F0 Feature Extraction by Polynomial Regression Function for Monosyllabic Thai Tone Recognition

Patavee Charnvivit, Somchai Jitapunkul, Visarut Ahkuputra, Ekkarit Maneenoi, Umavasee Thathong, and Boonchai Thampanitchawong

Digital Signal Processing Research Laboratory, Department of Electrical Engineering,
Faculty of Engineering, Chulalongkorn University, Bangkok 10330, THAILAND
e-mail : jsomchai@chula.ac.th

Abstract

This paper presents a monosyllabic Thai tone recognition system. The system is composed of three main processes, fundamental frequency (F0) extraction from input speech signal, analysis of F0 contour for feature extraction, and classification of each tone using the extracted features. In the F0 feature extraction, the polynomial regression functions are employed to fit the segmented F0 curve where its coefficients are used as a feature vector. In tone recognition, we used the maximum a posteriori probability classifier (MAP) to classify a tone by assuming that the feature is a multidimensional Gaussian random variable. The hypothetical words used in this paper are composed of numerical words and monosyllabic Thai words. The vocabulary set is composed of the short vowel words, the long vowel words and have the effect of initial and final consonant on the shape of F0 contour. The experimental results show that by using the system as a speaker-dependent system, the maximum recognition rate is 96.20% using three-dimension feature vector. The speaker-independent recognition rates are 79.99% for male and 82.80% for female using four-dimension feature vector.

1. Introduction

Thai is a tonal language. There are five tonemes in Thai, the mid, the low, the falling, the high and the rising. The feature of speech that was used to classify the tone is the shape of fundamental frequency (F0) contour, which shown in Figure 1. There are several parameters that also have the effect on the shape of F0 contour such as the gender and the age of speaker, the initial consonant, the final consonant and the duration of vowel (short or long). In this paper, we used the hypothetical words that consist of several effects. In our process, the F0 contour of input speech was automatically smoothed and segmented by the proposed algorithm in section 3.1. Then they were fit by the polynomial regression function, which we used its coefficients as the features of F0 contours. In the recognition process, we used the maximum a posteriori probability classifier (MAP) to classify the tones by assuming that the feature vectors are the multidimensional Gaussian random variables.

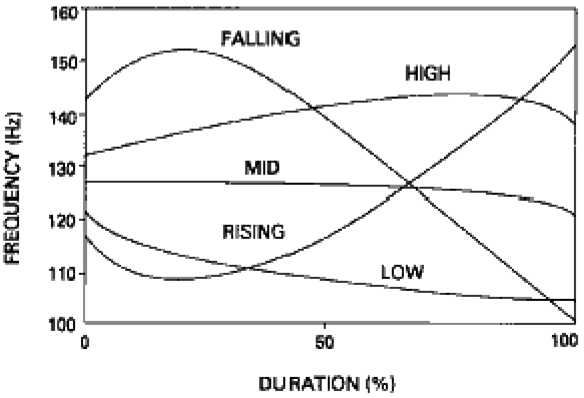

Figure 1: Average F_0 contours of the five Thai tones produced in isolation by a male speaker [3].

2. Thai Syllable Structure

The Thai syllable structure is composed of three different sound systems as follows. [2]

1. The system of consonants consists of 33 consonantal units, 21 consonants and 12 consonant clusters.
2. The system of Thai vowels consists of 18 monophthongs and 6 diphthongs. The monophthongs are qualitatively 9 different vowels, each of which has two members, short and long. Each of three different diphthongs also has 2 quantitatively different members.
3. The system of tones consists of 5 tones. There are 3 kinetic or relatively leveled tones, the high (H), the mid (M), and the low (L), and 2 dynamic or contour tones, the falling (F) and the rising (R).

The smallest construction of sounds or syllables in Thai is composed of one vowel unit or one diphthong, one two, or three consonants, and a tone. The construction can be represented with the structure as illustrated in Figure 2,

$$\mathbf{S = C_i(C_i)V^T(V)(C_f)}$$

Figure 2: Thai Syllable Structure

Where $\mathbf{C_i}$ is initial consonant, $\mathbf{C_f}$ is final consonant, $\mathbf{V}$ is vowel, and T is tone respectively.

