Visual Cues in Thai Tone Recognition

*Hansjörg Mixdorff *) & Patavee Charnvivit**)*

*) Faculty of Computer Science, TFH Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de
**) CRSLP, Chulalongkorn University, Bangkok, Thailand
patavee@chula.com

Abstract

The current paper presents preliminary experiments on the exploitation of visual cues in the perception of Thai tones. The lower half of a female speaker's face was recorded on digital video as she uttered a set of monosyllabic tokens covering the five different tones of Thai. The trajectories of 12 light points on the speaker's face were traced frame-by-frame and converted into velocity tracks which were then time-aligned with the speech signal. In parallel, a perception experiment was conducted in which the silent movies were shown to native speakers of Thai who had to decide which tone they perceived. Results, so far, are inconclusive. Whereas the identification results by the speaker herself being the subject are clearly above chance level, results from the other subjects are not and suggest that visual cues only are not sufficient for reliable identification. Since, however, subjects in Thailand were presented a compressed movie at a reduced resolution this circumstance might have also flawed their results.

1. Introduction

It is commonly known that syllabic tones in tone languages are connected with distinct *F0* patterns (rising, falling etc.). As has been shown in earlier works these patterns in Thai and other tone languages can be associated with underlying tone commands of the Fujisaki model [1][2][3][4]. Works on audio-visual data, however, suggest that speakers also exploit visual cues when identifying tones[5]. The current study presents a preliminary experiment using digital video data and examines whether articulatory cues might be present in the facial image. This kind of approach is also of interest because of its relatively limited hardware requirements as compared to *Optotrak*, for instance.

There has been so far relatively limited research with respect to the integration of acoustic and visual information in the perception of tone. An early EMG study [6] suggests that each tone of Thai is connected with distinct enervation patterns of the muscles involved. In the context investigated in the current study, the behavior of extrinsic muscles controlling *F0* is of special interest, as these so-called strap muscles are directly connected with the articulatory system, that is, the muscles of the jaw and tongue. More recent physiological studies [6] also suggest certain restrictions with respect to the coordination of the laryngeal and articulatory systems which might also be responsible for visual cues of tones. The current paper first discusses the method of analysis of the production data and then presents an associated perception test.

2. Speech Material and Method of Analysis

The set of monosyllabic tokens chosen for the current experiment contains the syllables [waan], [maai], [loon] and [raang] which were uttered by a female native speaker of Thai with the five different tones four times each. It must be stated that not all of the tokens present meaningful words. A subset of meaningful tokens only was later used in the perception study. The image of the lower half of the speaker's face was directly frame-grabbed onto a PC, as well as the associated audio which was recorded through a head-worn microphone.

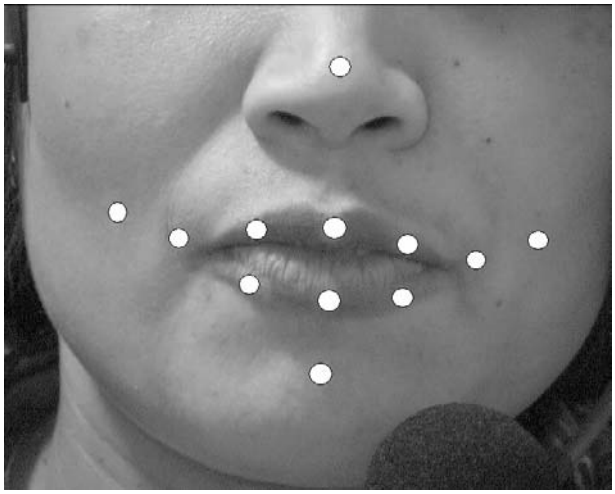

Figure 1: *Light points as marked in blue on the speaker's face.*

The video was grabbed at DV standard (720 x 576 pixels, 25 frames per second). In order to trace the movements on the face, twelve blue light points were marked on the chin, the edges of the mouth; the upper and lower lips etc. (see Figure 1). The video was cut using *Adobe Premiere 6.0* and output frame-by-frame

