Chapter 3

Configurations of Tone Features

As shown in Chapter 2, we extracted pitch contour on vowel part of syllable, and assumed that the final consonant should also carry the tone information. At least it must have some effects on tone. In this Chapter we present our experiments on this issue. The extracted pitch contour is not good enough for tone classification. The pitch contour will carry more information than tone information. The other information carried by the pitch contour is the noise information for tone classification in a way. What the people called the interactive factors of tone. Some factors that we know are speakers, context. Here we concentrate in reducing the interaction between the speakers. In order to reduce the effect of tone variations due to the context and the speaker's emotion, the feature setting is also an important issue for tone classification.

In this Chapter, the background of related topics are presented first. Then it is the implementation of the techniques that used in this research work. After that it is the experiments and discussions. Finally the summary of this chapter are concluded.

3.1 Background

Here the scaling techniques in speech processing are presented first. The normalization techniques are shown next.

3.1.1 Scaling

The hearing system performs a spectrographic analysis of any auditory stimulus. The cochlea can be regarded as a bank of filters whose outputs are ordered tonotopically, so that a frequency-to-place transformation is effectuated. The filters closest to the cochlear base respond maximally to the highest frequencies and those closest to its apex respond maximally to the lowest.

The hearing system can also be said to perform a temporal oscillo-graphic analysis of the set of neural signals that originate in the cochlea in response to an auditory stimulus.

It is often convenient to measure just frequencies when pitch perception is studied. Thus, the pitch of a sound may be specified by the frequency of a pure tone whose pitch is judged to be the same as the pitch of that sound, but auditory scales of frequency representation are required in models of auditory perception.

Generally, the auditory scales include semi-tone scale, mel-frequency scale, critical band scale (bark), and ERB-rate scale (Thubthong *et al.*,2002). The perceived semi-tone scale of complex tones is generally proportional to the logarithm of frequency, with only minor deviations. This is true over a wide range of frequencies up to about 5 kHz. For complex tones, the just noticeable difference (JND) for frequency is approximately constant on this scale. The conversion equation is shown in 3-1:

$$semitone = 69 + 12\log_2\left|\frac{f}{440}\right| \tag{3-1}$$

The mel-scale of auditory pitch was established on the basis of experiments with simple tones (sinusoids) in which subjects were required to divide given frequency ranges into four perceptually equal intervals or to adjust the frequency of a tone to be half as high as that of a tone given for comparison. One mel was defined as one thousandth of the pitch of a 1 kHz tone. The mel scale is now mainly used for the reason of its historical priority only. It is closely related with the critical-band rate scale. The conversion equation is as follows:

$$m = 1127\ln\left(1 + {}^{f}\!/_{700}\right) \tag{3-2}$$

For critical band rate (bark), Measurement of the classical "critical bandwidth" (CB) typically involves loudness summation experiments. Different summation rules have been found to hold for auditory stimuli, depending on whether their frequency components are separated by more or less than the CB. The critical band rate scale differs from Stevens' mel-scale mainly in that it uses the CB as a natural scale unit. The relation between frequency f and CB-rate z has been described by Zwicker (Cited by http://www.ling.su.se/staff/hartmut/bark.html) in form of a table, but for most applications it is more convenient to use the conversion equation listed below.

$$z = \left[26.81/(1 + 1960/f)\right] - 0.53, \tag{3-3}$$

The "notch-noise method" involves the determination of the detection threshold for a sinusoid, centered in a spectral notch of a noise, as a function of the width of the notch. On the basis of results obtained with this method, auditory frequency selectivity can be described in terms of an "equivalent rectangular bandwidth" (ERB) as a function of center frequency. Both spectral and temporal

analysis contributes to the detection of the sinusoid. The conversion equation is in the following:

$$ERB\_rate = 11.17 \ln \left| \frac{f + 312}{f + 14675} \right| + 43.0, \tag{3-4}$$

The CB and the ERB have been found to be proportional for center-frequencies above 500 Hz. For lower frequencies, the ERB decreases with decreasing center-frequency, while the CB remains close to constant. The discrepancy can be explained by the assumption that the temporal fine structure of the signal is not resolved in loudness summation, while it contributes substantially to frequency resolution for f < 500 Hz. Due to differences in bandwidth definition, the ERB is narrower than the classical critical band at all frequencies.

