Using Tone Information in Thai Spelling Speech Recognition

Natthawut Kertkeidkachorn
[1]Department of Computer Engineering, Chulalongkorn University Bangkok, Thailand
[2]Department of Informatics The Graduate University for Advanced Studies, Tokyo, Japan
Natthawut@nii.ac.jp

Proadpran Punyabukkana
Department of Computer Engineering, Chulalongkorn University Bangkok, Thailand
Proadpran.p@chula.ac.th

Atiwong Suchato
Department of Computer Engineering, Chulalongkorn University Bangkok, Thailand
Atiwong.s@chula.ac.th

Abstract

Spelling recognition is a workaround to recognize unfamiliar words, such as proper names or unregistered words in a dictionary, which typically cause ambiguous pronunciations. In the Thai spelling task, some alphabets cannot be differentiated by only spectral cues. In such cases, tonal cues play a critical role in distinguishing those alphabets. In this paper, we therefore introduce Thai spelling speech recognition, in which a tonal score, which represents a tonal cue, is adopted in order to re-rank N-best hypotheses of the first pass search of a speech recognition system. The Hidden Conditional Random Field (HCRF)-based Thai tone recognition, which was reported as the best approach for Thai tone recognition, is selected to provide tonal scores. Experimental results indicate that our approach provides the best error rate reduction of 23.85% from the baseline system, which is a conventional Hidden Markov Model (HMM)-based speech recognition system. Besides, another finding is that exploiting tonal scores in Thai spelling speech recognition could significantly reduce the ambiguity among some alphabets.

1 Introduction

A spelling speech recognition system plays an important role in many kinds of applications, of which a domain contains unfamiliar words such as proper names. Since those words might not be pronounced straight-forwardly, an automatic speech recognition (ASR) system would have difficulty to recognize such words correctly. A practical efficient solution for handling such words in an ASR system is to pronounce them letter by letter.

Nevertheless, in tonal languages, especially Thai, a spelling recognition task is a challenging task because merely consonantal sound and vowel sound cannot perfectly distinguish Thai alphabets. For example the “ข” alphabet and the “ค” alphabet are pronounced as \k^{h}Ɔ:\. Although the consonantal sound and the vowel sound of those alphabets are similar, their tones are significantly different. For the “ข” alphabet, its tonal sound is the rising tone, while the tonal sound of the “ค” alphabet is the mid tone. In Thai, tone information therefore not only expresses prosody as usual but also transmits explicit information, which characterizes lexical meanings of words (Luksaneeyanawin 1998).

In this paper, we therefore introduce a Thai spelling speech recognition employing tonal scores, which can represent tonal information, in order to re-rank N-best hypotheses according to the first pass search of an ASR system. The Hidden Conditional Random Field (HCRF)-based Thai tone recognition, which had been reported as the state of the art for Thai tone recognition

28th Pacific Asia Conference on Language, Information and Computation pages 178–184

(Kertkeidkachorn et al. 2014), is selected to provide tonal scores.

The rest of the paper is organized as follows. In Section 2, background knowledge on Thai spelling system is introduced and related works are reviewed and discussed in the following section. Section 4 presents our Thai spelling recognition approach. Then, the HCRF-based approach for Thai tone recognition is described in the next section. Experiments and results are presented in Section 6 and experimental results are discussed in Section 7. Eventually, we conclude our work in the last section.

2 Thai Spelling

In the Thai spelling task, a sequence of Thai alphabets, which can be consonantal alphabets, vowel alphabets, tone symbols, or punctuation symbols, is pronounced. The pronunciation of consonantal alphabet has two possible variations: a consonantal alphabet and a consonantal alphabet with its extension. The alphabet extension is a word or a phrase which follows that alphabet in order to distinguish that alphabet from others. For example, the “ข” (kh-@@-z^-4) alphabet is followed by the extension word “ไข่” (kh-a-j^-1) as “ข. ไข่” (kh-@@-z^-4 kh-a-j^-1), while the extension word of the “ฃ” (kh-@@-z^-4) alphabet is “ขวด” (kh-uua-t^-1) pronounced as “ฃ. ขวด” (kh-@@-z^-4 kh-uua-t^-1). This characteristic is similar to uttering “A alpha” or “B beta” in English (NATO phonetic alphabet 2014) but occurs much more frequently. For Thai vowel alphabets and tone symbols, there are also two possible pronunciation patterns which come from the presence or the absence of indicative words, “สระ” and “ไม้”, before vowel alphabets and tone symbols respectively. Punctuation marks are uttered by their actual names. In Table 1, Thai Alphabet patterns and their examples are presented.

