PERCEPTION OF TONE AND VOWEL QUANTITY IN THAI

Hansjörg Mixdorff, Sudaporn Luksaneeyanawin**, Hiroya Fujisaki*** and Patavee Charnvivit***

*Faculty of Computer Science, Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de
**Center for Research in Speech and Language Processing, Chulalongkorn University, Thailand
sudaporn.l@chula.ac.th; patavee@chula.com
***Professor Emeritus, University of Tokyo, Japan
fujisaki@alum.mit.edu

Abstract

The current study examines the interaction of syllable tones and vowel quantity in the production and perception of mono-syllabic words of Thai. A speech corpus containing groups of words differing only as to tone type and vowel quantity was designed. These were embedded in a short carrier sentence of five mid tone syllables, with the target word being the center syllable. The utterances were analyzed with respect to the tonal and segmental features of the target words and *F0* contours modeled using the Fujisaki model. Analysis shows that all mid tone sequences can be modeled using the phrase component only whereas the remaining tones require either single tone commands of positive or negative polarity, or a command pair. Based on the analysis results, a perception experiment was designed to explore the perceptual space between words of tone/vowel quantity contrasts. Results indicate, inter alia, that vowel quantity is perceived as shorter when words are presented in isolation than when embedded in a carrier sentence. Confusions generally occur more frequently between words of different vowel quantity than of different tones.

1. INTRODUCTION

The Thai language has five different lexical tones, 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, a phonemic distinction exists between long and short vowels [1]. As a consequence, there exist groups of words which stand in tone/vowel quantity opposition, that is, either share the tone or vowel quantity as shown in the following example (long vowels are indicated by vowel symbol doubling):

Word	Thai script	tone	vowel quantity	Translation
loon0	โลน	mid	long	crab louse
loon3	โล้น	high	long	to be bald
loon4	โหลน	rising	long	great great grandson of daughter
lon0	ลน	mid	short	to singe
lon3	ล้น	high	short	to overflow
lon4	หลน	rising	short	a kind of food

Hence these groups of words present a 'worst case' for production as well as for perception, since in theory the correct lexical access should be possible based on the tone and vowel quantity only, if the words are uttered in isolation.

In the current study we examine the segmental and tone properties of a selection of highly confusable mono-syllabic words following the example above. We analyze the durational properties of the phones in the syllable as well as the tonal features as described by the analysis using the Fujisaki model [2] which has already been successfully applied to tone languages such as Mandarin [3].

It has also been shown in principle by Potisuk et al. [4], that the Fujisaki model is applicable to Thai, though no attempt was made to relate the commands yielded to the tone properties of individual syllables. The study, however, indicated that Thai requires tone commands with negative polarity.

2. SPEECH MATERIAL AND METHOD OF ANALYSIS

A speech corpus was designed which contains 17 groups of highly confusable words embedded in the carrier sentence [tʰəə0 ʔaw0] X [ma:0 du:0], "you brought X to look at." The carrier sentence contains mid tone syllables throughout. We will see in the following that mid tones do not interfere with any of the other tones with respect to tone coarticulation. The following table gives the complete list of word groups which were recorded by one male and one female native speaker of Thai five times each.

word group	tonal contrast	word group	Tonal contrast
ta(a)j, wa(a)n	0:1	kho(o)t, pa(a)	1:3
kha(a)j, ma(a)n	0:2	sa(a)ng	1:4
ra(a)w, ra(a)ng	0:3	ra(a)j	2:3
sa(a)j	0:4	kho(o)n, ma(a)j	2:4
kha(a)j, pa(a)n	1:2	kha(a)n, lo(o)n	3:4

The utterances were digitized at 16 kHz/16 bit and auditorily checked for correct tone and vowel quantity. Phone and syllable boundaries were determined by means of automatic alignment and then manually adjusted.

The *F0* values were extracted at a step of 10 ms. A semi-automatic procedure for estimating the parameters of the Fujisaki model was applied which is based on a modified version of [5], but requires a pre-segmentation of the utterance into syllables. Parameter configurations were checked and if necessary corrected. The model constants *alpha, beta* and *Fb* for the male speaker to whose data will be referred in the following discussion were set to 2/s, 20/s and 110 Hz.

