6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

CLASSIFICATION OF THAI CONSONANT NAMING USING THAI TONE

Umavasee Thathong, Somchai Jitapunkul, Visarut Ahkuputra, Ekkarit Maneenoi, 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 proposes the novel technique for separation of Thai consonant naming or consonant spelling using its tones. Consonant spelling is used for many applications such as a voice-actuated typewriter that helping to correct the confusable word in sound. Because fundamental frequency (F0) can be suitably used in tone classification for Thai speech recognition, which is tonal language of five patterns: mid, low, falling and rising. Consequently, classification of Thai consonant naming algorithm used F0 for distinguishing rising tone from mid tone. From the experiment result, we found that not only the level of F0 indicates tonality of sound but also considering flattening, rising, oscillation and continuity of F0 that are necessary for Thai tonal language. This paper shows performance of algorithm that classifies 996 sounds and yielding 1.72% error-rate.

1. INTRODUCTION

Recognition of spelled letters is essential for many real-world applications that deal with arbitrary names or addresses such as car navigation, automated directory assistance, call-routing devices [1] and the ticket booking. Moreover, spelling letter recognition is also used as complement to the existing speech recognition system. In Thai language, the name of persons, places and other organizations are specific names that can be written in many styles, but there have the same sounds. That's why spelling recognition is indispensable for specifying correct name, which are applied in many applications and available for hand-busy, eye-busy, handicapped and novice person [2] also.

Thai syllables are composed of consonants, vowels and tone [3]. The smallest structure of sounds or syllables in Thai is composed of one vowel unit or one diphthong, one, two, or three consonants, and a tone. The structure can be represented with the structure as illustrated in Figure 1,

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

Figure 1: Thai Syllable Structure

Where C_i is initial consonant, C_f is final consonant, V is vowel, and T is tone respectively.

Using neural networks (NN) for Thai vowels recognition is yielding 85.92% recognition rate [4]. For tones of Thai vowels recognition using hidden Markov model (HMM) is yielding 91% recognition rate [5]. But both of NN and HMM is not proper for consonant recognition consequently the data analysis is used for finding the causes of misclassification. According to Figure 1 if tone is changed, meaning is change too. Hence tone classification is the first tasks for consonant recognition. Firstly, the detail of Thai language structure is explained in the first section then the algorithm will be proposed with respect to the problem of classifying tone by F0.

2. STRUCTURE OF THAI LANGUAGE

Thai syllables are composed of consonants, vowels and tones [3]. There are five tones in Thai, mid (/0/), low (/1/), falling (/2/), high (/3/) and rising (/4/), whose characteristics are shown in Figure 2. Thai consonant naming can be spelled in many ways and it contains /@@/ vowel in every sound. This experiment performed using shortest names of Thai consonant speaker independent isolated word. All of these pronunciations of Thai consonants are composed of several pronunciations in rising tone (/4/) and the other are in mid tone (/0/). Thus, this paper proposes an idea of classify these consonants into 2 groups according to their tones by separating rising tone from mid tone. The feature is fundamental frequency (F0), which is computed from cepstral technique. Cepstral coefficients were extracted from every frame of consonant.

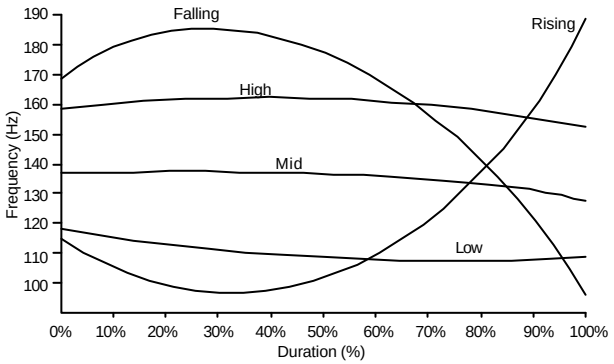

Figure 2: Five Tonal Levels in Thai language [7]

Database consists of 21 mid tones and 7 rising tones, which are listed in Table 1. Spelled by 20 males and 7 females comprised of 996 consonants name.