Type	Pattern	Example
Consonantal	Base name	ก
	Base name + Extension	ก ไก่
Vowel	Base name	อา
	(s-a-z^-1 r-a-z^-1) + Base name	สระอา
Tone	Base name	เอก
	(m-a-j^4) + Base name	ไม้เอก
Punctuation	Base name	จุลภาค

Table 1: Thai Alphabet Patterns and their examples

3 Related Work

In tonal languages, tone information has been investigated and exploited in many research works in order to improve performances of ASR systems. In Chinese, Lee et al. (2002) expanded syllable lattices via recognized tone patterns to improve the performance of Cantonese large-vocabulary continuous speech recognition (LVCSR). Their results indicated that reliable tone information could improve the overall performance of Cantonese LVCRS. Later, Lei et al. (2006) then utilized tone models for improving Mandarin broadcast news speech recognition. With exploiting tone information, their experiment significantly indicated the improvement of the ASR system. Wei et al. (2008) also explored Conditional Random Field (CRF)-based tone modeling to re-rank hypotheses generated from the first pass search of an ASR system. Their results showed that tone information could really help to improve the performance of the ASR system. In Vietnamese, which is also one of tonal languages, Quang et al. (2008) succeeded in improving the performance of Vietnamese LVCSR by utilizing tone information. In Thai, Chaiwongsai et al. (2008) proposed HMM-based isolated-word speech recognition with a tone detection function. With the tone detection function, tone results were considered together with word results in order to compute the final result. Their experiment reported that the performances of Thai isolated-word speech recognition were improved. Pisarn and Theeramunkong (2006) investigated tone features and these features were incorporated into their HMM-based Thai system in order to improve Thai spelling recognition.

Based on discussed works, in tonal languages exploiting tone information to an ASR system had directly contributed to its performances. We therefore aim to exploit reliable tone information in order to improve the performance of Thai spelling recognition.

4 A Thai Spelling Recognition Approach

In our approach, Thai spelling speech recognition incorporating a tone recognizer providing tone information, which can help to recognize alphabets more accurately, is proposed as shown in Figure 1.

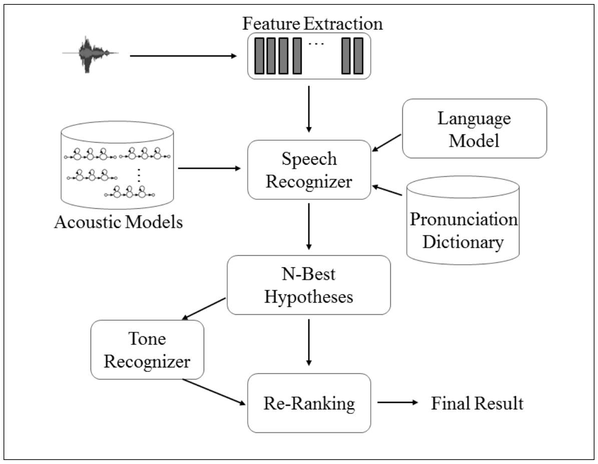

Figure 1: A Thai Spelling Recognition Approach

Acoustical feature vectors is extracted from an speech signal as an acoustic observation sequence (O) capturing spectral shapes of the signal via the feature extraction process and then these acoustical feature vectors are conveyed to the speech recognizer in order to recognize the word. With acoustical feature vectors, trained acoustic models, a language model and a pronunciation dictionary, the speech recognizer is to generate N-best hypotheses for the input speech signal. Generally, when N-best hypotheses are generated, the best hypothesis will be selected as the result. Nevertheless, in our approach the best hypothesis is not immediately decided yet. N-best hypotheses are fed to a tone recognizer in order to compute tonal scores. After that those N-best hypotheses are re-ranked according to their acoustic score (log($P(O/W)$)), which is the probability of the acoustic observation sequence O given the hypothesis W, their language score (log($P(W)$)), which is the probability of the hypothesis W, and their tonal score (log($P(T/W)$)), which is the probability of the tone sequence T given the hypothesis W. The best hypothesis (W') is computed as follows:

$$W' = \arg\max_{W} (\log(P(O \mid W)) + \log(P(W)) + w \log(P(T \mid W))) \tag{1}$$

, where W is a hypothesis from N-Best hypotheses and w is a weight for the tonal score.

