelling speech recognition. Although continuous speech utterances are quite different from spelling utterances, it would be beneficial to make use of the existing Thai speech resources. The second purpose is to investigate performance of spelling speech recognition when a relatively small spelling speech corpus is applied. The third objective is to study the effect of using a higher-order gram (i.e., trigram) on recognition performance.

In this work, the spelling utterances are recorded in an office environment. The recognition task is performed on speaker-independent and open-test basis. That is, the system is expected to recognize speech belonging to someone whose speech is never used for training the system, and the character sequences of the utterances being recognized are unseen beforehand.

This paper is organized as follows: In Section 2, Thai phonetic characteristics, alphabet system and spelling methods are presented. Our Thai spelling speech recognition approach is introduced in Section 3. The experimental results and analysis of spelling speech recognition are reported in Section 4. This section compares the

recognition performance gained from two different speech corpora, i.e., a small spelling speech corpus and a large normal speech corpus. The method that applies trigram, instead of bigram, is also evaluated, as are the number of Gaussian mixtures. Finally, a conclusion and future work are given in Section 5.

2. Thai phonetic characteristics, alphabet system and spelling methods

This section briefly explains the Thai phonetic characteristics and alphabet. It also provides an analysis of possible methods to spell Thai words.

2.1. Thai phonetic characteristics, alphabet system and spelling methods

Like most languages, a Thai syllable is composed of four components: (1) initial consonant, (2) vowel (3) final consonant and (4) tone. The phonetic representation of one syllable can be expressed in the form of $/C_i - V^T - C_f/$, where C_i is an initial consonant, V is a vowel, C_f is a final consonant and T is a tone which is phonetically attached to the vowel part. Some initial consonants are consonant clusters. A consonant cluster is a group of consonants that have no intervening vowel. In Thai, some principle consonants can be attached with one of the phones *r*, *l* and *w* to form consonant clusters. Each of the formed consonant clusters has a similar sound to the principle consonant that it is derived from. For example, the sounds of *pr* and *pl* are similar to that of their principle consonant *p*. Composed of 21 initial consonants, 17 consonant clusters, 18 vowel phones, 6 diphthongs and 15 final consonants, the set of 77 phonetic symbols are shown in Table 1. In addition to phonetic symbols, the five phonemic tones are also listed in the table.

Normally, phones are quite different in their durations, especially those in the vowel class. In Thai language, most vowels are arranged in short/long vowel pairs. The set of short/long vowel pairs is *a-aa, i-ii*, *v-vv*, *u-uu*, *e-ee*, *x-xx*, *o-oo*, *@-@@*, *q-qq*, *ia-iia*, *va-vva*, and *ua-uua*. Intuitively, these pairs are easily confusing in the recognition process.

2.2. The Thai alphabet system

Theoretically, the Thai language has 69 letters, which can be grouped into three classes by phonetic expression, i.e., 44 consonants, 21 vowels and 4 tone markers. Some Thai consonant letters share a phonetic sound. Therefore, only 21 consonant phones exist. On the other hand, some vowels can be combined with other vowels, yielding 32 possible phones. Furthermore, Thai has four tone markers for five phonemic tones (middle, low, falling, high, rising) since the middle tone has no tone marker. However, in practice, only 18 letters in the vowel class out of 21 are used. Therefore, 66 letters in total have been used as shown in Table 2.

Table 1
List of Thai phonetic symbols (initial consonants, vowels and final consonants) and phonemic tones

Initial consonant (C_i)		Vowel (V)		Final consonant (C_f)	Tone (T)
Base (21)	Cluster (17)	Base (18)	Diphthong (6)	Final (15)	Tone (5)
p,t,c,k,z,ph,th,ch,k,h,b, d,m,n,ng,r,f,s,h,w,j	*pr,phr,pl,phl,tr,thr,kr,khr, kl,khl,kw,br,bl,kh,fr,fl,dr*	*a,aa,i,ii,v,vv,u,uu,e,ee,x, xx,o,oo,@,@@,q,qq*	*ia,iia,va,vva,ua,uua*	*p∧, t∧, k∧, n∧, m∧, n∧, g∧, j∧, w∧, f∧, l∧, s∧, ch∧, jf∧, ts*	0 Middle 1 Low 2 Falling 3 High 4 Rising

Table 2
Three Thai alphabet classes: consonant, vowel and tone

Basic classes	Letters in each class
Consonant (44)	ก, ข, ฃ, ค, ฅ, ฆ, ง, จ, ฉ, ช, ซ, ฌ, ญ, ฎ, ฏ, ฐ, ฑ, ฒ, ณ, ด, ต, ถ, ท, ธ, น, บ, ป, ผ, ฝ, พ, ฟ, ภ, ม, ย, ร, ล, ว, ศ, ษ, ส, ห, ฬ, อ, ฮ
Vowel (18)	อั, อะ, อ็, อา, อิ, อี, อึ, อื, อุ, อู, เอ, แอ, โอ, อำ, ไอ, ใอ, ฤ, อ์
Tone (4)	อ่, อ้, อ๊, อ๋

Table 3
Two types of vowels

The vowels of the first type	อะ, อา, อิ, อี, อึ, อื, อุ, อู, เอ, แอ, โอ, อำ, ไอ, ใอ
The vowels of the second type	อั, อี๋, ฤ, อ์

Table 4
Pronunciation methods for each alphabet class

Alphabet class	Pronunciation methods
Consonant	1. Consonant core sound + representative word of consonant 2. Consonant core sound
Vowel of the first type	1. /*s*-*a*1//*r*-*a*1/ + vowel core sound 2. Vowel core sound
Vowel of the second type	1. The vowel name
Tone	1. The tone marker name

2.3. Basic pronunciation of Thai alphabet and word spelling methods

There are various styles in pronouncing the Thai alphabet. The consonants can be uttered in either of the following two styles. The first style is simply pronouncing the core sound of a consonant. The letter '**ก**', for example, has a core sound represented by the phonetic sound/*k*-@@0/. Some consonants share an identical core sound. For example, '**ค**', '**ต**', and '**ฆ**' have the same phonetic sound /*kh*-@@0/. In such case, the listener may encounter letter ambiguity. To solve this issue, the second style is generally applied by uttering a core sound of the consonant followed by the representative word of that consonant. Every consonant has a corresponding representative word. For example, the representative word of the letter '**ก**' is "**ไก่**" (meaning: "chicken", sound: /*k*-*a*1-*j*∧/), and that of the letter '**ข**' is "**ไข่**" (meaning: "egg", sound: /*kh*-*a*1-*j*∧/). To express the letter '**ก**' using the second style, the syllable sequence /*k*-@@0/+/*k*-*a*1-*j*∧/is uttered.

