ENGINEERING JOURNAL

Article

A Hidden Conditional Random Field-Based Approach for Thai Tone Classification

Natthawut Kertkeidkachorn[a], **Proadpran Punyabukkana**[b], **and Atiwong Suchato**[c]

Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
E-mail: [a]Natthawut.K@student.chula.ac.th, [b]Proadpran.P@Chula.ac.th (Corresponding author), [c]Atiwong.S@Chula.ac.th

Abstract. In Thai, tonal information is a crucial component for identifying the lexical meaning of a word. Consequently, Thai tone classification can obviously improve performance of Thai speech recognition system. In this article, we therefore reported our study of Thai tone classification. Based on our investigation, most of Thai tone classification studies relied on statistical machine learning approaches, especially the Artificial Neural Network (ANN)-based approach and the Hidden Markov Model (HMM)-based approach. Although both approaches gave reasonable performances, they had some limitations due to their mathematical models. We therefore introduced a novel approach for Thai tone classification using a Hidden Conditional Random Field (HCRF)-based approach. In our study, we also investigated tone configurations involving tone features, frequency scaling and normalization techniques in order to fine-tune performances of Thai tone classification. Experiments were conducted in both isolated word scenario and continuous speech scenario. Results showed that the HCRF-based approach with the feature F_dF_aF, ERB-rate scaling and a z-score normalization technique yielded the highest performance and outperformed a baseline using the ANN-based approach, which had been reported as the best for the Thai tone classification, in both scenarios. The best performance of HCRF-based approach provided the error rate reduction of 10.58% and 12.02% for isolated word scenario and continuous speech scenario respectively when comparing with the best result of baselines.

Keywords: Thai tone classification, hidden conditional random fields, tone feature, frequency scaling, normalization technique and fundamental frequency.

ENGINEERING JOURNAL Volume 18 Issue 3
Received 4 November 2013
Accepted 20 January 2014
Published 10 July 2014
Online at http://www.engj.org/
DOI:10.4186/ej.2014.18.3.99

DOI:10.4186/ej.2014.18.3.99

1. Introduction

Most of the modern approaches for constructing an automatic speech recognition (ASR) system are statistical machine learning-based approaches with spectral-based features. Although the approaches achieve highly results in general [1, 2, 3], the results still are not satisfying in case of tonal languages. Generally, an ASR system generates candidate hypotheses of possible words based on spectral information without considering tone information, which provides tonal sound in tonal languages. Consequently, ignoring tone information degrades a performance of ASR system evidently for tonal languages since some candidate hypotheses do not have a unique meaning. For example /ka:/ in Thai, it has five different meanings, including 'a kind of grass', 'galangal', 'to kill', 'to trade' and 'a leg'. In order to differentiate the meaning of word, therefore, tone information plays a key role because in tonal languages tone information not only expresses prosody or pragmatic information as usual but also conveys information, which characterizes the lexical meaning. As a result, incorporating tone information into ASR systems is widely attended by many researchers [4, 5] because it can improve a capability of ASR system for tonal languages. Nonetheless, it is still challenged that how we can obtain tone information, which is highly accurate. Therefore, a studying of tone classification is necessary to provide tone information correctly. In this article, we focus on studying a tone classification in Thai scenario.

In Thai, there are several studies investigating the tone classification problem [6, 7, 8, 9, 10]. Most of them rely on non-sequential discriminative classifier approaches, such as Logistic Regression-based approach, Artificial Neural Network (ANN) based approach and Support Vector Machine (SVM) based approach, in which acoustic feature vectors at various position of speech segment are extracted dependently in order to optimize classifier's parameters. As reported in [6, 7, 8], fundamental frequency (F_0) values and their derivative usually are selected to represent acoustic feature vectors, which use for identifying tones. Even though acoustic features are represented by F_0 values and their derivative in many studies, considering study in [6, 7, 8], we found that individual positions in speech segments for extracting acoustic feature vectors in each study are different. It leads to the individual positions in speech segment selection problem that which positions in speech segments is suitable for a tone classification task. Therefore, the non-sequential discriminative classifier approach is highly complicated to find appropriate acoustic feature vectors [11]. There is another study avoiding the individual positions in speech segment selection problem by using a Hidden Markov Model (HMM)-based approach [10]. A Hidden Markov Model (HMM)-based approach relied on a Hidden Markov Model (HMM)-based classifier. A HMM-based classifier is a sequential-based classifier so it can efficiently support sequential feature vectors. Consequently, a Hidden Markov Model (HMM)-based approach does not need to consider that which positions in speech segments are appropriate for extracting acoustic features. However, a HMM-based approach still has some drawbacks especially due to the independence assumption problem, as reported in study [12] since a Hidden Markov model is a generative model. A generative model generally needs to model relations between states, between observations and between those states and those observations for describing how those observations were generated by those states so a generative model is very complicating for training and inference. In order to make the computation of training and inference feasibly, some assumptions were assigned for reducing complexities of model. In case of a Hidden Markov model, two assumptions were assigned. The first assumption is that an observation was generated from a corresponding state only. The second assumption is a Markov property assumption. The Markov property assumed that a current state depended on a previous state. Based on these two assumptions, they leads to that each observation was generated independently. However in a HMM-based system for a tone classification task, two adjacent observations usually have overlapping regions when we extract features so the independence assumption that each observation was generated independently is wrong.

Based upon our review, we therefore choose a Hidden Conditional Random Field (HCRF)-based approach proposed by [12] in our study. There are three advantages of HCRF-based approach beyond other approaches in previous works; 1) a HCRF-based classifier is a sequential model which can easily apply to sequential data, especially F_0's value and their derivative, so feature selection methods were simplified, 2) the HCRF-based classifier does not need to assume the independence assumption like HMM-based classifier since the HCRF-based classifier use feature functions to capture relations among observations instead of defining the independence assumption as reported by [12] and 3) the HCRF-based classifier has hidden states, which can capture an intrinsic sub-structure of acoustic cues that dramatically thoroughly changes in each syllable.

DOI:10.4186/ej.2014.18.3.99

In this article, we introduced a study on Thai tone classification using a HCRF-based approach and also investigated three tone configurations regarding tone features, frequency scaling and normalization techniques. Note that, despite widely known knowledge that tone perception relates to information beyond F_0's values, tone features in our study are scoped such that they are limited to features involving with F_0's contours, F_0's values and their derivatives. Those three configurations are widely studied in Thai tone classification since there are some evidences in studies [6, 7, 8] showed that tuning those three configurations directly affected performance of Thai tone classification. We therefore had taken those three configurations into account. The rest of article is organized as follows. We briefly presented the overview of Thai tone in section 2 and the related literatures in section 3. Section 4 introduced HCRF. In the following section, we reported our speech corpora in the experiments; Section 6 presented the data preparation method. In section 7 and 8, we described the experiments and the results as well as the discussion respectively. Eventually, we concluded our study in the last section.

2. Background Knowledge

Components of sound system in Thai comprise of consonantal sound, vowel sound and tonal sound, which could compose to be a syllabic unit. Syllable structure in Thai is $C_i(C_i)V(:)(C_f)^T$ where 'C_i' is an initial consonant, 'C_f' is a final consonant, 'V' is a vowel, ':' is a length of vowel and T represents a tone.

There are 33 consonants, which are 21 consonants (C_i) and 12 cluster consonants (C_iC_i) as well. Those 33 consonants can be an initial consonant. However, in case of final consonants (C_f), there are only 9 possible consonants, which can appear at the final consonant position.

Vowels in Thai consist of 18 monophthongs and 6 diphthongs. In monopthongs, there are 9 qualitative differences and each of them has two quantitative differences of which duration are short and long. Diphthongs also have two quantitative differences, which are similar to monophthongs, but there are only three qualitative differences.

In Thai, there are five different tones: the mid tone, the low tone, the high tone, the rising tone and the falling tone. By their F_0 movements, these tones were categorized into two groups which are static tones and dynamic tones according to Luksaneeyanawin's study [14]. In static tone group, there are three tones: mid tone, low tone and high tone. They are considered as static tones because F_0 movements are relatively stable whereas the F_0 movements of the other group which consists of falling and rising tones are relatively dynamic. F_0 movements of the falling tone change from upward direction to downward direction while F_0 movements of the rising tone are reverse. Figure 1 shows the average of F_0 movements of five Thai tones in citation form, which were normalized by syllable duration, from a male Thai native speaker as reported in [13].

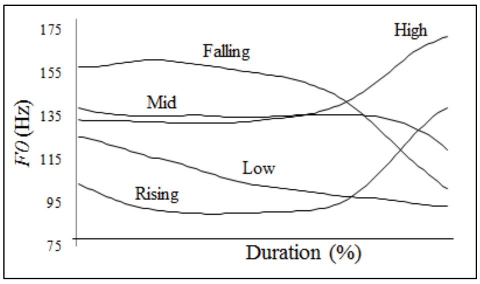


Fig. 1. The F_0 contours of five Thai tones in citation form [13].

In case of the continuous speech, the F_0 contours in Fig. 1 were shortly degraded their shapes because of adjacent syllables, especially the preceding and the following syllables. Figure 2 showed F_0 movements of the five tones in the continuous speech, which were uttered by the same speaker in citation form. The deformation of F_0 movements leads to the difficulty of tone discrimination in continuous speech. Considering Figs. 1-2, we can see that F_0 movements in isolated speech and continuous speech are different. Therefore, we investigated Thai tone classification tasks in both scenarios.

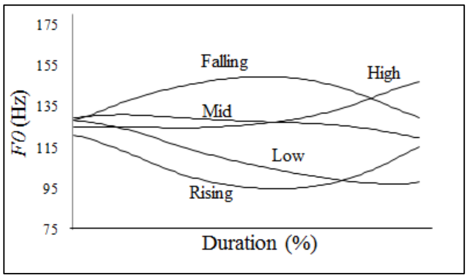


Fig. 2. The F_0 contours of five Thai tones in continuous speech [13].

3. Literature Reviews

In this section, we reviewed and discussed literatures of tone classification and tone recognition in tonal languages, especially in Chinese Vietnamese and Thai.