3. F0 Feature Extraction

The F0 feature extraction process has two procedures. The first is F0 smoothing and segmentation procedure. The second is polynomial curve fitting procedure.

3.1. F0 smoothing and segmentation procedure

F0 from the F0 extraction process will be smoothed in the smoothing procedure by using median filtering. In the segmentation procedure, there is algorithm that was used to segment the smoothed F0. This algorithm will determine the beginning and the ending frame of the longest time that F0 at each frame has the value differ from the neighboring frame no more than ΔF_{max} = 17 Hz.

3.2. Polynomial regression

The objective of this procedure is to determine the coefficients b_k of a polynomial that fits the segmented F0 contour. Let $\mathbf{F} = (F_1, F_2, ..., F_L)^T$ be a sequence of segmented F0 of length L, $\hat{\mathbf{F}} = (\hat{F}_1, \hat{F}_2, ..., \hat{F}_L)^T$ be an estimated vector of $\mathbf{F}$. A d-dimension feature vector $\boldsymbol{\beta} = (\beta_{0,}\beta_1, ..., \beta_{d-1})^T$ is the coefficient of (d-1)-order polynomial regression function

$$\hat{F}_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \ldots + \beta_{d-1} t_i^{d-1} \qquad (1)$$

where $t_i = \dfrac{i}{L}$ is a normalized time respect to F_i. Equation (1) can be expressed in matrix form as

$$\hat{\mathbf{F}} = \mathbf{T}\boldsymbol{\beta}, \qquad \begin{bmatrix} \hat{F}_1 \\ \hat{F}_2 \\ \vdots \\ \hat{F}_L \end{bmatrix} = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{d-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{d-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_L & t_L^2 & \cdots & t_L^{d-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{d-1} \end{bmatrix} \qquad (2)$$

A solution for $\boldsymbol{\beta}$ in a least-squares sense (minimize the Euclidean distance between vectors $\mathbf{F}$ and $\hat{\mathbf{F}}$) is obtained via forming the pseudoinverse of T, that is,

$$\boldsymbol{\beta} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{F} \qquad (3)$$

However, if $\mathbf{T}^T\mathbf{T}$ is nearly singular, the numerical errors incurred in forming $\mathbf{T}^T\mathbf{T}$, and then forming the inverse, spawn a need for alternate approaches that are not plagued by numerical sensitivities. One solution, known as QR decomposition, can be used for this case.

4. Classification Algorithms

A decision rule that used in the tone recognition process is a maximum a posteriori probability classifier (MAP). Before using this classifier, the input feature vector $\boldsymbol{\beta}$ will be sent to the automatic gender identification procedure to determine the gender of speaker. This procedure will help the MAP process to select the proper model for each gender.

4.1. Automatic gender identification

The fundamental frequency of men is usually found somewhere between the two bounds 50 - 250 Hz, while for women the range usually falls somewhere in the interval 120 - 500 Hz [1]. So, the parameter that possible to be used to identify the gender is the F0 levels. The average value of the segmented F0 is used in this paper. The identification algorithm is that if the average is more than the threshold, the recognized gender is female otherwise the gender is male. Because the feature vector is composed of the coefficient of the regression polynomial that fit the segmented F0, the average value can be determined by these coefficients as shown in 5.1.1.

4.1.1. Estimating the mean of the segmented F0

The estimated F0 can be expressed as a continuous time function, that is,

$$\hat{F} = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_{d-1} t^{d-1} \qquad (4)$$

where $t \in [0,\ 1]$.

The average value of this function, that used to approximate the mean of the segmented F0, can be determined by the integral from zero to one of $\hat{F}$ with respect to t as follows:

$$\begin{aligned} \hat{m} &= \int_0^1 \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_{d-1} t^{d-1} dt \\ &= \beta_0 + \frac{1}{2}\beta_1 + \frac{1}{3}\beta_2 + \ldots + \frac{1}{d}\beta_{d-1} \\ &= \mathbf{w}_d^T\boldsymbol{\beta} \end{aligned} \qquad (5)$$

where $\mathbf{w}_d = (1,\ \frac{1}{2},\ \frac{1}{3},\ \ldots,\ \frac{1}{d})^T$.