3.1.2 Normalization

F0 is a highly variable acoustical feature. Even for the same speaker's voice, the actual range of F0 changes from time to time because of a variety of physical, emotional, semantic or stylistic factors. Therefore, the absolute F0 values of the lexical tones in Thai may not be directly comparable if they are produced under different context and different speaker. The goal of F0 normalization is to reduce the undesirable variation caused by some irrelevant factors. In the context of this research, such undesirable variation refers to the change of the speaker's F0 range from one utterance to another, and the intonation movement within a long utterance. Usually, F0 normalization is done by, at each time instant, dividing the absolute F0 value by a normalization factor. This factor is expected to be a good indicator of the F0 range at that time instant. In paper (Lee *et al.*, 1995), the normalization factor was computed on per speaker basis for speaker-independent isolated tone recognition. To deal with the change of F0 from one utterance to another, the utterance-wide mean of F0 can be used. Although the absolute F0 level of a particular tone may vary greatly, its relative height with respect to each of the other tones remains largely invariant. Such invariance is preserved locally, i.e. between neighboring syllables, because of the requirement of communication accuracy and the continuous muscle movement of vocal cords. It was proposed to use a moving window approach to better capture the timely change of F0 within an utterance in the paper (Ying *et al.*, 1996).

3.2 Implementation

3.2.1 Framework

Here is the framework of the experiment for tone feature configurations. First the label information are edit through HTK toolkit, Dos command and Excel. Then the pitch extraction program will extract pitch contour for each tone in the speech database from the wave file and related label information. The extracted pitch contour will write into the corresponding pitch file for further processing. After this it's the feature extraction and adaptation part for getting tone feature of classification. The feature adaptation includes scaling, normalization and feature setting. Finally it's the classification part that implemented by three layer feed-forward neural network. The initial weight of NN is randomly chosen when training. So the NN is trained three times for each experiments in order to find the relatively good configurations for classification. Different number of hidden neurons is tested, say 10, 15, 20, 25, 30 etc. and it was found that 25 hidden neurons gave the best performance for our speech database.

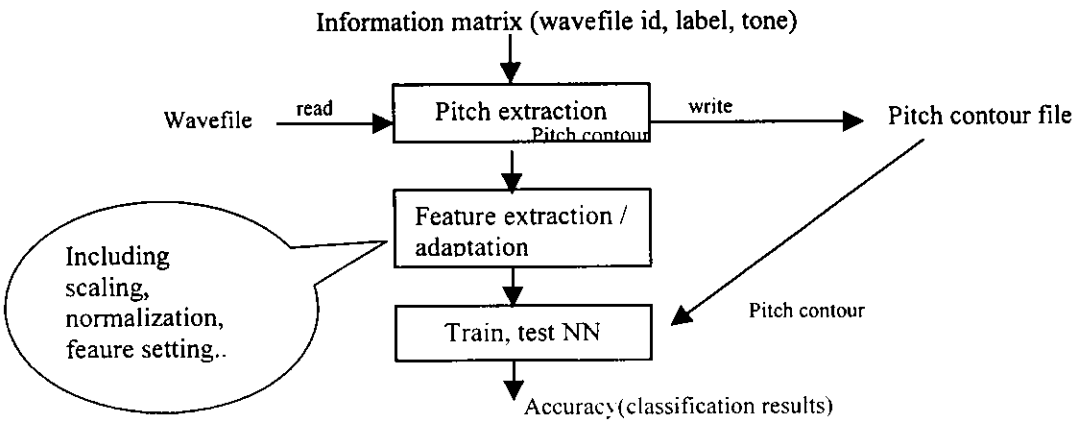
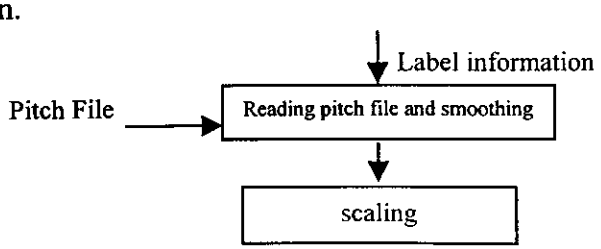