Pronunciation of a vowel letter depends on the vowel type. There are two different types of vowels. The first type can be pronounced in two alternative ways. One way is to pronounce the word "**สระ**" (meaning: "vowel", sound: /*s*-*a*1//*r*-*a*1/), followed by the core sound of the vowel. The other is to simply pronounce the core sound of the vowel. The second type can be pronounced by speaking their names. The vowel letters of both types are listed in Table 3. As the last class, tone markers are pronounced by speaking their names. Table 4 summarizes the pronouncing methods stated above.

Spelling a word is to utter each individual letter in the word sequentially. Spelling is a combination of pronouncing each letter in the word. There are four commonly used spelling methods in Thai. For all methods, the vowels of the second type and tones are pronounced by speaking their names. The differences among the methods are in spelling a consonant and a vowel of the first type. To investigate the concept of spelling speech recognition using HMM models, this paper focuses on the first spelling method, which is the most prevalent method in Thai spelling. In this method, the representative word of a consonant is pronounced after its core sound, and a vowel of the first type is pronounced by uttering the word "**สระ**" (sound: /*s*-*a*1//*r*-*a*1/) and then the core sound of the vowel.

3. Thai spelling speech recognition approach

In this work, HMMs are employed as an engine for recognizing a continuous spelling utterance. As a statistical approach, the HMM is widely used in speech recognition research, especially for continuous speech since it has the capability to capture and handle a set of continuous data [37]. Fig. 1 illustrates our HMM-based ASR system, which consists of two major components: the language model and the acoustic model.

In general, the language model can be formulated as a rule-based model or a statistical model. In a speech recognizer for a limited vocabulary environment, the language model can be coded as a set of simple rules. The larger the scope, the harder it is to write a complete set of rules that covers all probabilities. In such situation, a

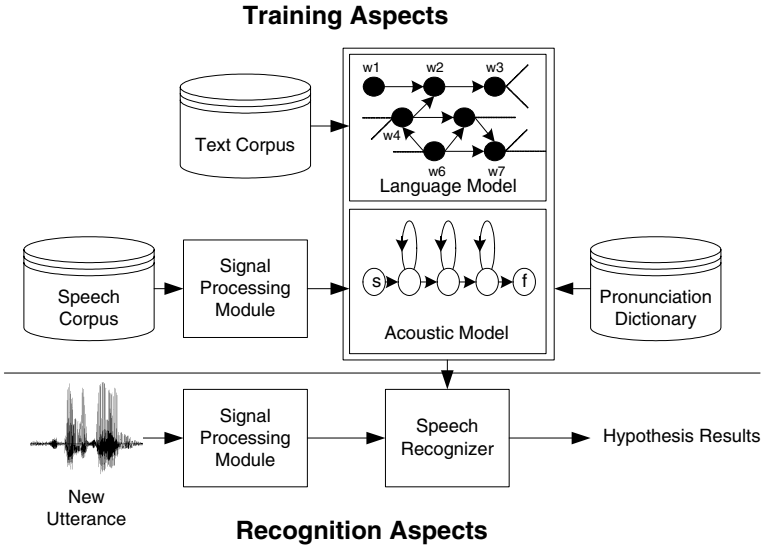

Fig. 1. Typical ASR framework.

statistical model, such as n-gram, can be applied. In Thai language, applying n-gram, especially bigram, can help to constrain the recognition model by pruning a lot of impossible pairs of contiguous characters. For example, a front vowel (a vowel that appears at the beginning of a syllable) cannot be followed by any front vowel or non-front vowel. For example, the front vowel letter '◌ิ' cannot follow the front vowel 'เ' to form ''เเ'' and the non-front vowel '◌า' cannot follow the front vowel 'เ' to form ''เา''. The set of front vowels are {'ใ', 'ไ', 'เ', 'แ'} while the set of non-front vowels are {'◌ั', 'อะ', 'อา', 'อิ', 'อี', 'อึ', 'อื', 'อุ', 'อู', 'อำ', 'อ็', 'อ็'}. Superior to a rule-based model, an n-gram can also help us make a soft decision to determine the most probable successor letter of a given letter. Usually, the larger n is the stronger constraint we have, but we need more training examples to have a reliable estimation.

For the acoustic model, the conventional phone-based HMMs are used to represent phones; one acoustic model for one phone. A series of features extracted from input speech waveform are used to compute probabilities of an acoustic model. The two main parameters that characterize a model are transition probability (a_{ij}) and emission probability ($b_j(o)$) where i and j are any states and o is the observation feature. Training of the acoustic model is performed to assign optimal values to model parameters. To obtain the most plausible hypothesis sequence of letters given a spelling utterance in the domain of spelling speech recognition, it is necessary to search among all possibilities for the letter sequence with the maximum probability. Theoretically the probability of a letter sequence is derived from the product of acoustic probability and language probability. The latter indicates the probability of how often the letter sequence is generated. Setting a weight between these two different probabilities varies the recognition result. As a common method [37], the weight w can be set in the calculation of emission probability $b_j(o)$ based on Gaussian approach as shown in the Eq. (1). Here, it is possible to introduce a Gaussian mixture of multiple probability models to the emission probability.

$$b_j(o) = \left[\sum_{m=1}^{M_i} c_{jm}\aleph\left(o;\mu_{jm},\sum_{jm}\right)\right]^w \tag{1}$$

where w is the weight, M is the number of mixture components in state i, $\aleph(o;\mu_{jm},\sum_{jm})$ is a multivariate Gaussian with the mean μ and covariance matrix $\sum$ (for details, see [37]). Applying a small weight ($w < 1.0$) lowers the emission probability and then makes the recognition decision more dependent on language probability (e.g., the bigram p('ร'|'ก') and p('ย'|'ก') in Fig. 2).

