International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems
Vol. 0, No. 0 (1993) 000—000

TONE RECOGNITION OF CONTINUOUS THAI SPEECH UNDER TONAL ASSIMILATION AND DECLINATION EFFECTS USING HALF-TONE MODEL

NUTTAKORN THUBTHONG and BOONSERM KIJSIRIKUL
Machine Intelligence & Knowledge Discovery Laboratory
Department of Computer Engineering, Chulalongkorn University
Bangkok, 10330, Thailand

Received March 2001
Revised October 2001

This paper presents a method for continuous Thai tone recognition. One of the main problems in tone recognition is that several interacting factors affect F_0 realization of tones. In this paper, we focus on the tonal assimilation and declination effects. These effects are compensated by the tone information of neighboring syllables, the F_0 downdrift and the context-dependent tone model. However, the context-dependent tone model is too large and its training time is very long. To overcome these problems, we propose a novel model called the *half-tone* model. The experiments, which compare all tone features and all tone models, were simulated by feedforward neural networks. The results show that the proposed tone features increase the recognition rates and the half-tone model outperforms conventional tone models, i.e. context-independent and context-dependent tone models, in terms of recognition rate and speed. The best results are 94.77% and 93.82% for the inside test and outside test, respectively.

Keywords: Thai tone; Tone recognition; Half-tone model; Total assimilation effect; Declination effect.

1. Introduction

During the past decade, speech recognition technology has undergone significant progress. Several applications of speech recognition to human-computer interface have been developed since speech is the most natural way of human communication and interaction. Most existing methods for speech recognition are developed mainly for spoken English, and some of them have been adapted to be applicable to Thai language. However, unlike English, Thai is a tone language. In such a language, the referential meaning of an utterance is dependent on the lexical tones[1]. Therefore, a tone classifier is an essential component of a speech recognition system of a tone language.

Many methods of tone recognition have been proposed for both isolated and continuous speech in Mandarin and Cantonese. They include the methods based on multi-layer perceptron for four-tone-recognition of isolated Mandarin syllables[2], for five-tone-recognition of continuous Mandarin speech[3,4] and for nine-tone-recognition of isolated Cantonese syllables[5,6]. The method based on a hidden Markov model

(HMM) for four-tone-recognition of isolated Mandarin syllables[7] and the fuzzy C-means-based method for four-tone-recognition of isolated Mandarin syllables[8] have also been proposed.

In Thai, there are five different lexical tones as follows: mid /M/, low /L/, falling /F/, high /H/, and rising /R/. The following examples show the effect of tones on the meaning of an utterance[9] : M /khāa/ ("a kind of grass"); L /khàa/ ("galangale"); F /khâa/ ("to kill"); H /kháa/ ("to trade"); and R /khǎa/ ("a leg"). The tone information is superimposed on the voiced portion of each syllable. The identification of a Thai tone relies on the shape of the fundamental frequency (F_0) contour. Fig.1 shows the average of F_0 contours of five different tones when syllables are spoken in isolation by a male speaker.

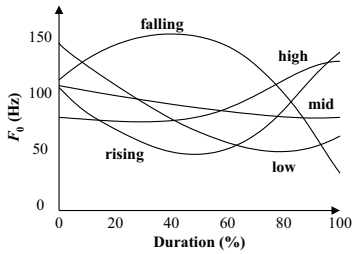

Fig. 1. F_0 contours of the five Thai tones.

In the past few years, some methods for five-tone-recognition of isolated Thai syllables have been proposed. Thubthong et al.[10,11] proposed a set of tone features and used neural networks to recognize the five tones. They concentrated on the effect of initial consonants, vowels and final consonants on tone recognition. Tungthangthum[12] used raw F_0's and hidden Markov models to classify the tones. His study shows that tones and vowels are independent from each other but his result was done on single-speaker tone recognition only. Kongkachandra et al.[13] proposed two techniques, intonation flow analysis and voiced identify calculation, for Thai tone recognition. The data used in their experiment are only ten Thai digits.

