Thai Spelling Recognition Using a Continuous Speech Corpus

Chutima Pisarn
Sirindhorn International Institute of Technology
131 Moo 5 Tiwanont Rd., Bangkadi,
Muang, Phathumthani, Thailand, 12000
chutimap@siit.tu.ac.th

Thanaruk Theeramunkong
Sirindhorn International Institute of Technology
131 Moo 5 Tiwanont Rd., Bangkadi,
Muang, Phathumthani, Thailand, 12000
thanaruk@siit.tu.ac.th

Abstract

Spelling recognition is an approach to enhance a speech recognizer's ability to cope with incorrectly recognized words and out-of-vocabulary words. This paper presents a general framework for Thai speech recognition enhanced with spelling recognition. In order to implement Thai spelling recognition, Thai alphabets and their spelling methods are analyzed. Based on hidden Markov models, we propose a method to construct a Thai spelling recognition system by using an existing continuous speech corpus. To compensate the difference between spelling utterances and continuous speech utterances, the adjustment of utterance speed is taken into account. Assigning different numbers of states for syllables with different durations is helpful to improve the recognition accuracy. Our system achieves up to 79.38% accuracy.

1 Introduction

Nowadays, several works on automatic speech recognition (ASR) for continuous speech are being developed, not only system that rely on dictionary, but also the recognition on out of vocabulary circumstances. In a situation of misrecognition and out-of-vocabulary words, a practical and efficient solution that would assist the ASR is to equip the system with a spelling recognition subsystem, where users can spell out a word letter by letter. Spelling recognition is a challenging task with a high interest for directory assistance sevices, or other applications where a large number of proper names or addresses are handled. Many works that focus on spelling recognition were widely developed in several languages, for instance, English, Spanish, Portuguese and German. In (San-Segundo et al., 2001) the hypothesis-verification Spanish continuous spelled proper name recognition over the telephone was proposed. In this work, several feature sets were investigated by using models of neural networks. In their succeeding work (San-Segundo et al., 2002), three different recognition architectures, including the two-level architecture, the integrated architecture and the hypothesis-verification architecture, are analyzed and compared. In (Rodrigues et al., 1997), a Portuguese speaker -independent system for recognizing an isolated letter was introduced. The system dealt with speech utterances over a telephone line using Hidden Markov Model (HMM). A number of experiments were made over four different perplexity language models. Mitchell and Setlur (1999) proposed a fast list matcher to select a name from a name list that was created from an *n*-best letter recognizer on spelling over a telephone line recognition task. In (Bauer and Junkawitsch, 1999), an approach is proposed to combine word recognition with spelling recognition in a user-friendly manner as a fall back strategy. As a German city name recognizer, the system was applied to directory assistance services.

Unlike other languages, spelling in Thai has several styles. One of them is similar to spelling in English, i.e., /d-ii//z-oo//g-ii/ for "dog". There are three more methods in Thai spelling, where some syllables are inserted to make it clearer for the hearer. One is to spell out a letter followed by its representative word's utterance. Another way is to mix the former two types. The third method is to spell out a set of letters that form a syllable, followed by its corresponding utterance. So far spelling recognition for Thai language has not been explored yet. One of the main reasons is that there is no standard corpus for this purpose. Creating a corpus of spelled utterances is a time comsuming task. In this work we use the NECTEC-ATR Thai Speech Corpus, a standard continuous Thai speech corpus, for our spelling recognition system. Another objective of this work is to examine how a spelling system can be implemented using a normal Thai continuous speech corpus. That is, as the preliminary stage, we investigate the effects of spelling using such existing corpus.

This paper is organized as follows. In section 2, language characteristics in Thai are introduced. Section 3 presents our recognition framework. The spelling styles for Thai words are discussed in section 4. The experimental results and analysis are shown in section 5. Finally, the conclusion and future works are given in section 6.

2 Thai Language Characteristics

In this section, Thai alphabets, phonetic symbols and the phone components of Thai syllable are described.

2.1 Thai Alphabets

Theoritically, Thai language has totally 69 alphabets which can be basically grouped into three classes of phone expression; consonant, vowel and tone. There are 44, 21, and 4 alphabets for consonants, vowels, and tones, repectively. Some Thai consonant alphabets share the same phonetic sounds. There are only 21 phones for Thai consonants. Since some vowels can be combined with others, there are possible 32 phones. However, in practical spelling manner, only 18 alphabets in the vowel class are mostly used. There are 5 tones in Thai, including one without an alphabet. In conclusion, there are totally 66 alphabets actually used. They are shown in Table 1.