In Chinese, there are many dialects, such as Mandarin, Cantonese and Taiwanese. Therefore, tone classification and tone recognition in Chinese are widely studied and attended by many researchers. Lee et al. [15] proposed tone recognition for isolated Cantonese syllables. They introduced acoustic features relating to relative pitch level, temporal pitch variation and duration. The relative pitch level is represented by F_0 values at initial position and final position of syllable's duration. For temporal pitch variation, the average changing of F_0 values is measured. Energy drop rate was used to represent duration features since duration and energy drop rate are highly correlated as reported in [15]. Acoustic features were used to optimize statistical parameter for constructing tone classification relied on an ANN-based approach. Their results showed that in case of speaker-dependent they achieved accuracy of 89% and in case of speaker-independent they acquired accuracy of 87.6%. Jian [16] investigated effect of energy movements for Taiwanese tone classification. Although F_0 values alone can identify most of Taiwanese tones, some tones in Taiwanese still cannot differentiate. Considering Taiwanese tones, especially between short falling and long falling, he found that F_0 movements alone cannot distinguish the tones correctly. Hence he introduced energy movement model, which based upon the ADSR (Attack, Decay Sustain and Release) model. His analysis results showed that energy movements which based on ADSR model can significantly improve Taiwanese tone classification successfully. Xu et al. [17] investigated Mandarin tone recognition using an ANN-based approach. F_0 values at middle of each segment which obtained from dividing tone segment into small segments varied from 1 to 16 were selected to represent acoustic features. They achieved accuracy of over 90%. Tian et al. [18] studied tone recognition in Chinese using a HMM-based approach and a Gaussian Mixture Model (GMM) based approach with Fractionalized model of F_0 values and their derivative. The best result of their experiments obtained from a GMM-based with Fractionalized model of F_0 values and their derivative which was estimated from only 4 subsections of tone segment. As a result, they concluded that to identify Chinese tones, the approximate some parts of tone contour are sufficient enough. Dong et al. [19] conducted a study to find the best configuration for Mandarin tone recognition system. Their study is interested in 3 configurations: 1) F_0 detection, 2) F_0 Extraction and 3) Classifier. In the first configuration, three methods: autocorrelation method, cross-correlation method and cepstrum method, for detecting F_0 were compared. In the second configuration, they inspected the feature extraction consisting of the feature selection and the feature normalization. In the last configuration, they examined classifiers, between Multi-Layer Perceptron (MLP) based classifier and SVM-based classifier, that which one is appropriate to classify Chinese tone. The results showed that autocorrelation method outperformed the other methods and z-normalization techniques yield better performance than other normalization methods as well as SVM with RBF kernel is the best classifier.

In Vietnamese tone recognition; Nguyen et al. [20] proposed a system for identifying Vietnamese tones. A HMM-based approach with F_0 values and their derivative and energy of speech signal is used to construct a Vietnamese tone recognition system. Moreover, normalization techniques also were taken into account carefully in order to reduce variation of speakers. The accuracy of system was 70.44%. In case that the gender of speakers was known, the accuracy of system increased up to 72.83%. Besides, Nguyen et al. [5] continued their study in [20]. The Vietnamese tone models from study in [20] are applied to a Vietnamese speech recognition system. As a result, the performance of Vietnamese speech recognition system is significantly improved.

In Thai, there are many approaches [6, 7, 8, 9, 10] regarding Thai tone classification. Tungthangthum [10] introduced a tone recognition system in Thai. His investigation is to study an effect of vowels to Thai tones. In his experiment, F_0 values were extracted from vowel interval by an autocorrelation method and then were used to optimize parameters of tone recognition system based on a HMM-based approach. The best result of his experiment was accuracy of 91%. Consequently, he concluded that tonal sound and vowel sound are independent and a tone in one's vowel can utilize to identify tone in another one. Nevertheless, his study was conducted on one male speaker and only some vowels. After that, Thubthong et al. [6] continued to deeply investigate in Thai tone recognition from Tungthangthum [10]. Thubthong's study is to examine Thai tone in isolated syllables and further explore an effect of vowels and consonants as well. They found that F_0 values in some vowels, especially between monophthong and diphthong, are slightly different and final consonants also hold critical information which evidently influences F_0 contour. Moreover, they introduced Thai tone recognition system using an ANN-based approach. They also proposed three feature sets for recognizing Thai tones. In the first and the second feature sets came from study in [10] and the third feature set consisted of F_0 values at initial and final position of syllable segment and delta F_0 (dF_0), which is the rate at which the F_0 values changes with time, at position 20%, 40%, 60% and 80% of syllabic duration were also proposed. In contrast with Tungthangthum [10], they conduct a study with speaker independence so normalization techniques should be considered in order to eliminate variation among speakers. They therefore used average value of F_0 values in each speaker to reduce variation on speakers. Their result showed that the third feature set give the best performance. Thubthong et al. continued their empirical study involving constructing Thai tone recognition system for continuous speech scenario in [7]. Three acoustic feature sets are proposed. In the first acoustic feature set, four coefficients of 3rd degree polynomial obtaining from fitting F_0 values through syllable segment are selected. The second acoustic feature set is similar to the third acoustic feature set in [6], which was reported as the best in their previous work. The final acoustic feature set is the combination of the first and the second feature set of their previous work in [6], which consisted of five F_0 values and five dF_0 values. By the way, the positions of syllabic duration are slightly different. The selected positions for the third acoustic feature set in study [7] are at 0%, 25%, 50%, 75% and 100% of syllabic durations. In this study, they also used an ANN-based approach which is similar to their previous study in [6]. Furthermore, they investigated effect of frequency scales and normalization techniques. Hertz, Semi-Tone and ERB-rate are chosen to study frequency scaling effect. Mean normalization technique and Z-score normalization technique are compared for normalizing F_0 values in order to reduce variation among speakers. The best result that they obtained was the third acoustic feature set with ERB-rate scale and Z-score normalization technique. However, Tan et al. [8] studied the tone classification as same as in Thubthong et al. [7]. They conducted the same study with different speech corpus but their finding and Thubthong are not consistence. The best result in Tan's experiment was the third acoustic feature set with Semi-tone scale and Mean normalization technique. Their system acquired accuracy of 72.21%. Besides, another finding is that final consonant information can improve a performance of tone classification system. When the final consonant information was given, they achieved accuracy of77.13%. Therefore, they concluded that although tone information relied in vowel part, consonant part still has some significant role in enhancing a Thai tone classification system. Their results also conform to Thubthong et al. [7] that final consonants hold crucial information about Thai tone. Recently, Maleerat et al. [9] conduct experiments on classifying isolated Thai tones by using a MLP-based approach. Their experiments are similar to Thubthong et al. [6] but there is a difference that they combine three feature sets which proposed in [6] into one acoustic feature set. Their system yield accuracy of 91.4%.

Based upon reviewed list above, we can roughly group methods to classify tones into 2 groups: non-sequential-based discriminative classifier and sequential-based generative classifier. In non-sequential-based discriminative classifier, many machine learning approaches such as ANN-based approach, approach, MLP-based approach and SVM-based approach were applied to a tone classification task. Nevertheless, because of the feature selection and feature extraction procedure, it is still complicated to construct an appropriate acoustic feature vector which can perfectly represent the speech data. The feature selection and feature extraction procedure for a non-sequential-based discriminative classifier are to transform the speech data into an acoustic feature vector by some criteria so many various criteria are proposed for extracting an acoustic feature vector from the speech data. However, this method had some drawback reported in [11] that the speech data, which is time series data, is represented by one acoustic feature vector. Therefore, it will not only cause some information lost if the criteria are not tightly defined for constructing the acoustic feature vector but also the selected acoustic feature vector may not reflect the true nature of speech data. For example, in the studies [6, 7, 8, 9], F_0 values and their derivative are used but positions of syllabic segment for extracting F_0 values and their derivative are different. Therefore, it led to a question which position in syllabic segment is better for extracting F_0 values and their derivative. Besides, adding other acoustic features are difficulty possible due to the same problem. In sequential-based generative classifier, the popular method is to use a HMM-based approach. Although the feature selection and feature extraction procedure for HMM-based approach are easier than non-sequential-based discriminative classifier since criteria using to select acoustic features are simplified by characteristic of model which can handle time series data well, there is still an assumption, in particular the independence assumption, that is not suitable for classification task. Based on mathematical model of HMM-based approach, The HMM-based approach is a generative model so they have a strong independence assumption, which assumed that observations in different times are independence as reported in [3, 12, 21]. In fact, the independence assumption is not correct since there is some overlapping between each observation as we can notice from overlapping time between speech frames. However, adding complicated acoustic features to an HMM-based approach is possible by concatenating all acoustic features selected into a long acoustic feature vectors. Therefore a HMM-based approach is easier to add other acoustic features than non-sequential-based discriminative classifier but it will violate the independence assumption caused by model since the actual correlation between each feature is unknown [21]. Note that the independence assumption will not occur in a discriminative classifier.

In this study, we therefore selected a sequential discriminative classifier which acquired both advantages of sequential classifier and discriminative model called Hidden Conditional Random Fields (HCRF) based approach. Although the HCRF-based approach can easily incorporate other acoustic features into feature vectors without causing a problem as report in [22], we limit acoustic tone features to F_0 values and their derivative in other to fairly evaluate the performance of Thai tone classification when comparing with a HMM-based approach and an ANN-based approach.

4. Hidden Conditional Random Fields

A hidden conditional random field (HCRF) model is a sequential discriminative model introduced by Gunawardana et al. [12], and usually use for classifying a class or label of sample sequence. The HCRF model is extended from Conditional Random Fields (CRF) model, which firstly were proposed by Lafferty [21], by incorporating hidden states into CRF model. Consequently, the HCRF model can capture intrinsic relations, which abrupt change in speech signal, by hidden states. There are many successful cases in applying the HCRF model to speech recognition, in particular a phoneme classification task [12] and a phoneme recognition task [3]. In both studies, the HCRF-based approach significantly outperforms any state of the art techniques. Comparing the HCRF model with the previous model for tone classification and tone recognition based on our reviews, we found that there are four advantages for applying a HCRF-based approach for tone classification. The first advantage is that the HCRF model is a sequential model supporting feature vector sequences. As a result, it is not worried about which positions of speech segment are appropriate for extracting acoustic features. In the next advantage, the HCRF model is a discriminative model, which modeled probability distributions by the conditional probability directly. Therefore, the model was relaxed the independence assumption among state sequences, which was usually occurred in generative models, in particular a HMM model. The third advantage of the HCRF model is that the model does not get stuck with the label bias problem as introduced by Lafferty [21]. The label bias problem is that the state, which has few outgoing transition, is likely to be selected to be the best choice without considering actual observation sequences. However, unlike other sequential discriminative models, especially Maximum Entropy Markov Model (MEMM) model, the HCRF model do not suffer from the label bias problem because the HCRF model modeled transition weights by single partition without taking every transition probability of states into account. The last advantage is the HCRF model was augmented by hidden states. Therefore, the HCRF model can capture intrinsic variation change of speech signal that the CRF-based model cannot capture.

