Speaker-Independent Automatic Classification of Thai Tones in Connected Speech by Analysis-Synthesis Method

Siripong Potisuk and Mary P. Harper
School of Electrical Engineering
and
Jackson T. Gandour
Department of Audiology and Speech Sciences
Purdue University, West Lafayette, IN 47907 USA

ABSTRACT

Tone classification is a crucial component of any automatic speech recognition system for tone languages. It is imperative that tonal information be incorporated into the word hypothesization process because patterns of pitch (or tones) contribute to the lexical identification of the individual words. In this paper, we present a novel algorithm for automatically classifying Thai tones in connected speech using an analysis-synthesis method based on an extension to Fujisaki's model. We have successfully incorporated into the model two major factors affecting the phonetic realization of tones in connected speech: tonal coarticulation and declination. Also addressed is an F_0 normalization procedure for achieving speaker-independence. In our preliminary experiment, we were able to achieve 89.1% classification accuracy.

1. INTRODUCTION

Prosody, which is defined as the change in the fundamental frequency (F_0), timing, and intensity of a speech signal, provides crucial information for decoding that signal. It can be used to signal linguistic information at the syllable, word, or sentence level, or it can be used to indicate non-linguistic information such as attitude or emotion. This paper focuses only on linguistic prosody. To build a system which effectively utilizes linguistic prosody requires a better understanding of stress, intonation, and tone.

The language chosen for the current investigation is Thai, a tone language. Thai has five tones: mid (M), low (L), falling (F), high (H), and rising (R). Fig. 1.1 shows average F_0 contours of the five Thai tones in isolation. Tones, which are indicated by contrastive variations in F_0 at the syllable level, are used to signal differences in lexical meaning. As a result, it is imperative that tonal information be incorporated into a Thai speech recognition system.

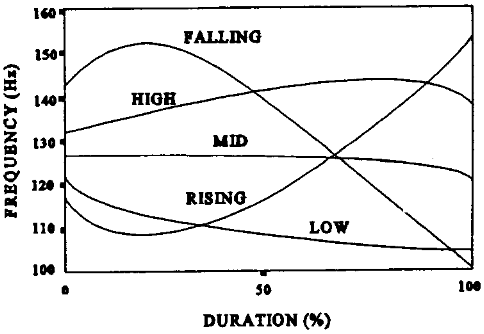

Fig. 1.1. Average F0 contours of the five Thai tones in isolation (adapted from Abramson [1]).

Automatic tone classification has been investigated by several researchers [2,3,4]. Unfortunately, their studies were based on words in isolation. To our knowledge, no attempt has been made to classify tones in connected speech. In this paper, an algorithm using an analysis-synthesis approach based on an extension to Fujisaki's model is proposed to automatically classify Thai tones in connected speech. The model attempts to take into account many factors affecting the phonetic realization of Thai tones including tonal coarticulation, stress, and declination. Also discussed are important considerations for normalizing F_0 contours of Thai tones to achieve speaker-independence.

2. THE PROPOSED ALGORITHM

The problem of tone classification in Thai can simply be stated as finding the best sequence of tones, T_1, T_2, ... , T_n, given an input speech signal. Because the primary acoustic correlate of tone is F_0 and Thai has five distinct F_0 contour patterns, the problem is to find the best possible combination of F_0 contour patterns that closely match the given input F_0 contour.

The general design of this pattern recognition system involves two major steps: feature extraction and pattern

0-7803-2431-5/95 $4.00 © 1995 IEEE

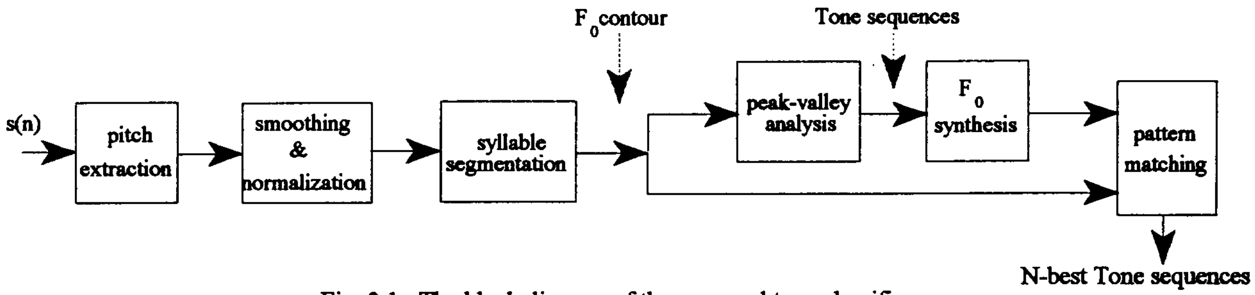

Fig. 2.1. The block diagram of the proposed tone classifier.

