Tone Classification for Isolated Thai Words using Multi-Layer Perceptron

Maleerat S., Supot N., and Choochart H.

***Abstract*— This paper proposed a tone classification for isolated Thai words using Multi-Layer Perceptron (MLP). According to the syllable structure effect of initial consonants, vowels, and final consonants on tone recognition, three tone feature set were extracted from F_0 contour to capture the characteristic of five Thai tones and feed it into the MLP to classify five different tones of target corresponding to five Thai tones. Two dataset of Thai monosyllable were used for training and testing. The data were collected from 5 male and 5 female of Thai native speakers. The results show that these tone feature set which extracted from F_0 can be achieved a high performance of Thai tone recognition.**

***Index Terms*— Tone Recognition, Tone Language**

I. Introduction

Thai language is one of a tone language like Mandarin Chinese. The identification of Thai tone relies on the shape of pitch frequency or the fundamental frequency (F_0) contour. Thai has five canonical tones: mid, low, falling, high, and rising. Fig. 1 shows the fundamental frequency (F_0) of five different tones in standard Thai language. Tones are indicated by contrastive variation in F0 at the syllable level which make different meaning of word or sentence of speech. The variation in F0 that is not related to the syllable or word, but depend on the structure of sentence which can be changed the meaning of sentence from affirmative sentence to interrogative sentence so called “intonation”. In addition to the intonation of a statement, there is another aspect of speech that indicates meaning such as phrasing of a conversation.

There are some researches related to Thai tone recognition and parameter that effected to the precision of tone recognition such as syllable structure, coarticulation, intonation, stress, speaking rate, dialect, sex, age and emotion [1]. The researchers [2], [3], [4], [5], [6], [7] studied about the effect of syllable structure, coarticulation, intonation and stress on tone recognition in continuous speech and spontaneous speech to improve the performance of speech recognition.

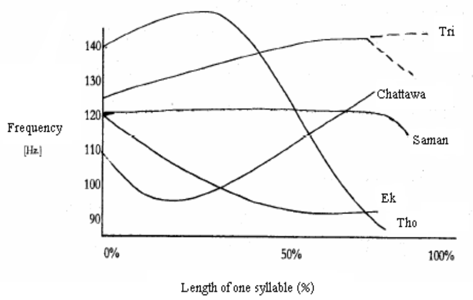

Fig. 1. F_0 contour of five tones in standard Thai language

This paper is organized as follows. In section 2, we first present the tone in Thai language. Section 3 describes tone recognition with tone feature extraction and tone classification model. Section 4 describes the experiments and gives a conclusion in Section 5 respectively.

II. TONE IN THAI LANGUAGE

A. *Thai syllable structure*

Thai is a tonal language which makes it very different to Western languages. The phonetic structure of Thai is based primarily upon the monosyllable. Thai syllable structure consists of /C(C)V(:)(C)^T/ where C,V,: and ^{T} represent a consonant, vowel, vowel length and lexical tone, respectively [10]. Each syllable has a choice between five distinct tones: low, mid, rising, high and falling with 4 mark symbols respectively. Tone is a supra-segmental that lies on a group of voiced segment and usually associated with vowels. However, tone is a high-low phone or pitch that useful to identify the different meaning of words which have the same phoneme of consonants and vowels for example: “มัน” /man^0/ and “หมั้น” /man^2/. It means that a high-low phone for each word has significant in term of linguistic. However, tone will be occurred with all voiced (all vowels are voiced). The level of tone depends on frequency of vocal cords vibration. If high frequency of vibration the level of tone will be high, on the other hand the tone level is low if the frequency of vibration is low. When a tone of vowel is occurred in the syllable, the other tone in that syllable will be changed to the same as vowel tone. Normally, Tone is occurred together of vowel. There are 2 groups of standard Thai language tone: level tone and contour tone [9].

Level tone means the level of tone is stable from beginning to the end. There are three phones of this tone:

1. Low tone: is the phone tone “Ek” which started at level of 118 Hz and then decrease a bit until close to the end syllable around 110 Hz. This is a level tone with no inflection but lower in pitch than common tone. The symbol for this phone tone is /^{1}/ for example: ป่า /Pa^1/ หมาก /Mak^1/ สับ /sap^1/.

2. Mid Tone: is the phone tone “Saman” which started at level of 120 Hz and then decrease a bit until close to the end

Manuscript received July 26, 2009. Tone Classification for Isolated Thai Words using Multi-Layer Perceptron

Maleerat S. Author is with of the Information Technology, King Mongkut’s University of Technology North Bangkok, Thailand (phone: +662-9132500; fax: +662-9122019; e-mail: msn@ kmutnb.ac.th).