An HCRF model [12] is defined as follows:

$$p(w \mid o; \lambda) = \frac{1}{z(o; \lambda)} \sum_{s \in w} \exp\{\lambda \cdot f(w, s, o)\} \tag{1}$$

where w is a class of input sequence and o represents an observation sequence of input sample, which corresponds to class w. s is hidden state sequences and λ represents a parameter of model. $f(w,s,o)$

denotes a feature function represented the relation of model and $z(o;\lambda)$ is normalization portion used to normalized probability as defined in Eq. (2).

$$z(o;\lambda) = \sum_{w,s\in w} \exp\{\lambda \cdot f(w,s,o)\} \tag{2}$$

Feature functions directly affected the performance of HCRF model because they defined the graphical structure of model. If the graphical structure of model is not suitable for the task, we cannot achieve the reasonable performance of HCRF model. Therefore, selecting feature functions are very important. In our study, feature functions, which proposed by Gunawardana [12], are chosen because the feature functions are successful to classify phoneme as report in [3, 22]. The feature functions comprised of transition function and state function. In Eq. (3) and Eq. (4), transition feature functions were defined. Eq. (5) to Eq. (7) denoted state feature functions.

$$f_w^{LM}(w,s,o) = \delta(w = w') \tag{3}$$

$$f_{ss'}^{Tr}(w,s,o) = \sum_{t=1}^{T} \delta(s_{t-1} = s)\delta(s_t = s') \tag{4}$$

$$f_s^{Occ}(w,s,o) = \sum_{t=1}^{T} \delta(s_t = s') \tag{5}$$

$$f_s^{M1}(w,s,o) = \sum_{t=1}^{T} \delta(s_t = s)o_t \tag{6}$$

$$f_s^{M2}(w,s,o) = \sum_{t=1}^{T} \delta(s_t = s)o_t^2 \tag{7}$$

where w denoted class of input sequence and o represented an observation sequence, which corresponded to class w. s_t was a state at time t in hidden state sequences s and o_t was an observation at frame t. $\delta(\cdot)$ was an indicator function, which was activated to equal 1 when the condition in the function was true. Otherwise, indicator function returned 0.

5. Speech Corpora

In our study, we conducted our experiments on two speech corpora, the isolated word speech corpus and the continuous speech corpus. Each corpus was used for studying a Thai tone classification in each condition. The detail of each corpus was described in the following sub-section.

5.1. Isolated Word Speech Corpus

To study Thai tone classification in isolated word scenario, the isolated word speech corpus was constructed. The isolated word speech corpus was collected from twelve Thai native speakers, six males and six females whose age 21 to 22 years old. Speakers were invited to provide their speech audio in the recording room. Speakers were asked to utter 22 words with 5 different Thai tones as shown in Table 1. Hence, the isolated word speech corpus has totally 1,320 tone samples and there were 264 samples for each tone. The Shure SM58 Unidirectional Dynamic Microphone was used to record all utterances. The sampling rate was set at 44.1 kHz with 16 bit PCM and all files were saved in the MS wav format. In order to obtain the high quality of audio, some criteria are assigned. If Signal to Noise Ratio (SNR) of audio was less than 30dB or the maximum amplitude of audio was clip, a speaker was asked to re-record until the conditions were completed. After the recording process, the speech data was phonetically segmented and transcribed carefully by hand.

Table 1. List of words in isolated speech corpus.

Words	Mid	Low	Falling	High	Rising
1 (อี)	zī:	zì:	zî:	zí:	zĭ:
2 (อู)	zū:	zù:	zû:	zú:	zŭ:
3 (เออ)	zɤ̄:	zɤ̀:	zɤ̂:	zɤ́:	zɤ̆:
4 (อา)	zā:	zà:	zâ:	zá:	ză:
5 (แอ)	zē:	zè:	zê:	zé:	zĕ:
6 (ดา)	dā:	dà:	dâ:	dá:	dă:
7 (คา)	k^{h}ā:	k^{h}à:	k^{h}â:	k^{h}á:	k^{h}ă:
8 (ที)	t^{h}ī:	t^{h}ì:	t^{h}î:	t^{h}í:	t^{h}ĭ:
9 (แม)	mē:	mè:	mê:	mé:	mĕ:
10 (โห)	hō:	hò:	hô:	hó:	hŏ:
11 (รอ)	rɔ̄:	rɔ̀:	rɔ̂:	rɔ́:	rɔ̆:
12 (เตีย)	tī:a	tì:a	tî:a	tí:a	tĭ:a
13 (เสือ)	zuū:a	zuù:a	zuû:a	zuú:a	zuŭ:a
14 (บอก)	bɔ̄:k	bɔ̀:k	bɔ̂:k	bɔ́:k	bɔ̆:k
15 (รูป)	rū:p	rù:p	rû:p	rú:p	rŭ:p
16 (ปาด)	pā:t	pà:t	pâ:t	pá:t	pă:t
17 (รีม)	rī:m	rì:m	rî:m	rí:m	rĭ:m
18 (แจง)	tɕē:η	tɕè:η	tɕê:η	tɕé:η	tɕĕ:η
19 (ยาว)	jā:w	jà:w	jâ:w	já:w	jă:w
20 (งาย)	ηā:j	ηà:j	ηâ:j	ηá:j	ηă:j
21 (กาน)	kā:n	kà:n	kâ:n	ká:n	kă:n
22 (แบน)	bē:n	bè:n	bê:n	bé:n	bĕ:n

5.2. Continuous Syllable Speech Corpus

The large vocabulary Thai continuous speech recognition corpus called LOTUS [23] was selected for investigating tone classification in the continuous speech scenario. LOTUS is the most popular Thai continuous speech for developing many Thai speech recognition applications [24, 25]. LOTUS corpus consists of four sets of which purpose are different. In the study, phonetically distribution (PD) set was chosen. The PD set generally used for primary training acoustic models because it covered most of sound unit patterns in Thai.

6. Data Preparation

There are three steps for data preparation as follows: 1) syllable segmentation, 2) F_0 extraction and 3) F_0 smoothing. The process of preparing data is shown in Fig. 3. The speech data were segmented into syllable unit and then F_0 values were extracted. Finally, to reduce the formant structure's effect, which strongly influences F_0 values, the F_0 smoothing process was applied. The detail of data preparation in each step is described in the following sub-sections.

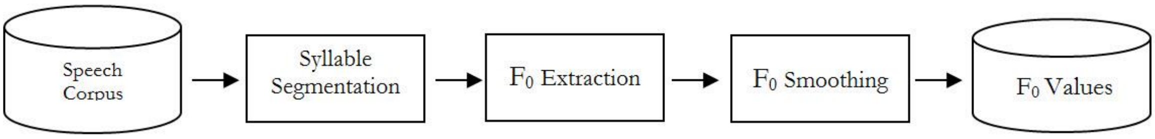


Fig. 3. The data preparation process for a tone classification experiment.

6.1. Syllable Segmentation

In a Thai tone classification, the syllable segmentation is necessary to obtain the actual boundaries of syllables so that the feature extraction completed without information lost. As we mentioned in the corpora section, the boundaries of syllables were provided by the tone transcription with time alignment.

For isolated speech corpus the tone transcription with time alignment was completed manually. The analysis speech tool, WaveSurfer [26], was used to visualize boundaries of speech segment. The boundaries of each syllable are determined and considered by the speech specialist. To make the two corpora, isolated word corpus and continuous speech corpus, consistence, the speech specialist also used the same criteria, which are similar to syllable segment criteria in LOTUS.

For continuous speech corpus, the PD set in LOTUS was chosen in the study because the phonetic transcription with the time alignment was available and they were completely done by expert linguists. This information is quite certainly correct. However, there are no tone labels in the phonetic transcription. In order to obtain tone labels, we therefore matched the phonetic transcription with the sentence transcription in LOTUS's database since the sentence transcription had already provided tone labels for the PD set. Unfortunately, there still are some problems because the sentence transcription is not exactly correct. Therefore, we re-checked the sentence transcription and corrected them before using in the data preparation. After the matching process, we got the tone labels for the PD set. The tone detail of PD set was reported in Table 2. Note that, the PD set which was recorded at NECTEC was chosen in experiment.

Table 2. The number of tones in the corpus counted from sentence transcription in LOTUS.

Tone	Syllable	Count
Mid	4,198	8,662
Low	2,953	5,277
High	2,373	4,966
Falling	2,080	4,202
Rising	1,098	2,150
Total	12,702	25,257

6.2. Fundamental Frequency Extraction

Most tone classification systems rely on F_0 and their derivative for distinguishing tones because F_0 values are closely related to tones. Therefore the F_0 Extraction is essential. In the study, F_0 values were extracted by the well-known phonetic analysis tool called Praat [27]. Studies in [19, 20] also utilized Praat as tool for extracting F_0 values. For extracting F_0 values, Praat uses a modified autocorrelation method [28], which is highly accurate.

Based upon our reviews, we found that Nguyen et al. [20] also investigated configurations of Praat for extracting F_0 values. Their analysis showed that "Pitch (ac)" function, which used autocorrelation method in [27], was better than other methods provided by Praat. Their finding is consistent with Dong [19]. Therefore, we choose "Pitch (ac)" function for extracting F_0 values and customized the parameters based upon their study in [19, 20, 28]. The configuration parameters of Praat in our study are listed in Table 3.

Table 3. The configuration parameters of Praat for extracting F_0 values.

Parameter	Value
Method	Pitch (ac)
Time step (second)	0.01
Pitch floor (Hz)	100.0
Pitch ceiling (Hz)	400.0
Very accurate	True
Silence Threshold (Pascal)	0.02

6.3. Fundamental Frequency Smoothing

Although the autocorrelation method in [28] highly accurate, there were still some errors due to the formant structure [8]. The F_0 Smoothing method provided by Praat was chosen for reducing the formant's structure effect. The F_0 Smoothing method is to convolute F_0 values with Gaussian curve in time domain. Before the convolution, the F_0 values in unvoiced parts were estimated by the linear interpolation in order to obtain all F0 values of syllable. The convolution then between F_0 values and Gaussian curve were computed. After the convolution, unvoiced parts were removed from F_0 values as they were. Note that there is a bandwidth parameter of Gaussian curve. The bandwidth parameter was set at 10 Hz as a default value by Praat.

7. Experiment

In the Experiment, we investigate a Thai tone classification based on a HCRF-based approach and then we compare a HCRF-based approach with two baselines, an ANN-based approach, which had been report as the best for Thai tone classification and a HMM-based approach, which had been claimed as the state of the art for ASR systems. We firstly introduced acoustic tone feature set in the following subsection. After that we presented frequency scaling for studying frequency scale's effect on Thai tone classification and then explained some normalization techniques for reducing variations among speakers. Before the experiment, we also described the experiment setting. Eventually, we reported and discussed the results.