matching. Fig. 2.1 shows the block diagram of the proposed tone classifier. The first three blocks represent the feature extraction step necessary to produce relatively reliable, normalized F_0 contours. The last three blocks represent the pattern matching step based on the analysis-synthesis method. This approach can be regarded as a kind of bottom-up or data-driven approach. Each component of the system will be described in detail below.

2.1 Pitch Extraction.

The raw F_0 contour is automatically extracted from the input speech signal. In this paper, we used a time domain approach to pitch analysis (modified autocorrelation with center clipping) with nonoverlapping variable frame length.

2.2 Smoothing

Since pitch extraction from the speech signal is a difficult task to perform automatically, the resulting raw F_0 contours often contain "drop-outs" during voiced speech segments or spurious pitch values in regions of unvoiced speech segments. Lengthy stretches of aperiodicity due to creakiness also contribute to a phenomenon in pitch extraction called "double pulsing" in which the extracted pitch values are twice the actual value. In order to reduce these undesirable discontinuities in the extracted F_0 contour, smoothing techniques are employed. We used median filtering and linear interpolation.

2.3 Normalization

For a multiple-speaker system, the normalization process is introduced to neutralize variability from one F_0 contour to the next. Sources of variability include speaker's physiological differences, the kinetics of vocal fold vibration, consonantal perturbations on F_0, and speaking rate.

The raw F_0 contour is first converted into an equivalent-rectangular-bandwidth-rate (ERB) scale. An ERB scale is a psychoacoustic scale that gives equal prominence to excursions in different pitch registers based on the human's perception of F_0 [5]. Then, a z-score normalization is employed to account for pitch range differences across speakers based on the precomputed mean and standard deviation from all utterances in the training set. This method has the effect of making the first- and second-order moments of the pitch distributions the same [6]. To neutralize the declination effect in the F_0 contour (see 2.6 for a definition of declination), we subtract a time-varying mean F_0 value from the input F_0 contour. A time-varying mean F_0 value is computed by fitting an exponential curve to the overall contour according to Fujisaki's model.

2.4 Syllable Segmentation

Since tones are properties of syllables, it is logical to segment the smoothed and normalized F_0 contour into syllabic units. Syllable boundary information can be provided by an automatic syllable segmentation algorithm based on energy contours and spectral information, or by segmentation information from a phone recognizer unit. Since neither is readily available for Thai, syllable boundaries were obtained from hand-labeled segmentation information.

2.5 Peak-and-Valley Analysis

Given the smoothed, normalized and segmented F_0 contour, local extrema (peaks and valleys) are detected by using first and second derivatives. The derivative at any point in the contour, except for the first and last two points, is computed by calculating the linear regression coefficients of a group of five F_0 values consisting of the current point, and its preceding and following two points. The locations of these extrema coupled with syllable boundary information are then used to determine all possible tone sequences based on some specified rules. For example, between two syllable boundaries, only the falling tone can occur if a maximum occurs, and only the rising or the high tone can occur if a minimum occurs. Also, if a maximum occurs at or in the vicinity of a syllable boundary, the preceding tone can either be a high or a rising tone. If a minimum occurs at or in the vicinity of a syllable boundary, the preceding tone can either be a mid or a low tone, or a sequence of two falling tones.

2.6 Fundamental Frequency Synthesis

Based on the extension to Fujisaki's model for synthesizing F_0 contours of tone languages, the input tone sequences are used to generate predicted F_0 contours. These predicted F_0 contours are basically reference templates to be used for pattern matching against the input contour. The model is trained by empirically estimating its parameters from hand-labeled utterances comprising the training set. During the training of the model, we can incorporate linguistic constraints, such as tonal coarticulation, declination, and stress effects.