Supot S. Author is with of the Information Technology, King Mongkut’s University of Technology North Bangkok, Thailand (phone: +662-9132500; fax: +662-9122019; e-mail: sns@ kmutnb.ac.th).

Choochart H. Author is with the National Electronics and Computer Technology Center (NECTEC), Pathumthani, Thailand (e-mail: choochart.haruechaiyasak@nectec.or.th).

syllable around 112 Hz. This is spoken in the speaker’s ordinary tone of voice without any inflection. It is the tone used in English for ordinary conversation. The symbol for this phone tone is / ^{0}/ for example: กา /kaa^0/ คน /kon^0/ บิน /bin^0/.

3. High Tone: is the phone tone “Tri” which started at level of 125 Hz and then increasing around 134-140 Hz until close to the end syllable and decrease a bit when close to the end syllable. This is a uniform tone pitched well above the level of the speaker’s normal voice. The symbol for this phone tone is /^{3}/ for example: ฟ้า /faa^3/ น้ำ /naa^3/ วัด /wat^3/.

Contour tone means the level of tone is change from the beginning. It can be from high to low and vice versa between the periods of utterance for each syllable and high frequency changing between the initial and final syllable. There are two phones of this tone:

1. Rising tone: is the phone tone “Chattawa”. It starts from low level around 110 Hz, decreasing a bit and then change to high rapidly and at the end of syllable the frequency is around 140 Hz. This as the name implies has a rising inflection. The symbol for this phone tone is /^{4}/ for example: หมา /maa^4/ สาว /saaw^4/ ฉัน /chan^4/.

2. Falling tone: is the phone tone “Tho”. It starts from high level around 140 Hz, increasing a bit and then change to low rapidly till lowest at the end of syllable around 100 Hz. This is an emphatic and heavily accented tone with a falling inflection. The symbol for this phone tone is /^{2}/ for example: พ่อ /ph^2/ แก้ว /kaaw^2/.

Intensifying tone, it changes the tone level with the same syllable. It starts with high level before falling a bit to the end of syllable. This tone almost appears in the repeated word which stress on the first syllable just to determine a special meaning such as ดี๙ ดี /diː^5-diː^0/, ชอบ๙ ชอบ /chɔːp^5-chɔːp^2/.

For tone recognition, F_0 is considered to use tone classification as an input. Several interacting factors affect F_0 realization of tones [10]. Therefore, it’s difficult to identify tones from different speaking style of speakers. However, in this paper we proposed the way to extract a distinguish features of tones and also a tone model which suitable for Thai tone recognition.

B. *Factors determining the tone*

There are two different types of Thai syllable: stressed and unstressed syllable. Stressed syllable is the syllable that can be stressed individually in normal speaking. Its consist of at least three parts: initial consonant vowel and tone as (CV^T)or maximum five parts of two initial consonants (cluster consonant) one vowel one final consonant and tone (CCVC^T).

TABLE 1
THAI PHONEME SET (IPA)

Initial consonants	p t c k ? ph th ch kh b d f s h m n ɳ w y r l
Cluster consonants	pr pl tr kr kl kw phr phl thr khr khl khw br bl fr fl dr
Final consonants	p t k m n ɳ w y
Short vowels	i ʝ u e ə o æ a ɔ ua ʝa ia
Long vowels	i: ʝ: u: e: ə: o: æ: a: ɔ: ua: ʝa: ia:
Tones	ˉ ˋ ˆ ˊ ˇ

Unstressed syllable is the syllable that appears in the unstressed position in normal speaking. The Thai phoneme set consists of 21 consonantal phonemes, 17 consonantal cluster phonemes, and 24 vowels as shown in Table 1. To determine the tone of any syllable the following three factors have to be considered.

1) Class of the initial consonant: The Thai 44 consonants are divided up into three groups known respectively as High, Middle and Low class consonants and the first thing to look at in determining the tone of a word or syllable is the class of the initial consonant. There are many cases where the letter ห as an initial consonant is silent and there are a few cases where the letter อ as an initial consonant is also silent, but this makes no difference to the rule, the tone is still governed by the class of the initial consonant even though it can be a silent consonant.

