Improving Thai Spelling Recognition with Tone Features

Chutima Pisarn and Thanaruk Theeramunkong

Sirindhorn International Institute of Technology
131 Moo 5 Tiwanont Rd., Bangkadi, Muang, Pathumthani 12000, Thailand
{chutimap, thanaruk}@siit.tu.ac.th

Abstract. Spelling recognition has been used for several purposes, such as enhancing speech recognition systems and implementing name retrieval systems. Tone information is an important clue, in addition to phones, for recognizing speeches in tonal languages. In this paper, we present a method to improve accuracy of spelling recognition in Thai, a tonal language, by incorporating tone-related acoustic features to a well-known front-end feature named Perceptual Linear Prediction Coefficients (PLP). The proposed method makes use of three kinds of tone information: fundamental frequency (pitch), pitch delta and pitch acceleration, to enhance the original features. Compared to the baseline result gained from the original feature, our HMMs-based recognition model shows improvement of 1.73%, 2.85% and 3.16% of letter accuracy for close-type, mix-type and open-type language models, respectively.

1 Introduction

Spelling recognition is one of specific speech recognition tasks the scope of which is the domain of recognizing spelled utterances. Not only applied to assist a telephone directory system, a spelling recognition system can also be a practical way to enhance a speech recognition system. Normally, when the system cannot predict a word due to an out-of-dictionary word or signal distortion, it should request the user to spell out the word in order to recognize it. Several works on spelling recognition have widely been developed for many languages such as English [1], German [2, 3] Spanish [4] and Portuguese [5]. Most of them concentrated on how to retrieve a correct name from a telephone directory using various techniques. Unlike spelling in other languages, Thai spelling has special characteristic in the sense that one can spell Thai in several ways. In the past, there were few works on Thai spelling recognition [6].

Like Chinese, Vietnamese and some oriental language, Thai is a tonal language. For those tonal languages, tone information can be considered as a potential source in improving recognition rate. As an early stage of tonal-language speech recognition, several methods have been proposed to use tonal characteristics in speech recognition in both isolated and continuous speech of Mandarin and Cantonese [7-9]. Some works [8-11] indicated that these tonal characteristics were very helpful

T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 388 – 398, 2006.

in increasing speech recognition. In [8], Chen proposed a method to incorporate pitch information of only the main vowels of some syllables into feature vectors. The result showed dramatic reduction in word recognition error rate when demi-syllable pitch information was applied to the conventional method. Instead of using syllables as processing units, Chang [9] decomposed a syllable into two parts; syllable initial and syllable final as basic acoustic processing units. The basic acoustic units with a same phoneme but different tones are treated as different phonemes. The pitch information and its first-order and second-order derivatives are smoothed and then applied to feature vectors. However, this work limited the experiments to the evaluation based on individual gender. As another work, Wong [11] reported that the integration of tone-related information, such as frame energy, probability of voicing and pitch period, in addition to its derivatives to the feature vector could reduce Chinese speech syllable error rate. For Thai [12], by incorporating tone features, i.e. fundamental frequency (pitch), pitch delta and pitch acceleration, the accuracy of Thai continuous speech recognition has been improved. Unfortunately, so far there has been no work related to exploiting tone information in spelling recognition.

In this work, to improve accuracy of spelling recognition, we propose a method to incorporate tone information to the classical front-end feature vector. All experiments are performed based on three environments (i.e. close, mix, open) of bigram language model. This paper is organized as follows. In section 2, Thai phonetic characteristics, alphabet system and spelling methods are presented. Section 3 describes the pitch extraction method. The spelling recognition framework with tone incorporation is introduced in section 4. The experimental results and analysis of spelling recognition are reported in section 5. Finally, a conclusion and some future works are given in Section 6.

2 Thai Phonetic Characteristics, Alphabet System and Spelling Methods

2.1 Thai Phonetic Characteristics

Like most languages, a Thai 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. Some initial consonants are cluster consonants. Each of them has a phone similar to that of a corresponding base consonant. For example, *pr* and *pl*, are similar to their corresponding base consonant *p*. In the vowel part, there are 18 vowel phones and 6 diphthongs. Following the concept in [12], there are totally 76 phonetic symbols and 5 tone symbols in Thai, as shown in Table 1.