Basic Classes	Alphabets in each class
Consonant	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	อั, ะ, อ็, า, อิ, อี, อึ, อื, อุ, อู, เ, แ, โ, อำ, ไ ,ใ, ฤ, อ์
Tone	อ่, อ้, อ๊, อ๋

Table 1. Thai Alphabets: Consonants, Vowels and Tones.

2.2 Thai Syllable Characteristics and Phonetic Representation

Initial Consonant (C_i)	Vowel (V)	Final Consonant (C_f)	Tone (T)
p,pr,phr,pl,phl,t,tr,thr,c,kr,khr,k,z,ph,th,ch,k,kl,khl,kw,khw,h,b,br,bl,d,dr,m,nr,ng,r,f,fr,fl,s,h,w,j	*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^,ng^,j^,w^,f^,l^,s^,ch^,jf^,ts^*	0,1,2,3,4

Table 2. Phonetic Symbols Grouped as Initial Consonants, Vowels, Final Consonants and Tones.

In Thai language, a syllable can be separated into three parts; (1) initial consonant, (2) vowel and (3) final consonant. The phonetic representation of one syllable can be expressed in the form of $/C_i\text{-}V^T\text{-}C_f/$, 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. Following the concept in (Pisarn and Theeramunkong, 2003) there are 76 phonetic symbols and 5 tone symbols applied in this work as shown in Table 2.

3 Our Framework

The recognition framework illustrated in Figure 1 presents our overall framework designed for Thai continuous speech recognition system that incorporates a conventional recognizer with a spelling recognition subsystem. The whole process can be divided into two modules; (1) training module and (2) recognition module.

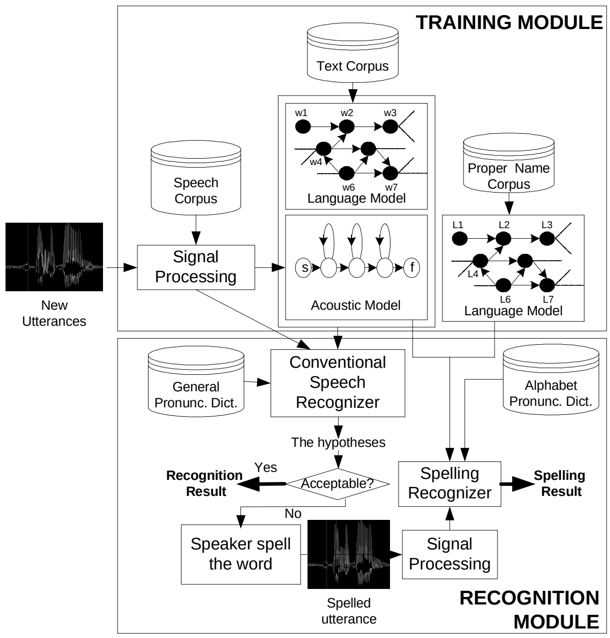

Figure 1. The Recognition Framework

In the training module, waveforms of continuous speech utterances in a corpus are transformed to feature vectors by using a signal quantization technique. The derived feature vectors are used for training a set of acoustic models. In the system, two language models are equiped; one stands for traditional word recognition, whereas the other one is used for spelling recognition. The traditional language model is trained by transcriptions in the text corpus, while the spelling language model is trained by sequences of letters in a proper name corpus.

In the recognition module, the two well-trained models; the acoustic model and the traditional language model, together with a pronunciation dictionary are applied to recognize a new utterance, yeilding a set of hypothesis results. The hypotheses are then verified whether it is valid or not. If it is not, the system will turn to the spelling recognition subsystem.

At this stage, the user is asked to spell the word letter-by-letter. The utterance of spelling is then fed to the signal-processing module for converting the waveform to feature vectors. In this work, as our preliminary stage, we used the acoustic models trained by normal continuous speech utterances because of a lacking spelling corpus. Working with well-trained spelling language model and alphabetic pronunciation dictionary, the spelling results could be obtained.

