Modeling Rhythmic Variation in Thai and its Application to Speech Synthesis

Hansjörg Mixdorff†, Sudaporn Luksaneeyanawin‡, Patavee Charnvivit‡ and Nuttakorn Thubthong‡
† Berlin University of Applied Sciences, Germany
‡ Chulalongkorn University Bangkok, Thailand
E-mail: mixdorff@tfh-berlin.de, sudaporn.l@chula.ac.th, patavee@chula.com, tnuttako@chula.ac.th

ABSTRACT

This study concerns a preliminary experiment on modeling the duration of Thai syllables. It is based on a corpus of minimal pairs of sentences only differing as to their stress patterns. Following a factor analysis of syllabic durations in the corpus a simple duration model was developed. This model was used for re-synthesizing the utterances by manipulating speech from a Thai TTS system by adjusting syllable durations and monotozing the F0 contour. A perception experiment was conducted with respect to the discrimination of members in each minimal pair. Although the results show that natural utterances are identified more easily, the synthetic utterances were at least correctly identified well above chance level. In prosodically ambiguous cases subjects tend to select the semantically 'more plausible' interpretation.

1. INTRODUCTION

Thai is widely known as a tone language having five different tones denoting lexical contrasts and which have been intensively studied, namely three static tones, mid (0), low (1) and high (3), and two dynamic tones, falling (2) and rising (4) (tone indices commonly used given in brackets). Furthermore, there exist different types of sentence intonations which, inter alia, are associated with sentence mode distinctions [1]. There has been, however, so far little research in the stress patterns of the Thai, and with respect to their interaction with the melodic features. As the default position for stress is the last syllable in a lexical item there do not exist any words differing as to the lexical stress position. Nevertheless, in the context of a sentence, contrasts marked exclusively by stress can occur. This study examines a corpus of 28 pairs compiled by S. Luksaneeyanawin containing segmentally and tonally identical sentences which can convey different meanings when the stress pattern varies.
Most of the pairs are constructed like the following example ('[', ']' denoting syntactic boundaries, 'ˈ' denoting word stress, 'ˈˈ'denoting sentence stress, numbers indicating the syllabic tones):

(1)	[ruk3 ˈsaa4 ˈkhon0]	[con0	ˈˈhaaj4]
	(she) nursed the man	until he recovered	
(2)	[ruk3 ˈsaa4 khon0	ˈcon0]	[ˈˈhaaj4]
	she nursed the poor man		(and) he recovered

It should be noted that the Thai script for the two sentences in a pair is identical. A few main patterns of possible confusion can be identified. One concerns the sequence of a noun and a verb that can also be interpreted as a compound noun as in the following example:

(1)	[naa0 li3 ˈkaa0]$_N$	[ˈpluk1]$_{VB}$	ˈˈchan4
	the clock	woke up	me
(2)	[naa0 li3 kaa0	ˈpluk1]$_N$	ˈˈchan4
	alarm-clock		my

In a similar fashion, a sequence of a verb and a noun can be reinterpreted as a compound noun:

(1)	[sam4 ˈrap1]	[ˈr@@ng0]$_{VB}$	[ˈthaaw3]$_N$	ˈˈkhraj0?
	for	supporting	feet	whose
(2)	[sam4 ˈrap1]	[r@@ng0	ˈthaaw3]$_N$	ˈˈkhraj0?
	for	shoes		whose

A third frequent pattern involves proper names:

(1)	[naa3	ˈkhaaw4]	[dii0 ˈˈcaN0]
	Auntie Khaw		is very good
(2)	[ˈnaa3]	[ˈkhaaw4]	[dii0 ˈˈcaN0]
	Auntie	white	is very

meaning: "Auntie has a very fair complexion." A fourth pattern of confusion will be discussed in Section 4.