Training the system requires two kinds of corpora; a speech corpus for learning the acoustic model and a text corpus for training the language model. Constructing a speech corpus is a time-consuming task, while it is easy to acquire a large-scaled text corpus for spelling purpose. In Thai, there is no public speech corpus of

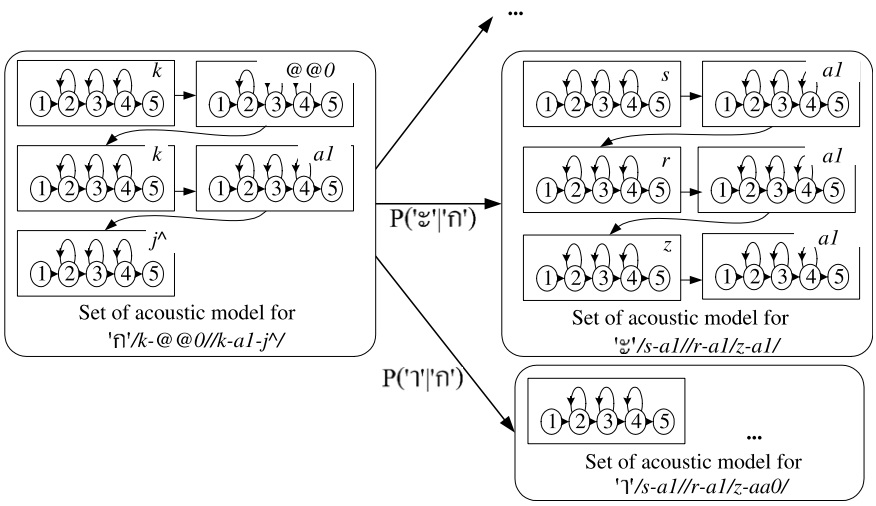

Fig. 2. State transition when the speech is recognized as "**กะ**" or "**กา**".

spelling utterances available, but we have a well-known corpus of general speech utterances called the NECTEC-ATR corpus [12]. So two alternative approaches to learn the acoustic models are (1) to reuse the existing general speech corpus for spelling speech recognition and (2) to construct a corpus of spelling utterances. In the former, we need to consider the difference in utterance speed between normal speech and spelling speech. Starting from scratch, the latter consumes a lot of time and labor in creating a large corpus. With limited resources, we may use a small-scaled spelling speech corpus instead. In this work, these two approaches are compared and evaluated.

A text corpus is used as a source for assigning probabilities to letter sequences to train the language model for spelling speech recognition. Compared to speech data, text data are much easier to obtain. Moreover, different approaches may be suitable for different domains of spelling speech recognition. For example, if a domain is limited to a small set of some proper names, a language model can be trained from that set and the system can yield high accuracy. As a more flexible domain, a general recognizer, which accepts any spelling utterance, usually needs a larger text corpus. However, the larger the text corpus for training is, the more ambiguity the system has to cope with. In this work, both limited and flexible domains are investigated as shown in the next section.

4. Experimental results and analysis of spelling speech recognition

To evaluate the performance of the proposed Thai spelling speech recognition approach, four experiments are conducted. The first experiment is to adjust the ratio weight parameter between acoustic and language models in our spelling speech recognizer. Both spelling speech and normal continuous speech corpora are explored in this experiment. The second experiment is to investigate recognition performance gained when the speed of utterances in the continuous speech corpus are adjusted before using them as the training material. The third experiment is performed to evaluate the method using trigrams, compared to bigrams as the language model. The last experiment is performed to investigate the effect of the number of mixtures on the recognition performance.

4.1. Experimental environment

Two alternative speech corpora are provided as training sets in the experiments. One is a small set of Thai spelling utterances collected from six subjects (three males and three females) by assigning them to spell 150 proper names, resulting in 900 utterances in total. We denote this set as SPELL. The other is a large set of 3900 Thai speech utterances gathered from five males and five females (390 utterances each). Denoted as

NECTEC-ATR, the corpus is part of the NECTEC-ATR corpus [12], which contains normal continuous Thai speech utterances, not spelling utterances. For both corpora, the subjects are requested to spell naturally with a prepared script. To construct the test utterances, another six subjects (three males and three females) are requested to spell 136 proper names, consisting of shop names, company names and person names. The 150 spelled proper names of the training set (SPELL) are different from the 136 spelled proper names of the test set. For the training set, the number of characters is 1122, the average number of letters per name is 7.48, and the number of phonemes is 7665. For the test set, the number of characters is 1149, the average number of letters per name is 8.45, and the number of phonemes is 6654. Each spelling utterance is a continuous speech. However, a short pause between two letters may occur in some parts of the utterance due to the nature of spelling. The speech signals are digitized with a 16-bit A/D converter under the frequency of 16 kHz. The feature vector we applied is the widely used 39-PLP-feature vector that consists of 12 PLP coefficients and the 0-th coefficient, as well as their first and second order derivatives. The PLP (Perceptual Linear Predictive) feature is set to imitate the behavior of the human ear, and is more robust in speaker-independent conditions with computational efficiency and compact representation [9].

More precisely, the experiments are performed under three different domains; closed-type, opened-type and mixed-type language models. The closed-type language model, henceforth called LM1, is constructed from the test transcription, i.e. the 136 proper names. The opened-type model, later denoted by LM3, is trained by a holdout corpus not being used as the test transcription. In this experiment, as the holdout corpus, we use 5971 location names including Thai provinces, districts and sub-districts. The mixed-type model, denoted by LM2, is generated from the combination of the test transcription and the holdout corpus, i.e., 136 proper names plus 5971 location names.

The HTK toolkit [37] is used as the recognition engine in the experiments. The acoustic model used is a set of phone-based HMMs, each of which represents an individual phone. They are context-independent in the sense that the recognition of a phone in an utterance is independent of its preceding and following phones. To implement the HMMs with continuous observations, the two important factors, i.e., the model topology and the number of mixtures, are considered. For the model topology, a phone model is set to a three-state left-to-right model without skip in all experiments. The first experiment finds the optimal weights in cases of SPELL and NECTEC-ATR respectively when the number of mixtures is set to 1. The second experiment examines the recognition performance when the utterance speed of the NECTEC-ATR corpus is adjusted. The third experiment investigates the recognition performance when trigram is applied instead of bigram. The last experiment explores the performance when the number of mixtures ranges between 4 and 16, in both bigram and trigram models. The more mixtures we have in this experiment, the better predictive model we will obtain. However, a higher-mixture models leads to an over-fitting problem and needs more computational time for training.

Due to the length limitation of this paper, only the results of LM2 are displayed in all experiments since the LM2 is considered to be the most natural environment. Generally, LM1 gains the highest performance and LM3 usually obtains a similar figure as LM2 but slightly lower. More detail can be found in [18].