4 Spelling Style for Thai Word

4.1 Basic Pronunciation of Thai Alphabets

As refered in section 2.1, there are three basic classes of Thai alphabets. Pronouncing Thai alphabets in different classes has different styles. The consonant class alphabets can be uttered in either of the following two styles. The first style is simply pronouncing the core sound of a consonant. For example, the alphabet ‘ก’, its core sound can be represented as the syllable phonetic /*k-@@0*/. Normally, some consonants share a same core sound, for instance ‘ค’, ‘ฅ’, ‘ฆ’ have the same phonetic /*kh-@@0*/. In such case, the hearer may encounter an alphabet 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 its representative word. For example, the representative word of ‘ก’ is “ไก่” (chicken), with the phonetic sound /*k-a1-j^*/, and that of ‘ข’ is “ไข่” (egg, /*kh-a1-j^*/). To express the alphabet ‘ก’ using this style, the sound /*k-@@0*/+/*k-a1-j^*/ is uttered.

Expressing alphabets in the vowel class is quite different to that of the consonant class. There are two types of vowels. The first-type of vowels can be pronounced in two ways. One is to pronounce the word “สระ” (meaning: “vowel”, sound: /*s-a1*//*r-a1*/), followed by the core sound of the vowel. The other is to pronounce by simply pronouncing the core sound of the vowel. The second-type of vowels are uttered by calling their names. The vowel alphabets of each type are listed in Table 3. As the last class, tone symbols can be pronounced by calling their names.

Type	Vowels
The first-type	ะ, า, อิ, อี, อึ, อื, อุ, อู, เ, แ, โ, อำ, ไ,ใ
The second-type	อั, อึ็, อ็, ฤ

Table 3. Two Types of Vowels

4.2 Thai Word Spelling Methods

Spelling out a word is the way to utter each alphabet in the word in order. It refers to combinations of the pronunciation style of each alphabet in the word. Only the four Thai most commonly used spelling methods have been addressed. For all methods, the second-type vowels and tones are pronounced by calling their names. The differences are taken place in spelling consonants and the first-type vowels. For the first spelling method, the consonants are spelled out by using only their core sounds, and the first-type vowels are pronounced by their core sound without the word “สระ” (/*s-a1*//*r-a1*/). This spelling method is similar to the spelling approach in English language.

In the second method, the representative word of each consonant is pronounced followed by its core sound while pronounce a first-type vowel is to utter the word “สระ” and then its core sound. In the third method, the way to pronounce a consonant and vowel are varied. For instance, the word can be spelled a consonant by using only core sound together with a vowel beginning with the word “สระ”. The last method is to spell out a set of letters that form a syllable and then followed by its corresponding utterances. The spelling sequence of alphabets in each syllable starts with initial consonant, vowel, and followed by final consonant (if any) and tone (if any), and then, the sound of that syllable is inserted at the end of the sequence.

The examples of these methods in spelling the word “สิงห์” are depicted in Figure 2.

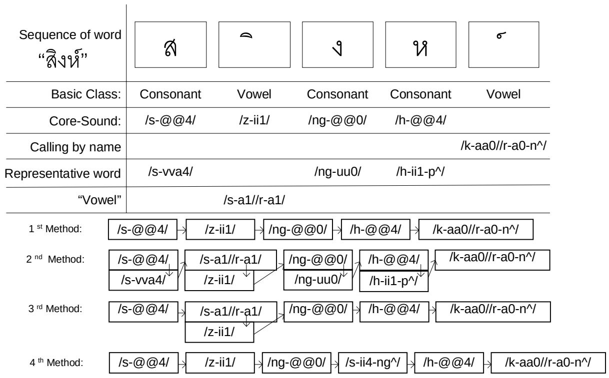

Figure 2. Four spelling methods for the word “สิงห์”

In this paper, we concentrate on the second method as the first step, since this is the most popular spelling method in Thai language.

5 Experimental Results and Analysis

5.1 Experimental Environment

As mentioned above, the corpus for a spelling recognition task is unfortunately not available at this time. Therefore, this work applies the NECTEC-ATR Thai Speech Corpus, constructed by NECTEC (National Electronics and Computer Technology Center) incorporated with ATR Spoken Language Translation Laboratories. In Thai language speech recognition, this corpus is normally used for a continuous speech recognition task. This speech corpus is used as the training set for our spelling recognition system. The corpus contains 390 sentences gathered by assigning 42 speakers (21 males and 21 females) to read all sentences for a trail. So, there are totally 16,380 read utterances.

In the first place, by the reason of computation time, only utterances of 5 males and 5 females, are used, i.e., totally 3,900 trained utterances. In addition, as our preliminary work, the effects of spelling result with a normal continuous training corpus are investigated. Even though, the training corpus has quite different characteristics compared to the test utterances, we can expect a reasonable result. The test utterance is constructed by recording the spelling of 136 proper names by a female participant.

