# Are there Facial Correlates of Thai Syllabic Tones?

*Hansjörg Mixdorff*, Denis Burnham **, Guillaume Vignali** and Patavee Charnvivit****

*Faculty of Computer Science and Media, TFH Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de

**MARCS Auditory Labs, University of Western Sydney, Australia
denis.burnham@uws.edu.au; guillaume@vignali.net

***CRSLP, Chulalongkorn University, Bangkok, Thailand
patavee@chula.com

## Abstract

This paper deals with the influence of tones on syllabic articulation in Thai. Motion capturing of 24 facial points in the face of a female speaker was performed using an *Optotrak* system as she uttered 24 sets of syllabic tokens covering the five different tones of Thai 12 times each. After rigid and non-rigid-movements had been separated, a PCA was conducted on the non-rigid data. In order to determine the influence of the tones on the facial movement the first PC reflecting the jaw opening was analyzed by aligning the derivatives of the first PCs with respect to the point of maximum velocity and averaging over all tokens of a syllable/tone combination. Analysis showed great similarities in the shapes of the resulting mean velocity contours. In some syllable sets, however, certain tones exhibited specific temporal alignments that were strongly correlated with the underlying syllable duration. This outcome suggests that certain syllable/tone combinations require a specific temporal alignment of articulatory and tonal gestures, though a consistent physiological explanation remains yet to be found.

## 1. Introduction

Syllabic tones in tone languages have distinct *F0* patterns. Thai has five different lexical tones: three static tones - mid (0), low (1) and high (3); and two dynamic tones, falling (2) and rising (4) (commonly used tone indices are given in brackets). Research has shown that tone contours patterns in Thai and other tone languages can be associated with underlying tone commands of the Fujisaki model [1][2]. While this might suggest that tone is a purely acoustic phenomenon, there are now auditory-visual studies that suggest that speakers also exploit visual cues when identifying tones [3][4]. In addition a preliminary study by two of the current authors [5] suggested that tone identification in a video-only condition was difficult and depended on the syllables employed and the tones contrasted. Nevertheless, under certain cases identification was much better than chance. This observation plus the similar results found previously [8] provided the impetus for the current study.

There has been so far relatively limited research with respect to the influence of tones on articulation. An early EMG study [6] suggests that each tone of Thai is connected with distinct enervation patterns of the muscles involved. In the context of 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. A recent production study [7] also suggests certain restrictions with respect to the coordination of the laryngeal and articulatory systems which might be responsible for visual cues of tones. In the associated realm of prosody, it has been shown that there is a strong correlation between head movements and F0 [8]. These correlations are continuous and seem to be used by multimodal perceivers during auditory-visual perception [9], but direct studies on the perception of these movements are yet to be conducted.

The current paper focuses on production data in Thai that were collected in the context of a larger study of Thai, Vietnamese and Mandarin. For each of these languages mono-syllabic corpora were collected that consist of audio and video, as well as *Optotrak* recordings. Only the production data for Thai is presented here. Results on the perception of tones from associated video data are published in [10]. They suggest that listeners benefit from visual cues when performing tone identification on noisy speech, but yield only small gains when the tonal information is no longer present in the speech signal as in devoiced stimuli.

## 2. Speech Material and Method of Analysis

The corpus of mono-syllabic tokens compiled for the current experiment contains a total of 24 syllables which were uttered by a female native speaker of Thai with the five different tones. The 24 syllables were chosen based on the following criteria: (1) a maximum number of members in each syllable set should represent real words, (2) a good coverage of Thai vowels, as well as

articulatory trajectories (for instance, tongue movements from the back to the front, from the front to the back, etc.). All syllables are so-called "live" syllables with a sonorant coda. A list of the syllables used is shown in Table 1.

*Table 1:* List of syllables used in the study, Thai-SMPA notation.

| | | | |
|---|---|---|---|
| cun | lang | luuang | pvng |
| jaw | law | maaj | s@@n |
| jing | lim | man | seng |
| joong | liw | muj | siiaw |
| khiian | lom | muuaj | waaj |
| khlum | lon | ngaw | waan |

The tokens were randomized and recorded 12 times each. 24 facial points in the speaker's face were parked with active IR markers and the 3D-positions of the points captured at a rate of 60 frames per second, together with the associated audio at 48 kHz, 16 bit at MARCS Auditory Labs. The positions of the *Optotrak* markers are displayed in Figure 1.