in bitmap format. A *Matlab* program was written which detected the blue light points in the bitmaps and calculated the center of gravity of all pixels belonging to the same light point. From the coordinates of the light points in the first frame of a sequence their locations (nose, chin, etc.) was determined manually and then an automatic tracking routine was used for tracing the subsequent movements of the points. In order to time-align the 2D data with the speech signal, the velocity of the points was calculated and the principal direction of the associated movement. Predominantly downward and left-ward movements (with respect to the image) were assigned a negative sign and vice versa. In order to smooth the velocity tracks they were filtered with a low pass with a cut-off frequency of 2.5 Hz, stop frequency 5 Hz, and attenuation of 60 dB. The resulting tracks associated with the words were later analyzed using fourth order polynomial approximation.

The audio signals were output separately and down-sampled from 48 kHz to 16 kHz. The waveform was labeled on the phone level and *F0* contours were extracted at a step of 10 ms using *Praat*. Fujisaki parameters were estimated automatically and if necessary corrected. The fundamental frequency contours were extracted at a step of 10 ms and analyzed using the Fujisaki model employing the strategy developed in [3] 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.

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

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

3. Results of Analysis

Figure 2 shows results of analysis for the syllable [maai]. The figure displays from top to bottom: The speech waveform, the extracted F0 contour (+ signs), the Fujisaki model-based contour (solid line), the velocity patterns of three facial points (lower lip, centre: solid line; upper lip, centre: dashed line; left corner of lip: dotted line), the underlying phrase and tone commands of the Fujisaki model. The velocity is scaled in pixels per frame for the lower lip, whereas the pattern has been amplified by five for the two other points for better display.

As can be seen, the tones of Thai (from left to right: mid, low, falling, high and rising tones) are modelled using distinct combinations of Fujisaki tone commands as given in Table 1, with the exception of Tone 2: When this tone is uttered in isolation, an additional negative tone command is required to account for the quick drop in *F0* observed towards the offset of the utterance.

The velocity tracks suggest a close alignment of movements to the onset of the rhyme. Although the maximum speed of the points slightly differs across tones in the current example, this is not evidence for significant articulatory differences between the tones. Comparison of polynomial coefficients of tokens pertaining to the same group of word and tone does not yield consistent differences either.

4. Perceptual Test

In order to examine why the articulatory data had yielded dissatisfactory results, a perception experiment was devised which used all tokens in the data set pertaining to meaningful words as given in Table 2. Two randomized lists of these words were created, each containing every token twice and therefore a total of 24 items. The associated video sequences were then presented in silent mode. Subjects were given a numbered list of the words in the sequence in which they occurred and were asked to select the word which they thought had been uttered. During presentation, first the number of the word was displayed and then the associated video sequence. The video was stopped for a moment until the subject had made her/his choice.

Due to the size of the raw videos originally produced in Berlin (>500 MB each) they had to be reduced to half of the resolution (360 x 288) and compressed using the Indeo Video 5 Codec for transfer to Thailand. At Chulalongkorn University, Bangkok, eleven subjects (six male, five female) took part in the perception experiment. Each of the subjects watched the two movies once.

The statistical chance level for the current experiment (disregarding the possible distinction between common and less common words) is of 33%, that is, (4 x 25% + 6 x 33% + 2 x 50%)/12. On the average subjects reached 37% correct decisions on List1, and only 28% on List2, hence neither result was significantly different from chance. The best subject's average result was of 40% correct votes, though on individual lists there were cases in which 50% were reached. These results, however, are not significant. In order to examine how confusions are distributed and whether certain tendencies, that is, articulatory proximity or distance between tones can be observed, seven trials with 38% correct results and above were

evaluated and results pooled. The resulting confusion matrices are displayed in
Table **3**.

