Contribution of Spectral Shapes to Tone Perception

Natthawut Kertkeidkachorn, Surapol Vorapatratorn, Sirinart Tangruamsub, Proadpran Punyabukkana, Atiwong Suchato

Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand
Natthawut.K@student.chula.ac.th, Surapol.V@student.chula.ac.th, Sinart@gmail.com, Proadpran.P@Chula.ac.th, Atiwong.S@Chula.ac.th

Abstract

Tones in tonal languages are defined based on the characteristics of their fundamental frequencies. However, it has been shown that fundamental frequencies alone do not lead to good tone identification, either by human or machines. This paper reports a study based on tone perception experiments in which participants identified Thai tones from recorded mono-syllabic, bi-syllabic, and tri-syllabic stimuli as well as their modified counterparts. Stimuli were modified in various aspects so that effects of fundamental frequency, syllable energy envelope, and spectral shape to tone perception could be studied. Results indicate that spectral shape, especially in the vicinity of the first formant, contributes significantly to the participants' ability to correctly identify tones and the contribution is greater with more syllables in the stimuli.

Index Terms: tone perception, Thai tones, formant frequencies, fundamental frequencies, pitches

1. Introduction

In tonal languages, tone information is crucial for listeners to discriminate words of the languages. Therefore, it is unavoidable in most situations that developments of spoken language components such as automatic speech recognizers and text-to-speech systems targeting tonal languages must incorporate acoustic features related to tones in order to detect or reproduce them appropriately. Apart from such components, acoustic cues contributing to tone discrimination are also important to many computerized language education software packages.

Despite a long history of research works [1, 2, 3, 4] dedicated to finding acoustic cues related to tone perception in human, the only widely-adopted acoustic cues for discriminating tones are pitches or fundamental frequencies ($F_0$) of speech signals together with temporal changes in their values, which are quantities intrinsically used to define tones of various types. Consequently, most spoken language components, in which tones are explicitly modeled, only rely on fundamental frequencies ($F_0$) for tonal information. Despite some satisfying performances in detecting tones from speech signals, tone detections based on fundamental frequencies alone still need improvement, especially in the case of running speech. The fact that human can identify tones much more consistently than automatic procedures relying on fundamental frequencies motivates researchers to find acoustic cues that enhance tone discrimination.

In this work, we conducted a study on the perception of Thai tones by native speakers in order to conclude on evidences of possible additional acoustic cues that could contribute to tone discrimination. The paper is organized as follows. In section 2, we presented brief overview of Thai tones and their fundamental frequencies. Related literatures were reviewed and discussed in the following section. Then, we elaborated on the details of our study in section 4. Section 5 presented the results of the study. Finally, we concluded the work in Section 6.

2. Backgrounds on Thai Tones

There are five different tones in Thai. They are the mid tone, the low tone, the high tone, the rising tone, and the falling tone. By the tones’ linguistic definitions, Thai syllables with different tones are distinguished based on the values as well as the movements of their fundamental frequencies. The first three tones (mid, low, and high) are considered as static tones in which their fundamentals frequencies are relatively stable throughout the syllables, while the fundamental frequencies of the other two (falling and rising) tones are relatively dynamics. A syllable with the falling tone has its fundamental frequency changing in a downward direction while the fundamental frequency of one with the rising tone changes upward toward the end of the syllable. Figure 1 shows graphs of the fundamental frequencies of the five Thai tones versus normalized syllable duration, $F_0$ contours, as reported by Thubthong [5]. These graphs were generated based on Thai monosyllabic words uttered Thai native speakers. This result was widely cited and used for making research assumptions in many works on Thai tones [6].

![Figure 1: $F_0$ contours of the five Thai tones reproduced based on [5].](image)

Still, ones need to keep in mind that co-articulations from adjacent syllables can easily affect the $F_0$ contours of a syllable in continuous speech. Furthermore, value distributions (in Hertz) of fundamental frequencies are also dependent on the global intonation pattern of the sentences.
3. Literature Review