7.1. Tone Features

In the Experiment, there are three approaches. One is our HCRF-based approach and other approaches, an ANN-based approach proposed by Thubthong [7] and a HMM-based approach proposed by Tungthangthum [10], are baselines in the experiment. Firstly, we presented tone feature sets for two baselines. Begin with the baseline based on an ANN-based approach, Thubthong's study [7] was selected because Thubthong's tone feature sets widely applied in many Thai tone classification system based on a non-sequential-based discriminative approach, especially an ANN-based approach and a SVM-based approach, [6, 8]. Furthermore, Thubthong's study seems to be the state of the art for a Thai tone classification. In Thubthong's study, three tone feature sets are proposed. In the first tone feature set referred as "PCR" set, 4 coefficients of 3rd degree polynomial (ax^3+bx^2+cx+d) are selected. In 3rd degree polynomial, F_0 values through syllable segment are used to compute 4 coefficients, a, b, c and d. Then 4 coefficients, a, b, c, and d are used to be tone features. The second set of Thubthong's tone features used F_0 values at initial position of syllable segment and final position of syllable segment and dF_0 values at 0%, 25%, 50%, 75% and 100% of the duration in voiced part of syllable. dF_0 values were computed by 3rd degree polynomial, which was fitted by F_0 values in syllable segment and then derivation of F_0 values was calculated at each time point. The second set was referred as "F2_dF5" set. The third tone feature set was quite similar to the F2_dF5. However, Thubthong augmented tone features in the third by also adding F_0 values at 25%, 50% and 75% of the duration in voiced part of syllable. F_0 values and dF_0 values computed by 3rd degree polynomial, was similar to the F2_dF5 set. We called the third tone feature set as "F5_dF5" set.

The other baseline for a Thai tone classification is a HMM-based approach introduced by Tungthangthum [10]. Since a HMM-based approach is widely succeeded for constructing many ASR systems, the HMM-based approach is selected to be another baseline. In addition, the HMM-based approach is also used for a tone classification in other tonal languages, in particular Mandarin [20] and Vietnamese [18]. As we mentioned in the literature reviews section, a HMM-based approach is a sequential model, in which input is sequence of feature vectors with time changing. In Tungthangthum's study [10], therefore, tone features based on sequence of F_0 value are proposed. We noticed that Tungthangthum's tone features are limited to F_0 values without considering their derivative. Based upon our reviewed about tone classification in other tonal languages [16, 20], we found that most study generally augmented F_0 values' derivative, especially dF_0 values, into tone features. Furthermore, Thubthong's study [6] showed that dF_0 values evidently improve a Thai tone classification performance. We therefore enhanced tone features from Tungthangthum's tone features by adding sequence of dF_0 value. dF_0 values are computed by Eq. (8). In Eq. (8), d_t is dF_0 value at time t and θ represents the internal distance between two F_0 values and f_t is F_0 value at time t and T denots a length of speech data. We referred this tone feature set consisting of F_0 value sequence and dF_0 value sequence as "F_dF" set.

$$d_t = \begin{cases} \frac{f_{t+\theta} - f_{t-\theta}}{2\theta}, & \theta < t < T - \theta \\ f_{t+1} - f_t, & t < \theta \\ f_t - f_{t-1}, & t \geq T - \theta \end{cases} \tag{8}$$

Based on our reviewed list, we also found that the studies in Vietnamese tone classification [5, 20] usually used acceleration of F_0 values (aF_0) other than F_0 values and dF_0 values. We therefore construct tone features set, "F_dF_aF" set, by appending sequence of aF_0 value into F_dF set. aF_0 values are calculated by Eq. (9) where $a_t$is aF_0 value at time *t* and f_t denotes F_0 value at time *t*. θ represents the internal distance between two F_0 values and T is a length of speech data.

$$a_t = \begin{cases} \frac{d_{t+\theta} - d_{t-\theta}}{2\theta}, & \theta < t < T - \theta \\ d_{t+1} - d_t, & t < \theta \\ d_t - d_{t-1}, & t \geq T - \theta \end{cases} \tag{9}$$

For our HCRF-based approach, we aimed to utilize tone feature sets, which are similar to a HMM-based approach. Since a HCRF-based approach and a HMM-based approach are a sequential model, the way to represent feature vectors is similar. Although a HCRF-based approach is flexible for adding other acoustic features as we described in HCRF section, we still limited acoustic features to F_0 values and their derivative in order to fairly evaluate between a HCRF-based approach and two baselines in the Experiment. In Table 4, the detail of tone feature sets for the Experiment is summary

Table 4. The detail of tone feature sets for the Experiment.

Tone Feature set	Feature
PRC	a, b, c, d
F2_dF5	F_0, F_{100} dF_0, dF_{25}, dF_{50}, dF_{75}, dF_{100}
F5_dF5	F_0, F_{25}, F_{50}, F_{75}, F_{100} dF_0, dF_{25}, dF_{50}, dF_{75}, dF_{100}
F_dF	Sequence of F_0 values Sequence of dF_0 values
F_dF_aF	Sequence of F_0 values Sequence of dF_0 values Sequence of aF_0 values

7.2. Frequency Scaling

For Thai tone classification, Thubthong [7] and Tan [8] showed that changing frequency scaling from Hertz (Hz) scale to Semi-Tone scale , which is a musical scale, or Equivalent rectangular Bandwidth rate (ERB-rate) scale, is a psychoacoustic scale, can improve a performance of Thai tone classification system. Although Thubthong and Tan conducted the experiment with similar setting, excepting speech corpus, on changing frequency scaling for Thai tone classification, unfortunately, Thubthong's results and Tan's results were not consistent. Thubthong obtained the best accuracy by changing frequency scaling from Hz scale to ERB-rate scale while Tan acquired the highest performance by changing frequency scaling from Hz scale to Semi-Tone scale. We therefore conducted the experiment to confirm and investigate that which frequency scale is suitable for Thai tone classification using a HCRF-based approach. The frequency scale in Hz scale is converted into Semi-tone scale and ERB-rate scale by Eq. (10) and Eq. (11) respectively.

$$S(f) = 69 + 12\log_2 \left| \frac{f}{440} \right| \tag{10}$$

$$ERB(f) = 11.17 \ln \left| \frac{f + 312}{f + 14675} \right| + 43.0 \tag{11}$$

where f is a frequency in Hz scale, $S(f)$ is a function to convert a frequency in Hz scale to Semi-Tone scale and $ERB(f)$ function is to convert a frequency in Hz scale to ERB-rate scale.

7.3. Normalization Technique

One problem that challenged in a tone classification problem is speaker variations since F_0 values generally relied on speakers. Usually dynamic range of F_0 values, which was produced by each speaker, is different and the difference is significantly increased especially in case of across gender. Generally, dynamic range of F_0 values in male is around 90-180 Hz whereas dynamic range of F_0 values in female is approximately 150-240 Hz that is higher than dynamic range of F_0 values in male. Consequently, if we do not reduce variations among speakers, in particular between male and female, we will suffer from speaker variations and cannot get a reasonable performance for Thai tone classification. Therefore, normalization techniques were taken into consideration in order to reduce the individual dynamic range of F_0 values among speakers. Thubthong [7] and Tan [8] were investigated and studied two normalization techniques, Z-score normalization and Mean normalization. The Z-score normalization technique normalized F_0 values of each speaker by subtracting mean of F_0 values and then divided by standard deviation of F_0 values while the Mean normalization technique used only mean of F_0 values for normalizing variation of speaker. Note that, mean of F_0 values and standard deviation of F_0 values are computed by each utterance. Again, Thubthong's result [7] and Tan's result [8] are inconsistent. We therefore conduct the experiment with two normalization techniques in order to find the appropriate normalization technique for an ANN-based approach.

7.4. Experiment Setting

In the Experiment, there are three approaches. Two of them are baseline approaches and another one is our HCRF-based approach. The first baseline, which is an ANN-based approach, was proposed by Thubtong [7]. Three tone feature sets, PRC set, F2_dF5 set and F5_dF5 set, with two normalization techniques, z-score normalization technique and mean normalization technique, and three frequency scales, Hz, Semi-Tone and ERB-rate, are involved in constructing tone models for ANN-based approach. Based upon our review, we found that there are some inconsistence of results between Thubtong's study [7] and Tan's study [8]. In the studies, Thubtong found that tone features normalized by z-score normalization technique and changed frequency scale to ERB-rate gave the highest result while Tan concluded that tone features normalized by mean normalization technique and changed frequency scale to Semi-Tone yielded the best result for Thai tone classification. Therefore, in our experiment we also figured out that which normalization techniques and frequency scales are appropriate for the first baseline. In the first baseline, the configuration of ANN-based approach was set similar to Thubtong's study [7]. The ANN-based approach used a three-layer feed-forward neural network. The number of input nodes relied on the dimension of acoustic tone feature vectors. The number of hidden nodes and output nodes are fixed and set at 20 and 5 respectively. We train an ANN-based model by standard back-propagation, in which initial weights were randomly selected from value lied between -1.0 and 1.0. The Fast Artificial Neural Network (FANN) library [29] was selected to construct the model. Note that in the experiment, the PRC set does not apply normalization techniques since the PRC set is designed to reflect F_0's contours. Therefore normalization techniques are not necessary because there is no variation of F_0's contours in Thai due to speakers as reported by [14].

The second baseline is a HMM-based approach introduced by Tungthangthum [10]. Considering Tungthangthum's tone features, we found that Tungthangthum's tone features limited to F_0 values without their derivative. Based on our review, we noticed that in other tonal languages, not only F_0 values but also their derivative usually were applied for tone classification and F_0 values with their derivative yielded the better result than F_0 values alone. We therefore improve tone features proposed Tungthangthum's by adding F_0 derivative. In the second baseline approach, tone feature sets, F_dF set and F_dF_aF set, are selected. We also applied three frequency scaling, which are similar to the first baseline, but only z-score normalization technique is chosen. Based on our reviews [5, 20], a HMM-based approach usually applied z-score normalization technique. Furthermore, there is no inconsistence like an ANN-based approach, in which Thubtong's study [7] and Tan's study [8] are not consistence. In addition based on the experiment, we conduct the experiment on an ANN-approach first and we found that z-score normalization gave the better result. Therefore, there is no need to consider mean normalization technique. Configurations of

HMM-based approach are as follows. The model of HMM-based approach is HMM model with five states from left-to-right. An acoustic observation generated from states of HMM model represented by Gaussian component with diagonal covariance matrices. Each HMM model is to represent each tone, which is a context-independent phoneme (monophone). Building tone models are done by the Hidden Markov Model Toolkit (HTK) [30].