Table 1. Phonetic symbols grouped as initial consonants, vowels, final consonants and tones

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

2.2 Pronunciation of Thai Alphabet

In Thai language, there are 66 commonly used letters as shown in Table 2. These letters can be grouped into three alphabet classes by phone expression, i.e., consonant, vowel and tone. There are various styles in pronouncing Thai alphabet. An alphabet in each alphabet class may have more than one pronunciation styles. The consonantal letters 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 consonantal letter ‘ก’, its core sound can be represented as the phonetic sound */k-@@0/*. Normally, some consonants share a same core sound. For example, ‘ค’, ‘ฅ’, and ‘ฆ’ have the same phonetic sound */kh-@@0/*. In such case, the hearer may encounter with 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 its representative word. For example, the representative word of the letter ‘ก’ is “ไก่” (meaning: “chicken”, sound: */k-a1-j^/*), and that of the letter ‘ข’ is “ไข่” (meaning: “egg”, sound: */kh-a1-j^/*). To express the letter ‘ก’ using this style, the syllable */k-@@0/+/k-a1-j^/* is uttered.

Table 2. Thai alphabets: consonants, vowels and tones

Basic Classes	Alphabets 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
First-type vowel (14)	อะ, อา, อิ, อี, อึ, อื, อุ, อู, เอ, แอ, โอ, อำ, ไอ, ใอ
Second-type vowel (4)	อั, อ็, อ์, ฤ
Tone (4)	อ่, อ้, อ๊, อ๋

Expressing letters in the vowel class is quite different from that of the consonant class. There are two types of vowels. The first-type 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 simply pronounce the core sound of the vowel. On the other hand, for the second-type vowels, they are uttered by calling their names. As the last class, tone symbols are always pronounced by calling their names. Table 3 concludes how to pronounce a letter in each alphabet class.

Table 3. Pronouncing methods for each alphabet class

Alphabet Class	Pronouncing Methods
Consonant	1. the core sound of the consonant + representative word of the consonant
	2. the core sound of the consonant
First-type vowel	1. */s-a1//r-a1/* + the core sound of the vowel
	2. the core sound of the vowel
Second-type vowel	1. the name of the vowel
Tone	1. the name of the tone

2.3 Thai Word Spelling Methods

Spelling a word is to utter letters in the word one by one in order. We can refer to spelling as a combination of the pronunciation of each letter in the word. In [6], Thai spelling methods were analyzed into four spelling styles. In this paper, we focus on one of the most frequently used spelling methods. In this method, for consonant letter, we pronounce only a consonant core sound. While the first-type vowel is pronounced as */s-a1//r-a1/* and then a vowel's core sound. The second-type vowel and tones are pronounced by calling their name. As mentioned above, some consonantal letters may share a same core sound. However, there will be exactly one letter, which is the most frequently used letter for each core sound, later called a representative letter. We will call the other letters with the same core sound as subordinate letters. Table 4 indicates a set of core sounds with their representative letters and subordinate letters. In order to differentiate which letter it is, a representative letter is pronounced by its core sound while a subordinate letter is pronounced by its core sound followed by its representative word.

Table 4. A set of core sounds with their representative letters and subordinate letters (consonantal letters)

Core Sound	Representative letter	Subordinate letter	Core Sound	Representative letter	Subordinate letter
/kh-@@4/	ข	ฃ	*/n-@@0/*	น	ณ
/kh-@@0/	ค	ฅ, ฆ	*/ph-@@0/*	พ	ภ
/ch-@@0/	ช	ฌ	*/j-@@0/*	ย	ญ
/d-@@0/	ด	ฎ	*/r-@@0/*	ร	ฤ
/t-@@0/	ต	ฏ	*/l-@@0/*	ล	ฬ
/th-@@4/	ถ	ฐ	*/s-@@4/*	ส	ศ, ษ
/th-@@0/	ท	ฑ, ฒ, ธ			

3 Extraction of Tone Feature

Tone (or pitch) information is considered as a potential source for improving recognition accuracy for any tonal language. Even in a non-tonal language, such tonal effect may occur when one would like to put a stress on some words or to make an interrogative utterance. The pitch information can be extracted from speech automatically and used in the recognition process. The following subsection displays the standard method to extract pitch (or tone information) and its derivatives.