Although there are only five different tones, the tone behavior is very complicated in continuous speech. Fig.2 shows the comparison of F_0 realization of an FHRL sequence when each monosyllabic word is spoken in isolation (see top panel) and when all four tones are spoken naturally in running speech (see bottom panel). The tones produced on isolation words are very similar to those in Fig.1, while tones produced on words in continuous speech are much more difficult to identify. Several interacting factors affect F_0 realization of tones, e.g., syllable structure, stress, speaking rate, tonal assimilation and declination[14,15,16,17]. Syllable structure affects an F_0 contour in terms of the voiced/unvoiced property of contiguous phones and also consonantally-induced perturbations on the F_0 contour of the following vowel[14].

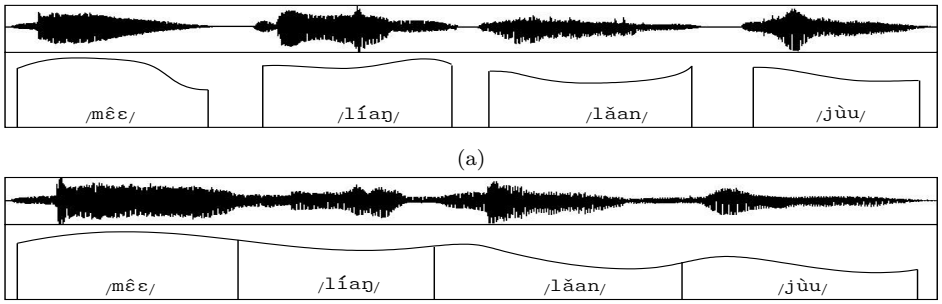

Fig. 2. The waveform and F_0 contours of tones in an FHRL sequence when (a) each word is spoken in isolation and (b) when the whole utterance is naturally spoken.

The F_0 contours of stressed syllables are generally quite different from unstressed ones[4]. For standard Thai, 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[14,15].

Speaking rate effects on F_0 contours in unstressed syllables are more extensive, both in terms of the height and slope, than those of stressed syllables[16]. The height of F_0 contours in unstressed syllables are generally higher in the fast speaking rate when compared to the slow, whereas the slope of F_0 contours in unstressed syllables varies depending on the range of F_0 movement.

The tone of a neighboring syllable may affect the shape and level of the F_0 contour[3]. The effects of the following syllable and the preceding syllable are called *anticipatory coarticulation* and *carry-over coarticulation*, respectively. Gandour et al.[17] studied these effects on Thai tones. The study shows that Thai tones are more influenced by carry-over coarticulation than by anticipatory coarticulation. In this paper, we refer to these effects as *tonal assimilation effect*.

The F_0 level will be adjusted to conform to the intonation pattern of the sentence[3]. This effect is called *declination effect* that refers to a gradual modification over the course of a phrase or utterance against which the phonologically-specified local F_0 targets are scaled[14].

Potisuk et al.[14] proposed an analysis-by-synthesis algorithm for recognizing Thai tones in continuous speech. This algorithm used an extension to Fujisaki's model for a tone language that incorporates tonal assimilation and declination effects. In this paper, we attack the same problem as in[14] by proposing tone features to compensate tonal assimilation and declination effects and investigate a series of experiments on tone recognition. The feedforward neural networks are employed to simulate our experiments. We also propose a novel model called *half-tone model* to improve the performance of tone recognition.

This paper is organized as follows. Section 2 presents the continuous Thai tone recognition method. In Section 3, experiments are simulated and the results are

analyzed in several directions. The conclusion is given in Section 4.

2. Continuous Thai Tone Recognition Method

This section presents a continuous Thai tone recognition method. First, we describe several tone features that include the baseline tone features, the tonal assimilation features, and the declination normalization. We then introduce two conventional tone models, i.e. context independent and context-dependent tone models, and propose a new tone model, called the *half-tone model* which is used to improve the performance of our method.

2.1. *Tone Features*

2.1.1. *The baseline tone features*

As tone information is superimposed on the voiced portion of a syllable, we first locate the voiced portion before extracting the F_0 feature. The portion is detected by using energy and zero crossing[6]. The Average Magnitude Different Function (AMDF) algorithm[18] is then applied for F_0 extraction with 60 ms frame size and 12 ms frame shift.

An F_0 is basically a physiologically determined characteristic and is regarded as being speaker dependent[6]. For example, the dynamic F_0 range of a male voice is much narrower (90-180 Hz) than that of a female voice (150-240 Hz). Therefore, for independent-speaker tone recognition that uses the relative F_0 of each syllable as the main discriminative feature, a normalization procedure is needed to align the range of the F_0 height for different speakers. In this paper, a raw F_0 is normalized by transforming the Hertz values to a z-score[16]. The precomputed mean and standard deviation are computed from raw F_0 values of all syllables for each speaker. Since not all syllables are of equal duration, F_0 contours of each syllable are equalized for duration on a percentage scale[17,15].