In our approach, we do not directly embedded tone features into the speech recognizer as reported in Pisarn and Theeramunkong's study (2006) due to the feature extraction problem. Typically fundamental frequency (F_0) movements are selected as the representation of tone information. Nonetheless, in unvoiced parts or silent parts, F_0 movements would be absent. Consequently, tone information might not be steady. Our workaround is to compute tonal scores only on voiced parts of words, which are provided by the speech recognizer, instead.

5 HCRF-Based Tone Recognition

Since our Thai spelling speech recognition approach depends on performances of Thai tone recognition, the HCRF-based Thai tone recognition (Kertkeidkachorn et al. 2014), which had been reported as the best approach for Thai tone recognition, is selected to calculate tonal scores. Given the hypothesis W, which is a sequence of syllables ($W = s_1s_2s_3...s_n$; s_i = the i^{th} syllable of the hypothesis W), the probability of the tone sequence (T) corresponding to the hypothesis W given the hypothesis W ($P(T|W)$) is computed through the following equation:

$$\log(P(T \mid W)) = \sum_{i=1}^{n} \log(P(t_i \mid s_i)) \tag{2}$$

, where t_i is the tone of the i^{th} syllable of the hypothesis W ($T = t_1t_2t_3...t_n$) and t_i is directly associated with s_i. Although $P(T|W)$ is a kind of measurement for the tone sequence T given the hypothesis W, its value is very small. We therefore take logarithm functions on its value and referred it as the tonal score.

Even though the HCRF-based Thai tone recognition reported by Kertkeidkachorn et al. (2014) outperformed other approaches, still, their work limited their acoustical features to F_0's values and their derivative. In the Thai tone perception study, Kertkeidkachorn et al. (2012(a)) found that spectral information could contribute to the tone perception of Thai native speakers. We therefore assumed that spectral information might contribute to the HCRF-based tone recognition as well. A preliminary experiment was conducted to prove our assumption. This preliminary experiment was conducted under the Thai tone continuous speech recognition scenario and all configurations in the preliminary experiment are also similar to Kertkeidkachorn's work (2014). Nonetheless, two further acoustical features, which were widely used in many ASR systems, were investigated by appending each of them into Kertkeidkachorn's tone feature in order to measure the improvement of the HCRF-based tone recognition. Mel-frequency Cepstral Coefficients (MFCC) and Perceptual Linear Predictive coefficients (PLP) were chosen to represent the spectral information of speech signals in the preliminary experiment. Results of the preliminary experiment are shown in Table 2.

Approach	Accuracy (%)
Kertkeidkachorn's work	71.01
Appending MFCC	74.91
Appending PLP	**75.04**

Table 2: % Accuracy results of the tone recognition in the preliminary experiment

According to the results of the preliminary experiment on the HCRF-based tone recognition, appending the PLP-based feature yields the best accuracy result. Besides, appending the PLP-based feature into tone features can provide an error rate reduction of 13.90% from what reported in Kertkeidkachorn's work (2014). We also notice that appending the MFCC-based feature gives better results than what reported in Kertkeidkachorn's work (2014) as well. The findings conform to our assumption in which spectral information could contribute to the performance of the HCRF-based Thai tone recognition as well. We therefore append the PLP-based feature into the tone feature of the HCRF-based Thai tone recognition.

6 Experiments and Results

6.1 Experimental Setting

In the experiment, the CU-MFEC corpus for Thai and English spelling speech recognition (Kertkeidkachorn et al. 2012(b)) is selected to evaluate our approach. The experiment is conducted on randomly selected speech data of 50 speakers from the alphabet with short pause set of the corpus. And, only Thai alphabets are considered in the experiment. Speech data of 40 speakers is randomized as the training data and the rest of the speech data is used as the testing data.

The speech recognizer in our approach is a traditional HMM-based speech recognizer of which models represented 135 Thai alphabets. Our models do not represent normal phoneme units because when tonal units are included, there are 375 model units which are more than 135 models of Thai alphabets. To represent speech frames, the standard 39-dimensional MFCC feature vectors are extracted at every 10 ms and each of the speech frames is windowed with 25 ms-Hamming window. Because a left to right HMM model was used to represent a context dependent Thai alphabet, of which duration is typically longer than usual phoneme duration, we also conduct another preliminary experiment to adjust a number of states of a HMM model and also fine-tune a number of appropriate Gaussian mixtures for our recognizer. Results are presented in Table 3.