The High class consonants are:
ข-kh ฉ-ch ฐ ถ-th ผ-ph ฝ-f ศ-ษ-ส-s ห-h
The Middle class consonants are:
ก-k จ-c ฎ-ด-d ฏ-ต-t บ-b ป-p อ-z
All the remainders are Low class consonants:
พ-ภ-ph ฟ-f ฑ-ฒ-ท-ธ- th ค-ฅ-ฆ-kh ซ-s ฮ-h ช-ฌ-ch ง-ng ญ-ย-j น-ณ-n ร-r ว-w ม-m ฬ-ล-l

2) The final sounded consonant: All words which do not end in a vowel sound must have either m, n, ng, k, p, or t as the final sound. Although this is strictly true, but in some conversation the final consonant is often slurred and particularly after a long vowel, the final “p” may sound more like a “b” and the final “t” more like a “d”. Where there is no tone mark, the tone of the syllable or word will depend on both the class of the initial consonant and on whether it ends with the m, n, ng sounds or the k, p, t sounds. It should be noted that a final consonant with the sign -์ over it is not sounded and hence can have no effect on the tone.

3) The type of final vowel: If the word has no tone mark and ends in a final vowel, the tone is dependent on whether this final vowel is a long or short one. The short vowels for tonal purposes are -ะ, -ั, -ิ, -ึ, -ุ, the inherent “a”, the inherent “o” and all vowels shortened by the sign -็ over the consonant or by the addition of the vowel -ะ at the end. All the others are long vowels.

Fig. 2 shows pitch of five tones of monosyllable. Each syllable has different meaning. The first syllable is mean throw away, the second means forest, the third means aunt and forth and fifth syllable mean father (like sound of Chinese).

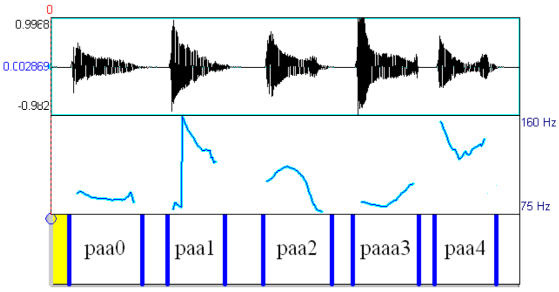

Fig. 2. Pitch of five tones of monosyllable

III. TONE RECOGNITION

The framework of tone recognition shows in Fig. 3. It starts from the input speech signal to voiced portion detection to extract F_0 contour, then the F_0 contour of speech signal is passed to tone feature extraction to extract the tone feature set and then all features are passed to tone classifier to classify and recognized tone respectively.

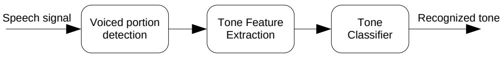

Fig. 3. Tone recognition frameworks

A. *Tone Feature Extraction*

F_0 extraction

The fundamental frequency (F_0 or pitch frequency) is correlated with the gender of the speaker. Normally, F_0 contour changes depend on type of tone as shown in Fig. 1 as a standard and Fig. 2 from the experiments. Since F_0 has found in voiced portion only. Therefore, an F_0 contour can be extracted by using root mean square energy and zero-crossing rate to detect voiced portion from speech signal. The definition of both function are shown in (1) and (2).

Root Mean Square Energy (RMSE)

The root mean square energy can be defined as:

$$RMSE(n) = \frac{1}{N}\sum_{m=0}^{N-1} s(m)^2, \qquad (1)$$

which is the square root of the average of the sum of the squares of the amplitude of the signal samples. A frame with high energy corresponds to a voiced portion of speech signal, while one with low energy to an unvoiced region.

Zero Crossing Rate (ZCR)

The short-time zero crossing rate is an indicator for the presence of voiced or unvoiced speech respectively. For unvoiced speech the zero crossing rates shows in general higher than voiced speech. A short-time zero crossing rate for a block of N speech samples starting at sample m is defined as:

$$ZCR_s(m) = \frac{1}{2(N-1)}\sum_{n=m+1}^{m+N-1} |sgn[s(n)] - sgn[s(n-1)]| \qquad (2)$$

where

$$sgn[s(n)] = \begin{cases} 1, & s(n) \geq 0 \\ -1, & s(n) < 0 \end{cases} \qquad (3)$$

An estimation of ZCR, we used hamming window of 25ms frame with 10 ms frame shift.