3.1 Pitch Extraction

In the past, it was known that tone features could be characterized by tracing pitch or fundamental frequency (F_0) in every time unit on the voiced part of a syllable, resulting in a line shape or contour. In our work, these pitch values can be added directly to the classical feature vector in each time frame. In the past, there were two well-known pitch detection algorithms based on the time domain method called autocorrelation and the average magnitude difference function [13].

To recognize the tone of a syllable, we need normalize extracted pitch values instead of directly utilizing the extracted pitch values themselves. The aim of this task is to compensate the variety of speakers. Here, the normalized pitch value $\bar{p}_t$ can be derived by the following formula.

$$\bar{p}_t = \frac{p_t}{p_{avg}} \tag{1}$$

where p_t is the pitch value at the time frame t, and p_{avg} is the average of pitch values in the utterances. By the autocorrelation method the voice part of an utterance can be processed in order to get a pitch. However, it is impossible to get any pitch from an unvoiced part. Then the pitch is set to zero in the case of unvoiced parts. To solve the problem, we pass the normalized pitch values ($\bar{p}_t$) to a smoothing process in order to flatten pitch values for continuous speech recognition [9]. The smoothed values of a voiced part and an unvoiced part are calculated using equation (2) and (3), respectively.
Voiced :

$$f_t = \log_{10}(\bar{p}_t) + x \tag{2}$$

Unvoiced :

$$f_t = \begin{cases} f_{t-1} + \lambda(fav_{t-1} - f_{t-1}) + x, & t > 0 \\ \lambda & , \quad t = 0 \end{cases} \tag{3}$$

$$fav_t = \frac{\sum_{i=0}^{t} f_i}{t} \tag{4}$$

where fav_t is the running average of pitches in the previous frames and, x and λ are small random values determined through the experiments. In this work, they are set to 0.01 and 0.05, respectively.

3.2 Pitch Delta and Pitch Acceleration

The model can be enhanced by adding time derivatives. To grasp differences among pitch contours, the time derivatives of pitches can be used as important tone information. The pitch delta at a time frame is computed by the following formula.

$$d_t = \begin{cases} \dfrac{f_{t+\theta} - f_{t-\theta}}{2\theta}, & \theta < t < T - \theta \\ f_{t+1} - f_t \ , & t < \theta \\ f_t - f_{t-1} \ , & t \geq T - \theta \end{cases} \tag{5}$$

In the second and third equations, the end-effect problem was solved by using a simple compensation of first-order differences at the start and the end of utterances [14], where θ is the internal distance between two pitch points, f_t is a smoothed pitch value at time frame t. The same formula is applied to the delta values to obtain acceleration values.

$$a_t = \begin{cases} \dfrac{d_{t+\theta} - d_{t-\theta}}{2\theta}, & \theta < t < T - \theta \\ d_{t+1} - d_t \ , & t < \theta \\ d_t - d_{t-1} \ , & t \geq T - \theta \end{cases} \tag{6}$$

4 The Spelling Recognition Framework

In this work, HMMs are employed as the engine for recognizing continuous spelling utterances. The HMM is widely used in several works on speech recognition, especially for continuous speech since it can capture and handle a set of continuous data which are input as a sequence. Figure 1 illustrates our HMM-based automatic speech recognition (ASR) system, which consists of three major components: signal processing module, training module and recognizing module.