Mid Tone	/k@@0/	/kh@@0/	/ng@@0/	/c@@0/
	/ch@@0/	/s@@0/	/j@@0/	/d@@0/
	/t@@0/	/th@@0/	/n@@0/	/b@@0/
	/p@@0/	/ph@@0/	/f@@0/	/m@@0/
	/r@@0/	/l@@0/	/w@@0/	/@@0/
	/h@@0/			
Rising Tone	/kh@@4/	/ch@@4/	/th@@4/	/ph@@4/
	/f@@4/	/s@@4/	/h@@4/	

Table 1: 28 Thai consonants

3. CLASSIFIED TONE ALGORITHM

Generally, F0 should gradually increase if consonant is rising tone and F0 should flatten if it is mid tone [6]. This paper shows the problems of tone separated by F0 in the experiment result. For example, F0 in some mid tone increases at the end of frame (Figure 7) in spite of being rising tone; some data lose F0 in a short time (Figure 4), etc. Accordingly, two cases were employed in this algorithm. The first case was general case and the second was respected to experiment (ambiguous case). Thus, the algorithm can be divided into 4 steps. Step 1 was used for choosing the begin and the end of frame by considering continuity of F0. Step 2, smoothing data with respect to frequency and standard threshold were computed. Finally, the tone was decided by standard threshold. If sound is included in general case then calculate in step 3. If sound is included in ambiguous case then it will continue to calculate in step 4.

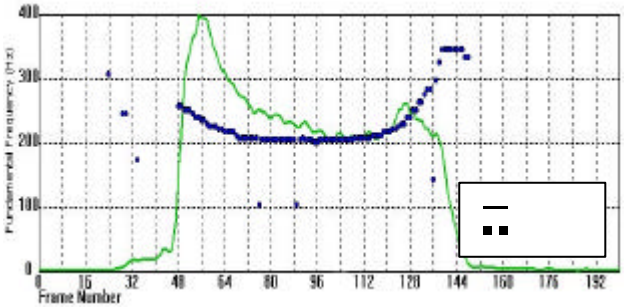

Figure 3: Strong rising in /s@@4/

STEP1: Finding the begin and the end of frame

F0 value will be segmented in three parts that are the beginning, the middle and the end of a segment respectively.

1.1) The begin of frame (*Begin frame)* is determined from the first frame to 5/11 of entire frames. The F0 reference value is computed from the middle of a segment and is used for arrangement the level of F0 in higher level, equal level or lower level. In general case, the beginning of a segment was used for supported the decision that sound does rising tone shown in Figure 3. In ambiguous case the beginning of a segment is not significant when comparing to the end of a segment. So, *Begin frame* is chosen when F0 does not contain zero value more than 4 frames.

1.2) Finding the end of frame (*End frame*) is the same as 1.1), but it started from the last frame backed to middle frame. The *End frame* is chosen when F0 does not contain the concatenation of zero value more than 5 frames. The length of chosen frame from *Begin frame* to *End frame* must be more than 60 frames because shorten data is difficult to arrange the level in rising or flattening level. In Figure 3, if data was ignored at the end of a segment then chosen frames will be less than 60 frames and result will be flattening tone, which is wrong decision. These reasons show that F0 was lost in a short time therefore smoothing data will be computed in next step.

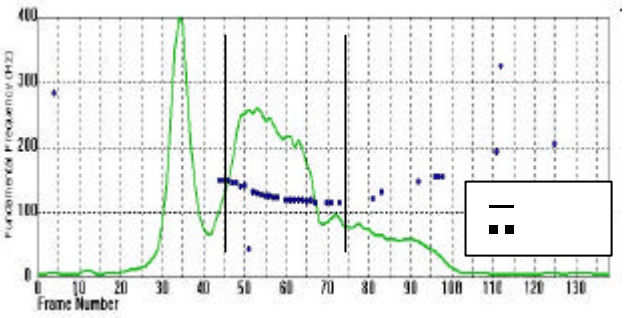

Figure 4: Lose of F0 in */ch@@4/*

STEP2: Smoothing data depend on frequency by finding the begin and the end of frequency.

First, project the data will be projected into the frequency axis and the density of frequency will be measured then the range of frequency will be searched in which contains maximum density and contains zero value no longer than 25 Hz. Unrequired frequencies that does not include in chosen frequency will be rejected. Rejected data are replaced by the next frame and the *End frame* is decreased. F0 that losing in a short time shown in Figure 4 will be rejected too. Last, standard threshold value will be set as eq(1)-(3).

$$\textit{Chosen length} = \textit{Begin frame} - \textit{End frame} \quad (1)$$

$$\textit{Reference frame} = 2/3 * [\textit{Chosen length}] \quad (2)$$

$$\textit{Reference value} = [\boldsymbol{Nr}(\textit{Reference frame})] \ / \ 10 \quad (3)$$

Where ***Nr***(*Reference frame*) are 10 frames summing values of F0 around the *Reference frame*.