In our experiment, it was observed that the set of phonetic units in SPELL and that in NECTEC-ATR are not exactly identical. The former has fewer phones than the latter due to the limited number of possibilities in spelling utterances compared to normal utterances. Table 5 illustrates the list of phonetic units in each corpus. In the case of vowels, the number in parentheses denotes the possible tone expansions of the vowel. For example, "$a(0-4)$" means the vowel 'a' occupies all five possible tones, that is 0 (middle), 1 (low), 2 (falling), 3 (high) and 4 (rising).

Following the standard evaluation, the recognition performance is usually evaluated in terms of word correction rate (WCR) and word accuracy (WA). However, since the task concerned is the spelling speech recognition (neither normal speech recognition nor word recognition), the original definitions of word correction rate and word accuracy are modified to letter correction rate (LCR) and letter accuracy (LA), respectively. The LCR is defined as the ratio of the number of correct letters to the total number of letters. Slightly different from the LCR, the LA is the ratio of the subtraction of the number of incorrectly inserted letters from the number of correct letters, to the total number of letters. It is obvious that the LA measure is more restrictive than the LCR measure. The following equations define LCR and LA. Here, H is the number of correctly recognized letters, I is the number of incorrectly inserted letters, and N is the total number of actual letters.

Table 5
Acoustic units existing in SPELL and NECTEC-ATR

Part	SPELL (72)	NECTEC-ATR (195)
Ini-Cons	*b, c, ch, d, f, h, j, k, kh, khw, l, m, n, ng, p, ph, pl, r, s, t, th, tr, w, z*	*b, bl, br, c, ch, d, dr, f, fl, fr, h, j, k, kh, khl, khr, khw, kl, kr, kw, l, m, n, ng, p, ph, phl, phr, pl, pr, r, s, t, th, thr, tr, w, z*
Vowel	*@@(0,4), a(0-4), aa(0-1,3-4), e1, ee(0-1), i(0-1,4), ii(0,4), o(0,3), oo(0,2), qq0, u(1,4), uu(0,2-4), uua3, v(1-3),vv0, vva(0,4), xxx(0,4)*	*@(0-4), @@(0-4), a(0-4), aa(0-4), e(0-4), ee(0-4), i(0-4), ia1, ii(0-4), iia(0-4), o(0-4), oo(0-4), q(0-3), qq(0-4), u(0-4), uu(0-4), uua(0-4), v(0-4), vv(0-4), vva(0,4), x(0-4), xx(0-4)*
Fin-Cons	*ch∧, j∧, k∧, m∧, n∧, ng∧, p∧, t∧, w∧,*	*ch∧, f∧, j∧, jf∧, k∧, l∧ m∧, n∧, ng∧, p∧, s∧, t∧, ts∧ w∧*

$$\text{Letter correction rate (LCR)} = \frac{H}{N} \tag{2}$$

$$\text{Letter accuracy (LA)} = \frac{H - I}{N} \tag{3}$$

In addition to LCR and LA, another measure called utterance correction rate (UCR) is also defined to explore how much better the approach can recognize the whole spelling utterance. It is formulated as follows: H_u is the number of correctly recognized utterances, and N_u is the total number of utterances. Note that one utterance corresponds to one word spelled.

$$\text{Utterance correction rate (UCR)} = \frac{H_u}{N_u} \tag{4}$$

4.2. Adjustment of the ratio weight between the acoustic and language models

This section shows the results of investigating the optimal ratio weights when the training speech corpus is either SPELL or NECTEC-ATR. In principle, the ratio weight defines the importance ratio between the acoustic and language models. Using SPELL as the training data, the acoustic model of the phone-based HMMs for Thai spelling speech recognition is trained by a relatively small corpus of 900 spelling utterances. The HMM used is a three-state left-to-right model without skip, with one Gaussian mixture. The applied language model is a bigram model, encoding the occurrence probability of letter pairs. The numbers of bigrams for opened-type and mixed-type models are 1614 and 1656, respectively. In this experiment, ratio weight is varied from 0.05 to 1.0 in order to find the most effective value. The result is shown in Fig. 3. The smaller the weight is, the less important role the acoustic model plays compared to the language model. The lower-right graph in the figure indicates 100-LA, which means the error rate.

The result shows that the weight of 0.2 achieves the best result. This fact indicates that the language model plays a more important role in gaining high performance than the acoustic model. This result is reasonable since the language model in spelling speech recognition is more limited than in general speech recognition. For example, in the case of using bigrams as the language model, the number of possible letter pairs is much smaller than the number of possible word pairs. The closed-type language model (LM1) achieves higher performance than the others, i.e., 93.4% LCR, 92.5% LA, and 53.3% UCR. The mixed-type model (LM2) obtains 89.9% LCR, 89.1% LA, and 37.6% UCR. Even with the hardest problem, the recognition performance of the opened-type model (LM3) is slightly lower than the result of the mixed-type model (LM2), i.e., 88.4% LCR, 87.6% LA, and 32.7% UCR, respectively.

In the case of NECTEC-ATR, the experiment is set up similar to the SPELL experiment. The result indicates that an optimal weight is located approximately between 0.1 and 0.2. However, the weight of 0.1 is selected for further exploration since it gains better performance than the weight of 0.2 in most cases. In this case, the system achieves 93.1% LCR, 92.9% LA, and 55.9% UCR for LM1 (closed-type), 85.7% LCR, 85.3% LA, and 27.6% UCR for LM2 (mixed-type), and 84.0% LCR, 83.3% LA, and 20.3% UCR for LM3 (opened-type). Compared to the results of SPELL (Fig. 3), the recognition performance employing NEC-