The speech signals were digitized by 16-bit A/D converter of 16 kHz. A feature vector used in our experiment is a 39-feature vector, consists of 12 PLP coefficients and the 0^{th} coefficient, as well as their first and second order derivatives. Therefore, there are totally 39 elements.

The language model used in this task is a bigram language model, trained from totally 6,107 proper names, i.e., 5,971 Thai province, district and subdistrict names, as well as 136 proper names from the test transcription.

A phone-based HMM is applied as the recognition system. The acoustic units used in this experiment are defined in the same manner as in (Pisarn and Theeramunkong, 2003). All experiments, including automatic transcription labelling, are performed using HTK toolkit (Young et al., 2002). The word correctness is given by the percentage of numbers of correct words divided by total number of words and the accuracy is computed by the percentage of subtracted the numbers of correct words by the number of insertion errors, which are then divided by total number of words.

5.2 Setting a Baseline

In the first experiment, we investigate the spelling results using the original training and testing data as they are. This will be a baseline for all of our experiment. In this initial stage, the context-independent method (CI), achieves 79.94 and 57.99 for correctness and accuracy, respectively. The system with context-dependent method (CD) gains 70.80 and 46.09 for correctness and accuracy respectively. In principle, low accuracy is triggered by a large number of insertion errors. Because of this figure, two possible assumptions can be made (1) there is in compatible duration between the training and the test set, and (2) Our HMM models are inappropriate.

5.3 Adjusting the Duration

To investigate the results of the first assumption, the utterance speed of the utterances from the training and testing are measured in the form of the number of phone per second. The speed can be computed by dividing the number of total phones in each utterance transcription by its utterances duration in seconds. As a result, the average utterance speed of the training set is 11.7 phones/sec while the average utterance speed of the test set is only 4.6 phones/sec. This indicates that the speed of test utterances are approximately 2.5 times slower than that of train utteraces. This difference may cause low accuracy.

To compensate for this duration difference among the training utterance and the testing utterance, a method to shrink and stretch a speech signal, by preserving pitch and auditory features of the original signal, is applied in our signal preprocessing. The experiments are done in two environments; stretching the training utterances and shrinking the test utterences. By adjusting the duration of the training and testing utterances, insertion errors could be reduced. Stretching the training utterances and shrinking the test utterances are performed using various scale factors in order to investigate the effectiveness. Table 4 shows the recognition results of stretched training utterances with various scale factors. Here, the original test utterances are used.

Duration	Model	%Correct	Accuracy
1.25Train	CI	81.91	62.49
	CD	82.05	66.36
1.43Train	CI	85.43	68.54
	CD	85.86	70.09
1.67Train	CI	86.42	63.34
	CD	84.59	63.97

Table 4. Recognition Results of Stretched Training Utterances with Various Scale Factors.

In principle, stretching training utterances causes the original utterances to be distorted. The more scale the utterances are stretched, the more distored the utterances we obtain. As stated in the previous section, utterances training are approximately 2.5 times faster than the test utterances. However, they are expected to achieve a very low accuracy. The experimental results show that by adjusting training utterances 1.43 times slower than the original one (1.43Train) can improve the correctness to 85.86 % and the accuracy to 70.09% in a context-dependent method. But with more stretching, the accuracy drops to 63.97%.

Reversely we also examine the system accuracy when the test utterances are shrinked on various scale factors. The original training utterances are used for training our system. The recognition results are shown in Table 5.

Duration	Model	%Correct	Accuracy
0.71Test	CI	86.28	74.88
	CD	82.41	73.12
0.43Test	CI	82.97	77.34
	CD	80.93	75.93

Table 5. Recognition Results of Shrinked Test Utterances with Various Scale Factors.

Shrinking test utterances can improve accuracy. Especially, the test utterances with 0.43 scaling factor can reduce the accuracy error to 19.35%, from 57.99% to 77.34%.

No.of states	Model	%Correct	Accuracy
3	CI	82.97	77.34
	CD	80.93	75.93
4	CI	80.01	76.78
	CD	79.73	76.85
5	CI	80.79	79.38
	CD	79.31	78.25
6	CI	78.04	76.99
	CD	76.92	76.21

Table 6. Recognition Accuracy with Various Numbers of States for a Long Vowel Phoneme.

