S. Potisuk, M. P. Harper, and J. Gandour, "Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method," IEEE Trans. Speech Audio Processing, vol. 7, pp. 95 - 102, January 1999.



[2] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1365–1376, Oct. 1987.

[3] J. E. Greenberg and P. M. Zurek, "Evaluation of an adaptive beamforming method for hearing aids," *J. Acoust. Soc. Amer.*, vol. 91, pp. 1662–1676, Mar. 1992.

[4] M. Hoffman and K. Buckley, "Robust time-domain processing of broadband microphone array data," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 193–203, May 1995.

[5] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78, pp. 1508–1518, 1985.

[6] R. Zelinski, "A microphone array with adaptive postfiltering for noise reduction in reverberant rooms," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1988, pp. 2578–2581.

[7] S. Affes and Y. Grenier, "A source subspace tracking array of microphones for double talk situations," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 1996, vol. 2, pp. 909–912.

[8] S. Fischer, K. Kammeyer, and K. U. Simmer, "Adaptive microphone arrays for speech enhancement in coherent and incoherent noise fields," in *Proc. 3rd Joint Meeting Acoustical Society of America and the Acoustical Society of Japan*, Honolulu, HI, Dec. 1996.

[9] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, July 1995.

[10] S. H. Jensen *et al.*, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 439–448, Nov. 1995.

[11] K. U. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," in *Proc. 2nd Cost 229 Workshop on Adaptive Algorithms in Communications*, Bordeaux-Technopolis, France, Sept. 30–Oct. 2 1992.

[12] K. M. Buckley, "Spatial/spectral filtering with linearly constrained minimum variance beamformers," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 249–266, Mar. 1987.

[13] M. W. Hoffman, T. D. Trine, K. M. Buckley, and D. J. Van Tasell, "Robust adaptive microphone array processing for hearing aids: Realistic speech enhancement predictions," *J. Acoust. Soc. Amer.*, vol. 96, pp. 759–770, 1994.

[14] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[15] A. H. Gray, Jr. and J. D. Markel, "Distance measure for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 380–391, Oct. 1976.

[16] M. J. Link and K. M. Buckley, "Pre-whitening for intelligibility gain in hearing aid arrays," *J. Acoust. Soc. Amer.*, vol. 93, pp. 2139–2145, Apr. 1993.

# Classification of Thai Tone Sequences in Syllable-Segmented Speech Using the Analysis-by-Synthesis Method

Siripong Potisuk, Mary P. Harper, and Jack Gandour

***Abstract*—Tone classification is important for Thai speech recognition because tone affects the lexical identification of words. An analysis-by-synthesis algorithm for classifying Thai tones in syllable-segmented speech is presented that uses an extension to Fujisaki's model for tone languages that incorporates tonal assimilation and declination. The classifier correctly identifies all of the tones in 89.1% of the test utterances.**

***Index Terms*—Analysis-by-synthesis, intonation, lexical tone classification, speech processing, spoken Thai, tonal assimilation.**

## I. Introduction

Tone, which is indicated by contrasting variations in $F_0$ at the syllable level, is an important part of a speech understanding system for tone languages (e.g., Chinese, Thai) because it signals differences in lexical meaning. This correspondence describes a tone classification algorithm for Thai tones. Thai has five contrasting lexical tones traditionally labeled mid (M), low (L), falling (F), high (H), and rising (R). The following examples illustrate the effect that tone has on meaning: M /k$^h$aa/ ("to get stuck"); L /k$^h$àa/ ("a kind of spice"); F /k$^h$âa/ ("to kill"); H /k$^h$áa/ ("to engage in trade"); and R /k$^h$ǎa/ ("leg"). Average $F_0$ contours of the five Thai tones, as produced in isolation, are presented in Fig. 1 [1]. Perceptual investigations have revealed that $F_0$ height and shape carry sufficient information for high intelligibility of Thai tones [10]. The problem of tone classification in Thai speech can be simply stated as finding the best sequence of tones given an input speech signal. Since tone is a property of the syllable, each tone is associated with a syllable of the utterance. Because the primary acoustic correlate of tone is $F_0$ and Thai has five distinct $F_0$ contours, the problem is to find the best possible combination of $F_0$ tonal contour patterns that closely match the input $F_0$ contour. Fig. 2 compares the $F_0$ realization of an FHRL sequence when each monosyllabic word is spoken in isolation (see top panel) to when all four tones are spoken naturally in running speech (see bottom panel). The tones produced on isolated words are very similar to those in Fig. 1. The tones produced on words in continuous speech are much more difficult to identify.

Several interacting factors affect $F_0$ realization of tones: syllable structure, tonal assimilation, stress, and intonation. Syllable structure affects tone classification due to the *continuity effect* on the $F_0$ contour [4] in terms of the voiced/unvoiced property of contiguous phones. A continuously voiced stretch of speech has a continuous $F_0$ contour; a stretch of speech with intervening voiceless obstruents has a discontinuous $F_0$ contour. There are also consonantally-induced perturbations on the $F_0$ contour of the following vowel [9].

Manuscript received June 29, 1996; revised May 27, 1998. This work was supported in part by the National Institute on Deafness and Other Communication Disorders, National Institutes of Health, under Grant 5 R01 DC 00515-08. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Douglas D. O'Shaughnessy.

S. Potisuk and M. P. Harper are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: harper@ecn.purdue.edu).

J. Gandour is with the Department of Audiology and Speech Sciences, Purdue University, West Lafayette, IN 47907 USA.

Publisher Item Identifier S 1063-6676(99)00179-0.