F_0-normalized data are then fitted with a third-order polynomial ($y = a_0 + a_1x + a_2x^2 + a_3x^3$) that has been proven to be successful for fitting F_0 contours of the five Thai tones[16]. To evaluate changes in the F_0 height and slope of each syllable, a time aligned F_0 profile is used. The profile is obtained by calculating the F_0 heights at five different time points between 0% to 100% throughout each syllable with the equal step size of 25%, see Fig 3(a). Moreover, the slopes at these five points are also computed by using the polynomial coefficients. The five F_0 heights and five slopes are used as the baseline features. Therefore, the tone feature vector of each syllable has the same dimension of 10.

2.1.2. *Tonal assimilation features*

Thai tones are influenced by carry-over and anticipatory coarticulations. As described in[17], carry-over effects extend forward to about 75% of the duration of the following syllable, while anticipatory effects extend backward to about 50% of the

duration of the preceding syllable. In order to alleviate tonal assimilation effect, the F_0 heights and slopes of neighboring syllables (both preceding and following syllables) are considered. The F_0 heights and slopes at 50% and 75% of the preceding syllable and at 25%, 50% and 75% of the following syllable are used as the tonal assimilation features, see Fig. 3 (b). In the case of the beginning or ending syllable, the F_0 height and slope are set to 0 and 10000, respectively.

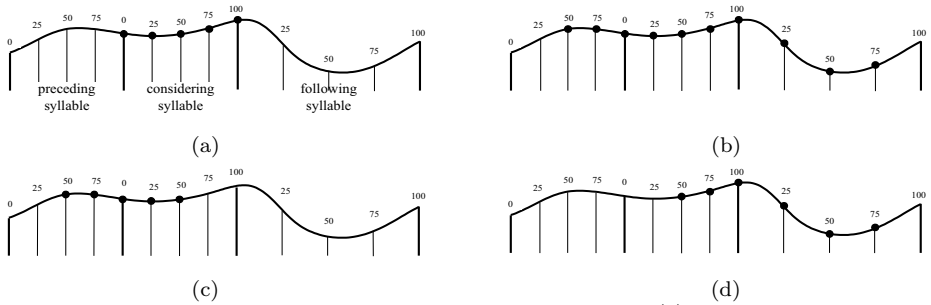

Fig. 3. The different time points of the proposed tone features when (a) the baseline tone features, (b) the baseline tone features + tonal assimilation features, (c) the first-half tone features, and (d) the second-half tone features.

2.1.3. *Declination Normalization*

According to the phonological approach to intonation[19], the intonation contour is a string of pitch accents and boundary tones, and there is an overall downstep trend of the F_0 height of the pitch accents. Like[21], we assume that all utterances in our speech data have a similar underlying intonation contour. Therefore, an F_0 contour can be viewed as a "constant" intonation component with additive "random" perturbation[21]. An averaging approach can be used to smooth out the "random" variations due to tones and the average can be obtained as the underlying intonation contour.

Fig.4 show the F_0 contours of all utterances. Since not all syllables are of equal duration, the time scale of each utterance is normalized by the utterance duration. The plot shows that there is a steady downdrift of the mean F_0 contour. The mean F_0 contour is then fitted with a first-order polynomial ($y = a_0 + a_1 x$). The straight line represents an F_0 downdrift. To neutralize the declination effect, we subtract the downdrift from each F_0. This method is referred as *declination normalization.*

2.2. *Tone models*

2.2.1. *Context-independent and context-dependent tone models*

In this paper, we used two conventional tone models as the baseline tone models.

(i) *Context-independent tone model* (CI-T)

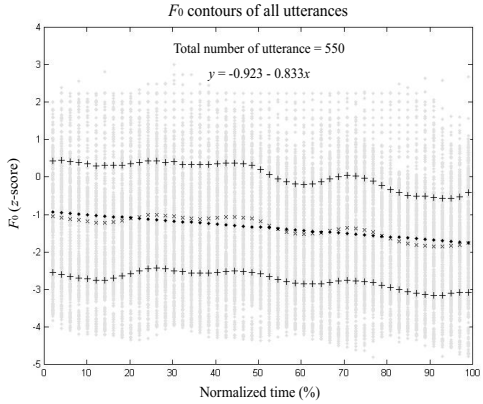

Fig. 4. F_0 contours of all utterances. The '×' line represents the mean F_0 contour, with the upper and lower '+' lines for standard deviation. The dot line is the first-order polynomial regression line for the average F_0 contour.