STEP3: General case of F0

Strong rising: For all frames in the chosen range, the current frame is compared with the next frame. If F0 value increases in the next frame, this frame is called *Increasing frame.* If F0 value equals to the next frame, this frame is called *Equality frame.* If F0 value decreases in the next frame, this frame is called *Decreasing frame.* While the length of *Increasing frame* is more than 20% of *Chosen frame length*, F0 value of every frame is higher than the *Reference value*, maximum value of F0 minus

Reference value is more than 30 Hz, then this will be detected for rising tone as shown in Figure 3.

Rarely rising: Data gradually increases more significant than decrease and equal to the next frame. If the length of *Increasing frame* is more than 40% of *Reference frame* and the maximum value in the length of *Increasing frame* is more than 15 Hz then result will be rising tone as shown in Figure 5.

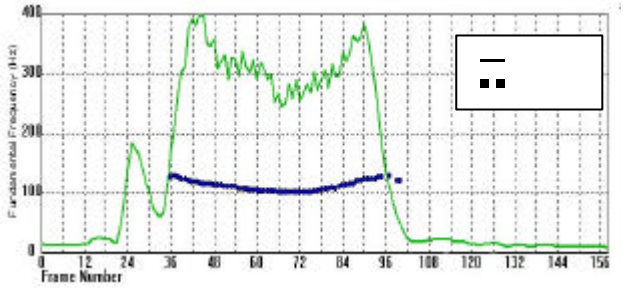

Figure 5: Rarely rising in /kh@@4/

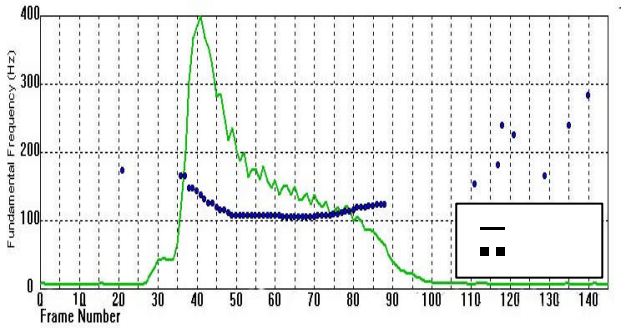

Figure 6: Rarely rising in /ph@@4/

STEP4: Decision on F0 ambiguity

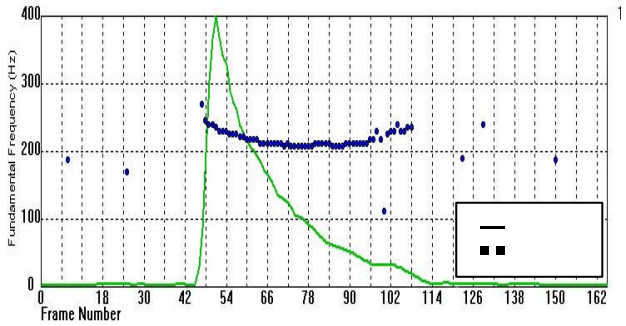

Figure 7: F0 not continuous in /t@@0/

In this case, the data are rearranged because the level of frequency always change and are not incessant as shown in Figure 7 and 8 correspondingly. As a result, considering the important data and rejecting undesired data are computed from *Reference frame* to *End frame.* The frame that F0' s value differs from concatenated frame more than 20 Hz is rejected (only keep the continuity of F0).

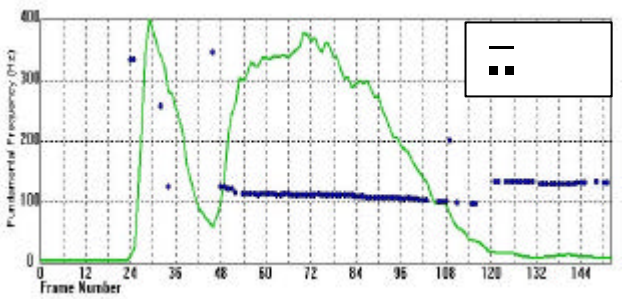

Figure 8: F0 not continuous in */ph@@0/*

For every frame, from *Reference frame* to *End frame*, if *Increasing frames* or *Equality frames* are more than *Decreasing frames* then its slope will be calculated as eq(4).

$$slope = \frac{}{End\ frame - Refernce\ frame} \quad (4)$$

If slope value is more than 2 Hz per frame, then it is classified as rising tone depicted as Figure 4, otherwise it is classified as mid tone.