Figure 3-1 Experiment's Framework

3.2.2 Feature Adaptation

Feature adaptation part is trying to find the best configuration of tone feature for Thai tone classification. The main technique we considered here is scaling, normalization, feature setting. First the original pitch data is read in and smoothed. After that it's scaling of pitch contour. Before the normalization, the normalization coefficients such as: mean and standard deviation are calculated. Then it's the normalization and feature setting. Finally it's the processed tone feature for classification.



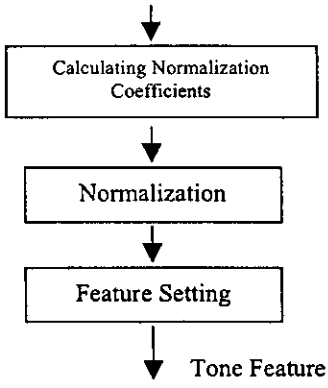


Figure 3-2 Feature Adaptation Framework

3.3 Experiments and Discussions

3.3.1 Experiments subject

In our experiments we use a large vocabulary continuous Thai speech corpus which is taken from the project described in (Thongpraset *et al.*, 2002). The corpus consists of 360 utterances. The data are collected from 18 native Thai speakers (9 male and 9 female speakers). Each speaker read 20 utterances which are randomly arrange from Thai words. Totally the corpus contain 5726 tones. The speech signals are digitized by a 16-bit A/D converter at 16 kHz. The speech data are manually segmented and transcribed at phoneme levels using the waveform analysis software.

3.3.2 Experiments Setting

Based on our speech data, we separate the speech data into two sets: training set, testing set. For training set, we collect the first 15 utterances of each speaker (270 utterances) as training set. And the last 5 utterances of each speaker is the testing set (90 utterances). The experiments are performed using a three-layer feed-forward neural network. The number of input depends on the number of features. The number of hidden is depends on the training performance during the process of testing. The number of output units is 5 corresponding five tones in Thai language. All feature parameters are normalized to lie between –1.0 and 1.0. Training algorithm is back-propagation method. The Matlab is used to implement all of the work presented here.

3.3.3 Experiments Results and Discussions

3.3.3.1 Tone-critical Segment

Tone-critical segment refers to the part containing critical information for tone classification. Based on the current research, the tone mainly lie on the vowel part of

the syllable is generally accepted. Through the previous research work, it concludes that the rhyme segment (including vowel and final consonant) takes advantage over the syllable segments (including initial consonant, vowel and final consonant). Here we first did the experiments only on vowel segment that described in Chapter 1 and we got the 63% accuracy from 373 tones. Based on the observation in Figure 3-3, we think the final consonant should also contain the tone information.