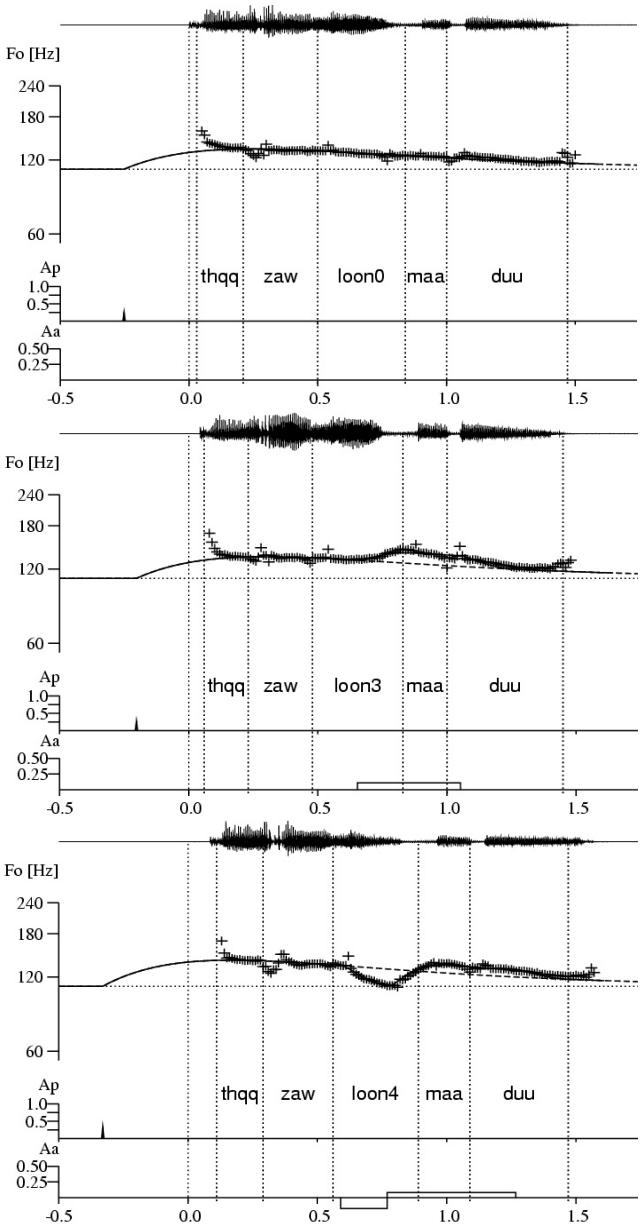

Figure 1:Examples of analysis of mid (top), high (center), and rising (bottom) tone words embedded into a context of mid tone syllables.

Table 1: Combinations of tone commands chosen for modeling the five syllabic tones.

0	mid tone	no tone command
1	low tone	single tone command, negative polarity
2	falling tone	single tone command, positive polarity, early in the syllable
3	high tone	single tone command, positive polarity, late in the syllable
4	rising tone	tone command pair of negative and positive polarity

3. RESULTS OF ANALYSIS

Examples of analysis are displayed in Figure 1 and Figure 2. Each panel shows from top to bottom: The speech waveform, the extracted (+) and model contours (solid), the text of the utterance, and the underlying tone and phrase commands.

