A STUDY OF TONE CLASSIFICATION FOR CONTINUOUS
THAI SPEECH RECOGNITION

Li Tan, Montri Karnjanadecha, Thanate Khaorapapong

Department of Computer Engineering, Faculty of Engineering,
Prince of Songkla University, Songkhla, THAILAND 90110
litan212@hotmail.com, montri@coe.psu.ac.th, kthanate@coe.psu.ac.th

Abstract

This paper presents a study of tone classification for continuous Thai speech recognition. A modified auto-
correlation algorithm was implemented with pitch detection, and the tone classifier utilized 3-layer feed-forward neural
network with back-propagation. The best performance
configuration of tone features was obtained with semi-tone
scaling and mean-normalization producing a classification
accuracy of 72.21%. Also, after considering the effects of
final consonants, the average performance of the tone
classifier improved to 77.13%. Experimental results showed
that the pitch value of a tone with no final consonants has
more variation than one with final consonants. Also the
classification for tones with voiced final consonant gave
better performance than tones with unvoiced final consonants.

1. Introduction

Most speech recognition technologies have been developed
for English with some adapted to Thai. However, unlike
English, Thai is a tonal language that has five different tones:
mid, low, falling, high, rising, making the referential meaning
of a syllable dependent on the lexical tones. Therefore, a tone
classifier is an essential component of a successful speech
recognition system.

A. Tunthangthum concluded that the effects of vowels
on tones can be ignored in tone classification when using a
speech database of vowels with tones [1]. N. Thubthong and
B. Kijisirikul studied context variation of Thai tones, and a
half-model for tone classification was proposed [2]. N.
Thubthong, B. Kijisirikul and S. Luksaneeyanawin studied the
configuration of a tone classifier based on three different
speech corporuses [3]. They concluded that the best
configurations of tone features use ERB scaling and Z-score
normalization. The Thai speech corporuses are Potisuk-1999
containing 11 sentences, each consisting of four tones. The
second is the Thai proverb corpus containing 30 Thai
proverbs made up of 10 four-syllabic, 10 five-syllabic and 10
six-syllabic proverbs, collected from 40 native speakers,
total 1200 utterances. The Thai animal story corpus contains
four animal stories, consisting of 50 different sentences. The
data was collected from 20 native speakers, and total 1000
sentence utterances.

2. Database and Experiment Setting

2.1. Database

Our speech data is taken from a continuous Thai speech
corpus described consisting of 360 utterances [4]. The data
was collected from 18 native Thai speakers (9 male and 9
female speakers) with each speaker reading 20 utterances
randomly chosen from Thai words, and total 5726 tones. The
speech signals were recorded in a quiet environment and
digitized by a 16-bit A/D converter at 16 kHz. The speech
data was manually segmented and transcribed at phoneme
level.

2.2. Experiment Setting

The speech data was separated into a training set and a
testing set. The first 15 utterances of each speaker were
placed in the training set, and the last 5 utterances were the
testing set. The complete training set has 4506 tones, and the
testing set 1220 tones. Among 4506 training tones, there are
1569 mid tones, 1019 low tones, 861 falling tones, 655 high
tones and 402 rising tones. This reflects the uneven
distribution of tone in Thai speech. The mid-tone is the most
common and the rising-tone least frequent. The experiments
were performed using a three-layer feed-forward neural
network with Back-propagation as the training algorithm.
The number of inputs depended on the specific feature setting,
and the unit number of hidden layer depended on the training
process. There were five output units corresponding to the
five tones in Thai. All feature parameters were normalized to
lie between −1.0 and 1.0.

3. Experiments, Results and Discussions

3.1. Experimental Framework

The speech data were phonetically segmented and the label
information collected for all the tones. Based on the label
information and wave data, the pitch contour was extracted
from wave data using a modified auto-correlation method [5].
This contour may still contain errors due to the formant
structure of the speech, so the smoothing was applied using a
3-point median filter followed by nonlinear smoothing [7].
Then the smoothed pitch contour was processed with a
scaling, normalization technique in order to reduce the
interaction factors between the tones. The resulting contour
was fitted into 3rd-order polynomial curve using orthogonal
polynomial approximation [6] and the baseline tone features
taken from the 4 polynomial coefficients. The extracted tone
features were feed into a 3-layer feed-forward neural network
for classification. The tone classification model is shown in
Figure 1.

3.2. Configurations of Tone features

As mentioned earlier, Thubthong et al. concluded that the best
configuration for tone features are ERB scaling, Z-score
normalization [3]. We utilized the same framework, but
applied to a different speech database. When only the four
polynomial coefficients were used as tone features, the
performance was 60.9%. The performance improved to
66.15% when a 10-feature setting was employed, including the slopes and the levels of the five equal sections in the pitch contour. The improvement was due to the addition of more intrinsic tone information and the reduction of interaction factors. The same effect was observed by Thubtong et al. [3].