Generally, tones are defined linguistically based on the values of fundamental frequencies and their changes within syllable boundaries. Many research works regarding the study of tones in languages dedicated to investigate the fundamental frequencies. Abramson [7] and Phanintra [8] attempted to characterize the five Thai tones based on their Fo contours. Abramson publicized average Fo contours in various cases and pointed out their variations due to the co-articulations from preceding syllables. Phanintra contradicted the former work by suggesting that the Thai high tone resembles a dynamic tone more than a static one. There were a few works [9, 10, 11] that utilized features related to Fo contours in automatic recognitions of Thai tones but the recognition performances were still significantly inferior to the human performance, especially in continuous speech where the accuracy was below 80%. Jian [12] added the energy of each speech frame to the recognition feature vector for a Taiwanese tone recognition task and obtained some improvements. Li [4] conducted an experiment in which subjects listened to speech stimuli whose fundamental frequencies were modified so that they only varied linearly in their corresponding syllables. The result indicated that subjects were still able to identify the original tones quite correctly. This showed that, even when much information about the fundamental frequencies was discarded, there could still be some significant information about tones in the spectral shapes. Liu [13] conducted a tone perception test in Mandarin Chinese and have found that, in many cases, subjects could correctly identify tones even when the fundamental frequencies were completely removed. Lv [14] and Chen [15] showed that formant structures were crucial for identifying tones in whispered Mandarin Chinese. Furthermore, Punyayodhin [16] reported some levels of correlation between the fundamental and the first formant frequencies in Thai syllables. However, in this work, speech data were uttered by multilingual but non-native Thai speakers.

Based upon the review listed above, it was our motivation in this work to further investigate acoustic cues that potentially carried information about tones, especially the spectral shapes, through tone perception experiments.

4. Details of the Study

Experiments were conducted to uncover whether fundamental frequencies are sufficient for a human listener to identify Thai tones, whether temporal changes in energy envelope during the interval of a syllable contribute to the perception of the tone of that syllable, and how spectral shapes might affect tone perception.

4.1. Stimuli

4.1.1. The original stimulus set

Speech stimuli were recorded from a male Thai native speaker in a recording chamber using a Rode NT1-A condenser microphone placed at approximately 15 cm. from the speaker lips. Speech signals were recorded at 44.1 kHz with 16 bit PCM and saved in the MS wav format.

The speaker was asked to utter speech stimuli in three groups. Each syllable of the stimuli was the /l a:/ syllable (which is pronounced similarly to the English word “La”). Different stimuli were different in terms of the number of syllable in each stimulus and the tone associated with each syllable of the stimulus. In the first group, each stimuli was a mono-syllabic /l a:/ with a unique tone. Therefore, there were 5 stimuli in this group. In the second group, each stimuli was a bi-syllabic token /l a: l a:/ with a unique tone combination of the two syllables. Therefore, 25 stimuli belong in this group. The third group consisted of 125 tri-syllabic tokens /l a: l a: l a:/, each of which was with a unique tone combination of the three syllables. Each stimulus started and ended with an easily noticeable period of silence. We will refer to this set of recordings as the “original stimulus set”.

Figure 2 (a), (b), and (c) show the average $F_o$ contours of all five tones calculated from all syllables in the mono-syllabic group, the bi-syllabic group, and the tri-syllabic group, respectively.

Figure 2: $F_o$ contours of the five Thai tones from the recorded stimuli. The horizontal axes span from 0% to 100% of the syllable duration.

4.1.2. Modified stimulus sets

In order to investigate effects of various acoustic cues on the tone perception of Thai speech, we modified speech signals in the original stimuli set with various procedures in order to use the modified signal as stimuli in our tone perception experiments. The acoustic cues of interest included: 1) the $F_o$ contours, 2) the energy envelope of the signal spanning the syllables, and 3) the spectral components of the signal in ranges divided based on the signals’ formant structures.

Fundamental frequency tracks were extracted from the speech signals in the original stimuli set using Praat [17]. Then, the values of the fundamental frequencies on the extracted tracks were used for creating sinusoidal signals whose fundamental frequencies were governed by the tracks. The stimulus tokens
produced by this procedure were collected in a modified stimulus set called the Fo set. The amplitudes of the sinusoidal signals were set to a constant throughout the non-silent parts of the original stimuli. In any cases when the fundamental frequency of a speech frame was absent or was failed to be computed which could be due to irregular glottal activities, the values were computed using the \( F_0 \) contour smoothing techniques utilized in [9].

The amplitudes of the stimuli in the Fo set were then modulated with the envelope of their corresponding waveforms in the original stimuli set. This resulted in sinusoidal signals not only whose fundamental frequencies varied according to the ones of the original counterparts but also their amplitudes. The modified stimuli were collected in a set called \( \text{Fo+E} \).