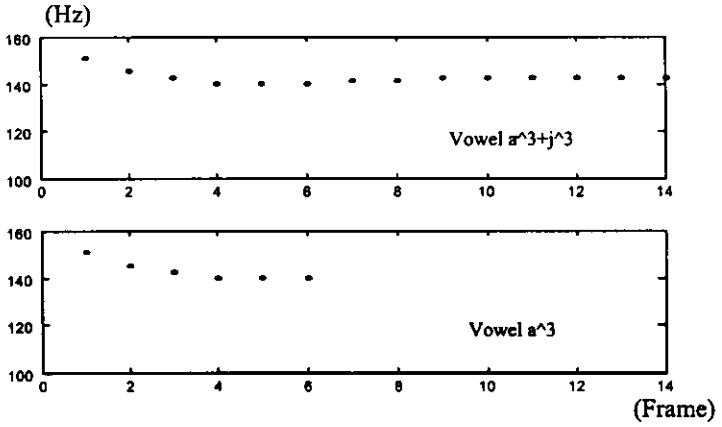


Figure 3-3 Pitch contour with final consonant and the vowel only

From the observation, for the rising tone, the pitch contour of vowel is 6 points but the pitch contour with vowel and final consonant together is 14 points. There are 8-point differences here. The classification experiment then is done on vowel aa. The experiment's results are shown in Table 3-1.

Table 3-1 Tone-critical segment classification results

	Vowel	Vowel+Stop Consonant
Percent	63%(237/373)	77.48%(289/373)

From the Table 3-1, after combining the final consonant into tone classification, the accuracy is increased from 63% to 77.48%. This shows that the final consonant should be considered into tone classification for Thai language.

3.3.3.2 Feature Setting

Tone is expressed by the shape of the pitch contour, such as: level and slope. For isolated tone recognition, the 3-rd order polynomial coefficients are generally used because of the instinct difference between the typical shape of pitch contour. But for continuous speech, the variation of pitch contour is affected by many other factors, such as: speakers, continuity. In order to catch more intrinsic difference among the

tones, here we use five F0 heights and slopes at 0%, 25%, 50%, 75% and 100% of pitch contour as the tone feature shown in Figure 3-4.

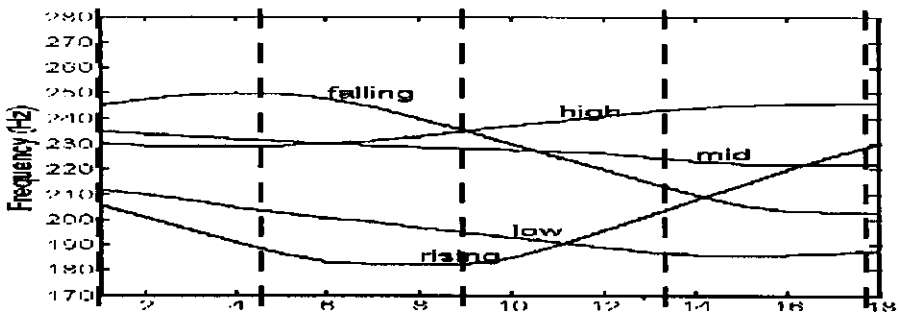


Figure 3-4 Pitch Contour and Tone Feature

Observing the shape of typical pitch contour, we can find that the level and slope of the beginning and end for five tones is absolutely different and also in another three point, 25%, 75% and 100%. Through this way, the more information is kept for tone classification and it reduced the effects of variation pitch contour. The classification results are shown in Table 3-2.

Table 3-2 Feature Setting Classification Results

	4 Coefficients	5 height+5 slopes
Percent	60.9%(743/1220)	66.15%(807/1220)

Here the speech data includes all of the speeches in the corpus. It includes 18 speakers and 20 utterances each speaker. The performance of 10 tone features is obviously improved from the 4 coefficients of 3-rd order polynomial about 6%. This shows the same the results presented in previous research work. Then the 10 tone-feature setting is better than the 4 coefficients tone feature can be concluded.