2. SPEECH MATERIAL AND METHOD OF ANALYSIS

The corpus was read by a male subject once and utterances were auditorily checked for conveying the intended meaning. The corpus was labeled on the phone and syllable level and with respect to the distinction of content vs. function word. The syllable 'con' in the example given above, for instance, would be classified as a function word ('until') in case (1) and as part of a compound content word ('poor man') in case (2). Furthermore, word boundaries were labeled. Since the Thai script does not supply means for marking word boundaries, that is blanks, sentences were analyzed with respect to the stressed syllables that mark the last syllable of a prosodic word and

 ISBN 1-876346-48-5

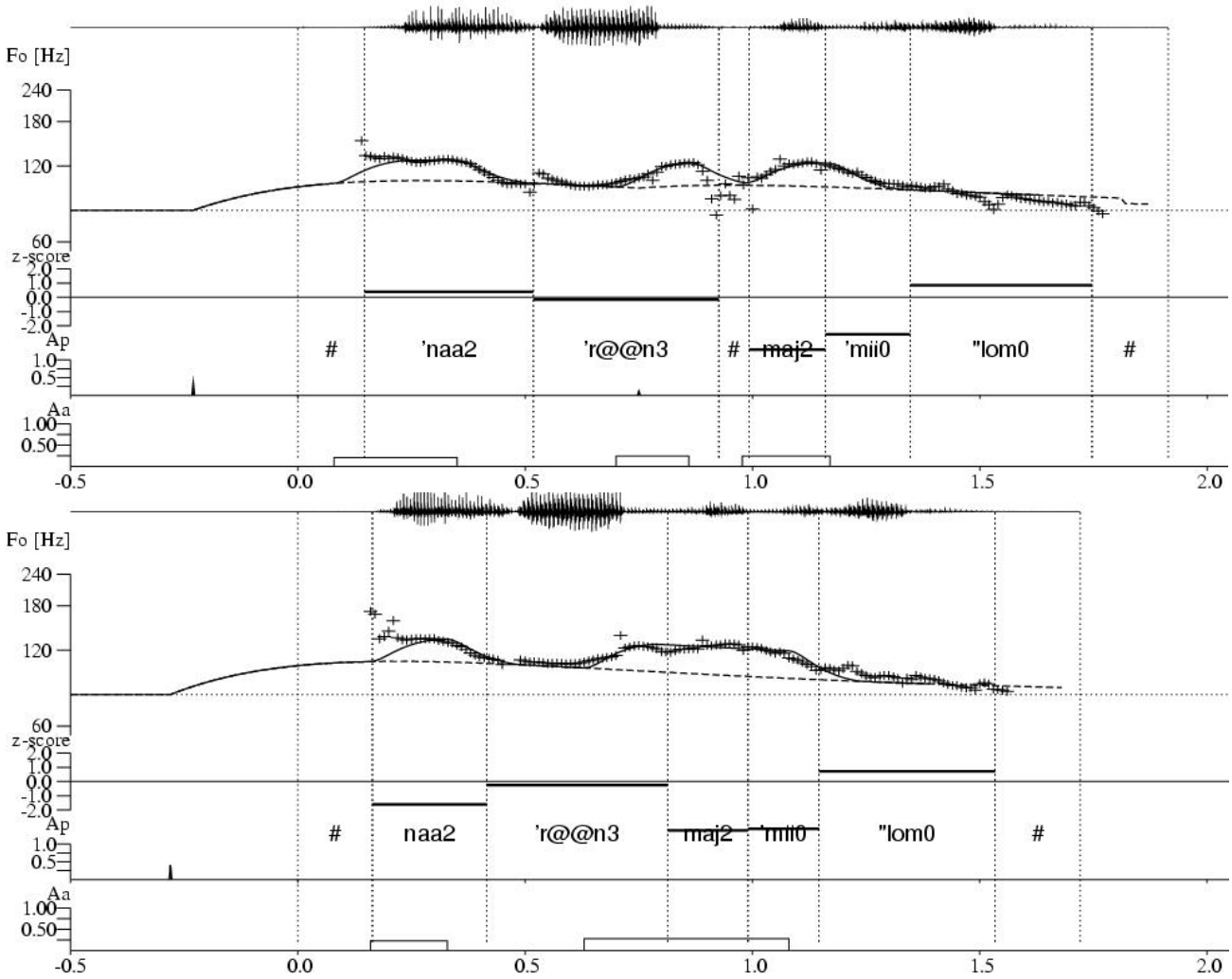

Figure 1: Examples of analysis of a minimal pair from the corpus. On the top the utterance [ˈnaa2 ˈr@@n3][maj2 ˈmii0 ˈˈlom0] – [it's likely to be hot][there is no wind] is displayed. On the bottom the utterance [naa2 ˈr@@n3 maj2 ˈmii0 ˈˈlom0] – [(in) the dry season there is no wind] is shown. Each panel displays from top to bottom: The speech waveform, the *F0* contour (extracted: +-signs, model-based: solid line), the syllable-based duration contour and the underlying phrase and accent commands.

hence delimit meaningful lexical units as explained above. The fundamental frequency contours were extracted at a step of 10 ms and analyzed using the Fujisaki model employing the strategy developed in [2] where analysis had shown that all mid tones could be modeled using the phrase component only whereas the remaining tones required either single tone commands of positive or negative polarity, or a command pair as shown in Table 1.

tone	code	tone commands: polarity and alignment
mid	0	none
low	1	negative
falling	2	positive early in the syllable
high	3	positive late in the syllable
rising	4	negative and positive

Table 1: Parametrization of Thai tones using Fujisaki model tone commands.

In the scope of the present study syllable durations were of key interest. The modeling of the *F0* contours was mainly performed in order to examine whether different stress patterns on tonally identical syllable sequences would still yield the same parametrization, that is, comparable sequences of tone commands.

3. RESULTS OF ANALYSIS