Average Magnitude Difference Function (AMDF)

There are several methods to extract F_0 or pitch from speech signal such as autocorrelation, simple harmonic motion analysis and average magnitude difference function etc. In this paper, one approach for extracting the F_0 of a speech signal is the average magnitude difference function (AMDF) [11] with 25 ms frame size and 10 ms frame shift. It used to measure periodicity in a time frame by summing up absolute differences between equidistant samples. The average magnitude difference function of a frame of speech signal is defined as:

$$AMDF(\tau) = \frac{1}{L}\sum_{n=1}^{L} |s(n) - s(n-\tau)| \qquad (4)$$

where $s(n)$: the samples of input speech signal
$s(n-\tau)$: the samples time shifted
L: the length of a frame of speech signal
$\tau = 1, \ldots \tau_{max}$

Average magnitude difference is a function alternative of auto-correlation. It measure the similarity of a time frame $s(n)$ of length L and it time shifted copy $s(n-\tau)$ by applying absolute difference instead of multiplication. An advantage of this method over auto-correlation is its computational simplicity, because subtraction and magnitude computational are much faster than multiply and add computational. It is applicable for real time application of speech communications. Since not all syllables are of equal duration, F_0 contours are equalized for duration on a percentage scale [4]. We obtain 11 different time points with the equal step size of 10% from 0% to 100% of the voiced portion. There are three tone feature sets that we can use as an input of MLP classifier [2].

Tone feature set 1:

The shape of F_0 contour represents a tone. Four slopes of F_0 and the average of the F_0 level are used to classify the tones [2]. A slope of F_0 is denoted as delta F_0 (dF0) which can be approximated as:

$$dF_0(n) = F_0(n+1) – F_0(n-1) \qquad (5)$$

where *n* is the index of the F_0 profile. In this thesis, four dF_0's, computed at 20, 40, 60, and 80% of the voiced portion are served as tone feature set 1.

Tone feature set 2:

The initial F_0 (F_{0I}), the final F_0 (F_{0F}) and four F_0 at 20, 40, 60 and 80% of the voiced portion are use to identify this tone feature set[2]. The initial F_0 (F_{0I}) and final F_0 (F_{0F}) are defined as follows:

$$F_{0I} = (F_0(0)+F_0(1))/2 \qquad (6)$$

$$F_{0F} = (F_0(9)+F_0(10))/2 \qquad (7)$$

Tone feature set 3:

It consists of the initial F_0 (F_{0I}), the final F_0 (F_{0F}) and four dF_0's at 20, 40, 60 and 80% of the voiced portion.

B. Tone Classification Model

Normalization

All feature vectors are normalized to lie between –1.0 and 1.0 using the following equation:

$$norm\ F_i = 2.0 * \left(\frac{F_i - min\ F_i}{max\ F_i - min\ F_i}\right) - 1.0 \qquad (8)$$

Classification

To classify five tones from three tone feature sets, a multi-layer perceptron neural network is used. An input layer is a number of normalized three tone features set with a hidden layer of 10 units and output of five units corresponding to Thai tones as shown in Fig 4.

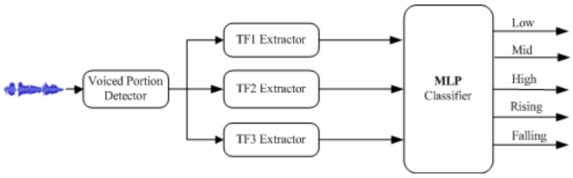

Fig. 4. Tone Extractor and MLP classifier of tone recognition system

IV. EXPERIMENTS AND RESULTS

A. A. Experiments

In this paper, a tone pair set was used to train and test the effect of phoneme on tone recognition. The data was collected from 10 native Thai speakers (5 male and 5 female speakers), aged from 24 to 35 years. First we test the similarity of tone from word pairs of monosyllabic as shown in Table 2. The speech signals were sampled at 22 kHz, digitized with a 16 bit A/D converter. The second set of data shown in Table 3. Five words with different tones were collected from 10 native Thai speakers (5 male and 5 female speakers), aged from 24 to 35 years. Tone pair set has 100 records and a sample words for all tones set has 625 records.