3.3.3.3 Frequency Scaling

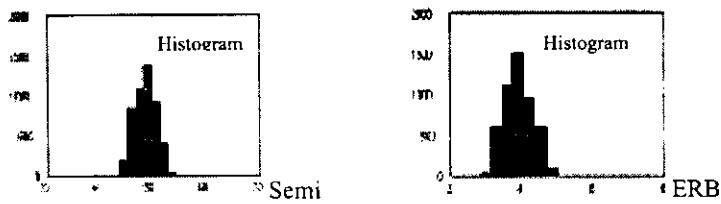
In physical, frequency is generally expressed in hertz. According to the research of the hearing and auditory system, there are several other units used for the speech, such as semi-tone, ERB-rate, Mel scale, Bark scale introduces in section 1.1. Here Mel scale is widely used for speech recognizing segments. For tone classification, the semi-tone and ERB-rate are often considered. The formula of semi-tone and ERB-rate are shown in equation 3-5 and 3-6.

$$semitone = 69 + 12\log_2\left|\frac{f}{440}\right|, \tag{3-5}$$

$$ERB\_rate = 11.17\ln\left|\frac{f+312}{f+14675}\right| + 43.0, \tag{3-6}$$

where f is the frequency in hertz. Semitone is a kind of log scale. It is used to define the music tone. According to some researches, the listeners judged F0 intervals to be equivalent if they were equal in semitones. So researchers often consider using semitone scaling to reduce the effects of tone variations. ERB frequency scaling is similar to the critical-band frequency scaling. Basically the ERB frequency defined according to the knowledge of human auditory system. So researchers consider using ERB-rate to improve the performance of classification through the way of modelling the human tone classification system. Figure 3-5 shows the histogram of the male speaker's pitch data before and after scaling.

(a) Histogram of Original Speech



(b) Histogram of Semi-tone Scaling and ERB scaling

Figure 3-5 Histogram of Scaling

From the histogram, we can see that after scaling the distribution basically keep unchanged. But the variation range of pitch is compressed. The classification results are shown in Table 3-3.

Table 3-3 Classification Results of Scaling

	5 height+5 slopes	Semi-tone	ERB
Percent	66.15%(807/1220)	66.89%(816/1220)	68.61%(837/1220)

From the table, both kind of scaling improved the performance. But the ERB scaling takes more advantages which follows the conclusions shown in previous research work.

3.3.3.4 Normalization

F0 is speaker dependent. The normalization is necessary for speaker independent tone classification. To observe the pitch variation among speakers, the pitch histogram of two speakers (one male and one female) are shown in figure 3-6.

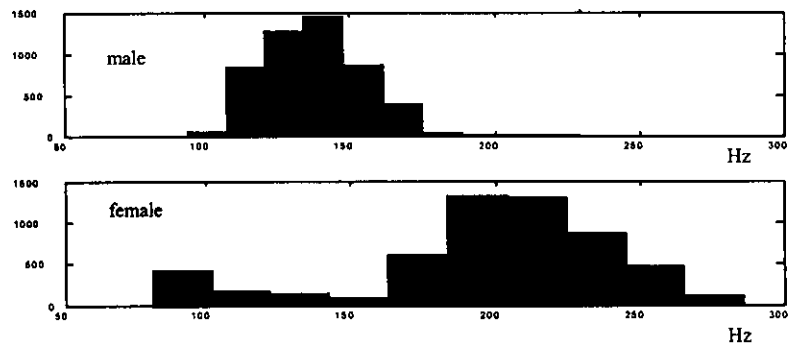


Figure 3-6 Histogram of Two Speaker's Pitch

From figure 3-6, the variation of pitch data between speakers is clearly expressed. The mean of F0 for female is clearly higher than the male. Also the distribution of pitch is different. This shows the normalization is a critical part for speaker independent tone classification. Generally the mean normalization and z-score normalization are used.

The mean normalization is shown in equation 3-7.

$$normf = \frac{f - m}{m}, \tag{3-7}$$

Where f is the original frequency, *norm-f* is the normalized frequency, m the mean value of the whole utterance. Mean-normalization normalizes the mean of total utterance to zero and keep the relative differences within the utterances.