Tonal coarticulation is defined as the change in the F_0 realization of tones depending on the influence from adjacent tones in connected speech. This blending occurs between F_0 contours of tones (similar to phone coarticulation effects) because of the continuous change in pitch levels. Sudden jumps from one pitch level to another seldom occur. Two types of tonal coarticulation are possible: anticipatory (right-to-left influences from the following tone) and perseverative (left-to-right influences from the preceding tone).

Declination refers to a gradual lowering of F_0 over the course of a phrase or utterance against which the phonologically specified local F_0 targets are scaled. The declination effect plays an important role in tone classification in terms of F_0 height adjustment of tones. For example, a falling tone at the beginning of an utterance is higher in pitch than at the end.

Stress effects can also be quantified in terms of the average height and shape of F_0 contours. The contours of stressed syllables more closely approximate the contours in citation forms than those of unstressed syllables. Also, the terminal portion of F_0 contours of unstressed syllables is "chopped-off", i.e., they do not reach their target pitch levels, in contrast to those of syllables uttered in isolation [7]. In this study, only tonal coarticulation and declination are incorporated into the model. A lack of concrete data on Thai stress prevents us from taking into account the stress effects.

Fujisaki's model is based on the idea of approximating F_0 contours as being the response of a second order linear system due to excitation commands. The model for tone languages consists of two components: phrase or utterance command and tone command. The phrase command is used to capture the declination effect; the tone command is used to capture tone types, tonal coarticulation, and stress effects. The response to each command is characterized by a time constant parameter, the command amplitudes and its temporal locations, and the damping coefficient (critical damping for phrase command). The phrase command will produce the "mid" line. The tone commands vary in both positive and negative directions with respect to the mid line to produce local contours corresponding to tone types.

2.7 Pattern Matching

The classification of input F_0 contours into likely sequences of tones is accomplished in this step by pattern matching against the predicted F_0 contours or reference templates generated by the F_0 model. Pattern matching techniques, such as a simple zero-lag crosscorrelation method or a one-stage dynamic programming search can be used. In both cases, some measure of goodness of fit must be established in order to rank the results so that N-best tone sequences can be obtained. For example, for the zero-lag crosscorrelation method, a correlation coefficient of 0.9 or higher could be used to indicate a relatively good fit. Thus, we can infer that a strong similarity exists between the input and the predicted F_0 contours. For a one-stage dynamic programming search, a distance measure might be more appropriate. In this paper, we used the zero-lag crosscorrelation method.

3. EXPERIMENT

In this section, we describe our preliminary experiment and the results.

3.1 Method

3.1.1 Subjects

Five native speakers of Thai (three males and two females) at Purdue University participated in this study. All five subjects were native speakers of the Bangkok Thai dialect and were naive with respect to the purpose of the experiment.

3.1.2 Stimuli

There were two sets of stimuli: the training and test set. For the training set, stimuli consisted of 125 possible three-tone sequences of five Thai tones superimposed on monosyllabic words in a carrier sentence. Each sentence contained four monosyllabic words. The last word (L) is held constant while the first three words were varied to give all 125 three-tone sequences. The stress pattern of the carrier sentence (- - - -) was invariant in order to eliminate the potentially confounding interaction between stress and tone. The last word is stressed by virtue of being in a sentence-final position. To enhance coarticulatory effects, all four words began and ended with a sonorant. Thus, the sentence was continuously voiced throughout. For the test set, stimuli consisted of 11 sentences with varying tone sequences (e.g., FFFF, RRRR, LFHR, FHRL, LHLL, FFHH, etc.). Tone sequences were chosen to represent varying degrees of movement in the F_0 contour. Each sentence contained four monosyllabic words and was also continuously voiced throughout.

3.1.3 Recording Procedure

Speakers were asked to read a target sentence typed in Thai script on a 4 X 6 in. card. They were instructed to produce the sentences naturally at a conversational speaking rate, i.e., at a rate they considered representative of their conversational speech. A random order of presentation was used to minimize changes in speaking rate. For each speaker, the total corpus contained 125 utterances (125 tonal combinations) for the training set and 11 utterances for the test set.