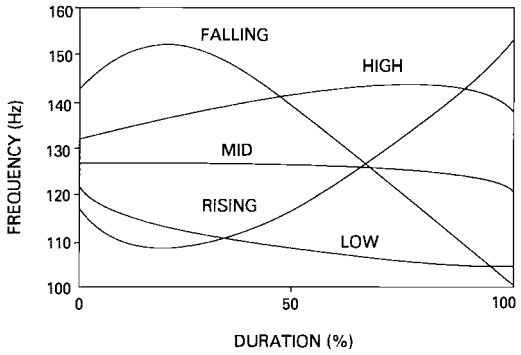



Fig. 1. Average $F_0$ contours of the five Thai tones produced in isolation by a male speaker (adapted from Abramson [1]).

Sequences of adjacent tones may influence each other phonetically. Changes in the phonetic manifestation of tones attributable to adjacent tones in continuous speech is called *tonal assimilation.* Research on tonal assimilation has received much attention in the Thai phonetics literature [2], [11], [21]. Gandour *et al.* [11] developed a quantitative procedure to continuously track the height and slope of $F_0$ contours to assess assimilation effects. In [21], Potisuk *et al.* used three-tone sequences in running speech to measure tonal assimilation effects. It was found that tonal assimilation affects both the height and shape of the tonal contours, that these effects do not extend beyond contiguous tones, and that the direction of the assimilation is primarily from left to right.

For standard Thai, stress has been consistently reported to be a rule-governed, phonetic characteristic of the language [26]. Duration appears to be the predominant cue in signaling stress in Thai [20]. Potisuk *et al.* [20] reported that, in continuous speech, despite systematic changes in $F_0$ contours, all five tonal contrasts are preserved in unstressed as well as stressed syllables. However, $F_0$ contours of stressed syllables more closely approximate the contours in citation forms than those of unstressed syllables.

Intonation, the pattern of pitch changes that extend over larger-sized linguistic units such as phrases or sentences may signal grammatical as well as attitudinal meaning. $F_0$ values of tones are affected by sentence intonation to varying degrees [7]. An important characteristic of intonation is declination [7], which refers to a gradual modification over the course of a phrase or utterance against which the phonologically-specified, local $F_0$ targets are scaled.

Concerning the interdependence of these factors, only the declination effect is hypothesized to be independent. That is, the declination effect is superimposed on the already-perturbed $F_0$ contour due to syllable structure, stress, and tonal assimilation. The temporal extent of assimilatory effects on tone are likely to be affected by syllable structure as well as stress. To systematically study the interactions among all of these factors would involve the collection of massive amounts of speech data. In this paper, we will take into account the effects of tonal assimilation and declination.

The design of a tone classifier involves $F_0$ extraction and pattern matching. Much of the tone classification research [5], [6], [14], [16], [18], [23]–[25], [27] has focused on Chinese (Mandarin, Cantonese, and standard Chinese). Several have considered only words in isolation [6], [14], [16], [18], [27]; others have hand-segmented the speech data into syllables [5]. Pattern matching is often based on hidden Markov models (HMM's); however, some use neural networks [5], [16]. No research on Thai tone classification, whether in isolation or in continuous speech, has ever been attempted. To construct a tone classifier for Thai, we adopt an abstract model of the speech perception process [12], *analysis-by-synthesis.* The major claim of this theoretical model is that listeners perceive (analyze) speech by implicitly generating (synthesizing) speech from what they have heard and then comparing the synthesized speech to the auditory stimulus. In Thai, the analysis module generates hypothesized tone sequences from the input $F_0$ contour; the synthesis module generates predicted $F_0$ contours to match against the input contour. The synthesis module of our tone classifier is based on an extension of Fujisaki's model [8] for synthesizing $F_0$ contours in tone languages.

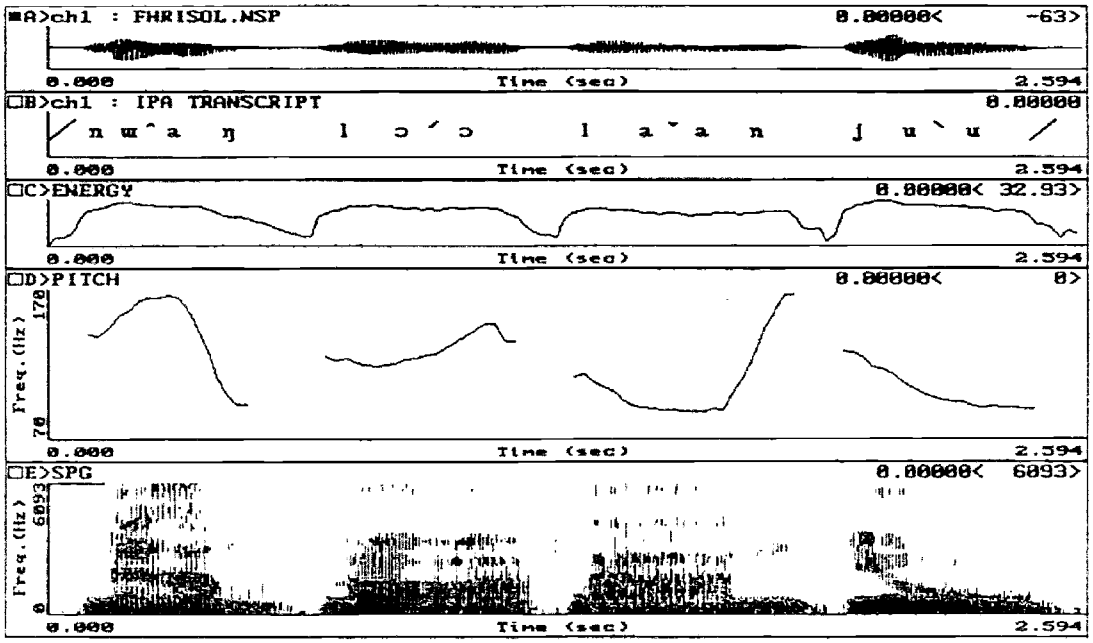
## II. An Extension of Fujisaki's Model of $F_0$ Contours to Tone Languages