In the signal-processing module, speech utterances (wave signal) in the training corpus are transformed into the form of a feature vector. The feature vector used is the combination of PLP (Perceptual Linear Prediction Coefficients) and pitch information. In our work, the PLP is selected since it works well our dataset according to a number of preliminary experiments. To incorporate tone information, pitch information are extracted and integrated to the original PLP feature. Three main types of pitch information are taken into account: fundamental frequency (pitch), pitch delta and pitch acceleration. The combined feature vectors are used as input through our acoustic models. In the training module, an acoustic model and a language model are trained. For the acoustic model, the conventional phone-based HMMs are used to represent phones; i.e. one acoustic model per phone. A series of feature vectors, PLP with pitch information, are used to compute probabilities of an acoustic model. To train the language model for spelling recognition, a text corpus is used as a source for

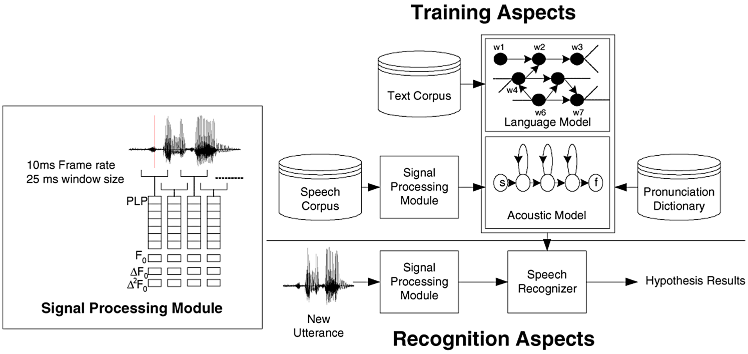

Fig. 1. The framework with a signal processing module for incorporating tone

assigning probabilities to a letter sequence. When 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 more flexible environment, 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, we investigate both limited and flexible environments. For the recognition module, an input waveform is transformed to a set of feature vectors that are the combination of PLP and pitch information. With three main sources, i.e., acoustic model, language model and pronunciation dictionary, the system searches among all possibilities, for the letter sequence with the maximum probability and returns it as the recognition result.

5 Experimental Result and Analysis

The section describes a set of experiments and their results. The purpose is to investigate the advantage of tone exploitation.

5.1 Experimental Environment

We have constructed two speech corpora, based on two sets of spelled proper names. The spelling style is the one shown in section 2.3. The first set (A) contains 150 spelled names recorded by three males and three females while the second set (B) contains 136 spelled names recorded by three other males and three other females. There is no overlap between proper names in the sets A and B. In this work, the set A is used as the training set while the set B is for a test set. The speech signals were digitized by a 16-bit A/D converter with frequency of 16 kHz. A set of feature vectors used for forming a baseline is a 39-PLP feature vector, which consists of 12 PLP coefficients and the 0th coefficient, as well as their first and second order derivatives. Therefore, there are 39 elements in total. In our proposed system, we construct a

42-component feature vector which consists of 39-PLP feature vectors as well as three components of tone information, i.e. pitch, pitch delta and pitch acceleration. Tone information components can be acquired by the method mentioned in Section 3. The layout of feature vectors used in our approach is illustrated in Figure 2.

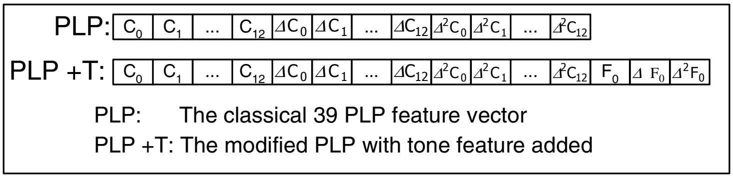
PLP:
C_0 | C_1 | ... | C_{12} | ΔC_0 | ΔC_1 | ... | ΔC_{12} | $\Delta^2 C_0$ | $\Delta^2 C_1$ | ... | $\Delta^2 C_{12}$
PLP +T:
C_0 | C_1 | ... | C_{12} | ΔC_0 | ΔC_1 | ... | ΔC_{12} | $\Delta^2 C_0$ | $\Delta^2 C_1$ | ... | $\Delta^2 C_{12}$ | F_0 | ΔF_0 | $\Delta^2 F_0$
PLP: The classical 39 PLP feature vector
PLP +T: The modified PLP with tone feature added