In our approach, tone models based on HCRF-based approach are trained by two acoustic tone feature sets, F_dF set and F_dF_aF set. As we mentioned in HMM-based approach, the preliminary experiment show that z-score normalization technique yields the best accuracy result. In addition, a HCRF-based approach generally applied only z-score normalization technique as reported in studies [3, 12, 22]. Therefore, only z-score normalization technique is taken into account. However, three frequency scaling, which are similar to baselines, still are applied in order to find appropriate setting for a HCRF-based approach. The HCRF Library [31] is selected in order to construct HCRF-based tone models. In the HCRF library, the Gaussian Hidden Conditional Random Field (GHCRF) toolbox implemented from Gunawardana's study [12] is chosen. We configured parameter of GHCRF toolbox by the following configuration. Considering Thai tone contours as shown in Figs. 1 and 2, we found that tone contours can be roughly divided into three parts, early changing part, stable part and late changing part, due to the characteristic of F_0 value movements. The early changing part is approximately started from the beginning of syllable until the early middle of syllable. The stable part begins at the early middle of syllable and ends up at the late middle syllable. The late changing part is started from the late middle syllable and ends at the end of syllable. In the early changing part and the late changing part, F_0 values are rapidly changed while in the stable part, the changing of F_0 values is quite stable. Note that, the characteristic of F_0 value movements can obviously find especially in the falling tone and the rising tone. In order to support the characteristic of F_0 value movements, the number of hidden states in the GHCRF toolbox therefore is set at 3 states for representing three parts of Thai tones. Although a HCRF-based approach has capability to represent and model relations among input samples by expanding window sizes, we limit this capability of HCRF-based approach in order to fairly evaluate with a HMM-based approach since the HMM-based approach doesn't have ability to model input sample dependently because of the independence assumption. The window size for each input sample therefore was set at 0 so that neighbor sample inputs were not taken into account as similar to a HMM-based approach. Initial weights were computed from mean and variances of each acoustic feature and limited the maximum weight and the minimum weight to 1.0 and -1.0 respectively. The optimization method for training tone model is one of the Quasi-Newton methods, namely Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) with L2 cache.

In the experiment, two speech corpora, isolated word corpus and continuous speech corpus, are involved. K-fold cross validation approach is chosen for evaluation. For isolated word scenario, k value of K-fold cross validation approach was set at 6. Speech data in isolated word corpus was randomly divided into six groups, in which each group consist of one male and one female. For continuous speech scenario, we set k-value of K-fold cross validation at 4. According to 4-fold cross validation, the speech data was randomly divided into four groups. In each group, the speech data were randomly selected from speech data of six male and speech data of six female.

7.5. Results

Table 5 showed results of the experiment in isolated word scenario while Table 6 listed all classification accuracies of the experiment in continuous speech scenario.

DOI:10.4186/ej.2014.18.3.99

Table 5. Accuracies of tone classification results in isolated speech scenario.

Approach	Tone Feature set	Normalization Technique	Accuracy (%)		
			Frequency Scale		
			Hertz	Semi-tone	ERB-rate
ANN-based	PCR	-	69.77	65.91	67.95
	F2_dF5		57.65	77.27	80.68
	F5_dF5		63.11	81.14	81.29
	F2_dF5	Mean	91.67	91.14	92.12
	F5_dF5		92.88	92.12	93.11
	F2_dF5	Z-score	91.74	91.89	92.27
	F5_dF5		92.95	93.11	93.48
HMM-based	F_dF	Z-score	85.45	85.45	85.91
	F_dF_aF		86.21	86.36	86.29
HCRF-based	F_dF	Z-score	92.88	93.03	93.48
	F_dF_aF		93.33	94.09	**94.24**

Table 6. Accuracies of tone classification results in continuous speech scenario.

Approach	Tone Feature set	Normalization Technique	Accuracy (%)		
			Frequency Scale		
			Hertz	Semi-tone	ERB-rate
ANN-based	PCR	-	47.32	49.17	47.40
	F2_dF5		46.15	55.66	58.09
	F5_dF5		51.71	54.86	58.24
	F2_dF5	Mean	64.95	64.11	64.99
	F5_dF5		65.58	65.26	66.69
	F2_dF5	Z-score	65.63	66.00	65.89
	F5_dF5		66.20	66.67	67.05
HMM-based	F_dF	Z-score	65.00	65.11	65.17
	F_dF_aF		65.31	65.58	65.67
HCRF-based	F_dF	Z-score	69.61	69.67	70.17
	F_dF_aF		70.32	70.62	**71.01**

According to experimental results in isolated word scenario, the results showed that the highest accuracy obtained from the HCRF-based approach with the feature F_dF_aF, z-score normalization and frequency scaling in ERB rate scale. The experimental results in continuous speech scenario are also consistent with the experimental results in isolated word scenario that the HCRF-based approach with the feature F_dF_aF, z-score normalization and frequency scaling in ERB rate scale outperformed other configurations.

Table 7 present the confusion matrix of the HCRF-based approach with the feature F_dF_aF, z-score normalization and frequency scaling in ERB rate scale in the isolated word scenario while Table 8 show the confusion matrix of HCRF-based approach with the feature F_dF_aF, z-score normalization and frequency scaling in ERB rate scale in the continuous speech scenario. Both confusion matrixes are obtained from the best configuration for isolated word scenario and continuous speech scenario as reported in Table 5. and Table 6. Since other confusion matrixes of other configurations in each scenario have similar trend as confusion matrix of best configuration in each scenario, we therefore reported only confusion matrix of the best configuration in both isolated word scenario and continuous speech scenario. For the confusion matrix of isolated word scenario, we can notice that the low tone frequently confused with the mid tone and rising tone while considering confusion matrix of continuous speech scenario, we found that accuracy results of high tone and rising tone are low when comparing with other tones.

Table 7. The confusion matrix of the best results in isolated word scenario.

Reference	Classify results					Accuracies (%)
	Mid	Low	Falling	High	Rising	
Mid	246	11	2	4	1	93.18
Low	13	240	1	0	10	90.91
Falling	3	0	259	2	0	98.11
High	0	0	1	258	5	97.73
Rising	0	19	1	3	241	91.29
Total						94.24

Table 8. The confusion matrix of the best results in continuous speech scenario.

Reference	Classify results					Accuracies (%)
	Mid	Low	Falling	High	Rising	
Mid	6557	1075	449	438	143	75.70
Low	825	3889	158	212	193	73.70
Falling	548	177	3755	473	13	75.61
High	751	389	386	2519	157	59.95
Rising	224	471	15	225	1215	56.51
Total						71.01

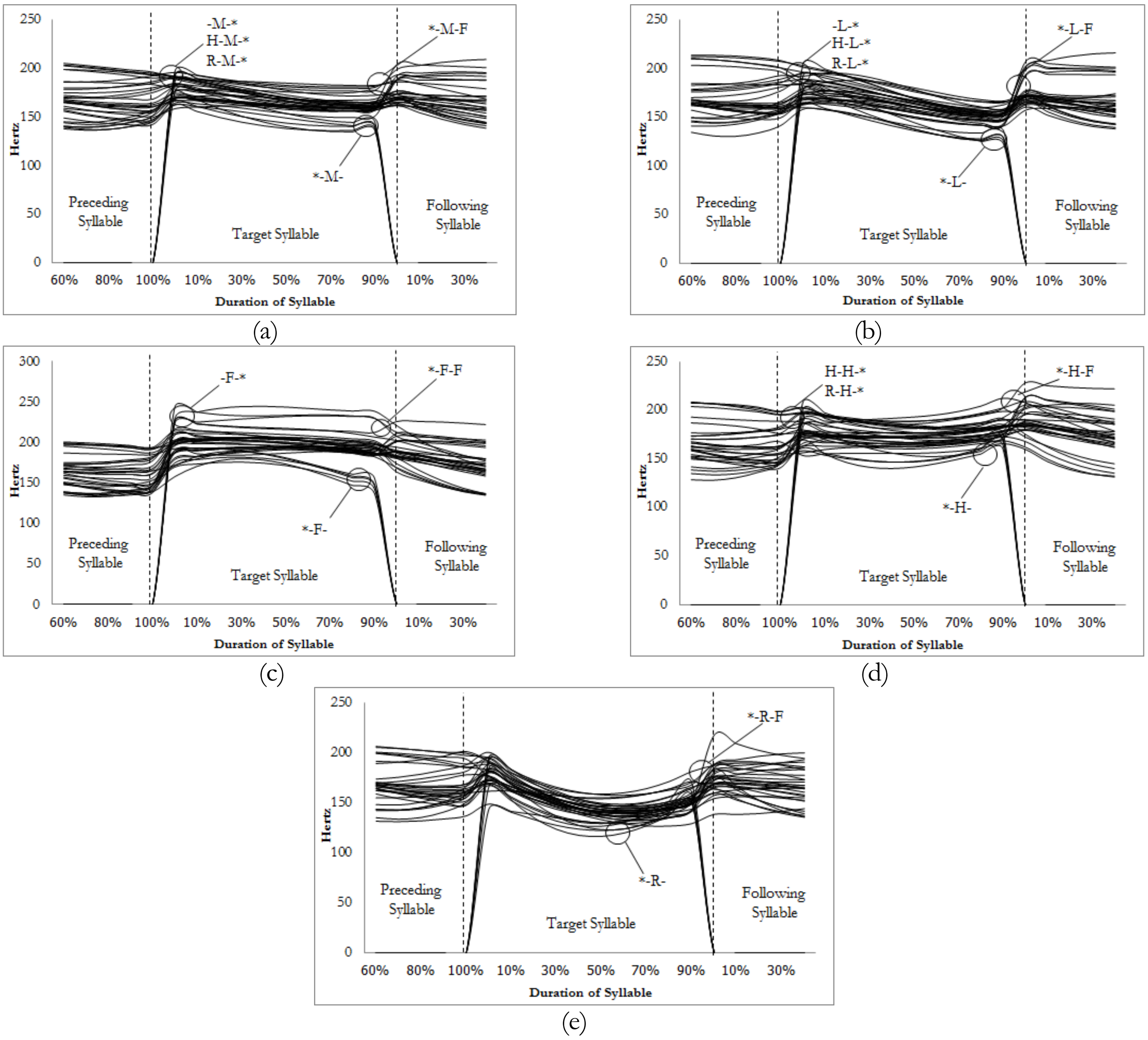
8. Discussions

In this section, we discussed experiments and results on how and why the results are like what we reported in the Experiment section. Since there are five aspects involving in experiments, we therefore discuss and analyze the results relied on those aspects. Five aspects are as follows: Effects of Contextual Variation, Classifier Approach, Frequency Scaling, Normalization Techniques and Tone Features. Details of each aspect are given in the following subsection.

8.1. Effect of Contextual Variation

According to results in Table 5 and Table 6, we found that the accuracy results between isolated word scenario and continuous speech scenario are quite different. One of important factors that affect accuracy results between isolated word scenario and continuous speech scenario is an effect of contextual variation. An effect of contextual variation occurred due to preceding and following tones. As we mentioned in the background knowledge section, both the preceding syllable and the following syllable influence F_0's contour of the target syllable. The effect of the preceding syllable called a carry-over co-articulation effect while the effect of the following syllable called an anticipatory co-articulation effect. Generally, the carry-over co-articulation effect influence F_0's contour of target syllable in starting period but the anticipatory co-articulation effect affect F_0's contour of target syllable in ending period.

In order to show effect of contextual variation, we therefore plot F_0's contours of target syllables with F_0's contours of preceding syllables and following syllables in continuous speech scenario as reported in Fig. 4. Each target tone consisted of all possible tone combinations of target tone with preceding tones and following tones. Consequently, there are 36 patterns for each target tone since tones in preceding syllables have 6 tone patterns, mid tone, low tone, falling tone, high tone, rising tone and non-preceding syllable and tones in following syllable also have 6 tone patterns as in preceding syllable. Note that in Fig. 4 mid tone, low tone, falling tone, high tone and rising tone were abbreviated as M, L, F, H and R respectively while * can represent any tone in five Thai tones.



