

Recognition of Intonation Patterns in Thai Utterance

Patavee Charnvivit¹, Nuttakorn Thubthong², Ekkarit Maneenoi¹, Sudaporn Luksaneeyanawin², and Somchai Jitapunkul¹

¹Digital Signal Processing Research Laboratory, Department of Electrical Engineering,

²Centre for Research in Speech and Language Processing, Department of Linguistics

Chulalongkorn University, Bangkok, Thailand

E-mail: patavee@chula.com Nuttakorn.T@chula.ac.th ekkarit@chula.com
Sudaporn.L@chula.ac.th jsomchai@chula.ac.th

Abstract

Thai intonation can be categorized as paralinguistic information of F0 contour of the utterance. There are three classes of intonation pattern in Thai, the Fall Class, the Rise Class, and the Convolution Class. This paper presents a method of intonation pattern recognition of Thai utterance. Two intonation feature contours, extracted from F0 contour, were proposed. The feature contours were converted to feature vector to use as input of neural network recognizer. The recognition results show that an average recognition rate is 63.4% for male speakers and 75.4% for female speakers. The recognizer can recognize the Fall Class from the others better than distinguish between the Rise Class and the Convolution Class.

1. Introduction

Like many languages, the information from F0 contour in Thai speech can be categorized into three types, linguistic, paralinguistic, and nonlinguistic information. Linguistic information in Thai is known as 5 lexical tones. Nonlinguistic information is composed of the factors such as physical and emotional states of the speaker [1]. Paralinguistic information represents different types of attitudinal meanings of the sentence. This type of information can be realized as intonation in Thai.

There are 3 classes of intonation patterns or tune in Thai continuous speech [2]. The class that the F0 at the beginning of the utterance is higher than the F0 at the end of the utterance was defined as “the Fall Class” or “the Downdrift”. This class of intonation pattern is found in utterance types of, for example, statement, citation form, submissive, concealed anger, and bored. The next class is “the Rise Class”, which can be found in sentence modes such as question, disagreeable, disbelieving, surprised, and unfinished. The last class, “the Convolution Class”, proposed in [2], has the shape of F0 contour that is the combination of Falls and Rises. This type of intonation can be found in utterances marked with emphatic, agreeable, interested, or believing attitudes.

This paper presents a method of extracting feature vectors from F0 contour. These feature vectors were used as input of neural network to classify the intonation shape of Thai utterance into three classes as mention above.

2. The Feature Contours

2.1. Feature contours extraction method

In order to extract paralinguistic information from F0 contour, we have to reduce (or eliminate) the effect of linguistic and nonlinguistic information. First of all, we view F0 contour as the superposition of three components in the same way as the Fujisaki-model of tonal language. These components are: the phrase component, which was directly affected by intonation type. The tone component, which corresponds to lexical tone type of each syllable. And the Fb, which is a constant for each speaker. So the mainly feature we have to extract is the phrase component, which may declines in declarative intonation and rises in interrogative intonation. The tone component, which seem to be directly correspond to linguistic information. It was, however, also affected by intonation type. This is because the range of F0 contour of the same tone is different in different intonation [2]. So we have to include the feature that represents the range of tone component but does not represent the shape of each tone. The Fb, the factor of F0 level, is also another feature that we have to include. This is because; the level of F0 of the utterances may different in different intonation although the same speaker spoke them.

So we present four steps to extract intonation features from speech as below:

1. Extract F0 contour from the utterance. In this work, we used PRAAT program (© P.Boersma) to do this task.

2. Apply error correction to F0 contour by using median filtering technique. Then, all unvoiced region in F0 contour will be filled by performing linear interpolation. We call the contour after passing this process as ‘connected F0’ or CF0, as can be seen in Figure 1.

3. In order to decompose the phrase component and Fb from the tone component, we use the technique, which is similar to the technique that [3] used to separate the accent component from the phrase component and Fb. That is, filter the CF0 by using FIR low-pass filter to get the slow movement component, which is the summation of the phrase component and Fb. In this work, we varied the cutoff frequency of the filter (LFC_Fc) from 0.5 to 3.0 Hz. The output of low-pass filter was called ‘low frequency contour’ or LFC. An example of the LFC compare with CF0 is shown in Figure 2.

4. Then the CF0 was subtracted by LFC to get the faster movement component, ‘high frequency contour’ (HFC), which is the tone component of F0 contour. As mention above, we do not want to get the shape of F0, which

corresponds to lexical tone. We just only want to find the feature that corresponds to the swinging range of the tone component. So the HFC is taken an absolute, then filtered by low-pass filter to get contour that is analogous to swinging range of tone component. Cutoff frequency of the filter, FVC_Fc, was varied in the range 0.5 to 2.5 Hz. This contour is called 'F0 variation contour' or FVC, as shown in Figure 3.