Figure 1 shows an example of analysis of a pair of utterances. This is yet another type of possible confusion which involves the word [naa2] originally signifying 'face' or 'front'. It can also mean 'it is likely to be' and be part of compound words such as [naa2 r@@n3] - 'the dry season'. In Figure 1, top panel, we see the two-phrase utterance [ˈnaa2 ˈr@@n]$_S$[maj2 ˈmii0 ˈˈlom0]$_S$ – [it's likely to be hot]$_S$ [(since)there is no wind]$_S$. In the bottom panel the single-phrase utterance [naa2 ˈr@@n maj2 ˈmii0 ˈˈlom0] $_S$ –[(in) the dry season there is no wind] is shown. Each panel displays from top to bottom: The speech waveform, the F0 contour ('+' signs indicating the extracted contour, solid lines indicating the Fujisaki model

based contour), the syllabic duration contour indicated by horizontal lines, and the underlying phrase and accent commands. The syllabic duration contour was calculated by relating the log duration of the syllable to the log added means and standard deviations of the phones of which the syllable is constructed, that is, the z-score. Phone-based duration means and standard deviations were available from the analysis of a different larger database produced by the speaker [3]. We are aware of the fact that this approach can only be regarded as an approximation since it presupposes the mutual independence of durations of neighboring phones. The duration contour, however, is only used for display purposes. In Figure 1 it is clearly seen that the syllable [naa2] is much shorter when it is part of a compound word. In the top example, the boundary between the two sentences is clearly marked by a pause, and also by a small additional phrase command at t=.75s. It can be seen that in the utterance at the bottom the positive tone commands assigned to the syllable [r@@n3] and [maj2] concatenate, whereas in the top utterance they are split by the boundary between the sentences. This shows that although the tone commands assigned to the different tones are similar, the resulting contours may differ.

Predictor factor	cross-correlation with log syllable duration
Sum of mean phone durations in the syllable from [3]	.333
Function/ content word	-.547
Number of syllables in prosodic word	-.125
Index of syllable in word counted from tail	-.318
Depth of phrase boundary to the right of syllable (0: intra-word, 1: inter-word, 2: inter-phrase)	.449
Stress level (1: unstressed, 2: word stress, 3: sentence stress)	.672

Table 2: Predictor factors of syllable duration and their respective cross-correlations.

Predictor Variable	Coefficient
Constant	2.887
Sum of mean phone durations	0.472
Function/ content word	-0.390
Index of syllable in word counted from tail	-0.231
Stress level	0.172
Number of syllables in prosodic word	0.049
Depth of phrase boundary right	0.048

Table 3: Coefficients of regression model.