3.1.4 Measurement Procedure

The stimuli were digitized at a sampling rate of 20 KHz using the KAY CSL (Computerized Speech Lab) Model 4300. F_0 was computed directly from the waveform using a time domain approach to pitch analysis (modified autocorrelation with center clipping) with nonoverlapping variable frame length. Syllable onset and offset were determined from a simultaneous display of a wide-band spectrogram with a scale from 0-8 KHz, energy and F_0 contours, and the speech waveform.

3.2 Results

The Fujisaki's model was first trained using 625 utterances from the training set to obtain the time constant parameters and the tone command amplitudes and temporal locations for all 125 tone sequences. Then, classification test is performed on each of the 55 utterances from the test set to obtain the crosscorrelation coefficients between the input contour and each of the predicted contours. Table 3.2.1 shows the crosscorrelation coefficients for the input sequence FHRL. Note that for speaker #3, the algorithm was able to identify the correct tone sequence despite low coefficient values.

Table 3.2.1. Correlation coefficients for the input sequence FHRL.

speaker	predicted tone sequence			
no.	FHRL	FHRM	FRRL	FRRM
1	0.8921	0.8165	0.6717	0.5982
2	0.8443	0.6131	0.7574	0.5174
3	0.5616	0.4009	0.5319	0.3693
4	0.9223	0.8329	0.7943	0.6778
5	0.9096	0.8195	0.7319	0.6488

All in all, the algorithm misclassified 6 of the 55 test utterances. Hence, the classification accuracy for this experiment is approximately 89.1%.

4. DISCUSSION

A bottom-up or data-driven approach to tone classification of Thai tones in connected speech has been described. The algorithm is relatively simpler to implement than the left-to-right HMM-based system implemented in [3]. Although tonal coarticulation was incorporated into the HMM model in [3], the system performance was only improved slightly. However, this preliminary study suggests that our novel algorithm is potentially simpler to use and more accurate. Also, we believe that the computational cost of our model is much less than the HMM-based system because it uses fewer parameters.

From this experiment, we observe that some of the misclassified tones belong to the sentence-final syllables. Having ignored the stress effects, this misclassification is to be expected because sentence-final syllables in Thai exhibit a higher level of prominence than the others in the sentence due to prepausal lengthening. Thus, three degrees of stress (unstressed, stressed, and sentence-final stressed) must be incorporated into the model. We further observe that misclassification occurs between a mid and a low tone. However, this misclassification is unidirectional (M → L). It is linguistically plausible because the mid tone has a greater "degree of freedom" to manifest itself in the tone space (see also Fig. 1.1) than the low tone.

ACKNOWLEDGEMENTS

This material is based upon work supported in part by NIH under Grant No. DC00515-07. The first author, who is currently on leave of absence from the Academic division of Chulachomklao Royal Military Academy, extends his gratitude to the Royal Thai Army.

REFERENCES

[1] A. S. Abramson, "The vowels and tones of standard Thai: acoustical measurements and experiments," *International Journal of American Linguistics,* vol. 28-2, Part II (Publication No. 20). 1962.

[2] X. Chen, C. Cai, P. Guo, S. Ying, "A hidden Markov model applied to Chinese four-tone recognition," in *1987 ICASSP,* Vol.2, May 1987 pp. 787-800.

[3] L. Liu, W. Yang, H. Wang, and Y. Chang, "Tone recognition of polysyllabic words in Mandarin Speech," *Computer Speech and Language.* vol. 3, pp. 253-264, 1989.

[4] R. Wu, J.A. Orr, S-K. Hsu, "Recognition of four tones in Chinese speech by parametric estimation of frequency trajectories," in *Second Biennial Acoustics, Speech and Signal Processing Central New England Miniconference,* 1989.

[5] D. Hermes and J. Van Gestel, "The frequency scale of speech intonation," *Journal of the Acoustical Society of America,* vol 9, pp. 97-102, 1991.

[6] P. Rose, "Considerations in the normalisation of the fundamental frequency of linguistic tone," *Speech Communication,* vol. 6, pp.343-351, 1987.

[7] S. Potisuk, J. T. Gandour, and M. P. Harper, "F_0 correlate of stress in Thai," *Linguistics of The Tibeto-Burman Area,* in press.