Fig. 2. Feature vector layout

The experiments are performed under three different bigram language models, LM1, LM2 and LM3. LM1 is a close-type language model constructed from the test transcription. LM3 is an open-type language model trained by another text corpus, which is not used as the test transcription. In this experiment we use 5,971 names of Thai provinces, districts and sub districts. LM2 is a mix-type language model generated from a corpus that includes both the test transcription and those 5,971 location names. In this work, Spelling recognizers are designed as phone-based HMMs. They are context-dependent in the sense that the recognition of a phone depends on its preceding and following phones. For each phone model, the topology is a 3-state left-to-right model with no skip. The number of phones is 56 as shown in Table 5. The numbers in parentheses denote the possible expansion of vowels using tones. For example "*a(0-4)*" indicates that the vowel phone '*a*' can be expanded by five different tones: 0 (mid), 1 (low), 2 (falling), 3 (high), 4 (rising).

Table 5. The list of possible acoustic models in the spelling corpora

Phonetic Types	Acoustic models (56 models)
Initial Consonant	*b,c,ch,d,f,h,j,k,kh,l,m,n,ng,r,s,t,th,tr,w,z*
Vowel	*@@(0,4),a(0-4),aa(0,1,4),e1,e(0,1),i(1,4),ii(0,4),* *o0,oo0, qq0,u1,u(0,2,3),uua3,v(1,3), vv0, xx0*
Final Consonant	*j^,k^,m^,n^,ng^,t^,w^*

All experiments, including automatic transcription labeling, are performed using the HTK toolkit [14]. The recognition performance is evaluated in the terms of correct rate and accuracy. Since the task concerned is spelling recognition not normal speech or word recognition, the definitions of word correct rate and word accuracy are modified to letter correct rate (LCR) and letter accuracy (LA). They are shown in equation (7) and (8) (see details in [14]). Here, *H* is the number of correct letters, *I* is the number of insertion errors, and *N* is the total number of letters.

$$\mathit{Letter\ Correct\ Rate\ (LCR)} = \frac{H}{N} \tag{7}$$

$$\mathit{Letter\ Accuracy\ (LA\ )} = \frac{H-I}{N} \tag{8}$$

5.2 Experimental Result

For comparison, we set up a baseline experiment, which employs the 39-PLP classical feature vector. By varying the grammar scale factor (GSF) [14] to adjust the appropriate weighting ratio between acoustic and language model, we can obtain the baseline result as shown in Table 6. In the table, it was observed that the best GSF for the close-type language model (LM1) is 25.0. It yields up to 80.08% LCR and 77.01% LA. In both the mix-type language model (LM2) and open-type language model (LM3), the appropriate GSF is 25.0. The best LCR and LA for LM2 are 74.18% and 60.47% while the best ones for LM3 are 73.47% and 59.48%, respectively. However, for further explanation, we will focus on the case that GSF equals to 25.0 since it gains higher accuracy in most cases. Therefore, the baseline results can be considered to 80.08% LCR and 77.01% LA for LM1 (close-type), 73.04% LCR and 68.17% LA for LM2 (mix-type), and 71.55% LCR and 66.36% LA for LM3 (open-type).

Table 6. Recognition performance for the classical PLP (the baseline)

Language		Grammar Scale Factor (GSF)					
model		5.0	6.25	10.0	25.0	50.0	100.0
LM1	COR	74.49	75.23	77.55	**80.08**	78.97	72.20
	ACC	55.17	58.28	65.76	77.01	**78.01**	71.01
LM2	COR	72.25	72.92	**74.18**	73.04	68.33	59.28
	ACC	51.72	54.36	60.47	**68.17**	66.02	57.08
LM3	COR	71.80	72.47	**73.47**	71.55	66.73	57.87
	ACC	51.14	53.81	59.48	**66.36**	64.02	56.11

Table 7. Recognition performance when tone information is incorporated to PLP (PLP+T)