This model treats the considering syllable as being independent of its neighboring syllables. Therefore, there are only five different tones.

(ii) *Context-dependent tone model* (CD-T)
In fact, the neighboring syllables affect the considering syllable. Potisuk et al.[14] used three-tone sequences to measure this effect. There are 125 possible three-tone sequences. However, they concentrated on only the syllable in the middle of a sentence. To enhance these sequences, we also consider the syllable at the beginning and the end of a sentence. Therefore, a total of 175 sequences will be needed, i.e. 5^3 (in the middle of a sentence) + 5^2 (at the beginning of a sentence) + 5^2 (at the end of a sentence).

2.2.2. *Half-tone model*

The number of 175 possible sequences in CD-T is too large and its training time is very long. Thus, we propose a novel model called *half-tone model* (H-T). This model is based on the "divide-and-conquer" principle. A syllable is separated into two parts at the center. For the first half, a total of 30 sequences will be needed, i.e., 5 (at the beginning of a sentence) + 5^2 (in the middle of a sentence). Also for the second half, a total of 30 sequences will be needed, i.e., 5 (at the end of a sentence) + 5^2 (in the middle of a sentence). The first half is trained by using one neural network and the second half is trained by the other network. The outputs of two networks are then combined to determine the classification result.

The tone features for this model are the baseline features incorporating the tonal assimilation feature and the declination normalization but the features are separated into two halves. For the first half, we obtain F_0 heights and slopes at three different time points at 0%, 25%, and 50% of the considering syllable and at two different time points at 50%, and 75% of the preceding syllable. In the same way, F_0 heights and slopes at three different time points at 50%, 75%, and 100%

of the considering syllable and at three different time points at 25%, 50%, and 75% of the following syllable are obtained for the second half (see Fig.3 (c) and (d)). Therefore, the tone feature vectors of the first half and the second half of each syllable have the dimensions of 10 and 12, respectively.

2.2.3. *Decision algorithms*

For the context-independent tone model, the output of the model is the classification result, but for the context-dependent and half-tone models, the outputs are not. They need a final decision process. In this paper, we employed two decision algorithms for context-dependent and half-tone models as follows.

(i) *Context-dependent tone model*: All output sequences are grouped into five groups depending on the tone of the middle syllable in each sequence. The *posteriori* probabilities of each sequence in each group are summarized as the group score. The group providing the highest group score is chosen as a classification result.
(ii) *Half-tone model*: Each output sequence pair of two classifiers (one for the first-half and the other for the second-half) are grouped into five groups depending on the tone of the middle syllable in each sequence. The *posteriori* probabilities of each sequence in each group are summarized as the group score. Then, the classification result is the group that provides the highest group score.

3. Simulation Experiments

In this section, we first describe the speech data used in the experiments. We then show a series of experiments which simulate our tone features and tone models. Finally, we report the results comparing our method with the previous ones.

3.1. *The Speech Data*

The speech data is based on Potisuk et al.[14]. The data is comprised of 11 sentences with varying tone sequences. Each sentence consists of four monosyllabic words. All four syllables began and ended with a sonorant, and the sentence was continuously voiced throughout. In order to eliminate the potentially confounding interaction between stress and tone, the stress pattern of the carrier sentence was invariant[20]. The data was collected from 10 native Thai speakers (five male and five female speakers), ranging in age from 20 to 22 years (mean=20.8 and SD =0.78). Each speaker read all sentences for five trials at a conversational speaking rate. Therefore, the corpus comprises 550 utterances. The speech signals were digitized by a 16-bit A/D converter of 11 kHz. These were manually segmented and transcribed at the syllable level using audio-visual cues from a waveform display.

3.2. *Experiments*

We conducted two main experiments. The first and second experiments are for simulating several tone features and tone models, respectively. Each experiment evaluates the recognition robustness against speaker variation by comparing the results on the inside and outside tests. The inside test uses identical speech data in both training and testing, while the outside test uses the data from different speakers for training and testing.