4. EXPERIMENTAL RESULTS

The tonal separation result achieved 99.28% correctness on 996 consonant names, 829 training consonants and 168 testing consonants, which are spelled by 20 males and 7 females. These comprised of 44 pronunciations by 15 persons and 28 pronunciations by 12 persons. However, this algorithm can be improved when the number of *Increasing frame, Decreasing frame* and *Equality frame* will be considered ensemble into inequality condition. If the inequality in equation (5) is true then this consonant is rising tone as shown in Figure 9. This condition decreased 0.72% error rate.

Increasing frame + *Decreasing frame* - *Equality frame* > 0.125**Chosen frame length* (5)

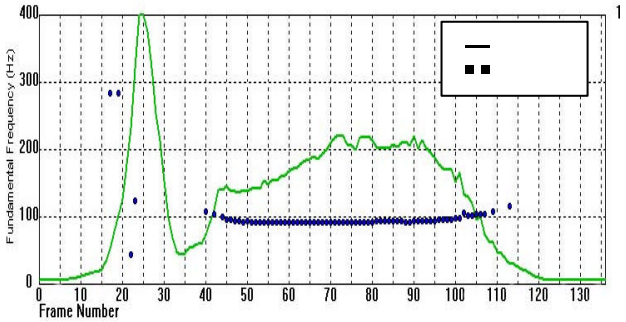

Figure 9: The other problem of rising tone in /ch@@4/

5. CONCLUSIONS

General case in step 3 was used for sound that have not problem. Contrast to step 4 that was used for ambiguous case, which always have many problems.

From misclassifying tone in ambiguous case, F0 does not obviously increases at the end of frame and does not oscillate in rising tone. In contrast to flatten tone, it often oscillates when there are rising. It was noticed that in female voices, the F0 value tends to increase in rising tone and increase with oscillation in mid tone. Quite the opposite, the F0 values of male voices are almost constant in rising tone and decrease in mid tone. From these results, the step 4 was used to calculate the periodic of oscillation, which is an important feature to distinguish the mid tone from rising tone.

Considering flattening, oscillation and rising of F0, the performance can be improved to 100%. From the results of this algorithm, only rising level of F0 could not indicate the type of tone correctly as shown in Figure 6. Flattening, oscillation and rising of F0 plays the major role to detect the tone of Thai consonants. In the future research, this algorithm will be improved the recognition rate for consonant recognition.

6. ACKNOWLEDGEMENT

The authors would like to acknowledge Digital Signal Processing Research Laboratory, Department of Electrical Engineering Faculty of Engineering and Asst. Prof. Dr. Sudaporn Luksaneeyanawin from faculty of arts for their supports of this research and also Chulalongkorn University for her some funding support.

7. REFERENCES

[1] Junqua, J. "SmarTspelL™:A multipass recognition system for name retrieval over the telephone" *IEEE Transactions on Speech and Audio Processing*. Vol. 5, No.2 1997.

[2] Loizou, P. "*High-Perfomance Alphabet Recognition" IEEE Transactions on Speech and Audio Processing*. Vol. 4, No.6, November 1996.

[3] Luksaneeyanawin, S. "Linguistics Research and Thai Speech Technology". *Paper read at the 5th International Conference on Thai Studies*. School of Oriental and African Studies, University of London, United Kingdom, July 1993.

[4] Maneenoi, E. "Thai Vowel Phoneme Recognition Using Neural Networks". *Proceeding of the 22nd Electrical Engineering Conference (EECON)*. 493-496, December 1999.

[5] Tungthangthum, A. "*Tone Recognition for Thai*" *Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems*. Chiangmai, Thailand, pp. 157-160, November 1998.

[6] Deller, J., Proakis, J., Hansen. J. *Discrete-Time Processing of Speech Signals*. Macmillian Publishing Company, New York, 1993.

[7] Luksaneeyanawin, S. "Linguistics Research and Thai Speech Technology". *Proceeding of the International Conference on Thai Studies*, School of Oriental and African Studies, University of London, United Kingdom, July 1993.