![Diagram](image1)

**Figure 1:** The tone Classification Model

The baseline using 10-feature setting, no scaling gave the performance 66.15%. The performance of semi-tone scaling, ERB scaling techniques are shown in Table 1.

**Table 1:** Classification results after scaling

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>10 features</th>
<th>Semi-tone</th>
<th>ERB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent</td>
<td>66.15% (807/1220)</td>
<td>66.89% (816/1220)</td>
<td>68.61% (837/1220)</td>
</tr>
</tbody>
</table>

Both semi-tone and ERB scaling improved the performance, and ERB scaling offers slightly better performance over semi-tone scaling. The effects of semi-tone scaling and ERB scaling on the distribution of pitch values are shown in Figure 2. The frequency axis is scaled, and the distribution of pitch values is modified. The less modification of ERB scaling to the distribution of pitch value can be noted that explained the higher performance given by ERB scaling.

The pitch values for ten different speakers (five males and five females) were collected to observe the speaker variation in tones. The histograms in Figure 3 show the speaker variation occurs in the pitch value with the female pitch value higher than that of the male. So speaker normalization is essential for a speaker-independent tone classifier.

![Histograms](image2)

**Figure 2:** Distribution of pitch values (ten speakers) before and after scaling

![Histograms](image3)

**Figure 3:** Histogram of pitch values for five males and five females

The effects of mean normalization and Z-score normalization on the pitch contours are shown in Figure 4.

![Histograms](image4)

**Figure 4:** Histogram of pitch value after normalization

Table 2 shows that the semi-tone scaling and mean-normalization gives the highest performance which does not agree with Thubtong et al.’s findings [3]. They concluded that ERB scaling and Z-score normalization gave the best performance. Based on the classification results and the histogram shown above, we conclude that the pitch data distribution does not absolutely follow a Gaussian distribution. However, Z-score normalization assumes the data does follow a Gaussian distribution, so modifies the
pitch information in such a way that the classifier gives the lower performance values.

Table 2: Classification accuracy of normalization

<table>
<thead>
<tr>
<th>Percent</th>
<th>Semi-tone</th>
<th>Semi+z-score</th>
<th>Semi+m-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>66.15%</td>
<td>69.84%</td>
<td>72.21%</td>
<td></td>
</tr>
<tr>
<td>68.61%</td>
<td>70.16%</td>
<td>72.05%</td>
<td></td>
</tr>
</tbody>
</table>

Consequently, the best tone feature configuration is semi-tone scaling with mean normalization giving a performance of 72.21%. The classification confusion-matrix for this configuration is shown in Table 3.

Table 3: Classification Confusion-matrix

<table>
<thead>
<tr>
<th>0(M)</th>
<th>1(L)</th>
<th>2(F)</th>
<th>3(H)</th>
<th>4(R)</th>
<th>Percent(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
<td>16</td>
<td>5</td>
<td>83.07</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>16</td>
<td>17</td>
<td>63.86</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>14</td>
<td>6</td>
<td>83.14</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>13</td>
<td>6</td>
<td>53.37</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>8</td>
<td>88</td>
<td>63.31</td>
<td></td>
</tr>
</tbody>
</table>

The confusion-matrix shows that the falling tone (F) provides the highest recognition rate, while the high tone (H) gives the poorest. Also, large confusions occur between the mid and low tones, the low and high tones, the low and rising tones, the falling and high tones. The typical pitch contour of Thai tones in Figure 5 shows that the similarity between the pitch contours for mid, low and high tones and the similarities between low and rising tones, and falling and high tones are the main reason for the confusion. Since the mid tone occurs much more often in speech compared with other tones, the classifier may be biased. Thus, the falling, high and rising tones may be also mis-classified as the middle tone.

3.3. Final Consonants and Tone Classification

Thubtong et al. concluded that he final consonants hold the critical tone informations [3], so their effects cannot be ignored in tone classification.

The final consonants in Thai include: plosive final consonants (p,t,k), nasal final consonants (m,n,ng), fricative final consonants (f,s), affricate final consonant (ch), approximant consonants (j,w), lateral approximant final consonant (l). Our speech database contains 5726 tones. There are 2084 with no final consonant, 1913 with nasal final consonants, 863 with plosive final consonants, 836 with approximant final consonants, 24 fricative consonants, and 4 with affricate consonant and lateral approximant consonant.