(a)

(b)

(c)

(d)

(e)

Fig. 4. F_0's contours of target tone with F_0's contours of preceding syllable and following syllable on continuous speech: (a) mid tone, (b) low tone, (c) falling tone, (d) high tone and (e) rising tone.

Considering F_0's contours of targets Fig. 4, we obviously see that F_0's contours of preceding syllables and following syllables influence F_0's contours of target. In our experimental setting, we do not consider effects of contextual variation in continuous speech scenario caused by preceding syllable and following syllable since our study focused on Thai tone classification in case of considering on a single target segment. Based on this reason, accuracy results between isolated word scenario and continuous speech scenario are quite different.

Furthermore, we also found that an effect of contextual variation from preceding tones extremely affected beginning period of target tone while the effect gradually decreased over time of syllable. In case that the preceding tones are high tone and rising tone, their F_0's contours lift up in the ending period of preceding syllable. Consequently, F_0's contours of target tones, especially static tones, mid tone low tone and high tone, in beginning period, that preceding syllable is high tone or rising tone, have higher level than other patterns since F_0's contours of preceding syllable in ending period directly influence levels of F_0's contours of target tones in beginning period. Moreover in case of non-preceding syllable, we noticed that the F_0's contours of target tones in beginning period are in high level. Since generally the non-preceding syllable occurred in the head of sentence, capabilities of sound production are higher than other parts. F_0's contours of target syllable with non-preceding syllable in starting period therefore are quite high. The contrary effect occurred in case of non-following syllable. As we can see from Fig. 4, F_0's contour levels of target tones that followed by non-following syllable are lower than other patterns in ending period because capabilities of sound production gradually decreased along syllable. Since patterns that followed by non-following syllable are usually happen in the end of sentence, capabilities of sound production are lowest at that position. Consequently, F_0's contour levels of target tones that followed by non-following syllable are quite low in ending period. In addition, for ending period we noticed that F_0's contours of target tones that followed by falling tone lifted up in the ending period since they were influenced by the following tone. Since in the beginning of falling tone F_0's contours changed upward, F_0's contours of target tones needed to lift up ending period in order to support shape of the falling tone of the following syllable. Consequently F_0's contours of target tones in the ending period that followed by falling tone are higher than other patterns.

8.2. Classifier Approach

In the experiment, there are three classifier approaches: 1) an ANN-based approach, 2) a HMM-based approach and 3) a HCRF based approach. Considering accuracy results of ANN-based approach in Table 5 and Table 6, we found that accuracy results of ANN-based approach highly depended on input features. There are three tone feature sets, PCR, F2_dF5 and F5_dF5, involving in experiment based on ANN-based approach. We obtained those three tone feature sets by literature review [6, 7, 8] which had been reported that they are suitable for Thai tone classification. Each tone feature set tried to represent F_0's contours of target syllable by different methods. The feature PCR represented F_0's contours of target syllable by considering a whole contour while the feature F2_dF5 and F5_dF5 represent F_0's contours of target syllable by selecting significant position of F_0's contours. As we mentioned about drawback of non-sequential-based classifier approach like ANN-based approach, feature selection for non-sequential-based classifier approach is very complicating. According to accuracy results of ANN-based approach, we can notice that although the feature F2_dF5 and F5_dF5 aim to select significant position of F_0's contours, accuracy results of ANN-based approach with different tone feature are considerably different. Consequently, if we would like to use ANN-based approach for Thai tone classification, the feature selection needs to carefully take into account in order to make reasonable results.

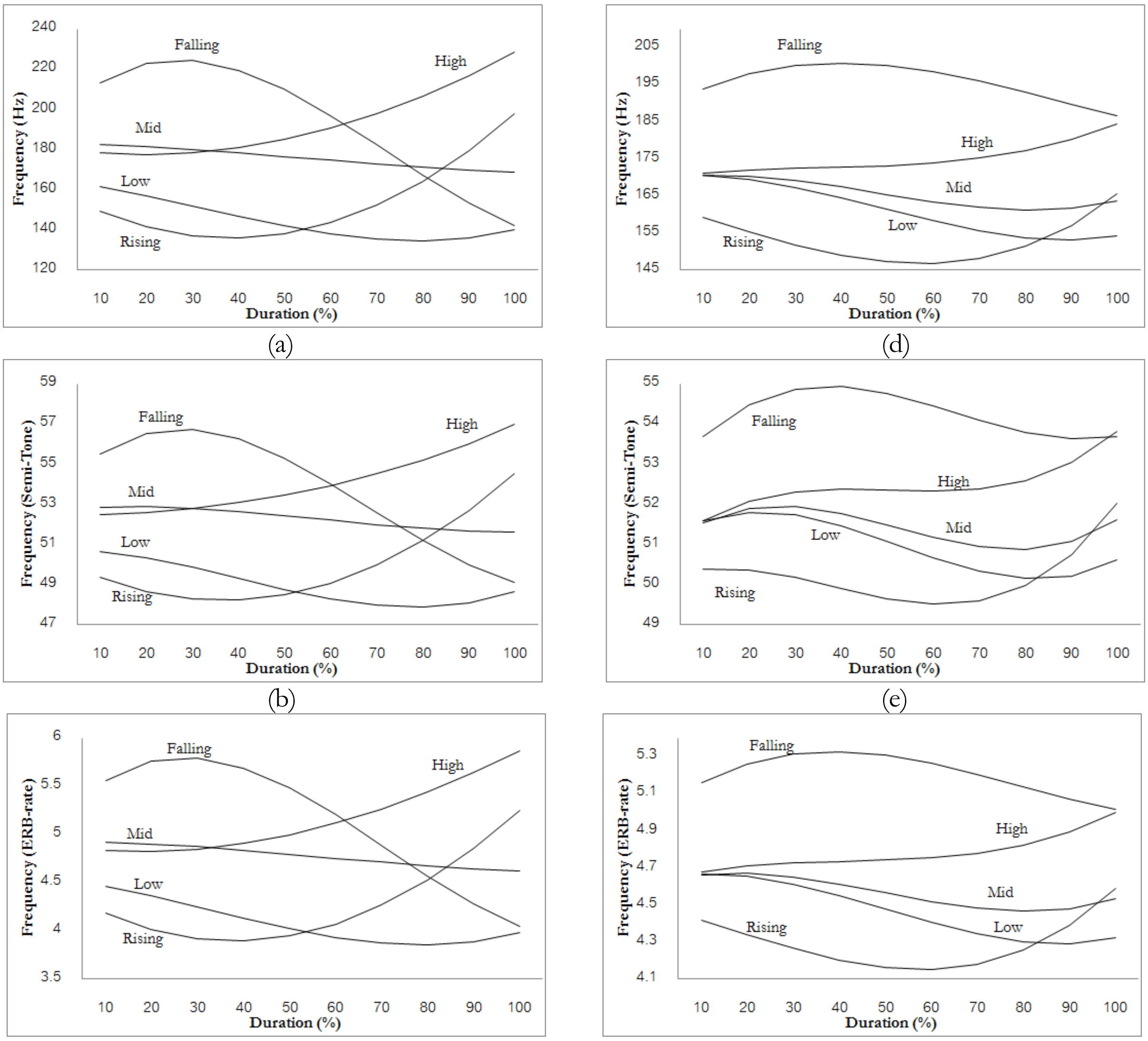
Comparing accuracy results of HMM-based approach with other approach, we can see that although a HMM-based approach gave reasonable performances, accuracy results of HMM-based approach were still lower than other approaches. The most important problem of a HMM-based approach is that a HMM-based approach assumed a strong independence assumption that observations are independent; however, the strong independence assumption is not true. According to fundamental frequency extraction, adjacent observation frames have overlapping region in order to keep high resolution in fundamental frequency analysis. According to our configurations, the window size of speech segments is 20 ms. with time shifting in 10 ms. for the next segments so it meant that there are 10ms of overlapping region between adjacent analyzed frames. Moreover to acquire the derivative of F_0's value for a considering observation frame, adjacent observation frames were used to compute the derivative of F_0's value via Eq. (8) - Eq. (9). Since the overlapping region between observation frames and the derivative of F_0's values were used, the independence assumptions assumed in HMM-based approach are not true in a tone classification task. This reason caused a HMM-based approach yielded lower performances than other approaches.

Accuracy results of HCRF-based approach in Table 5 and Table 6 give better results than other approaches. Since the HCRF-based approach used a sequential-based classifier, feature selection for the HCRF-based approach are less complicating than the ANN-based approach. Therefore, sequence of feature vectors can directly use to train the HCRF-based classifier. Furthermore, the HCRF-based approach also did not assume a strong independence assumption like the HMM-based approach. Consequently, we can use both the overlapping region between observation frames and the derivative of F_0's values without causing problem for the HCRF-based approach. Although the HCRF-based approach outperformed other approaches, there are still some drawbacks, especially time computation. Owing to our observation, we found that training time for the HCRF-based approach is more than other approaches, approximately five times. However, in classifying step, it did not consume observable time more than other approaches.

DOI:10.4186/ej.2014.18.3.99

8.3. Frequency Scaling

In the experimental setting, we also considered frequency scaling effect since there are some reports [7, 8] which show that changing frequency scale directly affect accuracy results. To investigate in details of frequency scaling effect, F_0's contours of each frequency scale are plotted in Fig. 5.