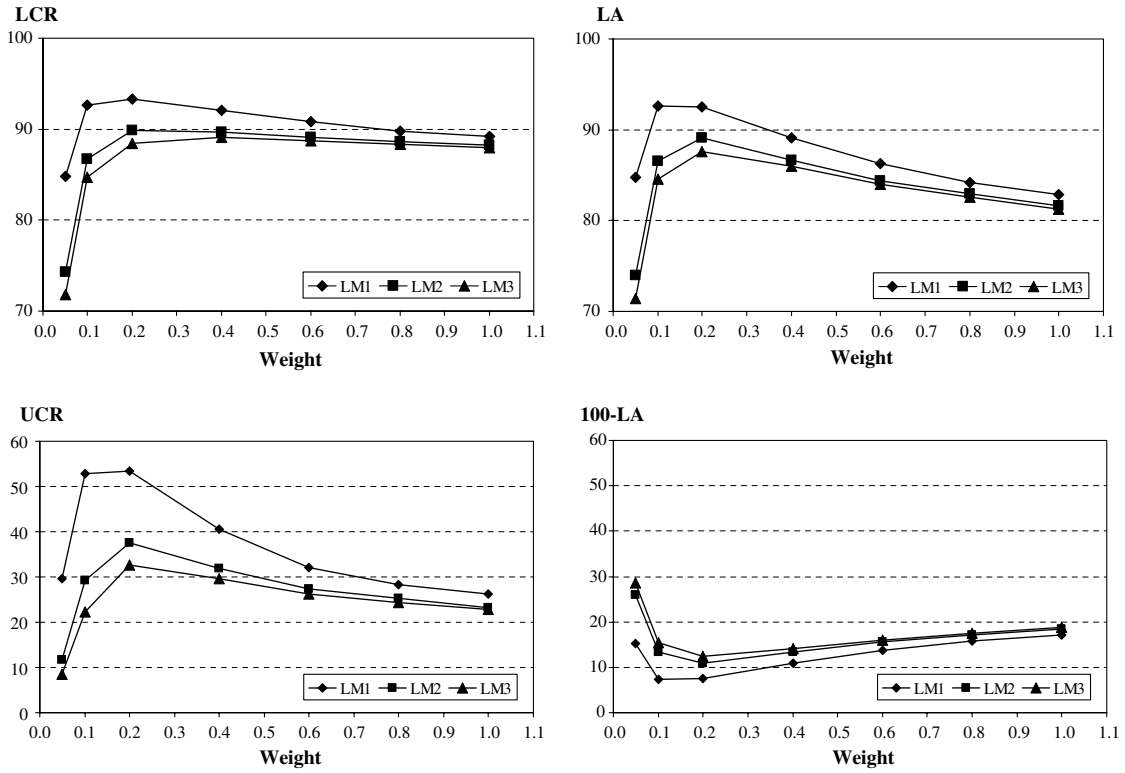

Fig. 3. Recognition performance for several weights (training with SPELL) (upper left = LCR, upper right = LA, lower left = UCR, lower right = (100-LA)).

TEC-ATR is worse in both LM2 and LM3, even though it is a larger corpus. Two possible reasons are (1) phone sequence and (2) utterance speed. For the former, spelling utterances in Thai usually have similar phone sequences since Thai spelling has fixed patterns as shown in Section 2. For the latter, the utterance speeds of a normal utterance and a spelling utterance are quite different. Although it is hard to provide solutions to the first problem, speed adjustment may be useful for the second problem as shown in the next section.

4.3. Duration adjustment for the NECTEC-ATR corpus

This section describes the experiments of adjusting utterance speed in the continuous normal speech corpus. A simple preliminary analysis shows that the utterance speed of a continuous speech and that of a spelling speech are quite different. Therefore, when we use normal continuous speech utterance to train a model for recognizing a spelling speech utterance, we may consider speed compensation to improve the recognition performance.

To this end, we measure the utterance speeds of both corpora by calculating the number of phones per second. To obtain this measurement, all utterances are automatically aligned yielding the information of phones and their durations. The acquired alignment information is then used to calculate the average speed of the utterances. At this step, we remove the silence and short pause duration from the total utterance duration and then divide the resultant duration by the number of phones. As the result, the spelling utterances are approximately 1.53 times slower than the NECTEC-ATR utterances. To compensate for this duration difference, the timing stretching method [5,17,33,34] is used. The method stretches a speech signal by preserving pitch and auditory features of the original signal. In the experiment, the original speech signals are stretched with three scaling factors; 1.25, 1.43 and 1.67 times. These three sets of stretched speech signals are used for training the system to recognize the spelling utterances. The results are compared with the system using the

Table 6
Performance comparison between the stretched utterances of three scaling factors, the original utterances (baseline=1.00ATR) and the spelling utterances

	Type of speech corpus				
	1.00ATR ($w = 0.1$)	1.25ATR ($w = 0.1$)	1.43ATR ($w = 0.1$)	1.67ATR ($w = 0.1$)	SPELL ($w = 0.2$)
LCR	85.7	87.4 (+1.7, +11.9%)	77.5 (−8.2, −57.3%)	85.4 (−0.3, −2.1%)	89.9 (+4.2, +29.4%)
LA	85.3	87.2 (+1.9, +12.9%)	76.9 (−8.4, −57.1%)	85.2 (−0.1, −0.7%)	89.1 (+3.8, +25.9%)
UCR	27.6	31.3 (+3.7, +5.1%)	12.1 (−15.5, −21.4%)	25.0 (−2.6, −3.6%)	37.6 (+10.0, +13.8%)

original speech utterance (1.00ATR), the baseline, as shown in Table 6. They are denoted by 1.25ATR, 1.43ATR and 1.67ATR. Furthermore, the result of using the spelling corpus (SPELL) is also given for comparison. To clarify the performance evaluation in this and later experiments, we use two measures called error reduction (ER) and error reduction rate (ERR). The former is calculated by subtracting the error of the baseline from the error of the proposed method and the latter is derived by dividing the obtained result (ER) by the error of the baseline. In Table 6, two numbers in each bracket are ER and ERR. Here, the baseline used is 1.00ATR. Due to length limitation, only the results of LM2 are provided.

The result shows that the recognition using the 1.25-times-stretched training utterances (1.25ATR) yields the best performance, compared with those of 1.43ATR and 1.67ATR. It was noted that stretching an utterance causes a distortion in the original utterance. The more an utterance is stretched, the more distorted utterance we obtain. From the result, the 1.25ATR achieves higher LCR, LA and UCR. It gains up to 87.4% LCR, 87.2% LA and 31.3% UCR, which are 11.9%, 12.9% and 5.1% ERR over the baseline (1.00 ATR), respectively.

In comparison of 1.25ATR against SPELL, an acoustic model created from the adjusted NECTEC-ATR, a larger corpus, cannot outperform SPELL. This result implies that even if we apply a larger corpus of general continuous speech utterances for training an acoustic model, we cannot gain higher performance since general speech utterances and spelling utterances are quite different in their acoustic characteristics.