Fujisaki [8] observed that an $F_0$ contour generally contains a smooth rise–fall pattern in the vicinity of the accented Japanese *mora* (a unit of syllable quantity). Differences in rise–fall patterns seem to be attributable to the accent type, and these rise–fall patterns appear to be superimposed on a base line that initially rises and gradually falls toward the end of the phrase or utterance regardless of the accent type. He hypothesized that the observed $F_0$ contour can be considered as the response of the phonatory system to a set of suprasegmental commands: the phrase (utterance) and the accent command. The phrase command produces the base line component while the accent command produces the accent component of an $F_0$ contour. Hence, Fujisaki proposed a functional model for generating an $F_0$ contour based on the idea of approximating $F_0$ contours as the response of a critically damped second-order linear system to excitation commands. The model is considered a superpositional model because it additively superimposes a basic $F_0$ value ($F_{\min}$), a phrase component, and an accent component together on a logarithmic scale. The logarithmic frequency scale is based on the biomechanical considerations of the speech apparatus. In short, the output $F_0$ contour is a linear combination of $F_{\min}$, a phrase component, and an accent component. A block diagram of the model is shown in Fig. 3. The control mechanisms of the two components are realized as critically damped second-order linear systems responding to rectangular functions. Mathematically speaking, an $F_0$ contour of an utterance generated from the model has the following functional form:

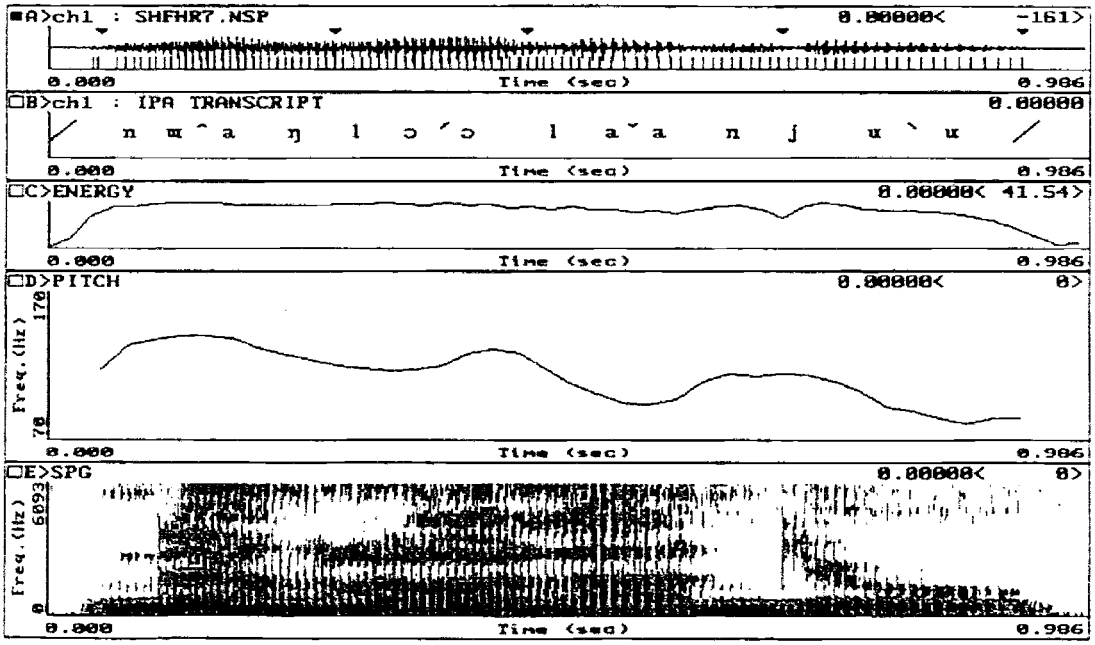
$$\ln F_0(t) = \ln F_{\min} + \sum_{i=1}^{I} A_{pi}[G_{pi}(t - T_{0i}) - G_{pi}(t - T_{3i})] + \sum_{j=1}^{J} A_{aj}[G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})] \quad (1)$$

where $G_{pi}(t) = \alpha_i t \exp(-\alpha_i t) u(t)$ and $G_{aj}(t) = [1 - (1 + \beta_j t) \exp(-\beta_j t)] u(t)$, $u(t) =$ unit step function, indicate the step response function of the corresponding control mechanism to the phrase and accent command, respectively. $F_{\min}$ is the lower limit of $F_0$ below which vocal fold vibration cannot be sustained in the glottis of a speaker. $A_{pi}$ and $A_{aj}$ are the amplitudes of the phrase and accent commands, respectively. $T_{0i}$ and $T_{3i}$ denote the onset and offset of the $i$th phrase command; $T_{1j}$ and $T_{2j}$ denote the onset and offset of the $j$th accent command. The $\alpha_i$ and $\beta_j$ are time constant parameters characterizing a second-order system. $I$ and $J$ are the number of phrases and accented mora, respectively, contained in the utterance. The damping coefficient, which also characterizes a second-order system, is unity in the case of a critically damped system.

The phrase component captures the global variation (declination effect) while the accent component captures local variations (accent effect) in the $F_0$ contour. The model is able to approximate naturally produced $F_0$ contours very accurately using only a small number of control parameters: the time constant parameters of the phrase and accent control mechanisms and the timing and amplitudes of




(a)




(b)

Fig. 2. Differences in the $F_0$ realization of tones in an FHRL sequence when (a) each word is spoken in isolation and (b) when the whole utterance is naturally spoken. From top to bottom in each display is the waveform, phonemic transcription, energy contour, $F_0$ contour, and wide band spectrogram.

the phrase and accent commands. The parameters can be empirically obtained by a curve-fitting method (i.e., minimizing the mean square error between the raw $F_0$ contour and that of the model) on a logarithmic scale. Fujisaki concluded from his experimental results that the time constant parameters and damping coefficients could be held constant without seriously affecting the resulting output $F_0$ contour.