Table 2: *Words chosen for the perceptual study, tones and English translations.*

word	English translation	word	English translation
waan0	to request	maai2	divorced
waan1	to scatter	maai3	wood
waan2	a kind of plant	maai4	to order
waan4	sweet	raang0	rail
loon3	bald	raang2	body
loon4	grand-grandson	raang3	deserted

Table 3*: Confusion matrices of pooled results from seven selected trials. Rows indicate intended tone, columns perceived tone rounded to full percents.*

	waan					loon		
	0	1	2	4			3	4
0	**36**	36	0	28		3	36	64
1	14	**36**	21	29		4	50	50
2	14	29	0	57				
4	14	21	7	**57**				
	maai					raang		
	2	3	4			0	2	3
2	**36**	7	57		0	21	36	43
3	14	**64**	21		2	43	**36**	21
4	21	36	**43**		3	21	14	**64**

As can be seen, results are not evenly distributed and must be discussed for each word individually. It should also be noted that the following discussion is rather speculative due to the poor overall recognition results.

Apparently there were tokens which were identified well above chance level (set in bold face), such as waan4, maai3 and raang3.

In the case of [loon] none of the tokens received a majority vote; therefore visual cues do not seem to have served disambiguation. This result could also suggest articulatory proximity between Tones 3 and 4, an assumption supported as well by the results for [maai]. Considering that the associated *F0* contours are relatively similar (see Figure 2) there might indeed be a closer relationship between these tones. On the other hand one could also argue that the words [maai3] (wood) and [waan4] (sweet) are the most commonly used members of their groups and therefore might have attracted more votes than the others.

As can be seen there also is a large number of instances of Tone 2 being confused with Tone 4, rather than with Tone 3, a confusion which does not occur as often in the reverse direction. Since Tone 2 has an *F0* pattern clearly different from Tone 4, this result is especially hard to interpret.

Table 4: *Confusion matrices of perception test results by the speaker of the data set. Rows indicate intended tone, columns perceived tone rounded to full percents.*

	waan					loon		
	0	1	2	4			3	4
0	13	0	13	0		3	63	38
1	0	**88**	0	25		4	50	50
2	0	0	**63**	13				
4	88	13	25	**63**				
	maai					raang		
	2	3	4			0	2	3
2	**38**	25	38		0	**63**	0	38
3	0	**63**	38		2	0	**63**	38
4	0	50	**50**		3	38	0	**63**

The female speaker who had produced the data set performed the perception experiment twice at a spacing of five days and scored significantly better than the group at Chulalongkorn University (on the average 56% correct, see Table 4 for details). Different from the other Thai subjects she was presented the original high quality DV recordings. The fact that she was an expert with respect to her own speech and a certain learning effect might explain her good results. Detailed analysis, however, revealed that also she was unable to distinguish [loon3] from [loon4], and consistently misclassified [waan0] as [waan4]. All other tokens were identified well above chance level. On the videos presented to the subjects at Chulalongkorn University, however, her performance dropped to only 38%. This suggests that the video compression and reduced resolution had had an adverse effect. Due to time limitations the experiment could not yet be repeated on the high quality video data, but will follow later.

5. Discussion and Conclusions

In the current paper a preliminary experiment concerning the visual perception of tones in Thai was presented. A method for tracking facial points on video was employed for calculating time-aligned velocity patterns. Although these patterns capture the movement of points on the mouth over time, they do not (yet) seem to permit the detection of specific visual cues of tones. The 3D-2D-1D data translation applied in the method as well as measuring inaccuracies might have flawed the results.

As follows from the outcomes of the perception experiment, visual cues only do not facilitate a reliable identification of tones by native subjects, though certain tendencies can be observed such as the similarity of Tones 3 and 4. The obvious deterioration of recognition results on the compressed video data as compared with the DV quality video suggests that cues are attenuated by the compression algorithm which poorly keeps track of the subtle frame-to-frame changes in the mouth region.

Articulatory gestures are in the first place a function of the word being uttered and only secondarily a function of tone, therefore results remain inconclusive as several factors (possible confusions in a word group, frequency of words involved, articulatory realization) come into play. Still the tendencies of occurring confusions provide tentative hints as to the articulatory proximity and distance between tones.