The z-score normalization is shown in equation 3-8.

$$normf = \frac{f - m}{sd}, \tag{3-8}$$

Here *norm-f* is the normalized frequency, m is the mean of the utterances and *sd* is the standard deviation of the pitch among the utterance. It normalizes the distribution of pitch within the utterance into the 0-1 normal distribution.

The figure 3-7 shows the histogram after normalization.

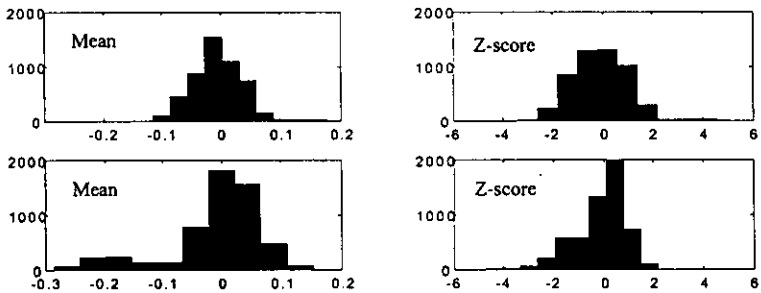


Figure 3-7 Histogram after Normalization

From figure 3-7, we notified that the variations between the speaker are reduced. For mean normalization, it keeps the distribution of original data. But for z-score normalization, the distributions of the pitch data are modified into Guassian distribution.

The classification results after normalization are shown in Table 3-4.

Table 3-4 Classification Results of Normalization

Percent	semi-tone	Semi+z-score	Semi+m-norm
	66.15%(807/1220)	69.84% (852/1220)	72.21%(881/1220)
Percent	ERB	ERB+z-score	ERB+m-norm
	68.61%(837/1220)	70.16%(881/1220)	72.05%(879/1220)

Here all of classification is using 10 tone features. From the table, the semi-tone scaling and mean-normalization give us the highest performance which is not the same as the results in previous research work. In previous work, it concludes that the ERB scaling and z-score normalization give the best performance. Based on the difference and the histogram of different speaker, we can conclude that the distribution of pitch data is not absolutely following the Guassian distribution. But the z-score normalization modified the distribution according to Guassian distribution. This should be the main reason make the lower performance. For the work described in N. Thubtong paper, the z-score give the higher performance. It may be because of the characteristics of speech data.

3.3.3.5 Confusion Analysis

From all of above, we got that using 10 tone features, semi-tone scaling and mean normalization give us the highest performance. Table 3-5 is the confusion-matrix with the best configurations.

Table 3-5 Confusion-matrix

	0(M)	1(L)	2(F)	3(H)	4(R)	Percent(%)
0	319	32	18	16	5	83.07
1	51	159	6	16	17	63.86
2	20	8	212	14	1	83.14
3	48	8	28	103	6	53.37
4	19	24	0	8	88	63.31
						72.21

From the confusion-matrix, it can be seen that the falling tone provides the highest recognition rate, while the high tone gives the poorest. Also we noted that there is big confusion between the mid and the low tone, the low and the high tone, the fall and the high tone, the low and the rise tone. Since the number of the speech with the mid tone is very large compared with other tones, the classifier is biased. This made the fall, the high and the rise tone mis-classified as the middle tone.

3.3 Summary

In this Chapter, we have presented a study of configuration of tone feature with respect to the tone-critical segment, feature setting, frequency scaling, normalize. A feed-forward neural network is implemented as tone classifier. 25 hidden neurons is found to give the best performance for 10 input features. Through all of the experiments, we conclude that it's necessary to consider the final consonant in Thai tone classification. The feature setting that use 5 heights and slopes gives 6% higher performance than 4 3-rd order polynomial coefficients. For frequency scaling, the ERB scale is better than semi-tone frequency. The mean-normalization takes advantages over z-score normalization because it doesn't change the distribution of tone data. Finally the best configuration based on the experiments is using 10 tone features, semi-tone scaling and mean normalization. The performance is 72.21%, which is 881 correct tones from total 1220 tones.