In Japanese, the $F_0$ realization of local pitch accents results only in rise–fall patterns in the $F_0$ contour. However, in the case of Thai, local $F_0$ variations due to tones result in a combination of both rise–fall patterns (e.g., F) and fall-rise patterns (e.g., R). As a result, a model for tone languages will consist of two components, the phrase and tone control mechanisms, driven by the phrase and tone commands. The phrase command and phrase control mechanism are used to capture the declination effect; the tone command and tone control mechanism are used to capture tone types. Instead of a base line, the phrase command will produce a midline. The tone commands in both positive and negative directions with respect to the midline will produce local contours corresponding to tone types, which are superimposed on the midline. As before, the model is characterized by time constant parameters and command amplitudes and their temporal locations. Critical damping is assumed for both the phrase and tone control mechanisms; hence, the damping coefficient is always unity. Our extension of Fujisaki's model to tone languages

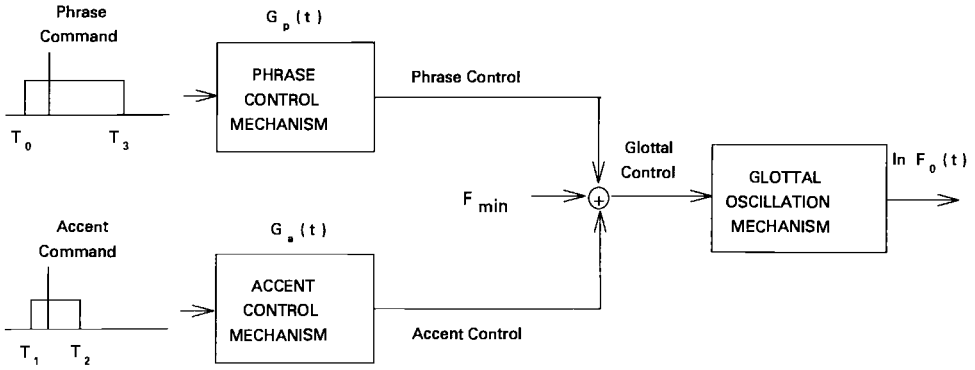



Fig. 3. Block diagram of the Fujisaki's model for synthesizing $F_0$ contours.

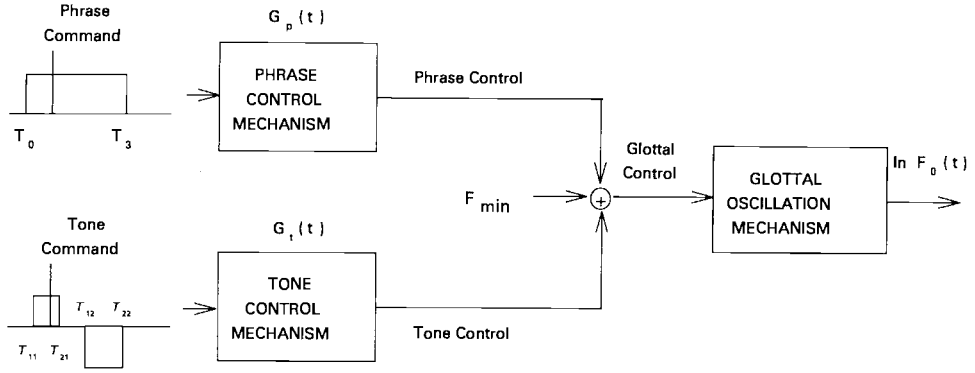



Fig. 4. Our extension of Fujisaki's original model of $F_0$ contours to tone languages.

is illustrated in Fig. 4. Analogous to that of Fujisaki's original model, the tone synthesis model has the following mathematical expression:

$$\ln F_0(t) = \ln F_{\min} + \sum_{i=1}^{I} A_{pi}[G_{pi}(t - T_{0i}) - G_{pi}(t - T_{3i})] + \sum_{j=1}^{J} \sum_{k=1}^{K(j)} A_{t,jk}[G_{t,jk}(t - T_{1jk}) - G_{t,jk}(t - T_{2jk})] \quad (2)$$

where $G_{pi}(t) = \alpha_i t \exp(-\alpha_i t)u(t)$ and $G_{t,jk}(t) = [1 - (1 + \beta_{jk}t)\exp(-\beta_{jk}t)]u(t)$, $u(t) =$ unit step function, indicate the step response function of the corresponding control mechanism to the phrase and tone command, respectively. $F_{\min}$ is the smallest $F_0$ value in the $F_0$ contour of interest. $A_{pi}$ and $A_{t,jk}$ are the amplitudes of the phrase and tone command, respectively. $T_{0i}$ and $T_{3i}$ denote the onset and offset of the $i$th phrase command. $T_{1jk}$ and $T_{2jk}$ denote the onset and offset of the $k$th component of the $j$th tone command. $\alpha_i$ and $\beta_{jk}$ are time constant parameters characterizing a second-order system. $I$, $J$, and $K(j)$ are the number of phrases, tones, and components of the $j$th tone, respectively, contained in the utterance. It is noted that the logarithmic scale will be replaced by an equivalent-rectangular- bandwidth-rate (ERB) scale, which is comparable to the logarithmic scale [13] and offers an advantage in that it gives equal prominence to excursions in different pitch registers. This is important in the analysis of, for example, the $F_0$ contours of male and female speech.

Based on this model, an input sequence of tones can be used to generate a contour for the sequence, which can then be used as a reference pattern for our tone classification algorithm. Tones in continuous speech are affected by stress, syllable structure, tonal assimilation, and declination. These linguistic factors can be incorporated into the model in terms of the tone command amplitudes and their temporal locations. To examine all of these factors jointly would require the collection of an extensive corpus of acoustic data to train the model. Here, we focus on *tonal assimilation* and *declination*.