Now the LFC and the FVC was used as the intonation feature contours to use in our experiment. However, for easily understanding of how the feature contours relate to the raw F0, LFC, LFC + FVC, and LFC - FVC were plotted in the same graph with the raw F0 (see Figure 4 - 7).

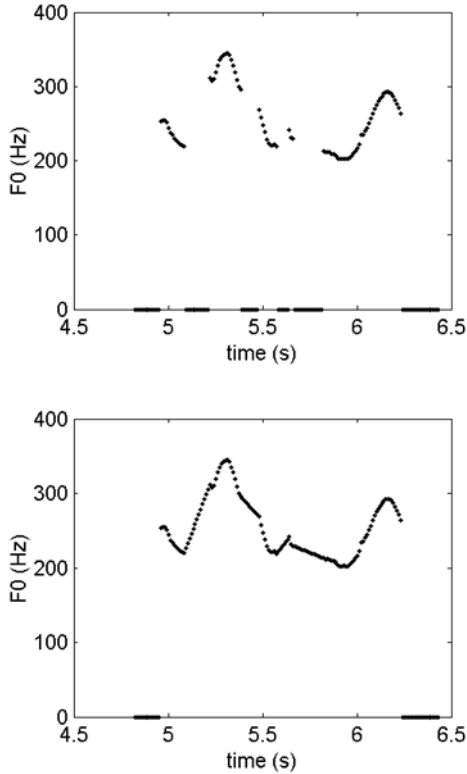


Figure 1: Top: Example of F0 contour after median filtering process Bottom: CF0

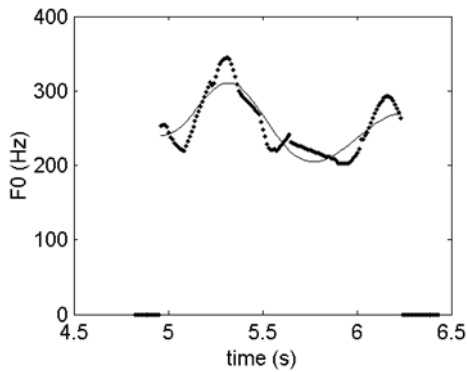


Figure 2: CF0 (dotted) and LFC (solid line) where LFC_Fc = 2.0 Hz

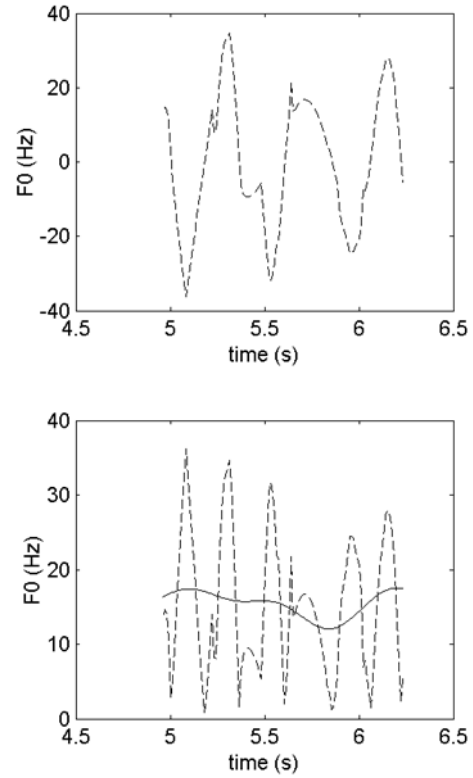


Figure 3: Top: HFC which is $CF0 - LFC$ Bottom: $|HFC|$ (dashed) and FVC (solid line) where FVC_Fc = 2.0 Hz

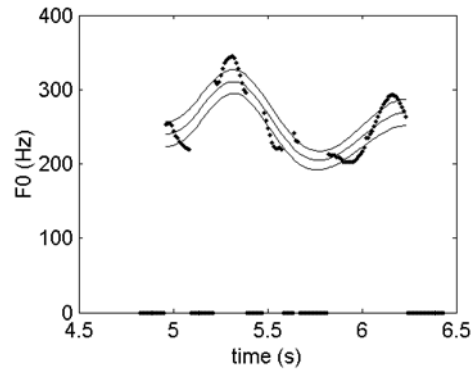


Figure 4: F0 contour (dotted) in Figure 1 compare with LFC (middle solid line), LFC + FVC (top solid line), and LFC - FVC (bottom solid line)

2.2. Examples of feature contour in different intonation class

As shown in Figure 5 – 7, the LFC is falling in the Fall Class, rising in the Rise Class, and both rising and falling in the Convolution Class. LFC + FVC and LFC – FVC are close to the LFC in the Fall Class, but they are far from the LFC in the Rise and the Convolution Class.