4.4. Exploiting trigram language models

In this experiment, first we calculate the perplexity of a language model against unseen test data in order to evaluate how predictive the model is. Similar to entropy, perplexity indicates the level of ambiguity [37]. Low perplexity of a language model means that the model is more predictive. In speech recognition, a language model with low perplexity on the test data tends to achieve better recognition performance, although it is not guaranteed [28]. In our corpus settings, the perplexity of the bigram model is calculated with results of 6.52, 25.41 and 23.96 for the closed-type, the mixed-type and the opened-type domains, respectively. They are 2.12, 12.80 and 18.71 for the trigram model. This implies that the trigram model is more predictive than the bigram model and should obtain better performance. Moreover, the distinct number of names used for training language models in the opened-type and mixed-type domains is merely 136. Therefore, the mixed-type and opened-type will obtain a similar figure of perplexity. It is also possible to get a result that may be contrary to our intuition, such as that we gain higher perplexity for mixed-type than the opened-type for the bigram model. We also investigate how the trigram model performs in spelling speech recognition and compare it to the bigram model. The models of the three training corpora; SPELL, 1.00ATR, and 1.25ATR are explored. The recognition performance is shown in Table 7. Here, the numbers in each bracket are bigram performance, ER (performance gap between bigram and trigram) and ERR.

The result indicates that the trigram model achieves higher performance than the bigram model. The LCR results are 93.1%, 90.5% and 91.1% for the SPELL, the 1.00ATR and the 1.25ATR, which are 31.7%, 33.6%

Table 7
Recognition rate using the trigram model (the numbers in each bracket = bigram performance, ER, ERR)

	Type of speech corpus		
	SPELL	1.00ATR	1.25ATR
LCR	93.1 (89.9, +3.2, +31.7%)	90.5 (85.7, +4.8, +33.6%)	91.1 (87.4, +3.7, +29.4%)
LA	92.5 (89.8, +2.7, +26.5%)	89.8 (85.3, +4.5, +30.6%)	90.8 (87.2, +3.6, +28.1%)
UCR	54.0 (37.6, +16.4, +26.3%)	47.8 (27.6, +20.2, +27.9%)	50.3 (31.3, +19.0, +27.7%)

and 29.4% ERR over the bigram performance, respectively. Moreover, the LA results are 92.5%, 89.8% and 90.8% for the SPELL, the 1.00ATR and the 1.25ATR, which are 26.5%, 30.6% and 28.1% ERR over the bigram performance, respectively. For the UCR result, the ERR are 26.3% (SPELL), 27.9% (1.00ATR) and 27.7% (1.25ATR). In conclusion, the trigram performs approximately 26–34% better than the bigram.

4.5. Exploring the effect of the number of mixtures in acoustic models

This section shows the exploration result of the number of Gaussian mixtures in acoustic models. The focused training corpora are 1.25ATR and SPELL since they obtained better performance in the previous experiments. With the mixed-type bigram and trigram language models, the number of Gaussian mixtures is explored in the range of 1 and 16. The results are shown in Table 8. Here, the baseline is the result of one Gaussian mixture.

The numbers in each bracket illustrate performance improvement over one Gaussian mixture (the baseline). For all conditions (bigram vs. trigram and 1.25ATR vs. SPELL), the model of 8 Gaussian mixtures achieves the best result. In the case of the trigram language model and the acoustic model trained by the SPELL corpus,

Table 8
Performance results of 1, 4, 8 and 16 Gaussian mixtures

Mixture		SPELL (weight 0.2)		1.25ATR (weight 0.1)	
		LM2	Tri-LM2	LM2	Tri-LM2
1	LCR	89.9	93.1	87.4	91.1
	LA	89.1	92.5	87.2	90.8
	UCR	37.6	54.0	31.3	50.3
4	LCR	96.6 (+6.7, +66.3%)	97.9 (+4.8, +69.6%)	91.1 (+3.7, +29.4%)	93.7 (+2.6, +29.2%)
	LA	96.3 (+7.2, +66.1%)	97.7 (+5.2, +69.3%)	91.0 (+3.8, +29.7%)	93.6 (+2.8, +30.4%)
	UCR	71.7 (+34.1, +54.6%)	82.1 (+28.1, +61.1%)	41.5 (+10.2, +14.8%)	59.7 (+9.4, +18.9%)
8	LCR	97.1 (+7.2, +71.3%)	**98.0** (+**4.9**, +**71.0**%)	91.9 (+4.5, +35.7%)	94.2 (+3.1, +34.8%)
	LA	97.0 (+7.9, +72.5%)	**97.9** (+**5.4**, +**72.0**%)	91.9 (+4.7, +36.7%)	94.1 (+3.3, +35.9%)
	UCR	77.1 (+39.5, +63.3%)	**82.8** (+**28.8**, +**62.6**%)	46.0 (+14.7, +21.4%)	61.6 (+11.3, +22.7%)
16	LCR	95.4 (+5.5, +54.5%)	96.4 (+3.3, +47.8%)	91.5 (+4.1, +32.5%)	94.1 (+3.0, +33.7%)
	LA	95.1 (+6.0, +55.0%)	96.2 (+3.8, +49.3%)	91.5 (+4.3, +33.6%)	94.0 (+3.2, +34.8%)
	UCR	61.4 (+23.8, +38.1%)	70.1 (+16.1, +35.0%)	42.5 (+11.2, +16.3%)	61.3 (+11.0, +22.1%)

it gains up to 98.0% LCR, 97.9% LA and 82.8% UCR which are 71.0%, 72.0% and 62.6% ERR over the baseline. With the bigram language model and the acoustic model trained by the SPELL corpus, the results are 97.1% LCR, 97.0% LA and 77.1% UCR which are 71.3%, 72.5% and 63.3% ERR over the baseline. In the cases of the 1.25ATR corpus, the models of 8 Gaussian mixtures also outperform the model of 1 Gaussian mixture. The trigram language model obtains 94.2% LCR, 94.1% LA and 61.6% UCR which are 34.8%, 35.9% and 22.7% ERR. The bigram language model 91.9% gains LCR, 91.9% LA and 46.0% UCR, which correspond to 35.7%, 36.7% and 21.4% ERR.