A factor analysis was performed on the syllable data yielding six main predictor factors for syllable duration given in Table 2.
The relatively small influence of the number of syllables can be explained by the fact that most of the 220 words in the database had only one (179) or two (37) syllable. Only three words had three and only one had four syllables.

Since the corpus was relatively small with a total number of 266 syllables, and therefore did not facilitate statistical approaches such as CARTs or neural networks, a regression model for log syllabic duration was calculated. The resulting linear regression model has the form displayed in Table 3.
As the default stress position in a prosodic group is on the last syllable there is a strong cross-correlation between factors *Boundary Depth* and *Stress Level* (.735).
A correlation of .775 was computed between measured and predicted durations for the corpus proper. Removing the last two factors from the model yielded an even higher correlation of .806.

4. PERCEPTUAL EVALUATION

A listening test was developed that used 1) the original recordings, and 2) synthetic stimuli with the syllabic durations supplied by the regression model and segments from the Thai TTS systems developed at Chulalongkorn University [4]. The utterances created by the TTS system were read into the *PRAAT* program (© P.Boersma) and modified with respect to their syllable durations on the *DurationTiers* by using the values calculated with the duration model. This yielded two variants of every sentence, that is, minimal pairs. Since the TTS systems concatenates syllables that had been originally recorded in isolation, modifying the durations resulted in strongly compressed and exaggerated F0 contours.

Therefore, the F0 contours of the utterances were reduced to an exponentially falling declination line (Fujisaki model phrase component), thus removing the tonal information. Figure 2 shows speech waveforms for the example explained in Section 1: In the two panels at the top the interpretations (1) [ruk3 ˈsaa4 ˈkhon0] [con0 ˈˈhaaj4]– [she took care of the man] [until he recovered] and (2) [ruk3 ˈsaa4 khon0 ˈcon0] [ˈˈhaaj4] – [she nursed the poor man] [(and) he recovered] as uttered by the speaker are displayed, in the two panels on the bottom those from modified synthetic speech. The most noticeable difference between the two interpretations are, that ‘con’ when uttered with interpretation (1) is a function word equivalent to a preposition and therefore considerably shorter than in interpretation (2) where it is the last syllable of the content word ‘khon con’ *‘poor man’* bearing the word stress.
Utterances were played back to two different groups of subjects: one group (13 subjects, 11 female/2 male) listened to the original recordings and one group (15 subjects, 13 female/2 male) to the synthetic ones. Subjects were undergraduate students at Chulalongkorn University.
A forced choice had to be made between the two possible interpretations associated with a minimal-pair which were supplied on the answer sheets. The two interpretations for every sentence were provided in the Thai script with occasional explanatory notes written in parentheses, along with the corresponding English translations. Departing from the usual writing conventions, also in the Thai script

 ISBN 1-876346-48-5

word boundaries were indicated by blanks.

5. RESULTS OF EXPERIMENT

Results showed that on the average the natural stimuli were wrongly interpreted in 26.4% of decisions (chance level=50%), whereas the synthetic stimuli were misclassified even at an average of 42.5%.