Language		Grammar Scale Factor (GSF)					
model		5.0	6.25	10.0	25.0	50.0	100.0
LM1	COR	75.13	75.82	77.87	**82.28**	82.27	77.03
	ACC	64.58	66.15	70.94	78.74	**80.32**	75.43
LM2	COR	73.45	73.88	74.41	**75.11**	71.65	61.98
	ACC	62.54	63.80	66.74	**71.02**	68.95	59.61
LM3	COR	73.16	73.54	**73.92**	73.81	69.51	60.08
	ACC	62.09	63.35	66.13	**69.52**	66.56	57.54

The recognition performance in Table 7 is obtained when tone information is incorporated to the classical PLP feature vector. Independent of GSF and language models, the improvement over the baseline is observed when tone information is incorporated into the PLP feature vector. With the most suitable GSF (25.0), the system can achieve up to 82.28% LCR and 78.74% LA for LM1 (close-type), 75.11% LCR and 71.02% LA for LM2 (mix-type), and 73.81% LCR and 69.52% LA for LM3 (open-type).

6 Conclusion and Future Work

This paper presented a method to improve accuracy in Thai spelling recognition by incorporating tone information into the classical PLP feature vector. Characteristics of Thai language and its spelling method were introduced. The proposed HMM-based method recognized spelling utterances of the most popular Thai spelling method. The system was examined under three language model environments; close-type, mix-type and open-type. With the 42-component feature vector (39-PLP with three tone features), the system outperformed the baseline system (39-PLP feature vector) with improvement of 1.73%, 2.85% and 3.16% for letter accuracy in close-type, mix-type and open-type language models, respectively. As further works, we plan to explore more tone features and study spelling recognition for other types of Thai spelling methods.

References

1. Mitchell, C.D., Setlur A.R.: Improved spelling recognition using a tree-based fast lexical match. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (1999) 597-600
2. Hild, H., Waibel, A.: Recognition of spelled names over the telephone. Proceedings of the International Conference on Spoken Language Processing, ICSLP 96, Philadelphia, (1996) 346-349
3. Bauer, J.G., Junkawitsch, J.: Accurate recognition of city names with spelling as a fall back strategy. Proceedings of EUROSPEECH (1999) 263-266
4. San-Segundom, R., Colas, J., Cordoba, R., Pardo, J.M.: Spanish recognizer of continuously spelled names over the telephone, Journal of Speech Communication 38, (2002) 287-303
5. Rodrigues, F., Rodrigues R., Martins, C.: An isolated letter recognizer for proper name identification over the telephone. Proceedings of 9th Portuguese Conference on Pattern Recognition (RECPAD'97), Coimbra (1997)
6. Pisarn, C., Theeramunkong, T.: Speed compensation for improving Thai spelling recognition with a continuous speech corpus. Intelligence in Communication System, LNCS 3283, IFIP International Conference, INTELLCOMM 2004, Bangkok, Thailand, (2004) 100-111
7. Lee, T., Ching, P.C., Chan, L.W., Cheng, Y.H., Mark, B.: Tone Recognition of Isolated Cantonese Syllables, IEEE Transaction on Speech Audio Processing (1988) 988-992
8. Chen, C. Julian, Recognize Tone Languages Using Pitch Information on The Main Vowel of Each Syllable, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, (2001) 61-64
9. Chang, E., Zhou, J., Di, S., Huang, C., Lee, K.: Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones, Proceedings of International Conference on Spoken Language Processing 2000

10. Thubthong, N., Kijsirikul, B.,: Improving Connected Thai Digit Speech Recognition using Prosodic Information, Proceedings of The 4th National Computer Science and Engineering Conference (2000) 63-68
11. Wong, P., Siu, M., Integration of Tone Related Feature for Chinese Speech Recognition, 6th International Conference on Signal Processing, (2002) 476-479
12. Pisarn, C., Theeramunkong, T.: Incorporating tone information to improve Thai continuous speech recognition, Proc. of International Conference on Intelligent Technologies, Chiangmai, Thailand, (2003) 84-89
13. Rabiner, L.R., et. al.: A Comparative Performance Study of Several Pitch Detection Algorithms, IEEE Transaction on Acoustics, Speech, and Signal Processing, Vol. ASSP-24 No. 5, (1976)
14. Young, S., et al, The HTK Book (for HTK Version 3.1), (2000)