5. Conclusion

This paper presented a detailed analysis of Thai spelling in order to develop a spelling HMM-based recognizer. Starting with an analytical introduction to the four methods for spelling Thai words, we proposed an HMM-based approach to recognize spelling utterances when the most common spelling method is used. Lacking a Thai spelling corpus, a small corpus (SPELL) was constructed to investigate our spelling speech recognition approach. As an alternative, we reused the existing Thai continuous speech corpus (NECTEC-ATR) in order to investigate the performance of recognizing spelling utterances using a larger set of Thai continuous speech utterances. Even though the experiments were performed under three different domains; closed-type, opened-type and mixed-type language models, the results of the mixed-type domain are focused since it was the most natural environment. For all experiments, the ratio weight of the acoustic model to the language model was adjusted to gain the optimal results. It was set to 0.2 and 0.1 for SPELL and NECTEC-ATR respectively, as tests showed these values yielded the best performance. A small ratio weight sets the language model to be more important than the acoustic model. It was found that the recognition rate using NECTEC-ATR was worse than SPELL due to the speed difference between these two kinds of speech corpora. Comparing the result of SPELL to that of NECTEC-ATR, the error reduction rates (ERRs) in the letter correction rate (LCR), the letter accuracy (LA) and the utterance correction rate (UCR) were 29.4%, 25.9%, and 13.8%, respectively.

By adjusting the utterance speed, it was possible to improve the recognition performance, but this still resulted in a lower performance than the methods using the small spelling corpus. Comparing the best speed-adjusted corpus (1.25ATR) against the original corpus (1.00ATR), the ERRs in LCR, LA and UCR were 11.9%, 12.9% and 5.1%, respectively. The results were lower than the error reductions of SPELL. It was hard to gain higher performance since general speech utterances and spelling utterances were quite different in their acoustic characteristics, even adjusting the utterance speed.

The trigram model was investigated as an alternative of the bigram language model. With small perplexity, the trigrams could improve the recognition rate over the bigrams in every type of training corpora; SPELL, 1.00ATR, 1.25ATR, especially for the mixed-type language model. The error reduction rates in LCR, LA and UCR range between 26% and 34%. The effect of the number of Gaussian mixtures in acoustic models was also investigated and compared with the results of one-Gaussian-mixture model. For all conditions (bigram/trigram and 1.25ATR/SPELL), the model of 8 Gaussian mixtures achieved the best result. For SPELL as the training corpus, the error reduction rate was up to 60–70%, compared to one Gaussian mixture. For 1.25ATR, we obtained even lower performance: the LCR, LA and UCR error reduction rate of approximately 20–35%. The best performance among all experiments was 98.0% LCR, 97.9% LA and 82.8% UCR, under the condition of the SPELL training corpus, 8-Gaussian-mixture model and the trigram model.

As our internal investigation, we found that letter substitution was the main source of the errors. The errors mostly came from the confusion of similar consonantal and vowel phones as well as the confusion between the pairs of short and long vowels in the spelling of those letters. As for further works, it is necessary to study techniques to recognize all possible four spelling methods simultaneously and explore a method to incorporate spelling speech recognition into the conventional speech recognition.

Acknowledgements

The authors would like to thank National Electronics and Computer Technology Center (NECTEC) for allowing us to use the NECTEC-ATR Thai speech corpus. This work has partly been supported by NECTEC

under project number NT-B-22-I5-38-47-04. The authors also deeply thank Prince of Songkla University for their financial support and encouragement during this research. Finally, the authors express appreciation to all members in the RDI research group at NECTEC for their valuable comments and advice. Last, but not least, the authors would like to thank Dr. Steven Donald Gordon for his precious comments on both contents and writing styles through this paper.

Appendix. Thai phonetic set (TPS)

The Thai phonetic set (TPS) used in this paper is referred from [29]. In order to clarify the notation, we compare the TPS with the IPA (the International Phonetic Alphabet), the notation widely used for the transcription of English and many other languages. The following shows the mapping between TPS and IPA [14].

Initial consonant				Vowel				Final consonant	
Base		Cluster		Base		Diphthong		TPS	IPA
TPS	IPA	TPS	IPA	TPS	IPA	TPS	IPA		
p	p	*pr*	pr	*a*	a	*ia*	i	p∧	p
t	t	*phr*	phr	*aa*	a	*iia*	i	t∧	t
c	tɕ	*tr*	tr	*i*	i	*va*		*k*∧	k
k	k	*kr*	kr	*ii*	i	*vva*		*n*∧	n
z		*khr*	khr	*v*		*ua*	u	*m*∧	m
ph	ph	*pl*	pl	*vv*		*uua*	u	*ng*∧	
th	th	*phl*	phl	*u*	u			*j*∧	j
ch	tɕh	*thr*	thr	*uu*	u			*w*∧	w
kh	kh	*kl*	kl	*e*	e			*f*∧	f
b	b	*khl*	khl	*ee*	e			*l*∧	l
d	d	*kw*	kw	*x*	ɛ			*s*∧	s
m	m	*khw*	khw	*xx*	ɛ			*ch*∧	tɕh
n	n	*br*	br	*o*	o			*jf*∧	p
ng		*bl*	bl	*oo*	o			*ts*∧	t
l	l	*fr*	fr	@					
r	r	*fl*	fl	@@			Tone		
f	f	*dr*	dr	*q*	*γ*		*0*	Middle	
s	s			*qq*	*γ*		*1*	Low	
h	h						*2*	Fall	
w	w						*3*	High	
j	j						*4*	Rising	

References

[1] V. Ahkuputra, S. Jitapunkul, N. Jittiwarangkul, E. Maneenoi, S. Kasuriya, A comparison of Thai speech recognition systems using hidden Markov model, neural network, and fuzzy-neural network, in: Proceedings of The 5th International Conference on Spoken Language Processing, vol. 3, Sydney, Australia, 1998, pp. 715–718.

[2] Y.A. Alotaibi, Investigating spoken Arabic digits in speech recognition setting, Information Sciences 173 (1–3) (2005) 115–139.

[3] J.G. Bauer, J. Junkawitsch, Accurate recognition of city names with spelling as a fall back strategy, in: Proceedings of the Sixth European Conference on Speech Communication and Technology, 1999, pp. 263–266.

[4] I. Bazzi, J. Glass, A multi-class approach for modeling out-of-vocabulary words, in: Proceedings of the International Conference on Spoken Language Processing, Denver, 2002, pp. 1613–1616.

[5] S.M. Bernsee, Time stretching and pitch shifting of audio signals: an overview, <http://www.dspdimension.com/html/time-pitch.html> (accessed 18.10.2006).