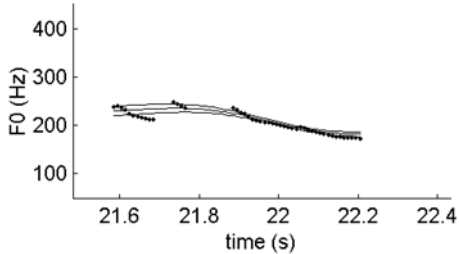


Figure 5: Example of raw F0 contour (dotted) compared with LFC (middle solid line), LFC + FVC (top solid line), and LFC – FVC (bottom solid line) of the Fall Class spoken by female speaker (LFC_Fc = 2.0 Hz, FVC_Fc = 2.0 Hz)

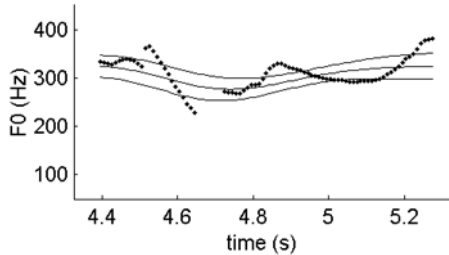


Figure 6: Example of raw F0 contour (dotted) compared with LFC (middle solid line), LFC + FVC (top solid line), and LFC – FVC (bottom solid line) of the Rise Class spoken by female speaker (LFC_Fc = 2.0 Hz, FVC_Fc = 2.0 Hz)

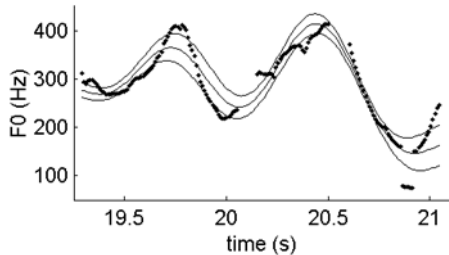


Figure 7: Example of raw F0 contour (dotted) compared with LFC (middle solid line), LFC + FVC (top solid line), and LFC – FVC (bottom solid line) of the Convolution Class spoken by female speaker (LFC_Fc = 2.0 Hz, FVC_Fc = 2.0 Hz)

3. Experiments

We used 61 sentence scripts from 6 spoken dialogues that were used in Thai intonation study [4]. There are 6 male speakers and 6 female speakers speaking these dialogues, so there are 732 sentences (366 sentences from male speakers and 366 sentences from female speakers) for our experiments. We treated the utterance from male and female speakers

separately in all experiments to discard the problem of frequency scale normalization. Although each sentence script was labeled by intonation type, some speaker may speak some sentence with the different intonation from the labeled intonation type. So we have to do a perception test to determine the perceived intonation of each sentence.

3.1. Perception experiment

Two listeners were asked to listen to all 732 sentences randomly for 2 times and label the intonation type of each sentence. So, each sentence was listened for 4 times. We found that some sentences were clearly categorized the intonation type, i.e. it was labeled as the same intonation type at least 3 times. However, some sentences were ambiguously classified the intonation type, i.e. they were labeled as the same intonation type less than 3 times. So we used only the clearly sentences for the next experiment. We labeled each clearly sentence as the most frequently intonation type chosen by listeners. Table 1 lists the number of the clearly sentences of each intonation type and gender.

Table 1: The number of the sentences that were clearly classified of each intonation type and gender by listeners

Intonation type	Gender	
	Male	Female
The Fall Class	98	106
The Rise Class	91	65
The Convolution Class	79	122

3.2. Automatic recognition experiment

We used three-layer feedforward neural network as a recognizer. The number of input nodes depended on the number of dimensions of feature vectors. The feature vectors were built from the value of LFC and FVC contour at 0%, 25%, 50%, 75%, and 100% time points of the duration of the contour. So there are 10 input nodes in this experiment. The number of hidden nodes is 20. Each output node represents each intonation class, so the number of output nodes is 3. LFC_Fc was set to 0.5, 1.0, 1.5, 2.0, 2.5 and 3.0 Hz. FVC_Fc was set to 0.5, 1.0, 1.5, 2.0, and 2.5 Hz. The NNs were trained and tested for all possible combinations of FVC_Fc and LFC_Fc. Average recognition rate of each combination of male and female speakers were shown in Table 2 and 3 respectively.