The sentences used for setting the parameters for the synthesis module consisted of the 125 possible three-tone sequences from the five Thai tones, superimposed on monosyllabic words in a carrier sentence with a fixed syntactic and prosodic environment (see [21] for more detail, including the list of sentences). The syntactic frame was: subject + verb + object + post-verb auxiliary. The last word, /jùu/, which carries a low tone, was held constant while the first three words were varied to give all 125 three-tone sequences. The stress pattern of the carrier sentence (all stressed) was invariant in order to eliminate the potentially confounding interaction between stress and tonal assimilation. The last word in the sentence was stressed by virtue of being in a sentence-final position. All four words began and ended with a sonorant, and so the sentence was continuously voiced throughout. A reading task was chosen to control the syntactic, prosodic, and segmental characteristics of the spoken sentences. A reading task provides a compromise between naturalness and the needed control for this prosodic research [17]. A total of five (three male and two female) monodialectal speakers from the Bangkok metropolitan area were instructed to read target sentences at a conversational rate. A total of 625 utterances were recorded (125 sequences $\times$ 5 speakers) for the training set.

The tape-recorded stimuli were low-pass filtered at 10 kHz and digitized at a sampling rate of 20 kHz by means of a 16-b A/D converter with a 5-V dynamic range using the Kay CSL (computerized speech

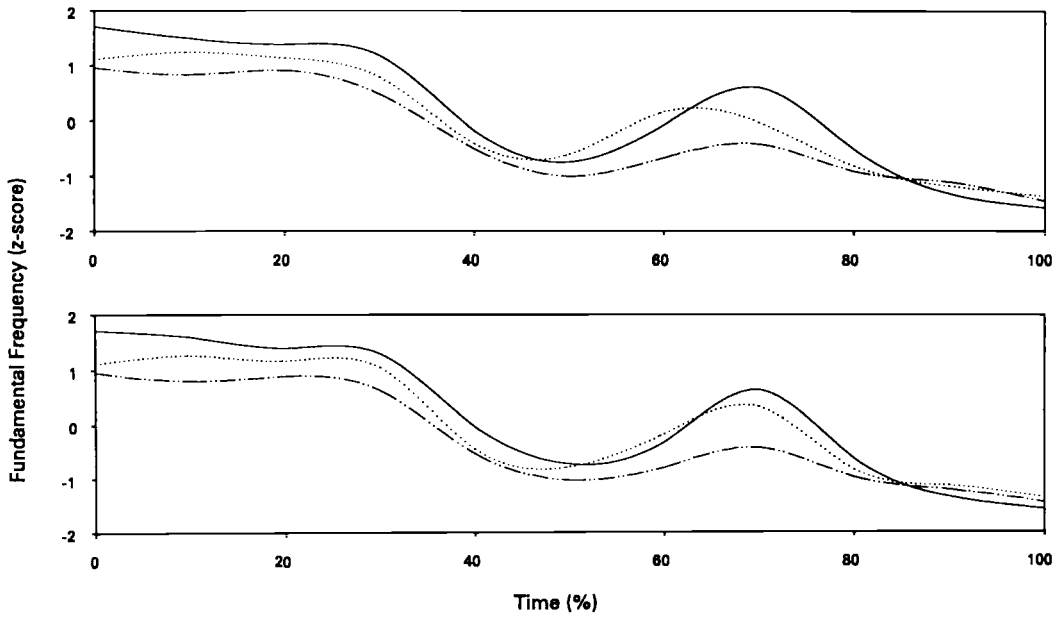



Fig. 5. $F_0$ contours of the same utterance by three different speakers, time-normalized without (top panel) and with (bottom panel) syllable by-syllable-temporal alignment. The vertical axis is in terms of a z-score scale which normalizes the $F_0$ values to neutralize interspeaker variability.