In all experiments, a 5-fold cross-validation approach[22] was used. The original utterances were partitioned into five disjoint sets of equal size. For the inside test, each set contains utterances collected from each trial of all speakers. For the outside test, each set contains utterances collected from one male and one female speakers. Then these five disjoint sets were used to construct five training sets. Five training sets were derived by overlapping the five disjoint sets and dropping out a different one systematically (totally 1760 tokens). The different sets, which were dropped, were used as test sets (totally 440 tokens). The experimental results are the average values of the five test sets.

The experiments were run using three layer feedforward neural networks. Each network has three layers, i.e. input, hidden, and output layers. The number of input and output units depend on the tone features and the tone models, respectively. The number of hidden units is 20 for the half-tone model and 60 for the other models. All feature parameters were normalized to lie between -1.0 and 1.0. The network was trained by the error back-propagation. Initial weights were set with random values between -1.0 and 1.0. The NICO (Neural Inference COmputation) toolkit[23] was used to build and train the network.

For the first main experiment, the context-dependent tone model was employed to simulate several tone features. The results are shown in Table 2. The results of the inside test and outside test were in the same direction. The recognition rates of the inside test were higher than those of the outside test about 1% for all experiments. This means that our tone features were less sensitive to speaker variation. The tonal assimilation features yielded better recognition rates than the baseline tone features. The declination normalization also improved the recognition rates, but provided lower recognition rates than the tonal assimilation features. When both tonal assimilation features and declination normalization were used together, the best recognition rates were obtained. The best error reduction rates (ERRs) were 40.32 and 35.32% for the inside test and outside test, respectively.

For the second main experiment, the context-independent tone model (CI-T-5), the context-dependent tone model (CD-T-175), and the half-tone model (H-T-30) were simulated on both the inside test and outside test. The features incorporating tonal assimilation features and declination normalization were employed. The results are shown in Table 3. The recognition rates of CD-T-175 outperformed those of CI-T-5. The recognition rates of H-T-30 was equal to and higher than those of CD-T-175 for the inside test and outside test, respectively. Comparing with the results of CI-T-5, H-T-30 provided best error reduction rates of 22.30 and 21.84%

for the inside test and outside test, respectively.

In addition, the training times for CD-T-175 were very long. They were longer than those for H-T-30 by more than an order of magnitude. These conclude that H-T-30 outperformed CI-T-5 and CD-T-175 in term of recognition rate and speed.

Table 1. Recognition rates with different tone features.

Feature	Inside test			Outside test		
	accuracy	SD	ERR(%)	accuracy	SD	ERR(%)
Baseline	88.73	0.24	-	87.77	0.36	-
+A	92.68	0.34	35.08	91.55	0.56	30.86
+D	90.41	0.25	14.92	89.36	0.30	13.01
+A+D	93.27	0.41	40.32	92.09	0.63	35.32

A, D and ERR are the tonal assimilation feature set, the declination normalization and error reduction rate, respectively.

Table 2. Recognition rates with different tone models.

Model	Inside test			Outside test		
	accuracy	SD	ERR(%)	accuracy	SD	ERR(%)
CI-T-5	93.27	0.41	-	92.09	0.63	-
CD-T-175	94.77	0.15	22.30	93.00	0.66	11.49
H-T-30	94.77	0.27	22.30	93.82	0.60	21.84

ERR is error reduction rate.

From Potisuk et al.[14], the best recognition rate is 89.10% for the inside test. Although we used the same list of 11 sentences for the test set, we cannot directly compare this result with our results (94.77% for the inside test). This is because the speech data were collected from the different speakers, the different number of speakers, and the different recording environment.

4. Conclusion

In this paper, we have demonstrated a method for Thai continuous tone recognition. The tonal assimilation and declination effects have been considered. We have proposed tone features and used the context-dependent tone model to compensate these effects. We have also proposed a novel model called *half-tone* model to alleviate the drawback of the context-dependent tone model. To simulate these tone features and models, feedforward neural networks were employed. The experimental results show that the proposed features outperformed the baseline features. The half-tone model provided the best recognition rates for both the inside test and outside test and surpassed the conventional tone models, i.e., context-independent and context-dependent tone models, in terms of recognition rate and speed. In the future, we plan to extend our works to cover the other effects on continuous Thai tone recognition.