In another modified stimulus set called \( \text{Lo}_F1 \), the signals in the original stimulus set were low-pass filtered with the cut-off frequency set at half of the average frequency of the first formants calculated among all stimulus tokens. This method roughly estimated the local minimum of spectrum between the fundamental frequency and the first formant. The cut-off frequency was approximately 450 Hz. Note that the variations of the formant structures were assumed to be small since the underlying phonemes of each syllable were the same and they were uttered by only one speaker, recorded in a single recording session.

The original signals were also low-pass filtered with another two different cut-off frequencies, including the midpoint between the first and the second formant frequencies, and the midpoint between the second and the third formant frequencies. The former was approximately at 1.2 kHz while the latter was approximately at 2.1 kHz. The modified stimulus sets associated with these two cases were called the \( \text{Lo}_F 1/\text{F}2 \) set and the \( \text{Lo}_F 1/\text{F}3 \) set.

4.2. Experiments

Two experiments involving perception tests were conducted.

4.2.1. Experiment #1

In the first experiment, fifteen Thai native speakers, 11 males and 4 females, without any hearing problems participated in a perception test. Each participant was given a test in order to verify their qualification in distinguishing the five Thai tones. They were asked to utter all five tones of the syllable \( /k\alpha l/ \) which was different from the base syllable used in the experiment and also to determine the tones of the syllables \( /k\alpha l/ \) uttered in person by the person conducting the experiment. All participants were qualified. The ages of the participants range from 21 to 27 years old with the average of 22.5 years old. They are all college students with the Computer Engineering major.

Each of the participants was presented with speech stimuli played back from a laptop computer via an OKER Dynamic Headphone OE-2688M.V. They were asked to identify the tones of all syllables of each of the stimuli presented to them by checking corresponding radio buttons located in the on-screen graphical interface. Participant can choose to listen to each stimulus as many times as they would like before the choices were made and they can proceeded to the next stimulus at their own pace.

Stimuli from the original stimulus set, the \( \text{Fo+E} \) set, and the \( \text{Fo} \) set were presented in random orders to the participants. The contents of speech stimuli in each set were the contents of all of the 5 stimuli from the mono-syllabic group, the contents of all of the 25 stimuli from the bi-syllabic group, and the contents of 25 stimuli selected from all of the 125 stimuli in the tri-syllabic group. Therefore, in this experiment, each participant listened to the total of 165 stimulus tokens.

4.2.2. Experiment #2

In the second experiment, eleven Thai native speakers, 9 males and 2 females, participated in a perception test. The screening of participants was the same as what was performed in the first experiment. Five of the male participants also participated in Experiment #1. The average age of the participants was 23.5 years old in this experiment. Also, the set up was similar to Experiment #1.

Stimuli from the original stimulus set, the \( \text{Lo}_F 1 \) set, the \( \text{Lo}_F 1/\text{F}2 \) set, and the \( \text{Lo}_F 2/\text{F}3 \) set were presented in random orders to the participants. The contents of the stimuli were obtained similarly to how they were in Experiment #1. Therefore, in this experiment, each participant listened to the total of 225 stimulus tokens.

4.3. Significant test

The McNemar’s tests [18] were used to test whether the abilities of participants to identify tones correctly under two different sets of stimuli were significantly different. The null hypothesis of each test was that the associated pair of stimulus sets did not contain any different cues for the perception of tones. The level of significance was set to \( \alpha = 0.05 \).

5. Results and Discussion

The results of the experiments were obtained from correct responses from the participants. Percentage of correctness implied number of times a participant was able to identify intended tone for a given stimulus. Shown in Table 1, Experiment #1 revealed highest percentage of correctness with mono-syllabic, and declined when more syllables were added. It also showed similar pattern for \( \text{Fo} \) and \( \text{Fo+E} \). The original stimulus set clearly showed highest correctness percentage. However, \( \text{Fo+E} \) and \( \text{Fo} \) results did not distinguish significantly, except for the mono-syllabic case. This fact could be interpreted that energy information failed to help recognizing the tone in the Thai language, which is contrary to the findings reported by Jian [12]. In addition, we observed large gap between the original stimulus set and \( \text{Fo} \), and therefore, we further implied that there must be other cues that contribute to identifying Thai tones.