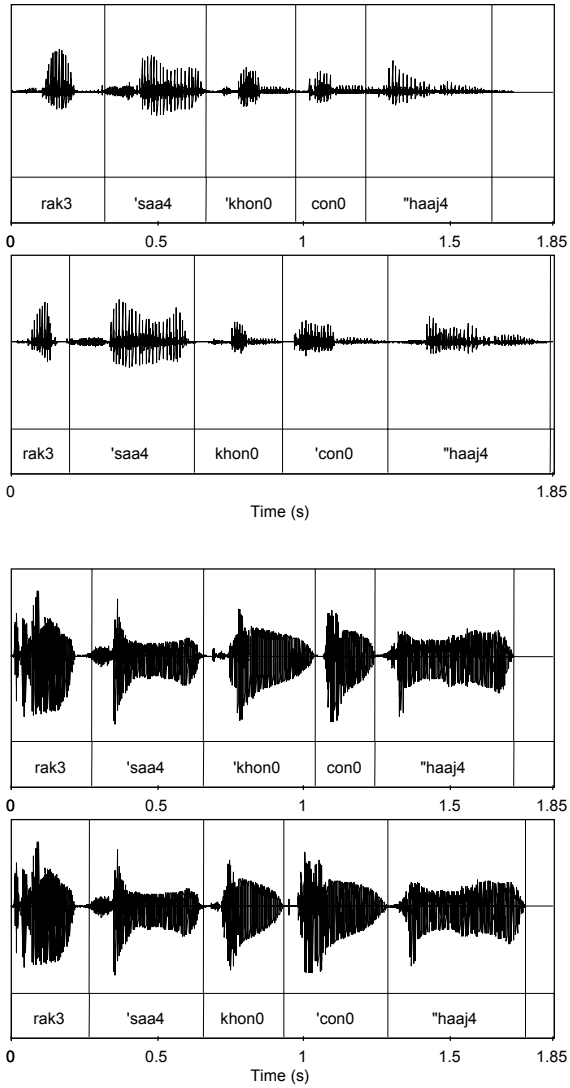

Figure 2: Speech waveforms of natural (top) and synthetic pairs (bottom) with underlying syllable tiers and boundaries.

If we assume that correct disambiguation means that each of the two alternatives in a pair is unanimously identified by all the subjects, only one pair in the set of natural utterances meets this requirement, but none in the synthetic set. If we, however, relax the criterion to a condition where both alternatives must have been identified with an error of less than 50%, we yield the following result: Of the 28 pairs of utterances, 22 pairs of the natural stimuli were reliably disambiguated, whereas the number is as low as seven pairs for the synthetic ones.

The correlation of .427 between classification errors on natural and synthetic utterances suggests that there were pairs of sentences in which one of the interpretations was more likely than the other. In order to test this hypothesis, interpretations were classified with respect to their likelihood. Interpretations in a pair that implied proper names or infrequent words, for instance, were rated with '0', those that didn't with '1'.

In the case of the synthetic stimuli, the correlation between the error rate of a particular stimulus and the 'likeliness parameter' was highly significant ($rho=-.373$, $p < .01$), for the natural stimuli still significant ($rho=-.238$, $p < .05$).

6. CONCLUSIONS

Despite the relatively high error rate in this particular experiment which can also be attributed to the lack of tonal cues and pause information in the synthetic stimuli, the duration model yields a significant improvement with respect to the TTS system which so far does not permit the modeling of stress-related duration changes. It must be noted, however, that the issue of determining word boundaries in Thai is not a trivial task, as the script only knows blanks between sentences. Hence, for instance, the decision whether [naa2] must be regarded as part of a compound word requires a sophisticated syntactic analysis of the whole sentence. Furthermore the problem whether certain serial verb constructions can be regarded as forming a prosodic word remains to be carefully examined. Hence, important future activities will involve the further development of the TTS front-end in order to yield the syntactic information needed for the duration model, that is, prosodic word boundaries and content/function word distinctions.

REFERENCES

[1] S. Luksaneeyanawin, "Intonation in Thai," in Hirst, D. and Di Christo, A. (Ed.), *Intonation Systems. A Survey of Twenty Languages.* Cambridge University Press, Cambridge, 1998.

[2] H. Mixdorff, S. Luksaneeyanawin, H. Fujisaki, and P. Charnvivit, "Perception of tone and vowel quantity in Thai," in *Proceedings of ICSLP2002*, pp. 753-756, Denver, USA, 2002.

[3] E. Maneenoi, V. Ahkaputra, S. Luksaneeyanawin, and S. Jitapunkul, "Acoustic modeling of onset-rhyme for Thai continuous speech recognition," in *Proceedings of the 9th Australian International Conference on Speech and Language Technology 2002*, Melbourne, Australia, 2002.

[4] *Thai Text-to-Speech System*, Centre for Research in Speech and Language Processing, Chulalongkorn University, Bangkok, Thailand, 2001.