4.1.2. F0 threshold

The threshold of F0, which used to classify the gender of speaker, can be determined by assuming that the probability density function (pdf) of $\hat{m}$ for each gender is Gaussian. We can apply the MAP classifier as shown in the next section but in the 1-dimension case to find this threshold.

4.2. Maximum a posteriori probability classifier (MAP)

A decision rule that used in the recognition process is a maximum a posteriori probability classifier (MAP). The MAP classifier will select tone i for the feature vector $\boldsymbol{\beta}$ of the unknown speech input if the posteriori probability

$$P(w_i \mid \boldsymbol{\beta}) > P(w_j \mid \boldsymbol{\beta}) \qquad , \forall j \neq i \qquad (6)$$

where w_0, w_1, w_2, w_3 and w_4 are the class of tone M, L, F, H and R respectively. This probability can be determined from

$$P(w_i \mid \boldsymbol{\beta}) = \frac{[p(\boldsymbol{\beta} \mid w_i)P(w_i)]}{p(\boldsymbol{\beta})} \qquad (7)$$

where $P(w_i)$, the priori class probability, is assumed to be equal for all tone.

The pdf of feature vectors is determined from the summation of all conditional pdf given each class, that is,

$$p(\boldsymbol{\beta}) = \sum_i p(\boldsymbol{\beta} \mid w_i) \qquad (8)$$

Notice that the quantity $P(w_i)$ and $p(\boldsymbol{\beta})$ are common to all class-conditional probabilities; therefore, it represents a scaling factor that may be eliminated. Thus the decision algorithm become

$$\text{select tone } i \text{ if} \qquad p(\boldsymbol{\beta} \mid w_i) > p(\boldsymbol{\beta} \mid w_j) \qquad , \forall j \neq i \qquad (9)$$

In this paper, we assume the class-conditional pdf to be d-dimensional Gaussian pdf. Therefore,

$$p(\boldsymbol{\beta} \mid w_k) = (2\pi)^{-\frac{d}{2}} \left|\boldsymbol{\Sigma}_k\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_k)\right] \qquad (10)$$

where $\boldsymbol{\beta}$ is d x 1 with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ for class k.

If we take the log function, a monotonically increasing function, to $p(\boldsymbol{\beta} \mid w_k)$ so the decision algorithm is

$$\text{select tone } i \text{ if} \qquad d_{ml}(\boldsymbol{\beta}, \boldsymbol{\mu}_i) < d_{ml}(\boldsymbol{\beta}, \boldsymbol{\mu}_j) \qquad , \forall j \neq i \qquad (11)$$

where we define the maximum likelihood distance as

$$d_{ml}(\boldsymbol{\beta}, \boldsymbol{\mu}_k) = (\boldsymbol{\beta} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_k) + \ln\left|\boldsymbol{\Sigma}_k\right| \qquad (12)$$

5. Tone Recognition System

The block diagram of the tone recognition system is shown in Figure 3. The first block is the F0 extraction process which a single syllable is its input. The speech was recorded at 11025 Hz sampling frequency and 16-bit quantization level. F0 was computed in this process from 256 samples speech frame with the overlapping of ¾ frame by using the modified short-term autocorrelation with center clipping method. The second block is the F0 feature extraction process, which determines the parameters that have sufficient information to describe the shape of F0 contour by the method of polynomial regression. The final block is the tone recognition algorithm that uses the parameters obtained from the previous process to determine the best matching tone for the input speech.