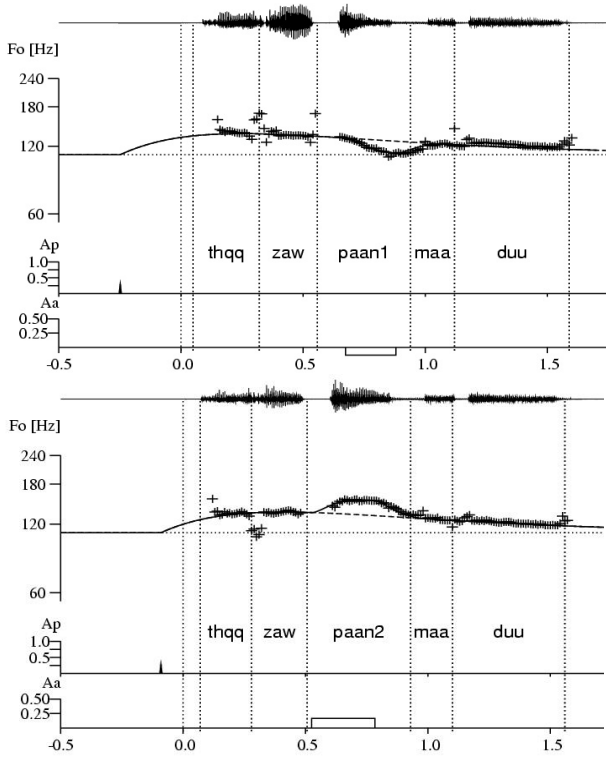

Figure 2: Examples of analysis of low (top), and falling (bottom) tone words embedded into a context of mid tone syllables.

Table 2: Average onset and offset times of tone commands with respect to the rhyme onset time in ms, and average accent command amplitudes for the five syllabic tones.

	$T11_{rel}$	$T21_{rel}$	*Aa1*	$T12_{rel}$	$T22_{rel}$	*Aa2*
mid tone	-	-	-	-	-	-
low tone	-3	220	-0.15	-	-	-
falling tone	-85	194	0.19	-	-	-
high tone	106	411	0.15	-	-	-
rising tone	-26	172	-0.19	172	417	0.10

3.1. Syllabic Tones

Figure 1 shows examples of the word 'loon' with mid (top), high (center) and rising tones (bottom), and Figure 2 utterances of the word 'paan' with low (top) and falling (bottom) tones. As can be seen in Figure 1, top, the mid tone sequence can be modeled using the phrase component only. Although tone command configurations chosen are tentative at this stage, for the current corpus the five syllabic tones can be accurately modeled using the combinations of commands listed in Table 1. As can be seen in Figure 1, center and bottom, the tone command may carry over to the following mid tone syllables.

Tone commands assigned to the tone types exhibit characteristic timing and amplitude properties which are given

in Table 2 listing onset and offset times with respect to the onset of the syllable rhyme which has been shown to be a reliable reference unit with respect to tone, along with average tone command amplitudes. Statistical analysis shows no significant correlation between the temporal characteristics of the tone commands and the quantity of the vowel, that is, the duration of accent commands is not affected by the syllabic duration. We therefore only illustrate examples of long vowel words.

3.2. Segmental Durations

Syllables containing a short vowel are only slightly shorter than those with a long vowel, as the coda is lengthened in the former ones. This effect can be seen in Table 3 listing segment durations for three pairs of long and short vowel syllables. Since statistical analysis does not yield any significant correlation between tones and segmental durations, segment durations are pooled across tone classes.

Table 3: Averaged segmental durations in long and short vowel syllables in ms.

syllable	onset	vowel	coda	total
loon	79	205	77	361
lon	79	131	117	327
saang	116	205	93	414
sang	116	139	123	378
waan	93	212	79	384
wan	110	116	113	339

4. PERCEPTION EXPERIMENT

A perception experiment was designed for examining the following issues:

- the validity of averaged Fujisaki parameter configurations assigned to each word in a highly confusable group with respect to the correct identification of this word
- the interaction between tone and duration cues
- the categorical boundaries between the highly confusable items in the perceptual space.

These issues are especially important for assessing the performance of a speech recognizer for Thai as compared to that of a human subject. Secondly, with respect to Thai speech synthesis, we hope to yield conclusions as to the accuracy with which tonal and durational features must be reproduced in order to facilitate correct word identification. As will be shown later, the experiment is also important as to the design of corpora for speech synthesis.

4.1. Stimuli Used

A subset of ten groups of words was selected for performing the perceptual study (tone contrasts given in brackets):

wa(a)n (0:1); ma(a)n (0:2); ra(a)ng (0:3); sa(a)j (0:4);
pa(a)n (1:2); kho(o)t (1:3); sa(a)ng (1:4); ra(a)j (2:3);
ma(a)j (2:4); lo(o)n (3:4)