(a) (d)

(b) (e)

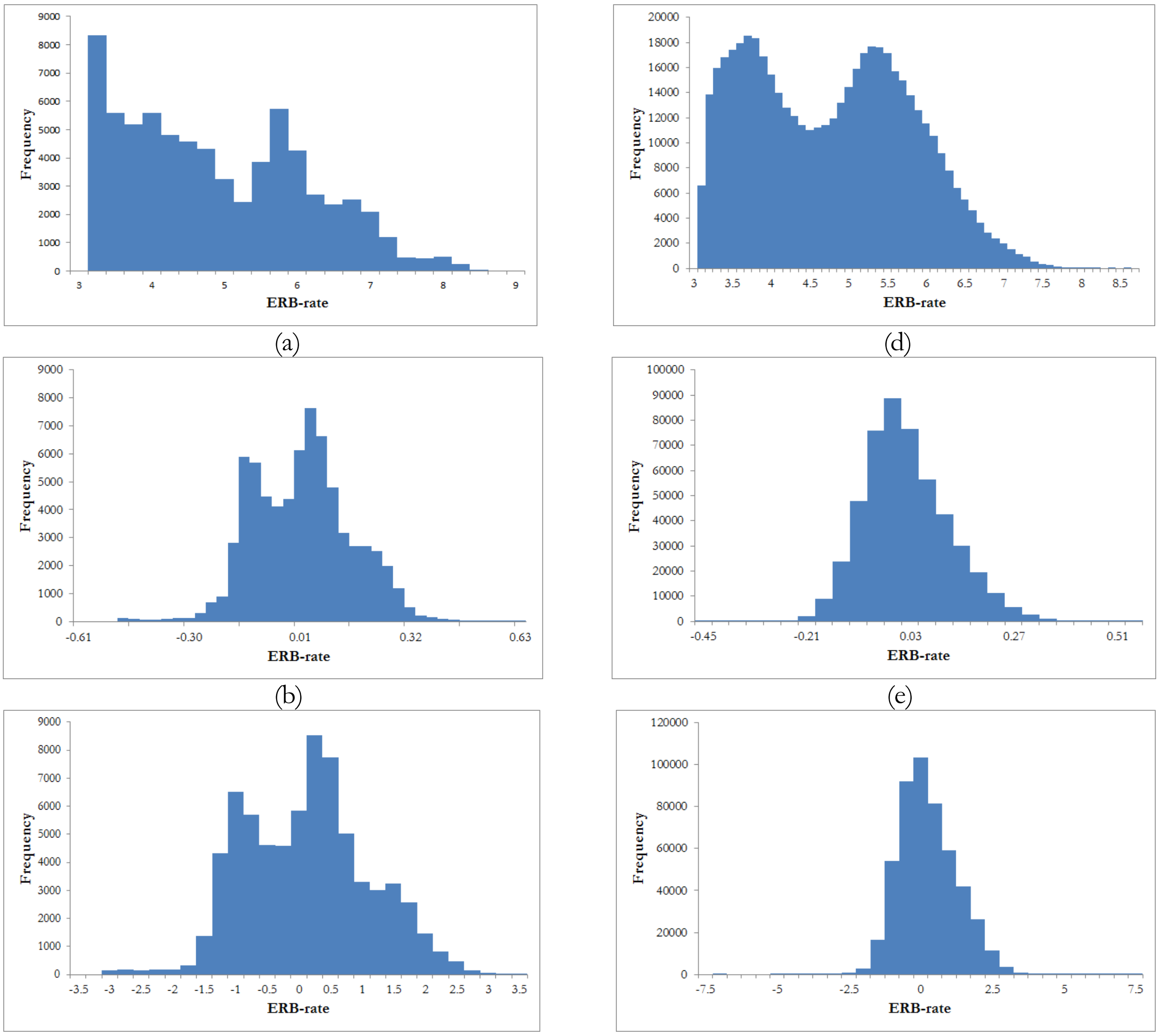
(c) (f)

Fig. 5. F_0's contours without normalization: on isolated word (a) in Hertz, (b) in Semi-Tone and (c) in ERB-rate and on continuous speech (d) in Hertz, (e) in Semi-tone and (f) in ERB-rate.

As shown in Fig. 5, we can see that F_0's contours of each frequency scale are quite similar to each other so this reason described why the results of ANN-based approach with PCR in difference frequency scales are similar. From the result in both scenarios, we also noticed that changing frequency scales, especially from Hertz to other frequency scales ,in particular ERB-rate, highly affect accuracy results and give better performances for ANN-based approaches with other tone feature sets, F2_dF5 and F2_dF5_aF5 than no changing frequency scales in case that normalization techniques are not applied. The results are consistent with reporting by Thubthong's study [7]. Therefore, changing frequency scales are worth for improvement Thai tone classification.

8.4. Normalization Techniques

In the experiment, one of important factors that affected accuracy results is a normalization technique. To understand in details of characteristic of each normalization technique, we presented histograms of F_0's values in both isolated word scenario and continuous speech scenario as shown in Fig. 6.



(a) (d)

(b) (e)

(c) (f)

Fig. 6. Histograms of F_0 values in ERB-rate: on isolated word (a) without normalization, (b) with mean normalization and (c) with z-score normalization and on continuous speech (d) without normalization, (e) with mean normalization and (f) with z-score normalization.

Based on accuracy results in Table 5 and Table 6, we also found that effects of changing frequency scales are become lower when normalization techniques, z-score normalization technique or mean normalization, are considered. Histograms of F_0's values in both scenarios showed that when normalization techniques are not applied, we can notice two peaks of histograms as shown in Fig. 6 (a) and (d) for isolated word scenario and continuous speech scenario respectively. Two peaks were caused by speaker variation especially due to gender so it led to difficulty for training acoustic model that can cover those variations. On the other hand, after z-score normalization technique or mean normalization are applied, according to Fig. 6 (b), (c), (e) and (f), we can see that two peaks are reduced to one peak. Consequently, results of ANN-based approach with normalization techniques highly improved significantly since both two of normalization techniques are effective for reducing speaker variation among speakers especially in case across gender. Note that only histograms of F_0's values in ERB-rate scale are shown in Fig. 6. Because histograms of F_0's values in other frequency scales also have similar trend as in histograms of F_0's values in ERB-rate scale, we therefore presented only histograms of F_0's values in ERB-rate scale.

8.5. Tone Features

To study effects of each tone features, we therefore conducted a statistical testing, paired t-test, so that we would understand that which tone features are appropriate for Thai tone classification. The t-test testing was conducted in each approach separately. We set the significant level for all testing at $\alpha = 0.05$. In the experimental setting, there are five tone feature sets: 1) PCR, 2) F2_dF5, 3) F2_dF5, 4) F_dF and 5) F_dF_aF. Each approach used different tone feature sets that are suitable for it.

For the ANN-based approach, three tone feature sets, PCR, F2_dF5 and F5_dF5 were selected. Testing significant differences between the feature PCR and the feature F2_dF5 and between the feature PCR and F5_dF5, we found that in case that normalization techniques were not applied there is no statistical significant difference between the feature PCR and the feature F2_dF5 and between the feature PCR and F5_dF5 while if normalization techniques are applied, there are statistical significant differences between the feature PCR and the feature F2_dF5 and between the feature PCR and F5_dF5. It meant that normalization techniques are necessary if F_0's values and their derivative are used. Based on this investigation, a HMM-based approach and a HCRF based approach take normalization into account by using z-score normalization. The z-score normalization technique is selected since the z-score normalization yield better result than mean normalization in an ANN-based approach. When conducting paired t-test between the F2_dF5 and F5_dF5, there is statistical difference between accuracy results of the feature F2_dF5 and F5_dF5 in isolated word while there is no significant difference at $\alpha = 0.05$ in continuous speech scenario. These results led to two issues. The first issue is that in isolated word scenario adding more F_0's values significantly improved the performance of an ANN-based approach while the second issue is that in continuous speech scenario adding more F_0's values did not enhance accuracy results significantly. The possible reason, that adding more F_0's values do not improve accuracy results significantly in continuous speech, is effect of contextual variation since F_0's values in case of continuous speech scenario are not stable due to influencing of preceding and following syllable as shown in Fig. 4. Consequently, adding more F_0's values do not help to enhance accuracy performances.

In the HMM-based approach, the paired t-test was conducted between the feature F_dF and the feature F_dF_aF. At the significant level at $\alpha = 0.05$, there are statistical significant difference between the feature F_dF and the feature F_dF_aF in both isolated word scenario and continuous speech scenario. This testing led to a conclusion that adding acceleration of F_0's values significantly improves performance of HMM-based approach. Furthermore, the HCRF-based approach also conducted the paired t-test between the feature F_dF and the feature F_dF_aF at the significant level at $\alpha = 0.05$. The results conform to the HMM-based approach that adding acceleration of F_0's values really help to improve performance significantly. To show the reason that why adding acceleration of F_0's values can improve performance of those approaches, we therefore reported F_0's of contours, delta of F_0's values of contours and acceleration of F_0's values of contours as shown in Fig. 7. As we can see from Fig. 7, contours of F_0's values, delta of F_0's value and acceleration of F_0's values can differentiate Thai tones since there are some different characteristics between each tone.

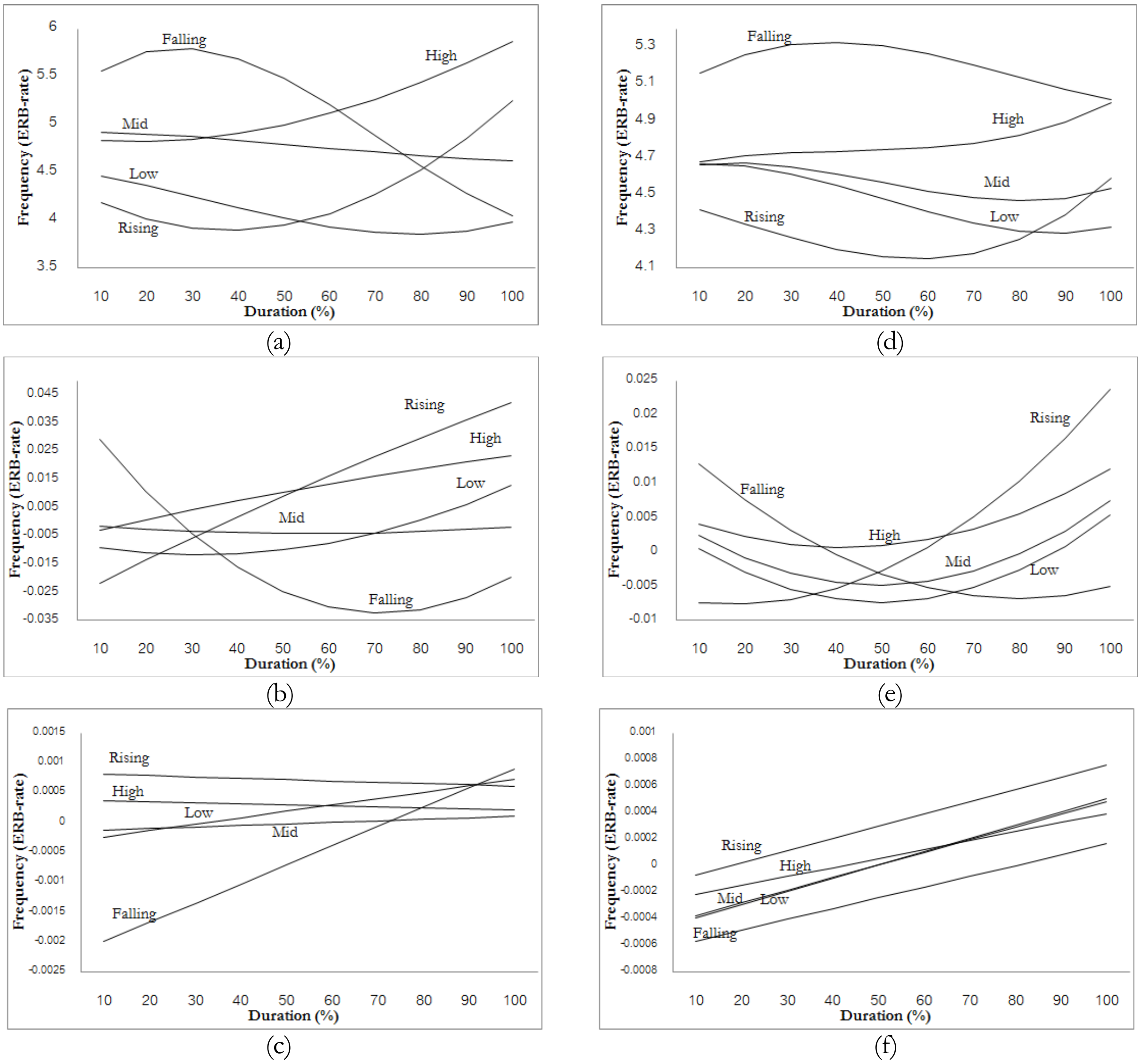


Fig. 7. F_0's contours and their derivative's contours: on isolated word (a) F_0 values, (b) delta of F_0 values (c) acceleration of F_0 values and on continuous speech (d) F_0 values, (e) delta of F_0 values and (f) acceleration of F_0 values.