[6] A. Deemagarn, A. Kawtrakul, Thai connected digit speech recognition using hidden Markov model, in: Proceedings of the International Conference on Speech and Computer, St. Petersburg, Russia, 2004, pp. 731–735.
[7] L. Delphin-Poulat, Robust speech recognition technique evaluation for telephony server based in-car applications, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2004, pp. 65–68.
[8] A. Ghobakhlou, M. Watts, N. Kasabov, Adaptive speech recognition with evolving connectionist systems, Information Sciences 156 (1–2) (2003) 71–83.
[9] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, Journal of the Acoustical Society of America 87 (4) (1990) 1738–1752.
[10] S. Kanokphara, Syllable structure based phonetic units for context-dependent continuous Thai speech recognition, in: Proceedings of the Eighth European Conference on Speech Communication and Technology, Geneva, Switzerland, 2003, pp. 797–800.
[11] M. Karnjanadecha, P. Kimsawad, A comparison of front-end analyses for Thai speech recognition, in: Proceedings of the International Conference on Spoken Language Processing, Denver, September 2002, pp. 2141–2144.
[12] S. Kasuriya, V. Sornlertlamvanich, P. Cotsompong, T. Jutsuhiro, G. Kikui, Y. Sagisaka, Thai speech database for speech recognition, in: Proceedings of the Oriental-COCOSDA Workshop, 2003, pp. 54–61.
[13] H. Klus, A. Rausch, A general architecture for self-adaptive AmI components applied in speech recognition, in: Proceedings of the 2006 International Workshop on Self-adaptation and Self-managing systems Shanghai, 2006, pp. 72–78.
[14] J. Laver, Principles of Phonetics, Cambridge University Press, New York, 1994.
[15] E. Maneenoi, V. Ahkuputra, S. Luksaneeyanawin, S. Jitapunkul, Acoustic modeling of onset-rhyme for Thai continuous speech recognition, in: Proceedings of the 9th Australian International Conference on Speech Science & Technology, Melbourne, Australia, 2002, pp. 462–467.
[16] C.D. Mitchell, A.R. Setlur, Improved spelling recognition using a tree-based fast lexical match, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, 1999, pp. 597–600.
[17] G. Pallone, Time-stretching and pitch-shifting of audio signals, Application to cinema/video conversion, <http://www.iua.upf.es/activitats/semirec/semi-pallone/index.htm> (accessed 18.10.2006).
[18] C. Pisarn, A study on Thai speech recognition: tone exploitation and spelling recognition, Ph.D. Dissertation, Sirindhorn International Institute of Technology, Thammasat University, 2006.
[19] C. Pisarn, T. Theeramunkong, Improving Thai spelling recognition with tone features, in: T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala (Eds.), Lecture Notes in Artificial Intelligence LNAI-4139, Springer-Verlag, 2006, pp. 388–398.
[20] C. Pisarn, T. Theeramunkong, Speed compensation for improving Thai spelling recognition with a continuous speech corpus, in: F.A. Aagesen, C. Anutariya, V. Wuongse (Eds.), Lecture Notes in Computer Science LNCS-3283, Springer-Verlag, 2004, pp. 100–111.
[21] C. Pisarn, T. Theeramunkong, Incorporating tone information to improve Thai continuous speech recognition, in: Proceedings of the International Conference on Intelligent Technologies, Chiangmai, Thailand, 2003, pp. 84–89.
[22] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (1989) 257–285.
[23] F. Rodrigues, R. Rodrigues, C. Martins, An isolated letter recognizer for proper name identification over the telephone, in: Proceedings of the Ninth Portuguese Conference on Pattern Recognition, 1997, <http://citeseer.ist.psu.edu/502582.htm> (accessed 18.10.2006).
[24] R. San-Segundo, J. Colas, R. Cordoba, J.M. Pardo, Spanish recognizer of continuously spelled names over the telephone, Journal of Speech Communication 38 (2002) 287–303.
[25] R. San-Segundo, J. Colas, R. Cordoba, J.M. Pardo, Detection of recognition errors and out of spelling dictionary names in a spelled name recognizer for Spanish, in: Proceedings of the Seventh European Conference on Speech Communication and Technology, Dinamarca Aalborg, 2001, pp. 2553–2556.
[26] R.M. Stern, F. Liu, Y. Ohshima, T.M. Sullivan, A. Acero, Multiple approaches to robust speech recognition, in: Proceedings of the Workshop on Speech and Natural Language, Human Language Technology Conference, Morristown, NJ, 1992, pp. 274–279.
[27] Y. Su, T. Bai, C.I. Watson, Design and development of a speech-driven control for an in-car personal navigation system, in: Proceedings of the Australasian language Technology workshop, Sydney, Australia, 2005, pp. 224–232.
[28] P. Taylor, R. Caley, A.W. Black, S. King, Edinburgh Speech Tools Library System Documentation Edition 1.2, for 1.2.0, <http://festvox.org/docs/speech_tools-1.2.0/x2921.htm>, 1999 (accessed 18.10.2006).
[29] R. Thongprasirt, V. Sornlertlamvanit, Standardization of Thai corpus development, Technical Report, NECTEC, Thailand, 2002.
[30] N. Thubthong, B. Kijsirikul, Tone recognition of continuous Thai speech under tonal assimilation and declination effects using half-tone model, International Journal of Uncertainty, Fuzziness and Knowledge-Based System 9 (6) (2001) 815–825.
[31] N. Thubthong, B. Kijsirikul, Improving connected Thai digit speech recognition using prosodic information, in: Proceedings the Fourth National Computer Science and Engineering Conference, 2000, pp. 63–68.
[32] A. Tungthangthum, Tone recognition for Thai, in: Proceedings of the IEEE Asia-Pacific Conference on Circuit and System, Beijing, China, 1998, pp. 157–160.
[33] W. Verhelst, M. Roelands, An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, Minneapolis, MN, 1993, pp. 554–557.
[34] Wikipedia: The free encyclopedia, Audio timescale-pitch modification, <http://en.wikipedia.org/wiki/Audio_time_stretching> (accessed 18.10.2006).

[35] J. Wilpon, L. Rabiner, C. Lee, E. Goldman, Automatic recognition of keywords in unconstrained speech using hidden Markov models, IEEE Transactions on Acoustics, Speech, and Signal Processing 38 (1) (1990) 1870–1878.
[36] C. Wutiwiwatchai, S. Furui, Pioneering a Thai language spoken dialogue system, Spring Meeting of Acoustic Society of Japan, 2003.
[37] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book (for HTK version 3.2.1), Cambridge University Engineering Department, 2002.