<table>
<thead>
<tr>
<th>Table 1.</th>
<th>%Correctness from Experiment #1.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mono-syllabic</td>
</tr>
<tr>
<td>Original</td>
<td>100.0%</td>
</tr>
<tr>
<td>( \text{Fo+E} )</td>
<td>97.3%</td>
</tr>
<tr>
<td>( F_0 )</td>
<td>92.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2.</th>
<th>%Correctness from Experiment #2.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mono-syllabic</td>
</tr>
<tr>
<td>Original</td>
<td>100.0%</td>
</tr>
<tr>
<td>( \text{Lo}_F 2/\text{F}3 )</td>
<td>100.0%</td>
</tr>
<tr>
<td>( \text{Lo}_F 1/\text{F}2 )</td>
<td>100.0%</td>
</tr>
<tr>
<td>( \text{Lo}_F 1 )</td>
<td>94.0%</td>
</tr>
</tbody>
</table>
Table 2 presents the results from Experiment #2. Again, the correctness is the highest for mono-syllabic set, while it is the lowest for the tri-syllabic case McNemar’s tests indicated that Lo_F1 cases are significantly different from the other cases. We noted from Table 2 that Lo_F1 yielded the lowest correctness percentage which means that some cues, including the formant structure, were lost during the low-pass filtering process. This finding is consistent with what Punyayodhin reported in [16] in an aspect that the structure of formant influences the identification of Thai tones. Also, it is interesting to see that, on average, eliminating frequency components higher than the first formant even improved the tone perception.

Table 3. Confusion Matrix of the original set from Experiment #1.

<table>
<thead>
<tr>
<th></th>
<th>Md</th>
<th>Low</th>
<th>Falling</th>
<th>High</th>
<th>Rising</th>
<th>%Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Md</td>
<td>380</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>98.96%</td>
</tr>
<tr>
<td>Low</td>
<td>14</td>
<td>355</td>
<td>2</td>
<td>4</td>
<td>13</td>
<td>91.49%</td>
</tr>
<tr>
<td>Falling</td>
<td>3</td>
<td>10</td>
<td>373</td>
<td>7</td>
<td>2</td>
<td>94.43%</td>
</tr>
<tr>
<td>High</td>
<td>8</td>
<td>1</td>
<td>31</td>
<td>355</td>
<td>5</td>
<td>88.73%</td>
</tr>
<tr>
<td>Rising</td>
<td>5</td>
<td>30</td>
<td>2</td>
<td>3</td>
<td>343</td>
<td>89.56%</td>
</tr>
<tr>
<td>All tones</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>92.62%</td>
</tr>
</tbody>
</table>

Table 4. Confusion Matrix of the original set from Experiment #2.

<table>
<thead>
<tr>
<th></th>
<th>Md</th>
<th>Low</th>
<th>Falling</th>
<th>High</th>
<th>Rising</th>
<th>%Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Md</td>
<td>259</td>
<td>9</td>
<td>241</td>
<td>1</td>
<td>1</td>
<td>99.62%</td>
</tr>
<tr>
<td>Low</td>
<td>7</td>
<td>1</td>
<td>248</td>
<td>4</td>
<td>0</td>
<td>95.38%</td>
</tr>
<tr>
<td>Falling</td>
<td>4</td>
<td>1</td>
<td>12</td>
<td>242</td>
<td>1</td>
<td>93.08%</td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td>1</td>
<td>232</td>
<td>89.23%</td>
</tr>
<tr>
<td>Rising</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>94.00%</td>
</tr>
</tbody>
</table>

Table 3 and Table 4 show the confusion matrices of the perception results on the original sets from both of the experiments. We can see that the most confusion pair was between the low tone and the rising tone. Considering the F0 contours of the two tones from Figure 1 and Figure 2, we can observe that their contours present the most similarity among all combinations. Figure 3 shows F0 contours of the low tone and the rising tone extracted from the stimuli used in our study. Note that similar observations could be made from confusion matrices in the cases associated with the other five sets.

Figure 3: F0 contours of the low tones (Left) and the rising tones (Right)

Also, we could make an observation that both tones possess the lowest average values of fundamental frequencies. From an aspect that takes the spectral shape into consideration, it is worth investigating further about its effects on tone perception when combined with the high first formant frequency of a low vowel like /a/.

6. Conclusion

This paper has investigated acoustic cues believed to affect the tone perception of Thai native speakers. It was found from the experiments that the F0 contours alone were not sufficient for human listeners to identify tones while some spectral information within the first formant frequency played some significant roles. However, temporal changes in syllable energy, as shown to be related to the tone perception in some languages, did not exhibit any significant contribution in tone perception in Thai. The findings encourage further investigations on the roles of the overall spectral shapes of syllables in discriminating tones in a language.

7. References