Based on the distribution and phonetics knowledge of the final consonant, the speech data can be placed into two groups: one group holds tones with no final consonant, with a plosive final consonant and other final consonants. The other group is for tones with a nasal final consonant and approximant final consonant, which is based on the voiced and unvoiced properties of the final consonants. Tone information mainly lies in the vowel part and exists in voiced sounds, the first group includes vowels with no final consonant and vowels with an unvoiced final consonant. The second group includes the vowels with a voiced final consonant. The classification results are shown in Table 5.

Table 4: Tone Classification Based on Voiced and Voiceless Final Consonants

<table>
<thead>
<tr>
<th>Group</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel &amp; Vowel+p,t,k,f,s,ch,l</td>
<td>69.6% (443/642)</td>
</tr>
<tr>
<td>Vowel+m,n,ng,j,w</td>
<td>82.01% (474/578)</td>
</tr>
<tr>
<td>Total</td>
<td>75.16% (917/1220)</td>
</tr>
</tbody>
</table>

About 70% of tone data lies in tones with no final consonants or a nasal final consonant. The data can be grouped in a different way, as three groups with different final consonants. The first group holds tones with no final consonant, the second is for tones with nasal consonants and the third for tones with a following final consonant based on the distribution information for final consonants. The classification results are shown in Table 5.

Table 5: Tone Classification Based on Different Main Final Consonants Group

<table>
<thead>
<tr>
<th>Group</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel</td>
<td>67.37% (286/429)</td>
</tr>
<tr>
<td>Vowel+m,n,ng</td>
<td>83.02% (313/377)</td>
</tr>
<tr>
<td>Vowel+p,t,k,j,w,f,s,ch,l</td>
<td>77.29% (320/414)</td>
</tr>
<tr>
<td>Total</td>
<td>75.57% (919/1220)</td>
</tr>
</tbody>
</table>

The last group of Table 5 can be divided into two groups: tones with a plosive final consonant and tones with a following final consonant which are approximant consonant. The classification results are shown in Table 6.

Tables 4, 5, 6 show that the best average classification performance is 77.13%. The performance for tones with no final consonant always gives the poorest performance comparing with those that have final consonants. Also, the final consonants help to stabilize the pitch contour. For the tones that have final consonants, different final consonants give us different performance. For example, tones with a

![Figure 5: Female’s Average F0 (fundamental frequency) Contour for Five Thai Tones (Tone types are indicated)
nasal final consonant and approximant final consonants perform as high as 83%, while tones with a plosive final consonant give 79%. Therefore, different final consonants have different effects on pitch contour. Also, the tone classification for tones with voiced final consonants should be emphasized over that with unvoiced final consonants.

### 4. Conclusions

We have studied the configurations of tone feature advantages for continuous Thai speech recognition. Tone features are affected by many interaction factors. So processing is necessary to reduce their effects. Techniques related to scaling and normalization were employed leading to a performance of 72.21% obtained from semi-tone scaling and mean normalization. ERB scaling is better than semi-tone scaling for tone classification, and mean normalization has advantages over Z-score normalization. The distribution of pitch values does not absolutely follow a Gaussian distribution, so Z-score normalization should not be utilized in continuous Thai tone classification.

The effects of different final consonants on tone classification are also considered. The final consonants for Thai can be placed into 3 groups: nasal final consonants, plosive final consonants and approximant final consonants. Nasal and approximant consonants are voiced, and plosive final consonants are unvoiced. Based on phonetic characteristics, classification was done on tones with voiced final consonants and those with unvoiced final consonants, leading to an improved performance of 75.16%. Based on the distribution of final consonants, classification was carried out on tones with no final consonants, tones with nasal final consonants, tones with plosive final consonants and tones with approximant consonants, improving the performance to 77.13%. Also, when comparing the classification results between groups, the performance for tones with no final consonant is lower than those with final consonants. Thus the final consonant is helpful for stabilizing the pitch contour. Tones with final consonant give different performance depending on the final consonant. The performance for tones that have voiced final consonants is higher than those with unvoiced final consonants. This shows that tone information mainly lies in vowels and the voiced parts of speech, but unvoiced parts still have some effects on tone classification.

### 5. Acknowledgements

The author would like to thank Dr. Andrew Davison for his review of this manuscript. This work is supported by National Electronics and Computer Technology Center (NECTEC), Thailand.

### 6. References


<table>
<thead>
<tr>
<th>Group</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel</td>
<td>67.37% (286/429)</td>
</tr>
<tr>
<td>Vowel+p,t,k</td>
<td>79.7% (161/202)</td>
</tr>
<tr>
<td>Vowel+j,w,f,s,gl</td>
<td>83.02% (313/377)</td>
</tr>
<tr>
<td>Total</td>
<td>77.13% (938/1220)</td>
</tr>
</tbody>
</table>

Table 6: Tone Classification Based on Further Grouping of Final Consonants in Table 5