These groups represent all possible combinations of tones. In terms of melodical properties, however, the perceptual distance between two tone types belonging to a pair cannot be assumed to be equal. If we compare low and rising tones (Figure 2, top, and Figure 2, bottom), they are much more similar than low and falling tone (Figure 2, bottom), for instance. Furthermore, if we move from a low tone to a falling one, we cross the region of the mid tone, when the *F0* contour basically follows the phrase component. Tone command configurations and segmental durations for each word in a highly confusable group were calculated by averaging over all five utterances of the respective word in the corpus. Prototypical resynthesis stimuli for all members in a group were then produced by prosodically manipulating a single utterance with long vowel/mid tone (Figure 1, top). The long vowel/mid tone version was chosen as it occupies a central position in the tonal space and we assumed that reducing the vowel duration would cause less segmental deterioration than increasing the duration of a short vowel.

First the durations of the phones in the target word were manipulated by modifying the *DurationTier* in the *PRAAT* program (© P.Boersma). Besides the long vowel and short vowel stimuli, three intermediate stimuli were produced by linearly interpolating between the two conditions. The resulting five mid tone stimuli of different durational characteristics were resynthesized. By manipulating the *PitchTier* of these utterances, the *F0* contour of the original mid tone syllable was modified to yield the tone types 1 to 4. In addition, intermediate stimuli were created by linearly interpolating between the tone command configurations underlying the two tones in each contrast.

4.2. Experimental Setting

Twenty-two phonetically untrained undergraduate and master students (14 male, 8 female) of Chulalongkorn University took part in the perception experiments. They were given numbered lists. Each line in the list contained all possible choices of words pertaining to one highly confusable group in orthographic form. The groups appeared in the same randomized order as the stimuli were played back later on. After a short introduction read by the speaker who had produced the corpus, the stimuli were presented by first stating the number of the stimulus, followed by a pause of one second, the stimulus proper, and a pause of two seconds before the next stimulus. Subjects were asked to mark the word in each group that matched the stimulus best (forced choice). A total number of 330 stimuli was presented.

4.3. First Results

Evaluation of subjects' judgments yielded a mean inter-subject correlation of 0.75. Corner stimuli associated with each of the words in a highly confusable group were correctly identified at a mean of 78.7% when embedded in the carrier sentence and 75.2% in isolation. Analysis shows that confusions mainly concern the distinction between long and short vowel words (93.0 % of errors, embedded condition) whereas tone confusion occurs relatively seldom (6.6% of errors, embedded condition), the rest being double tone/quantity confusions. When words are presented in isolation, tone confusions rise to 24% of errors, and (double) tone/quantity confusions to 9.5%. In Figure 3 we see pooled results for the word groups 'saang' and 'waan' embedded in the carrier sentence and in isolation. In the figures, the matrix of stimuli is indicated by the meshpoints in the grid, with the four corner stimuli indicated by black stars. There are three intermediate stimuli in the tone

direction and three in the vowel quantity direction. Lines of equi-probability of identification are drawn on the 90, 80, 70, 60, and 50 % levels around the corner stimuli.

Corner stimuli in the left-most figure, for instance, correspond to the words 'saang1' (left bottom corner), 'saang4' (right bottom corner), 'sang1' (left top corner), and 'sang4' (right top corner). The value of 86 written in the box next to the left bottom corner means that the corner stimulus was identified as pertaining to the word 'saang1' by 86% of the subjects. The contours of 80, 70, 60 and 50% drawn around the corner stimulus indicate the decrease of the vote 'saang1' as we move away from the corner stimulus, that is, gradually move from a low tone to a rising tone, and from a long vowel to a short vowel word condition. The categorical judgment shifts where the 50 % lines around the corner stimuli coincide. There is, however, a region where neither of the four words in the highly confusable group reaches 50%. The orientation of the contours indicates how the judgment is influenced by the tone and quantity properties, that is, vertical lines suggest mainly distinction by tone, and horizontal lines distinction by vowel quantity (see 50 % line of vote 'saang1'). Words when presented in isolation are more often identified as bearing a short vowel. The regions where none of the choices reaches 50% generally increases for the isolated condition. Due to space limitations an interpretation will only be given for the low tone/rising tone contrast in the left two panels of Figure 3. For the long vowel condition (bottom of panels), presentation in isolation shifts the categorical 50% boundary further to the left, favoring the rising tone. This can be explained by the fact that the low tone is defined with respect to the preceding relatively higher syllable which is not present in the isolated condition. The *F0* rise which is the perceptual cue of the rising tone, however, prevails as it occurs in the long vowel. In the short vowel word condition (top of panel), this rise occurs later towards the coda of the syllable and reaches its maximum in the (not present) following syllable. Here we observe the opposite effect: The low tone is favored over the rising tone.