In order to gain a better understanding regarding the articulatory closeness of tones, future experiments will employ audio-visual stimuli which feature residual acoustic information, that is, speech distorted by noise, for instance.

6. References

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

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

[3] Mixdorff, H., Luksaneeyanawin, S., Fujisaki, H. and P. Charnvivit, 2002. Perception of Tone and Vowel Quantity in Thai. *Proceedings of ICSLP 2002*, Denver, USA.

[4] Mixdorff, H., Hung, N. et al., 2003. Quantitative Analysis and Synthesis of Syllabic Tones in Vietnamese. *Proceedings of Eurospeech 2003*, Geneva.

[5] Burnham, D., Ciocca, V., & Stokes, S., 2001. Auditory-visual perception of lexical tone. In *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 395-398.

[6] Erickson, D., 1976. *A Physiological Analysis of the Tones of Thai*. PhD thesis, University of Connecticut.

[7] Xu, Y. and Sun, X., 2002. Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America* 111: 1399-1413.

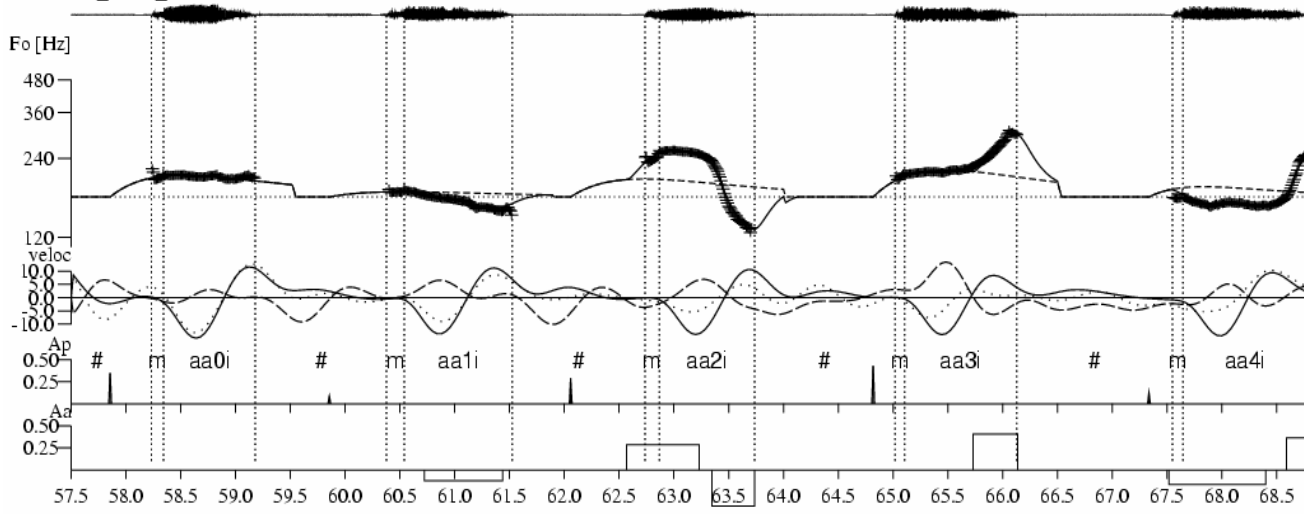

Figure 2: *Results of analysis,from top to bottom: The speech waveform, the extracted F0 contour (+ signs), the Fujisaki model-based contour (solid line), the velocity patterns of three facial points (lower lip, centre: solid line; upper lip, centre: dashed line; left corner of lip: dotted line), the underlying phrase and tone commands of the Fujisaki model. As can be seen, the tones of Thai (from left to right: mid, low, falling, high and rising tones) are modelled using distinct combinations of Fujisaki tone commands. The velocity tracks suggest a close alignment of movements to the onset of the rhyme. Although the maximum speed of the points slightly differs across tones - possibly due to measuring inaccuracies - this is not yet enough evidence for significant articulatory differences between the tones.*