No. of states	No. of Gaussian Mixtures					
	1	2	4	8	16	32
3	39.56	45.26	63.26	67.93	70.30	68.59
4	52.15	60.00	74.96	75.63	79.33	76.22
5	58.96	67.56	80.52	81.26	81.41	81.63
6	60.59	76.00	82.22	81.41	83.04	81.33
7	62.30	76.67	81.70	82.00	**84.15**	81.78
8	65.26	77.26	81.26	80.74	82.96	80.30

Table 3: % Accuracy results of the baseline varied by a number of states and a number of Gaussian mixtures

Based on results, a seven-stated HMM and 16-coponent Gaussian Mixtures with diagonal covariance matrices yields the best accuracy result at 84.15%. Therefore, this setting is set as the setting of the speech recognizer in our approach and also is referred as *baseline*.

After the first pass search of the speech recognizer, N-best hypotheses are generated. In the experiment, N is set at 135 equal to the number of Thai alphabets, so that possible hypotheses could be generated. To build the HCRF-based tone recognizer of which models represented five Thai tones, the HCRF Library (Morency et al. 2012) is used with the following setting. To represent speech frames, F_0 values, their delta and their acceleration together with the standard 39-dimensional PLP-based feature are combined as a tonal feature vector. Tonal feature vectors are extracted every 10 ms with 25-ms Hamming window. In the HCRF library, GHRF is set as the type of the model. A number of hidden states are set at 3 states due to the characteristic of Thai tones, which basically consist of three parts (Kertkeidkachorn et al. 2014), and initial weights of vectors are computed from mean and variances of each acoustic feature. The optimization method is configured as Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) with L2 cache. Testing on the testing data, our tone recognizer provides accuracy of 87.79%. In the experiment, the parameter *w* for weighting a tonal score in Equation 1 is adjusted in order to find the best setting and study effects of tonal weights on Thai spelling speech recognition.

6.2 Experimental Results

Results of adjusting the tonal weight *w* are shown in Figure 1. Our approach obtains the best accuracy of 87.93% and also provides 23.85% relative error rate reduction from the baseline, when *w* is at 52.

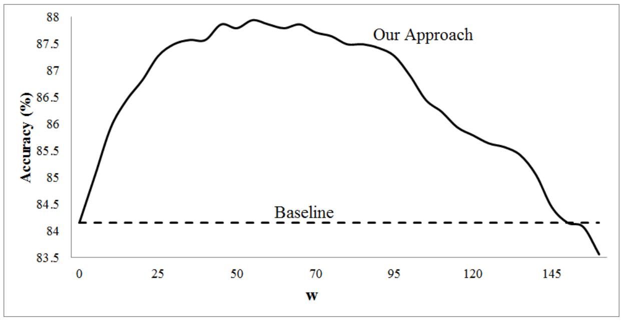

Figure 2: % Accuracy results when *w* parameter is adjusted

7 Discussion

In our approach, according to experimental results, adjusting the tonal weight *w* clearly affects the recognition accuracies of the Thai spelling task. At first, when *w* is increased, the recognition accuracy also tends to be increase. Nevertheless, when *w* becomes more than 52, the recognition performance is declined because acoustic scores and language scores are initially governed by tonal scores. Furthermore, after *w* is more than 150, tonal scores completely dominated results. The recognition accuracies of our approach become worse than the baseline. We therefore can conclude that tonal scores acquired from tonal cues could help improve Thai spelling speech recognition in case that the tonal weight *w* is set appropriately; however either acoustic scores or language scores are still far more important.

The significant testing is also conducted on experimental results to compare the recognition accuracy of Thai alphabets of the baseline with the best result of our approach, in which *w* is set at 52. The Mcnemar's test (Gillick and Cox 1989) is used to evaluate the statistical significance of accuracy results. The test result indicates that our approach, in which tonal cues had contributed to recognition results, statistically outperforms the baseline with p-value less than 0.01.

Paired alphabets are groups of alphabets, in which consonantal sound and vowel sound are similar but the tonal sound is different. For example, a group of the alphabets ช and ฌ pronounces as ch-@@-z^-0, while a group of the alphabet ฉ utters as ch-@@-z^-4. Without tone information, paired alphabets are difficult to differentiate their group from the other group. The error confusions between paired alphabets in the baseline and our approach are shown in Table 3. Error confusion is measured from the error that paired alphabets in the target group are misrecognized as the other group. Considering error confusion on Table 3, we found that our approach, in which tonal scores are adopted, could reduce the error confusion in any paired alphabet groups.