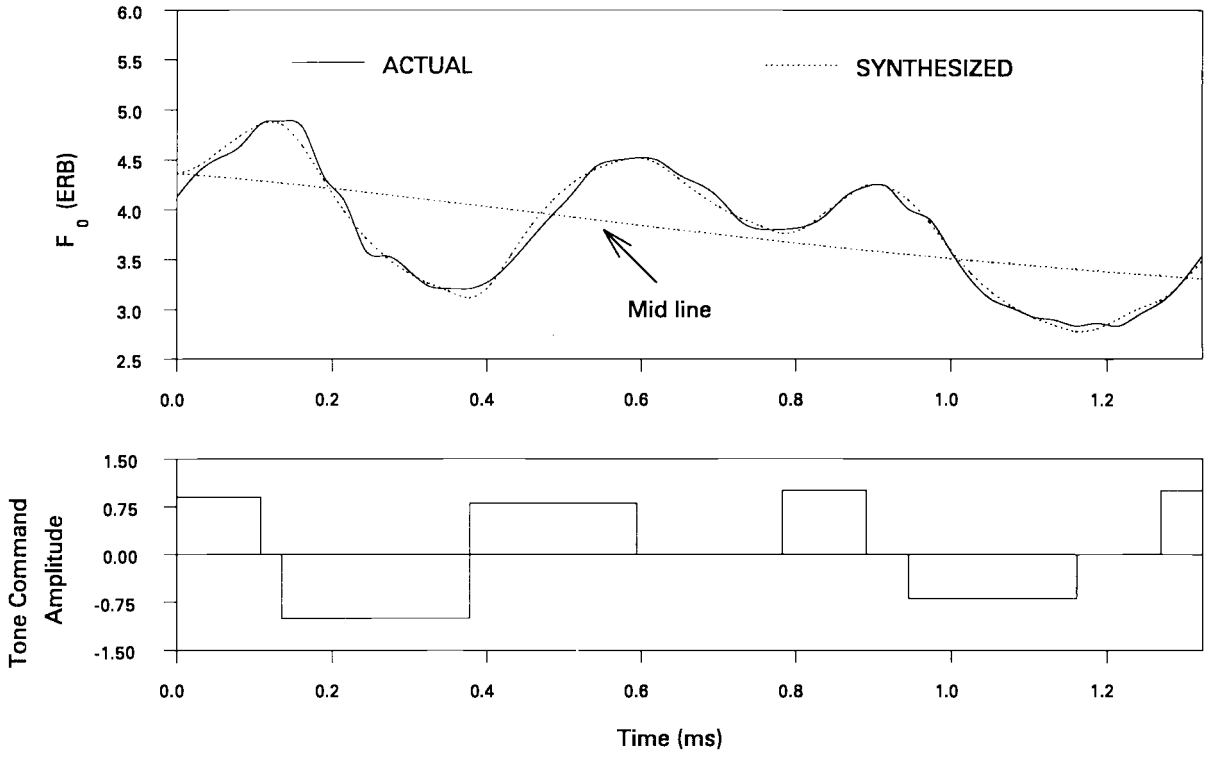



Fig. 6. Actual and the synthesized $F_0$ contours based on the extension to Fujisaki's model for a continuously voiced utterance with an HLFHR tone sequence (top panel). The bottom panel shows a plot of the amplitudes and temporal locations of the tone commands.

lab) Model 4300 installed on a Gateway 2000 P5-90 microcomputer. $F_0$ was computed directly from the waveform using a CSL algorithm that employs a time-domain approach to pitch analysis (modified autocorrelation with center clipping) with nonoverlapping variable frame length. For a particular speaker, frame length was determined by his/her pitch range to ensure that there were at least two complete cycles within a frame. The resulting raw $F_0$ contours were smoothed using median filtering and linear interpolation. Syllable onset and offset were determined from a simultaneous display of a wide-band spectrogram, energy contour, $F_0$ contour and audio waveform using conventional rules for segmentation of the speech signal [15].

A syllable-by-syllable temporal alignment procedure was used, instead of a linear time normalization, because the temporal variations of an utterance within and across speakers do not seem to be due to uniform stretching and shrinking of segments [20]. First, an average duration for every syllable in the utterance is obtained by averaging

TABLE I
FUJISAKI'S MODEL PARAMETERS FOR AN UTTERANCE WITH AN HLFHR SEQUENCE. NOTE THAT THE LAST TONE (R) HAS TWO COMPONENTS

| $F_{min}$ (ERB) | $i$ | $T_0$ (sec) | $T_3$ (sec) | $A_p$ | $\alpha$ (sec$^{-1}$) | tone | $j$ | $k$ | $T_1$ (sec) | $T_2$ (sec) | $A_t$ | $\beta$ (sec$^{-1}$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2.832 | 1 | -.809 | 1.323 | 4.3130 8 | 1.615 | H | 1 | 1 | .003 | .108 | .9 | 20.2 |
| | | | | | | L | 2 | 1 | .135 | .378 | -1.0 | 20.2 |
| | | | | | | F | 3 | 1 | .378 | .594 | .8 | 20.2 |
| | | | | | | H | 4 | 1 | .783 | .891 | 1.0 | 20.2 |
| | | | | | | R | 5 | 1 | .945 | 1.161 | -.7 | 20.2 |
| | | | | | | | | 2 | 1.269 | 1.323 | 1.0 | 20.2 |

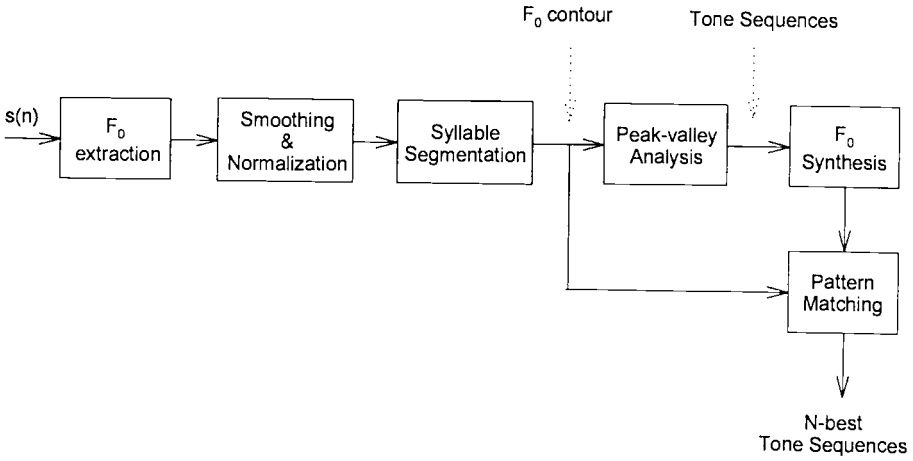



Fig. 7. Block diagram of the tone classifier.

across all corresponding syllable durations in all tonal sequences of all speakers. A ratio expressed in the percent of each average syllable duration to the average total utterance duration is then computed. Finally, a syllable-by-syllable linear time normalization is performed on each utterance based on the ratio for each syllable. A result of the syllable-by-syllable temporal alignment procedure is illustrated in Fig. 5. After time-normalization, differences in the excursion size of $F_0$ movements related to differences in voice range between speakers were normalized by converting raw $F_0$ values to an ERB scale [13]. A z-score normalization is then employed to account for pitch range differences across speakers based on the precomputed mean and standard deviation from all utterances. This method has the effect of making the first- and second-order moments of the pitch distributions the same [22].

Since our tone synthesis model is based on the principle of superposition, the phrase component parameters can be determined separately from the tone component parameters. The phrase command parameters, which account for the declination effect, are determined by fitting an exponential curve to the $F_0$ contour. To parameterize the tone command, the declination effect is removed by subtracting the exponential curve used as the response function of the phrase control mechanism. Then, each $F_0$ contour is processed from left to right, tone by tone, and the parameter values of the tone component are determined by a curve-fitting method [19]. The $F_0$ contour was optimized on a tone-by-tone basis so that the preceding tone commands will not be affected by the following optimization process.