Based on results and our discussions in many aspects for an ANN-based approach in both isolated word scenario and continuous speech scenario, the results and discussions led to three findings. The first finding is that F_0's contours alone are not sufficient enough to identify Thai tones correctly since the feature PCR gives quite low performances when comparing others. The first finding conforms to Tian's study [18] that F_0's contours lack ability to classify tones correctly in Mandarin. In the second finding, changing frequency scales are not necessary for tone feature set, PCR, while changing frequency scales are highly essential to improve Thai tone classification for tone feature sets, F2_dF5 and F2_dF5_aF5 in case that normalization techniques are not considered. The third finding is that necessities of changing frequency scales are alleviated if normalization techniques are applied. In addition, the best configuration for ANN-based approach obtained from tone feature set, F2_dF5_aF5, with z-score normalization and changing frequency scale to ERB-rate conforms to Thubthong's study [7] instead of Tan's study [8]. Therefore, we applied only z-score normalization for a HMM-based approach and a HCRF-based approach as we had already mentioned in experimental setting.

Results in both scenarios of HMM-based approach indicated that aF_0 can help to improve performances of HMM-based approach. Based on Thubthong's study [7], they had an assumption that a HMM-based approach is not good enough for Thai tone classification task. They therefore proposed an ANN-based approach for Thai tone classification. However, they still do not prove that assumption. In the experiment, we have already shown that an ANN-based approach outperform a HMM-based approach for Thai tone classification as Thubthong's assumption.

Considering our approach, HCRF-based approach, we can notice that the tone feature F_dF_aF gave better performances than the tone feature F_dF. After conducting paired t-test, we also found that there is statistical significant difference between the feature F_dF and the feature F_dF_aF in both isolated word scenario and continuous speech scenario at the significant level at $\alpha = 0.05$. We therefore can conclude that aF feature can enhance performances of HCRF-based approach significantly.

Furthermore, our results also indicat that a HCRF-based approach with F_dF_aF set in ERB-rate scale normalized by z-score normalization technique outperforms other baselines in both scenarios and yields the best result in the experiment.

9. Conclusion

This article has introduced a HCRF-based approach, which is a novel approach for Thai tone classification, with appropriate configurations for identifying five different Thai tones. In our study, the HCRF-based approach were reported that it give better results than other approaches, especially approach that had been reported as the best for Thai tone classification. Furthermore, this article presented empirical study of configurations of Thai tone classification in four aspects regarding classifier, frequency scaling, normalization technique and tone features. The study also further discuss about effect of contextual variation in continuous speech scenario. We compared a HCRF-based approach with an ANN-based approach and a HMM-based approach for Thai tone classification in both isolated word scenario and continuous speech scenario. Accuracy results in both scenarios indicated that a HCRF-based approach with the tone feature F_dF_aF, ERB-rate scaling and the z-score normalization technique outperforms other approaches. Moreover, in the experiments we found that necessities of changing frequency scales are alleviated if normalization techniques are applied. Besides, we also confirm Thubthong's configurations [7] that are suitable for an ANN-based approach in Thai tone classification which contrasted with Tan's finding [8]. We also found other findings that F_0's contours alone are not sufficient enough to identify Thai tones correctly. This finding is similar to Mandarin which reported by Tian's study [18].

For our future directions, we planned to study other acoustic features, which can improve performances of Thai tone classification since a HCRF-based approach had supported incorporating other acoustic features well. We also plan to applied Thai tone classification based on HCRF-based approach into Thai ASR systems in order to enhance performances for identifying Thai words that have many different meanings relied on tonal sound.

Acknowledgement

This research was sponsored by Graduate School Chulalongkorn University. Student fee and student's monthly allowance were provided by the Chulalongkorn University Graduate Scholarship to commemorate the 72nd Anniversary of His Majesty King Bhumibol Adulyadej.

References

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, pp. 257-286.

[2] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer, Speech, and Language*, vol. 17, no. 2-3, pp. 137–152, Apr-Jul 2003.

[3] Y.-H. Sung and D. Jurafsky, "Hidden conditional random fields for phone recognition," in *Proceedings of Automatic Speech Recognition Understanding Workshop*, Meran, Italy, 2009, pp. 107-112.

[4] H. Wei, X. Wang, H. Wu, D. Luo, and X. Wu, "Exploiting prosodic and lexical features for tone modelling in a conditional random field framework," in *Proceedings of Acoustics, Speech and Signal Processing*, Las Vegas, Nevada, USA, 2008, pp. 4549-4552.

[5] H. Q. Nguyen, P. Nocera, E. Castelli, and L. T. Van, "Using tone information for Vietnamese continuous speech recognition," in *Proceedings of Research, Innovation and Vision for the Future*, 2008, pp. 103-106.

[6] N. Thubthong, B. Kijsirikul, and A. Pusittrakul, "A method for isolated Thai tone recognition using combination of neural networks," *Computational Intelligence*, vol. 18, no. 3, pp. 313-335, 2002.

[7] N. Thubthong and B. Kijsirikul, "An empirical study for constructing Thai tone models," in *Proceedings of the 5th Symposium on Natural Language Processing and Oriental COCOSDA Workshop*, 2002, pp. 179–186.

[8] L. Tan, M. Karnjanadecha, and T. Khaorapapong, "A study of tone classification for continuous Thai speech recognition," in *Proceedings of 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, 2004, pp. 3033-3036.

[9] S. Maleerat, N. Supot, and H. Choochart, "Tone classification for isolated Thai words using multi-layer perceptron," in *Proceedings of the World Congress on Engineering and Computer Science*, 2009, pp. 1322-1325.

[10] A. Tungthangthum, "Tone recognition for Thai," in *Proceedings of IEEE Asia-Pacific Conf. Circuits and System*, 1998, pp. 157–160.

[11] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM Special Interest Group on Knowledge Discovery in Data Explorations*, vol. 12, no. 1, pp. 40-48, Jun, 2010.

[12] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification.", In Proceedings of 9th International Conference on Spoken Language Processing, Lisbon, Portugal, 2005, pp. 1117-1120.

[13] N. Kertkeidkachorn, S. Vorapatratorn, S. Tangruamsub, P. Punyabukkana, and A. Suchato, "Contribution of spectral shapes to tone perception," in *Proceedings of 13th Annual Conference of the International Speech Communication Association*, Portland Oregon, USA, 2012.

[14] S. Luksaneeyanawin, "Intonation in Thai," in *Intonation Systems a Survey of Twenty Languages*, Cambridge University Press, 1998, ch. 21, pp. 376–394.

[15] T. Lee, P. C. Ching, L. W. Chan, Y. H. Cheng, and B. Mark, "Tone recognition of isolated Cantonese syllables," *IEEE Transactions on Speech Audio Processing*, vol. 3, no. 3, pp. 204–209, May, 1995.

[16] F. H. L. Jian, "Classification of Taiwanese tones based on pitch and energy movement," in *Proceedings of International Conference on Spoken Language Processing*, 1998, pp. 329–332.

[17] L. Xu, W. Zhang, N. Zhou, C.Y. Lee, Y. Li, X. Chen, and X. Zhao, "Mandarin Chinese tone recognition with an artificial neural network," *Journal of Otology*, vol. 1, no. 1, pp. 30–34, Jan, 2006.

[18] Y. Tian, J.-L. Zhou, M. Chu, and E. Chang, "Tone recognition with fractionized models and outlined features," in *Proceedings of Acoustics, Speech, and Signal Processing*, 2004, pp. 105-108.

[19] J. Dong and C. Li, "A comparative study of the classification techniques in isolated Mandarin syllable tone recognition," in *Proceedings of the 49th Annual Southeast Regional Conference*, 2011, pp. 263-269.

[20] H. Q. Nguyen, P. Nocera, E. Castelli, and T. V. Loan, "Tone recognition of Vietnamese continuous speech using Hidden Markov Model," In *Proceedings of Communications and Electronics*, 2008, pp. 235–239.

[21] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labelling sequence data," In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 282–289.

[22] Y.-H. Sung, "Hidden Conditional Random Fields for Speech Recognition," Ph.D. Thesis, Stanford University, 2010.

[23] S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, S. Kanokphara, and N. Thatphithakkul, "Thai speech corpus for speech recognition," in *International Conference on Speech Databases and Assessments Oriental-COCOSDA*, 2003, pp. 54-61.

[24] C. Wutiwiwatchai and S. Furui, "Thai speech processing technology: A review," *Speech Communication*, vol. 49, no. 1, pp. 8-27, Jan, 2007.

[25] C. Wutiwiwatchai, K. Thangthai, and P. Sertsi, "Thai ASR development for network-based speech translation," presented at *International Conference on Speech Databases and Assessments Oriental-COCOSDA*, 2012.

[26] Speech, Music and Hearing part of School of Computer Science and Communication. WaveSurfer. [Online]. Available: http://www.speech.kth.se/wavesurfer/, [Accessed: 7 February 2012].

[27] P. Boersma, and D. Weenink, Praat 5. 3.11. (2011). A System for Doing Phonetics by Computer. [Online]. Available: http: http://www.fon.hum.uva.nl/praat/

[28] P. Boersma "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of 17 Institute of Phonetic Sciences*, University of Amsterdam, 1993, pp. 97-110.

[29] S. Nissen, Fast Artificial Neural Network Library (FANN). [Online]. Available: http://leenissen.dk/fann/wp, [Accessed: 7 February 2012].

[30] S. Young, G. Evermann, M. Galse, D. Kershaw, and G. Moore, Hidden Markov Model Toolkit–Speech Recognition Toolkit. [Online]. Available: http://htk.eng.cam.ac.uk, [Accessed: 7 February 2012].

[31] L. P. Morency, Hidden-state Conditional Random Field (HCRF) Library. [Online]. Available: http://sourceforge.net/projects/hcrf/, [Accessed: 7 February 2012].