Paired Alphabets	Error Confusion (%)	
	Baseline	Our Approach
(ช, ฌ) - (ฉ)	3.3	0.0
(ซ) - (ษ, ศ, ส)	5.0	0.0
(พ, ภ) - (ผ)	10.0	0.0
(ฑ, ฒ, ท, ธ) - (ฐ, ถ)	11.7	0.0
(ค, ต, ฆ) - (ข, ฃ)	24.0	2.0
(ฮ) - (ห)	25.0	5.0
(ฟ) - (ฝ)	25.0	15.0

Table 2: Error confusion comparing between paired alphabets in the baseline and our approach

Based on our discussion, we could conclude that the tone information is necessary for improving Thai spelling speech recognition, especially in case of confusions between paired alphabets.

8 Conclusion

Recently, in tonal languages, there are many researches utilizing tone information in many kinds of ASR systems, especially where the language modeling could partly help to recognize words, such as a spelling recognition task.

This paper introduces a Thai spelling speech recognition approach, in which tonal scores acquired from the HCRF-based Thai tone recognizer, which had been reported as the state of the art for Thai tone recognition, are employed. Furthermore, this paper also explores the performance of the HCRF-based Thai tone recognizer by applying the PLP-based feature representing spectral information to improve its performance so that more reliable tone information could be provided for our approach. Experimental results evidently show that tonal scores significantly contribute to the performance of Thai spelling speech recognition, when the weight of the tonal score is adjusted properly.

Still, further factors could definitely contribute to the recognition accuracies of the Thai spelling task beyond what reported in this paper.

References

Sudaporn Luksaneeyanawin, 1998. Intonation in Thai. In Intonation Systems a Survey of Twenty Languages, Cambridge University Press, 1998, ch. 21, pp. 376–394.

Natthawut Kertkeidkachorn, Proadpran Punyabukkana and Atiwong Suchato. 2014. A Hidden Conditional Random Field-Based Approach for Thai Tone Classification, In Engineering Journal, 18(3):99-122.

NATO phonetic alphabet. 2014. In Wikipedia, The Free Encyclopedia. Retrieved 08:43, August 3, 2014, from http://en.wikipedia.org/w/index.php?title=NATO_phonetic_alphabet&oldid=618764363

Tan Lee, Wai Lau, Y. W. Wong and P. C. Ching, 2002. Using tone information in Cantonese continuous speech recognition, ACM Transactions on Asian Language Information Processing, 1(1):83-102

Xin Lei, Manhung Siu, Mei-Yuh Hwang, Mari Ostendorf and Tan Lee. 2006. Improved tone modeling for Mandarin broadcast news speech recognition, In Proc. Interspeech 2006.

Hongxiu Wei, Xinhao Wang, Hao Wu, Dingsheng Luo and Xihong Wu. 2008. Exploiting Prosodic and lexical Feature for Tone Modeling in a Conditional Random Field Framework, In Proceedings of ICASSP 2008.

Nguyen Hong Quang, Nocera Pascal, Castelli Eric and Trinh Van Loan. 2008. Using tone information for Vietnamese continuous speech recognition, In Proceedings of RIVF 2008.

Jirabhorn Chaiwongsai, Werapon Chiracharit, Kosin Chamnongthai and Yoshikazu Miyanaga. 2008. An Architecture of HMM-Based Isolated-Word Speech Recognition with Tone Detection Function, In proceedings of ISPACS 2008.

Chutima Pisarn and Thanaruk Theeramunkong 2006. Improving Thai spelling recognition with tone features, Lecture Notes in Artificial Intelligence 4139: 388-398.

Natthawut Kertkeidkachorn, Surapol Vorapatratorn, Sirinart Tangruamsub,Proadpran Punyabukkana and Atiwong Suchato 2012(a). Contribution of spectral shapes to tone perception, In Proceedings of Interspeech 2012.

Natthawut Kertkeidkachorn, Supadaech Chanjaradwichai, Teera Suri, Krerksak Likitsupin, Surapol Vorapatratorn, Pawanrat Hirankan, Worasa Limpanadusadee, Supakit Chuetanapinyo, Kitanan Pitakpawatkul, Natnarong Puangsri, Nathacha Tangsirirat, Konlawachara Trakulsuk, Proadpran Punyabukkana and Atiwong Suchato. 2012(b). The CU-MFEC corpus for Thai and English spelling speech recognition, In Proceedings of Oriental-COCOSDA 2012.

Louis-Philippe Morency, Ariadna Quattoni, C. Mario Christoudias and Sybor Wang 2012. Hidden-state

Conditional Random Field (HCRF) Library, Online at http://sourceforge.net/projects/hcrf/.

L. Gillick and S.J. Cox, 1989. Some statistical issues in the comparison of speech recognition algorithms, In Proceedings of ICASSP 1989:532-535.