A test of the trained tone-based Fujisaki's model is shown in Fig. 6, which plots the actual and synthesized $F_0$ contour (see top panel) from the analysis of an utterance with an HLFHR tone sequence spoken by a single speaker. The step-wise plot of the amplitude and temporal locations of each tone command is presented in the bottom panel. The synthesized $F_0$ contour closely approximates the actual $F_0$ contour. Table I lists the values of the parameters for the Fujisaki model for the HLFHR tone sequence. Although the model was trained on three-tone sequences, the synthesized five-tone sequence closely matches the actual contour from a sentence containing five stressed and voiced monosyllabic words.

## III. THE TONE CLASSIFICATION ALGORITHM

The block diagram of our tone classifier is shown in Fig. 7. The first three blocks represent the feature extraction step which produces normalized $F_0$ contours; the last three represent the tone classification step. Feature extraction on the test sentences was the same as for the training sentences. The resulting smoothed and normalized $F_0$ contour for each of the test sentences was then segmented by hand into syllable units because of the unavailability of an implemented syllable segmentation algorithm for Thai.

The tone classification step is based on the analysis-by-synthesis method. The first step is to identify the possible tone sequences in the extracted $F_0$ contour. These sequences are used to synthesize $F_0$ contours using our trained Fujisaki model. The resulting contours are then compared to the smoothed and normalized input $F_0$ contour in the pattern matching module. To avoid generating all possible tone sequences to match against a test sentence, peak-and-valley analysis is used to reduce the number of reference templates. Given the smoothed, normalized and segmented $F_0$ contour for a test sentence, local extreme (peaks and valleys) are detected by using first and second derivatives. The derivative at any point in the contour, except for the first two 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.

TABLE II
A LIST OF TEST SENTENCES FOR THE AUTOMATIC TONE CLASSIFICATION ALGORITHM

| Tone seq. | Thai script | Phonemic transcription | English translation |
|---|---|---|---|
| 1. FFFF | น่านเล่นว่าวมั่ง | / nâan lên wâw mâŋ / | 'It's Nan's turn to fly the kite.' |
| 2. RRRR | หมอหนีหมีไหว | / mɔ̌ɔ nǐi mǐi wǎj / | 'The doctor can still run away from bears.' |
| 3. FHRL | แม่เลี้ยงหลานอยู่ | / mɛ̌ɛ liaŋ lǎan jùu / | 'Mother is taking care of her grandson.' |
| 4. LFHR | หน่องเลื่อยไม้ไหว | / nɔ̀ɔŋ luaj máaj wǎj / | 'Nong can still saw wood.' |
| 5. LHLL | ใหญ่ง้อหล่อนอยู่ | / jàj ŋɔ́ɔ lɔ̀ɔn jùu / | 'Yai is making up with her.' |
| 6. FFHH | แม่ย่างเนื้อแล้ว | / mêɛ jâaŋ núa léɛw / | 'Mother has already grilled the meat.' |
| 7. FHHH | แม่เลี้ยงน้องแล้ว | / mêɛ liaŋ nɔ́ɔŋ léɛw / | 'Mother has fed the baby.' |
| 8. RLLL | หลานแหย่หล่อนอยู่ | / lǎan jɛ̀ɛ lɔ̀ɔn jùu / | 'The nephew is teasing her.' |
| 9. LLLL | หน่องแหย่หน่อยอยู่ | / mɔ̀ɔŋ jɛ̀ɛ nɔ̀ɔj jùu / | 'Mong is teasing Noy.' |
| 10. HHHH | น้อยล้างแย้แล้ว | / nɔ́ɔj láaŋ jɛ́ɛ léɛw / | 'Noy has already cleaned the wood lizard.' |
| 11. MMMM | นายลางานมา | / naaj laa ŋaan maa / | 'The boss took a day off to be here.' |

The locations of these extreme coupled with syllable boundary information are then used to determine all possible tone sequences. Only a falling tone can follow if a maximum occurs on a given syllable, and a rising or the high tone can follow if a minimum occurs. 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. These rules reduce the number of possible reference templates generated by the synthesis module, and thus, reduces the running time of the pattern matching module. The tone sequences predicted by peak-and-valley analysis are then used to generate reference $F_0$ contours for pattern matching against the smoothed and normalized input contour. Zero-lag cross correlation was used as a measure of goodness of fit to rank the results.

A classification test was performed on a set of 55 test utterances consisting of 11 sentences with varying tone sequences (see Table II) spoken by the same five speakers as the training set. The sentences were chosen to represent varying degrees of movement in the $F_0$ contour. In particular, they were chosen to test the effectiveness of each of the peak-and-valley rules together with tonal assimilation (e.g., 3, 5), as well as the method of handling of the declination effect (e.g., 10, 11). As in the training set, each test sentence contained four stressed monosyllabic words that are continuously voiced throughout. The classification test was performed on each of the 55 test utterances to obtain the cross correlation coefficients between the input contour and the predicted contours. In this experiment, the peak-valley analysis module generated at most six tone sequences; hence, only six reference templates were generated by the $F_0$ synthesis module for matching. Table III shows the cross-correlation coefficients for the input sequence FHRL. For this sequence, the four-tone sequence with the greatest cross correlation coefficient is the target sequence. Note that for speaker 3, the algorithm was able to identify the correct tone sequence despite the somewhat low coefficient values.

The algorithm misclassified six of the 55 test utterances. Five of these errors resulted from the misclassification of the fourth tone in sequence 11 for each speaker (M was misclassified as L). The remaining error involved the misclassification of H, the fourth tone in sequence 6, as R. Of the six misclassified utterances, three would have been the second choice, one the third choice, and two the fifth choice out of a set of six possible tone sequences. This result emphasizes the importance of extracting N-best tone sequences for the word hypothesizer; the correct tone sequence may be recovered at a later stage by using other linguistic constraints. Overall 89.1% of the four tone sequences were correctly classified. The errors associated with the 11th test sentence suggest that the synthesis module did not produce reliable $F_0$ contours for the sentence-final syllables. Having ignored the stress effects, this 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. It is also possible that concomitant voice quality differences in tones, which were ignored in the synthesis module, might improve tonal classification in prepausal positions. We further observe that confusion occurs between M and L, and thus leads to a misclassification. However, this misclassification is unidirectional (M $\rightarrow$ L). This is phonetically plausible because of the relative positions of M and L in the Thai tonal space [3] (see Fig. 1). With the declination-induced narrowing of the tonal space in utterance-final position, mid tones are more likely to gravitate toward the floor of the tonal space. The low tone already occupies the floor of the tonal space.