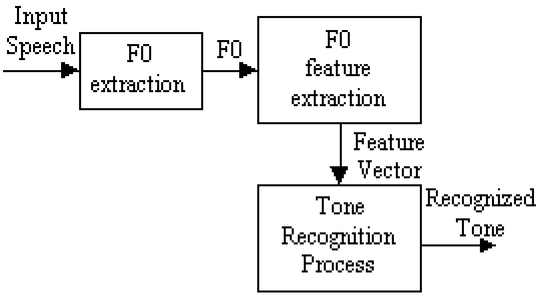

Figure 3: Block diagram of the system

6. Experiments

There are two experimental sets, the experiment on speaker-dependent tone recognition system and the experiment on speaker-independent tone recognition system. In the first one, there is no automatic gender identification system because the speaker of the training set and the testing set is the same. Both experiments use 30 hypothetical words as shown in Table 1.

Table 1: Hypothetical Words

Tone	Mid (M)	Low (L)	Falling (F)	High (H)	Rising (R)
Words	/'dqqn0/	/'paak1/	/'wing2/	/'nok3/	/'huu4/
	/'n@@n0/	/'pet1/	/'kluuaj2/	/'to3/	/'svva4/
	/'taa0/	/'kaj1/	/'som2/	/'nam3/	/'s@@ng4/
	/'mvv0/	/'hnvng1/	/'nang2/		/'saam4/
	/'thiian0/	/'sii1/	/'kxxw2/		/'suun4/
	/'kin0/	/'hok1/	/'haa2/		
	/'tiiang0/	/'cet1/	/'kaw2/		
		/'pxxt1/			

6.1. The automatic gender identification test

The average values of the segmented F0 for each gender in the training set are determined. The mean and the variance are also calculated. By assuming that the pdf of both genders are Gaussian, so we can determine the threshold for the minimum of error probability. The threshold is 161.5 Hz, which yield the accuracy of 100 % in the gender classification process of the testing set.

6.2. The speaker-dependent tone recognition test

In this case, the training set use utterances from three trials of all words in Table 1 and also using the other three trials from the same speaker as the testing set. By varying the dimension d of the feature vector from three to six, the recognition rate is shown in Figure 4. These results indicate that the recognizer has the maximum recognition rate at $d = 3$. The confusion matrix of this dimension is shown in Table 2.

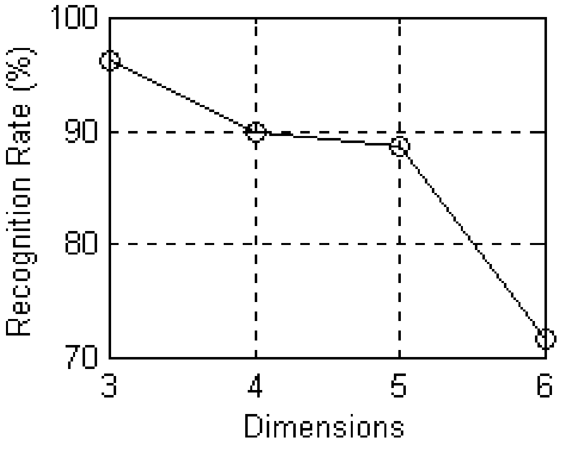

Figure 4: Recognition rate of the speaker-dependent tone recognition system

Table 2: Confusion matrix of the speaker-dependent tone recognition with $d = 3$

Desired Tone	Recognized Tone					Total	Accuracy (%)
	M	L	F	H	R		
Mid (M)	18	3	0	0	0	21	85.71
Low (L)	0	24	0	0	0	24	100.00
Falling (F)	1	0	20	0	0	21	95.24
High (H)	0	0	0	9	0	9	100.00
Rising (R)	0	0	0	0	15	15	100.00
						90	96.20

6.3. The speaker-independent tone recognition test

The training set in this experiment consists of the 30 hypothetical words similar to the previous section from four men and four women. The test set is composed of four men and women in the training set and the other four men and four women that do not have their speech in the training set. Similar to the previous section, the dimension of feature vectors is varied from three to six. Figure 5 shows the results compared between male and female speakers. These results differ from the case of speaker-dependent that for male, the recognition rate is increase when the dimension is increase until the dimension is four, for female, the recognition rate is approximately equal when the dimension is three and four. When the dimension is more than four, the recognition rate is decrease. The confusion matrices of male and female at the maximum accuracy are shown in Table 3 and 4 respectively.