Table 1. Thai tone pairs

Tone	**Set-1**	**Set-2**
Mid	/saj^{0}/ ไซ	/law^{0}/ เลา
Low	/saj^{1}/ ใส่	/law^{1}/ เหล่า
Falling	/saj^{2}/ ไส้	/law^{2}/ เล่า
High	/saj^{3}/ ไซ้	/law^{3}/ เล้า
Rising	/saj^{4}/ ใส	/law^{4}/ เหลา

Table 2. Sample words for each tone

Mid	**Low**	**Falling**	**High**	**Rising**
/chɤːj^{0}/ เชย	/paː^{1}/ ป่า	/khaː^{2}/ ค่า	/ruː^{3}/ รู้	/khaː^{4}/ ขา
/rɯm^{0}/ เรือ	/sum^{1}/ สุ่ม	/niː ^{2}/ หนี้	/chan^{3}/ ชั้น	/diaw^{4}/ เดี๋ยว
/ciɳ^{0}/ จริง	/siaɳ/ เสี่ยง	/jaːk^{2}/ ยาก	/khiaw^{3}/ เคี้ยว	/naːm^{4}/ หนาม
/lɯːm^{0}/ ลืม	/lut^{1}/ หลุด	/ruan^{2}/ ร่วน	/khit^{3}/ คิด	/suaj^{4}/ สวย
/jaːw^{0}/ ยาว	/ʔaːp^{1}/ อาบ	/baːn^{2}/ บ้าน	/nap^{3}/ นับ	/rɯː^{4}/ หรือ
/waːj^{0}/ วาย	/buak^{1}/ บวก	/sɔm^{2}/ ส้อม	/luʔ^{3}/ ลุ	/ɳaw^{4}/ เหงา

B. Results

According to our test set with almost monosyllable with different tones from different speakers. The results give a high accuracy of recognition both for tone pair and sample of each tone experiments. Table 4 shows the results of tone recognition accuracy with two sets of data. The highest accuracy is the falling tone. Table 5 shows tone recognition accuracy from sample words which falling tone is a highest accuracy for tone recognition. The recognition accuracy result shown in Table 5 is lower than in Table 4 because of different initial consonants within data set. The error of recognition is the duration of syllable and intonation.

Table 4. Result of tone recognition with different data set

Data Set	**Tone recognition accuracy (%)**				
	Mid	**Low**	**Falling**	**High**	**Rising**
Set 1	94	90	97	95	93
Set 2	94	89	96	95	94

Table 5. Tone recognition accuracy from sample words

Tone recognition accuracy (%)				
Mid	Low	Falling	High	Rising
91	90	94	92	90

V. CONCLUSION

Tone recognition is more important for speech recognition of tonal language like Thai. In this paper, we proposed a tone recognition method for Thai spoken language using Multi-Layer Perceptron (MLP). A fundamental frequency F_0 was used to extract a distinguish tone feature set and feed into the MLP for classification of different tones. The experimental results show that the proposed method gives a high accuracy of tone recognition with a dataset of monosyllable words.

REFERENCES

[1] Antonis Botinis, Bjorn Granström and Bernd Möbius **, "** Developments and paradigms in intonation research", Speech Communication 2001, pp.263–296.

[2] Thubthong, N "A method for isolated Thai tone recognition using a combination of neural networks". Computational Intelligence, Volume 18, Number 3,2002, pp. 312-335.

[3] Chao Wang and Stephanie Seneff, "Improving tone recognition by normalizaing for coarticulation and intonation effects," *International Conference on Spoken Language Processing*, 2000.

[4] Jian-lai Zhue, Ye Tian, Yi Shi, Chao Hung and Eric Chang, "Tone articulation modeling for Mandarin spontaneous speech recognition*",* ICASSP 2004 ,pp.997-1000.

[5] Ning Zhou, Wenle Zhang, chao-Yang Lee, and Li Xu, "Lexical Tone Recognition with an Artificial Neural Network",NIH Public Access, 2008.

[6] Siripong Potisuk and Mary O.Harper,"Speaker-Independent Automatic Classification of Thai Tones in Connected Speech by Analysis-Synthesis method**"**, Acoustics Speech and Signal Processing, 1995**,** pp 632-635.

[7] Siwei Wang, Gina0Anne Levow, "Mandarin Chinese Tone Necleus Detection with Landmarks",Interspeech 2008, pp 1101-1104.

[8] Luksaneeyanawin Anawin,"Intonation in Thai", In Intonation Systems:A Survey of Twenty Language, Edited by D.Hirst and A. D. Cristo, pp.376-394.

[9] Kanchana Naksakul, "*Thai Phonology System*", Chulalongkorn, 2008.

[10] Gandour, J.; Potisuk, S.; and Dechnongkit "Tonal coarticulation in Thai", Journal of Phonetics 22,1994, pp 477–492.

[11] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H.J.Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing,*vol. ASSP-22, pp. 353-362.