Table 2: Average recognition rates (%) of male speakers of all possible combination of LFC_Fc and FVC_Fc

FVC_Fc (Hz)	LFC_Fc (Hz)					
	0.5	1.0	1.5	2.0	2.5	3.0
0.5	56.7	54.9	57.5	60.1	63.4	60.1
1.0	61.9	57.5	60.8	60.8	60.8	59.0
1.5	58.6	63.4	62.3	61.9	61.6	59.7
2.0	61.6	61.6	60.1	60.1	62.7	58.6
2.5	62.3	59.7	62.3	61.9	60.1	57.5

Table 3: Average recognition rates (%) of female speakers of all possible combination of LFC_Fc and FVC_Fc

FVC_Fc (Hz)	LFC_Fc (Hz)					
	0.5	1.0	1.5	2.0	2.5	3.0
0.5	67.2	69.3	70.3	71.3	70.0	72.0
1.0	67.9	67.9	71.3	71.3	70.0	73.4
1.5	68.6	67.6	73.4	71.0	71.7	72.4
2.0	68.6	66.2	74.4	75.4	68.9	73.4
2.5	65.9	67.6	71.7	73.0	72.7	72.0

Table 2 and Table 3 show that the average recognition rates of female speakers are higher than the recognition rate of male speakers. The value of LFC_Fc and FVC_Fc that yield the maximum recognition rate are also different between male and female.

Then, from Table 2 and 3, we picked the combination that the pair of LFC_Fc and the FVC_Fc give the maximum recognition rate. For male, the maximum recognition rate occurs at LFC_Fc = 1.0 Hz, and FVC_Fc = 1.5 Hz. For female, it occurs at LFC_Fc = 2.0 Hz, and FVC_Fc = 2.0 Hz. Confusion matrices of the maximum recognition rate experiments of male and female speakers were shown in Table 4 and 5, where ‘‘F Class’’ is the Fall class, ‘‘R Class’’ is the Rise Class, and ‘‘C Class’’ is the Convolution Class.

Table 4: Confusion matrix of the highest recognition rate of male speakers from Table 2

Input intonation class	Recognized intonation class			Total utterances
	F Class	R Class	C Class	
F Class	77.6	12.2	10.2	98
R Class	17.6	56.0	26.4	91
C Class	15.2	30.4	54.4	79
Total				268

Table 5: Confusion matrix of the highest recognition rate of female speakers from Table 3

Input intonation class	Recognized intonation class			Total utterances
	F Class	R Class	C Class	
F Class	88.7	1.9	9.4	106
R Class	15.4	50.8	33.8	65
C Class	9.0	13.9	77.0	122
Total				293

From Table 4 and 5, we found that the recognizers were easy to distinguish the Fall Class from the other classes. We also found that the recognizers have a confusion to recognize between the Rise Class and the Convolution Class.

4. Conclusions

This paper presents a Thai intonation features extracting method from F0 contour. The method uses the fact that there are three types of information in F0 contour of Thai utterance. Extracting one type of information is to eliminate or reduce the effect of the other two types. The Thai intonation can be categorized as paralinguistic information. So we have to eliminate linguistic and nonlinguistic information. We

proposed a method of eliminating the linguistic information from F0 contour to get the LFC, and the FVC contour. These contours were used as feature contours to recognize intonation by neural network. The effect of nonlinguistic function was reduced by treating speech from male and female speakers separately. The results show that the maximum recognition rate of male utterances is 63.4%, which is lower than 75.4% of female utterances. The recognizers were easy to distinguish the Fall Class from the other classes, but difficult to recognize between the Rise Class and the Falling Class.

5. Acknowledgements

The authors would like to thank to Dr. Rachod Thongprasirt, and Dr. Satien Triamlamlert from National Electronics and Computer Technology Center (NECTEC) for helpful comments and suggestions. The authors would like to acknowledge the Thailand Graduate Institute of Science and Technology (TGIST) for partial financial support for the research to the first author, and Centre for Research in Speech and Language Processing, Department of Linguistics Chulalongkorn University (CRSLP) for providing facilities.

6. References

- [1] Fujisaki, H., ‘‘Prosody, Models, and Spontaneous Speech’’. *Computing Prosody (Sagisaka, Y., Campbell, N., and Higuchi, N., eds.)*, Springer-Verlag (1996) 27-42.
- [2] Luksaneeyanawin, S., ‘‘Intonation in Thai’’. *Intonation Systems, A Survey of Twenty Languages (Hirst, D., Cristo, A. D.)*, Cambridge University Press: 376-394, 1998.
- [3] Mixdorff, H., ‘‘A Novel Approach to The Fully Automatic Extraction of Fujisaki Model Parameters’’, *Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, Istanbul, Jun. 2000, pp. 1281-1284.
- [4] Luksaneeyanawin, S., ‘‘Intonation in Thai’’. *Ph.D. Dissertation, University of Edinburgh*, 1983.