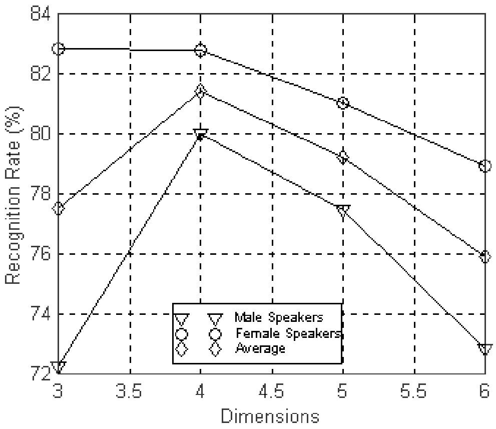

Figure 5: Recognition rate of the speaker-dependent tone recognition system

Table 3: Confusion matrix of the speaker-independent tone recognition with $d = 4$ for male speakers

Desired Tone	Recognized Tone					Total	Accuracy (%)
	M	L	F	H	R		
Mid (M)	40	6	8	2	0	56	71.43
Low (L)	17	44	1	1	1	64	68.75
Falling (F)	4	3	47	2	0	56	83.93
High (H)	1	2	1	20	0	24	83.33
Rising (R)	0	2	0	1	37	40	92.50
						240	79.99

Table 4: Confusion matrix of the speaker-independent tone recognition with $d = 4$ for female speakers

Desired Tone	Recognized Tone					Total	Accuracy (%)
	M	L	F	H	R		
Mid (M)	39	10	7	0	0	56	69.64
Low (L)	17	45	2	0	0	64	70.31
Falling (F)	8	0	48	0	0	56	85.71
High (H)	1	0	0	23	0	24	95.83
Rising (R)	0	2	0	1	37	40	92.50
						240	82.80

7. Conclusions

The system for monosyllabic Thai tone recognition has been proposed in this paper. We used the coefficient of polynomial regression function as a feature vector of the segmented F0 contour. In the training phase, the feature vector was used to determine the statistical parameters of the model for each gender. In the testing phase, the feature vector will be passed to the automatic gender identification to determine the gender of speaker and passed to the tone recognition process for each gender. The tone recognition process will determine the tone of the input speech by using the maximum a posteriori probability classifier. The results show that in the speaker-dependent system, the optimum dimension is three with recognition rate 96.2%. But in the speaker-independent system, the optimum dimension is four with recognition rate 79.99% for men and 82.80% for women.

8. Acknowledgement

The authors would like to acknowledge Digital Signal Processing Research Laboratory, Chulalongkorn University for the support of this research.

9. References

[1] Deller J. R., Proakis J. G., and Hansen J. H. L., “Discrete-Time Processing of Speech Signals”, Macmillan Publishing Company, a division of Macmillan, Inc., United States of America, 1993.

[2] Maneenoi E., Jitapunkul S., Ahkuputra V., Thathong U., and Thampanitchawong B., “Thai Monophthong Recognition Using Continuous Density Hidden Markov Model and LPC Cepstral Coefficients”, *Proceeding of International Conference on Spoken Language Processing (ICSLP 2000),* Oct. 2000.

[3] Potisuk S., Harper M. P., and Gandour J., “Classification of Thai Tone Sequences in Syllable-Segmented Speech Using the Analysis-by-Synthesis Method”, *IEEE Trans. On Speech and Audio Processing*, vol. 7, pp. 95-102, Jan. 1999.

[4] Schalkoff R. J., “Pattern Recognition: Statistical, Structural and Neural Approaches”, John Wiley & Sons, Inc., Singapore, 1992.

[5] Thathong U., Jitapunkul S., Ahkuputra V., Maneenoi E., and Thampanitchawong B., “Classification of Thai Consonant Naming Using Thai Tone”, *Proceeding of International Conference on Spoken Language Processing (ICSLP 2000),* Oct. 2000.

[6] Thubthong, N. “A Thai Tone Recognition system based on Phonemic Distinctive Features”. *Department of Computer Engineering Faculty of Engineering, Chulalongkorn University.*