5.4 Acoustic Models with Different Numbers of States

In fact, phone durations of each phoneme in Thai language do not have the same duration. Especially in the vowel class, there are vowels pairs, where one has a longer phone while the other has a shorter phone. For example, the vowel pair, *a* and *aa*, have a similar phone but different durations. The phoneme *a* has a shorter duration than the phoneme *aa*. The other vowel pairs are *i-ii*, *v-vv*, *u-uu*, *e-ee*, *x-xx*, *o-oo*, *@-@@*, *q-qq*, *ia-iia*, *va-vva*, and *ua-uua*. The shorter phone should not have the same number of state as the longer one. Therefore, we examined the recognition rate on different numbers of HMM states. The experiment is examined using the 0.43Test set since it is the best one in the previous experiment. The results are shown in Table 6. In this experiment, the number of states for a long vowel phoneme is varied from 3 to 6 states. However, the numbers of states for the other phonemes are set to 3 states.

Table 6 shows that a 5-state HMM for a long vewel phoneme and a 3-state HMM for the other phonemes achieve the highest recognition accuracy, i.e., 79.38. This is, 2.04% error rate reduction compared with the 3-state HMM.

6 Conclusion

In this paper, we present a general framework for Thai speech recognition enhanced with spelling recognition. Four styles for spelling Thai words were discussed. To recognize spelling utterances, HMMs were constructed using a continuous speech corpus. To achieve higher correctness and accuracy, we compensated the utterance speed among the training and test utterances by stretching the training utterances or shrinking the test utterances. The experimental results indicated promising performance of 79.38% recognition accuracy after this adjustment. With a good scaling factor, the system achieved 19.35% improvement compared with the baseline where the training and test utterances were used as they are. Assigning a larger number of states to a longer syllable (i.e., long vowel) could improve recognition accuracy by 2.04 %. Our further works include (1) to construct a system that deals with several kinds of spelling methods, and (2) to explore the incorporation of spelling recognition into the conventional speech recognition system.

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. We also would like to thank Dr. Virach Sornlertlamvanich for his useful suggestions through this work. This work has partly been supported by NECTEC under project number NT-B-22-I5-38-47-04.

References

A. Anastasakos, R. Schwartz and H. Shu. 1995. “*Duration Modeling in Large Vocabulary Speech Recognition*”. In, “Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing”, pages 628-631.

Carl D. Mitchell and Anand R. Setlur. 1999. *Improved Spelling Recognition using a Tree-based Fast Lexical Match*. In "Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing", volume2, pages 597-600.

Chutima Pisarn and Thanaruk Theeramunkong. 2003. *Incorporating Tone Information to Improve Thai Continuous Speech Recognition*. In "Proceedings of International Conference on Intelligent Technologies 2003".

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Frederico Rodrigues, Richardo Rodrigues and Ciro Martins. 1997. *An Isolated Letter Recognizer for Proper Name Identification Over the Telephone*. In "Proceedings of *9th Portuguese Conference on Pattern Recognition* (RECPAD'97)", Coimbra.

Josef G. Bauer and Jochen Junkawitsch. 1999. *Accurate recognition of city names with spelling as a fall back strategy*. In "Proceedings of EUROSPEECH 1999", pages 263-266.

Martin Betz and Hermann Hild. 1995. *Language Models for a Spelled Letter Recognizer*. In "Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing". Pages 856-859.

Nattakorn Thubthong and Boonserm Kijsirikul. 2001. Tone Recognition of Continuous Thai Speech under Tonal Assimilation and Declination Effects using Half-Tone Model. *Journal of International of Uncertainty, Fuzziness and Knowledge-Based System*, 9(6):815-825.

Ruben San-Segundo, J. Macias-Guarasa, J. Ferreiros, P. Martin and J.M. Pardo. 2001. *Detection of Recognition Errors and Out of the Spelling Dictionary Names in a Spelled Name Recognizer for Spanish*. In "Proceedings of EUROSPEECH 2001", Aalborg (Dinamarca).

Ruben San-Segundo, Jose Colas, Ricardo de Cordoba and Jose M. Pardo 2002. Spanish Recognizer of Continuously Spelled Names Over the Telephone. *Journal of Speech Communication*, volume 38, pp.287-303.

Steve Young, Gunnar Evermann, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. 2002. *The HTK Book (for HTK Version 3.2.1)*. Cambride University Engineering Department.

W.Verhelst and M.Roelands. 1993. *An overlap-add technique based on waveform similarit (wsola) for high quality time-scale modification of speech*. In "Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing", volume 2, pages 554-557, Minneapolis, Minnesota.