Although the present implementation of the tone classification algorithm produces promising results, more work is needed to incorporate additional linguistic constraints into the model. Automatic methods of performing syllable segmentation on the training and test sentences should be developed; automatic parameter extraction for training the synthesis module should be implemented; the training set should be expanded to model stress effects and varying syllable structures.

TABLE III
CORRELATION COEFFICIENTS FOR THE INPUT SEQUENCE FHRL. THE MAXIMUM COEFFICIENT, HIGHLIGHTED IN BOLDFACE FOR EACH SPEAKER IS THE BEST SEQUENCE OF TONES GIVEN THE MODEL

| speaker no. | predicted tone sequence | | | |
|---|---|---|---|---|
| | 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 | 0.369 |
| 4 | **0.9223** | 0.8329 | 0.7943 | 0.6778 |
| 5 | **0.9096** | 0.8195 | 0.7319 | 0.6488 |

## REFERENCES


[1] A. S. Abramson, "The vowels and tones of standard Thai: Acoustical measurements and experiments," *Int. J. Amer. Linguist.*, vol. 28-2, Part III, no. 20, 1962.

[2] ——, "The coarticulation of tones: An acoustic study of Thai," in *Studies of Thai and Mon-Khmer Phonetics in Honor of Eugenie J. A. Henderson*, V. Panupong, P. Kullavanijaya, K. Tingsabadth, and T. Luangthongkum, Eds. Bangkok, Thailand: Chulalongkorn Univ. Press, 1979, pp. 1–9.

[3] ——, "The Thai tonal space," in *Southeast Asian Linguistic Studies in Honor of Vichin Panupong*, A. Abramson, Ed. Bangkok, Thailand: Chulalongkorn Univ. Press, 1997, pp. 1–10.

[4] S. H. Chen, S. Chang, and S. M. Lee, "A statistical model based fundamental frequency synthesizer for Mandarin speech," *J. Acoust. Soc. Amer.*, vol. 92, pp. 114–120, 1992.

[5] S.-H. Chen and Y.-R.Wang, "Tone recognition of continuous Mandarin speech based on neural networks," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 146–150, 1995.

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

[7] B. Connell and D. R. Ladd, "Aspect of pitch realization in Yoruba," *Phonology*, vol. 7, pp. 1–29, 1990.

[8] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*, P. F. MacNeilage, Ed. Berlin, Germany: Springer-Verlag, 1983, pp. 39–55.

[9] J. T. Gandour, "Consonant types and tones in Siamese," *J. Phonet.*, vol. 2, pp. 337–350, 1974.

[10] ——, "Tone perception in Far Eastern languages," *J. Phonet.*, vol. 11, pp. 149–175, 1983.

[11] J. T. Gandour, S. Potisuk, and S. Dechongkit, "Tonal coarticulation in Thai," *J. Phonet.*, vol. 22, pp. 477–492, 1994.

[12] M. Halle and K. N. Stevens, "Speech recognition: A model and a program for research," *IRE Trans. Inform Theory*, vol. IT-8, pp. 155–159, 1962.

[13] D. Hermes and J. Van Gestel, "The frequency scale of speech intonation," *J. Acoust. Soc. Amer.*, vol. 90, pp. 97–102, 1991.

[14] X. Hu and K. Hirose, "Tone recognition of Chinese disyllables using hidden Markov models," *IEICE Trans. Inform. Syst*, vol. E78-D, pp. 685–691, 1995.

[15] D. H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *J. Acoust. Soc. Amer.*, vol. 59-5, pp. 1208–1221, May 1976.

[16] T. Lee *et al.*, "Tone recognition of isolated Cantonese syllables," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 204–209, May 1995.

[17] P. Lieberman *et al.*, "Measures of the sentence intonation of read and spontaneous speech in American English," *J. Acoust. Soc. Amer.*, vol. 77, pp. 649–657, 1985.

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

[19] S. Potisuk, *Prosodic Disambiguation in Automatic Speech Understanding of Thai*, Ph.D. dissertation, Purdue Univ., West Lafeyette, IN, 1995.

[20] S. Potisuk, J. Gandour, and M. Harper, "Acoustic correlates of stress in Thai," *Phonetica*, vol. 53, pp. 200–220, 1996.

[21] ——, "Contextual variations in trisyllabic sequences of Thai tones," *Phonetica*, vol. 54, pp. 22–42, 1997.

[22] P. Rose, "Considerations in the normalization of the fundamental frequency of linguistic tone," *Speech Commun.*, vol. 6, pp. 343–351, 1987.

[23] C. F. Wang, H. Fujisaki, and S. H. Chen, "The four tones recognition of continuous Chinese speech," in *Proc. ICASSP*, 1990, pp. 221–224.

[24] H.-M. Wang and L.-S. Lee, "Tone recognition for continuous Mandarin speech with limited training data using selected context-dependent hidden Markov models," *J. Chin. Inst. Eng.*, vol. 17, pp. 775–784, 1994.

[25] Y. R. Wang, J.-M. Shieh, and S. H. Chen, "Tone recognition of continuous Chinese speech based on hidden Markov model," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 8, pp. 233–246, 1994.

[26] U. Warotamasikkhadit, "Some phonological rules in Thai," *J. Amer. Orient. Soc.*, vol. 87-4, pp. 541–574, 1967.

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