5. DISCUSSION AND CONCLUSIONS

The current paper discussed a perceptual study on tone and vowel quantity distinctions in Thai. Analysis on a corpus of highly confusable mono-syllabic words showed that tone features and vowel quantity are statistically uncorrelated. The syllabic tones can be accurately modeled by using tone commands of positive and/or negative polarity with the exception of the mid tone which can be modeled by the phrase component only. Prototypical resynthesis stimuli corresponding to the highly confusable words were correctly identified with a probability of 78%, confusions usually occurring between long and short vowel words. This relatively low figure suggests that subjects usually rely on the (semantic) context for correct identification, especially with respect to vowel quantity, since the latter distinction is not strongly reflected by syllabic duration either. However, our results indicate that prosodically manipulating an inventory of mid tone/long vowel syllables could present a novel strategy to high-quality Thai speech synthesis. Besides activities to this effect, a more detailed evaluation of results from the perception experiment is in progress.

6. REFERENCES

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

[2] Fujisaki, H.; Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E)*, **5**(4), 233-241, 1984.

[3] Fujisaki, H., Hallé, P. and Lei, H., "Application of F_0 contour command-response model to Chinese tones," *Reports of Autumn Meeting, Acoustical Society of Japan,* **1***:* 197-198, 1987.

[4] Potisuk, S., Harper, M. P., and Gandour, J., "Classification of Thai tone sequences in syllable-segmented speech using the Analysis-by-Synthesis method, " *IEEE Trans. Speech and Audio Proc*., 7(1): 95-102, 1999.

[5] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters," in *Proceedings ICASSP 2000*, vol. 1, 1281-1284, Istanbul, Turkey, 2000.

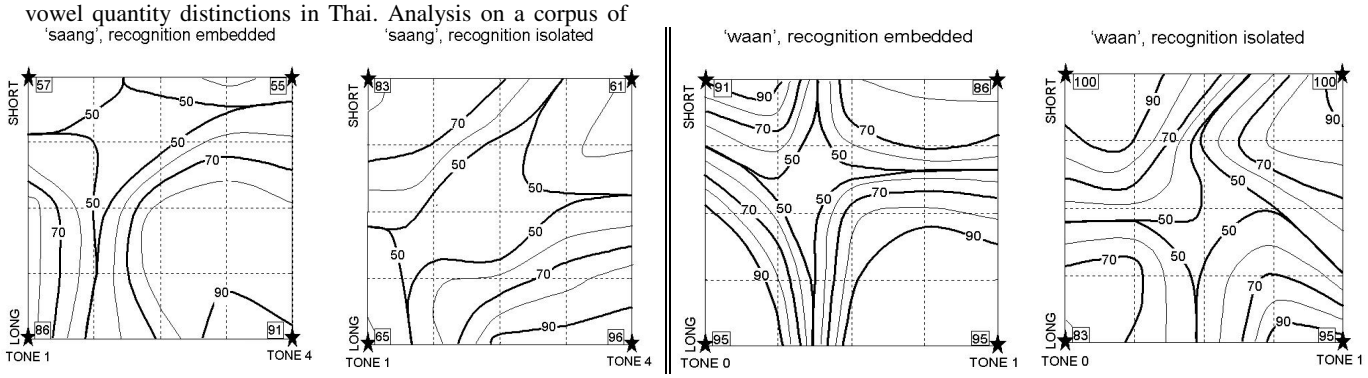

Figure 3: Results of perception experiment for syllable 'sa(a)ng' (left two panels) and 'wa(a)n' (right two panels) under embedded (left) and isolated (right) conditions. The black stars denote the corner stimuli of which the respective identification rate is given in percent. Intermediate stimuli are located at equal distance on the mesh points of the grid. Lines of equi-probability are marked at 50, 60, 70, 80 and 90 %. identification rate, respectively.