Acknowledgement

This work was supported in part by the Thailand-Japan Technology Transfer Project.

References

1. F. H. L. Jian, "Classification of Taiwanese tones based on pitch and energy movement", in *Proc. Int. Conf. Spoken Language Processing.* **2** (1998) 329–332.
2. P. C. Chang, S. W. Sue and S. H. Chen, "Mandarin tone recognition by multilayer perceptron", in *Proc. Int. Conf. Acoust., Speech, and Signal Processing.* **1** (1990) 517–520.
3. Y. R. Wang and S. H. Chen, "Tone recognition of continuous Mandarin speech assisted with prosodic information", *J. Acoust. Soc. Amer.* **96** (1994) 1738–1752.
4. S. H. Chen and Y. R. Wang, "Tone recognition of continuous Mandarin speech based on neural networks", *IEEE Trans. Speech Audio Processing.* **3** (1995) 146–150.
5. T. Lee, P. C. Ching, L. W. Chan, Y. H. Cheng and B. Mark, "An NN based tone classifier for Cantonese", in *Proc. Int. Joint Conf. Neural Networks.* **1** (1993) 287–290.
6. T. Lee, P. C. Ching, L. W. Chan, Y. H. Cheng and B. Mark, "Tone recognition of isolated cantonese syllables", *IEEE Trans. Speech Audio Processing* **3** (1995) 204–209.
7. W. J. Yang, J. C. Lee, Y. C. Chang and H. C. Wang, "Hidden Markov Model for Mandarin lexical tone recognition", *IEEE Trans. Speech Audio Processing.* **36** (1988) 988–992.
8. J. Li, X. Xia and S. Gu, "Mandarin four-tone recognition with the fuzzy C-means algorithm", in *Proc. IEEE Int. Conf. Fuzzy Systems.* **2** (1999) 1059–1062.
9. S. Luksaneeyanawin, "Intonation in Thai", in D. Hirst and A. D. Cristo, *Intonation Systems A Survey of Twenty Language.* (1998) 376-394.
10. N. Thubthong, "A Thai tone recognition system based on phonemic distinctive features", in *Proc. the* 2^{nd} *Symposium on Natural Language Processing.* (1995) 379–386.
11. N. Thubthong, A. Pusittrakul and B. Kijsirikul, "An efficient method for isolated Thai tone recognition using combination of neural networks", in *Proc. the* 4^{th} *Symposium on Natural Language Processing.* (2000) 224–242.
12. A. Tungthangthum, "Tone recognition for Thai", in *Proc. IEEE Asia-Pacific Conf. Circuits and System.* (1998) 157–160.
13. R. Kongkachandra, S. Pansang, T. Sripra and C. Kimpan, "Thai intonation analysis in harmonic-frequency domain", in *Proc. IEEE Asia-Pacific Conf. Circuits and System.* (1998) 165–168.
14. 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.* **7** (1999) 95–102.
15. S. Potisuk, J. Gandour and M. P. Harper, "Acoustic correlates of stress in Thai", *Phonetica*, **53** (1996) 200–220.
16. J. Gandour, A. Tumtavitikul and N. Satthamnuwong, "Effects of speaking rate on Thai tones", *Phonetica*, **56** 1999 123–134.
17. J. Gandour, S. Potisuk and S. Dechnongkit, "Tonal coarticulation in Thai", *J. Phonetics*, **22** (1994) 477–492.
18. M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg and H. J. Manley, "Average magnitude difference function pitch extractor", *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-22** (1974) 353–362.
19. R. D. Ladd, *Intonational Phonology* (Cambridge University Press, 1996).
20. S. Potisuk, M. P. Harper and J. Gandour, "Speaker-independent automatic classification of Thai tones in connected speech by analysis-by-synthesis method", in *Proc. Int.*

Conf. Acoustics, Speech, and Signal Processing. **1** (1995) 632–635.
21. C. Wang and S. Seneff, "Improved tone recognition by normalizing for coarticulation and intonation effects", in *Proc. Int. Conf. Spoken Language Processing.* (2000).
22. T. G. Dietterich, "Machine learning research: four current directions", *AI Magazine.* **4** (1997) 97–136.
23. N. Ström, "Phoneme probability estimation with dynamic sparsely connected artificial neural networks", *The Free Speech Journal.* **1** (1997).