*Figure 1:* Positions of *Optotrak* markers in the speaker's face. The microphone was positioned under the chin. The head rig was used for capturing rigid head movements.

In order to segment the long motion sequences into chunks of individual tokens, the audio tracks were downsampled to 16 kHz and annotated on the syllable elevel using *Praat TextGrid* [11]. A *Matlab* tool was written for linearly interpolating missing marker positons and chunking the motion data into sections pertaining to individual syllables based on the *TextGrids,* the associated soundtracks were written to individual wave files using a *Praat Script*. The resulting wave files were checked for correct syllable/tone combinations and erroneous tokens discarded. *F0* contours were extracted at a step of 10 ms using *Praat* default settings. After rigid and non-rigid movements in the *Optotrak* data had been separated, Principal Component Analysis was performed on the non-rigid part. It showed that five PCs were sufficient for accounting for 95% of the variation. Since we were mostly interested in the alignment of the jaw movement with the tonal contours we focused our further analysis on the first PC that is most strongly correlated with the jaw opening. We expected that if the tones had a measurable influence on articulation, the jaw trajectories associated with tokens of the same syllable, but different tones should as well differ.

Following the approach in [4] we calculated the derivative of the first PC and yielded contours of jaw velocity whose polarity reflects the direction of the movement, that is, negative polarity indicates jaw lowering, and positive polarity jaw raising. Since these contours are most strongly influenced by the articulatory gesture pertaining to the underlying syllable, each syllable set exhibits specific velocity contours. Most of these contours are characterized by conspicuous local extremes which mark the fastest point in the jaw movement. In some syllables these extremes occur during jaw lowering, in others during jaw raising. We therefore used the local extrema in each syllable set as a point of reference for aligning and averaging over velocity contours from tokens of the same syllable/tone combination, yielding five average velocity contours for every syllable set, one for each type of tone. As the local minimum can occur between two frames, and the excursion cycle is not necessarily symmetrical, two points around a local minimum were searched at which the velocity was 30% less than in the local minimum and the point in time halfway between the two points used as the alignment point for the contours.

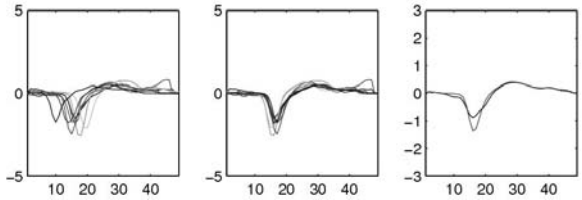


*Figure 2:* Example of contour alignment, syllable 'seng', before (left), after (center) and average contours (right).

Figure 2 shows an example of velocity contours before (left) and after (center) the alignment procedure for the syllable 'seng'. The right panel displays the averaged contours before (dark grey) and after (light grey). As can be seen, the average contour based on aligned contours captures the shape much more faithfully. The *F0* contours associated with the tokens were subjected to realignment using the alignment points from the velocity contours. Since the *F0* contour may contain unvoiced frames (*F0*=0 Hz), averaging was performed taking only voiced frames into account.

Subsequently five average velocity and *F0* contours were calculated for each syllable set. Since each syllable is distinct in its characteristics, comparison was performed within each syllable set. It should be

mentioned that some syllables like 'khlum' are not connected with any marked jaw movements and therefore the alignment fails. We will present example results from some of the syllable sets and discuss them individually.

## 3. Results of Analysis

Figure 3 and Figure 4 display examples of averaged *F0* contours (top panel of every sub-figure) and jaw velocity contours (bottom panel). The type of tone is indicated by the line styles used. Trajectories for tone 0 are represented by solid lines, for tone 1 by dashed lines, for tone 2 by + symbols, for tone 3 by dot-dashed lines, and for tone 4 by dotted lines.

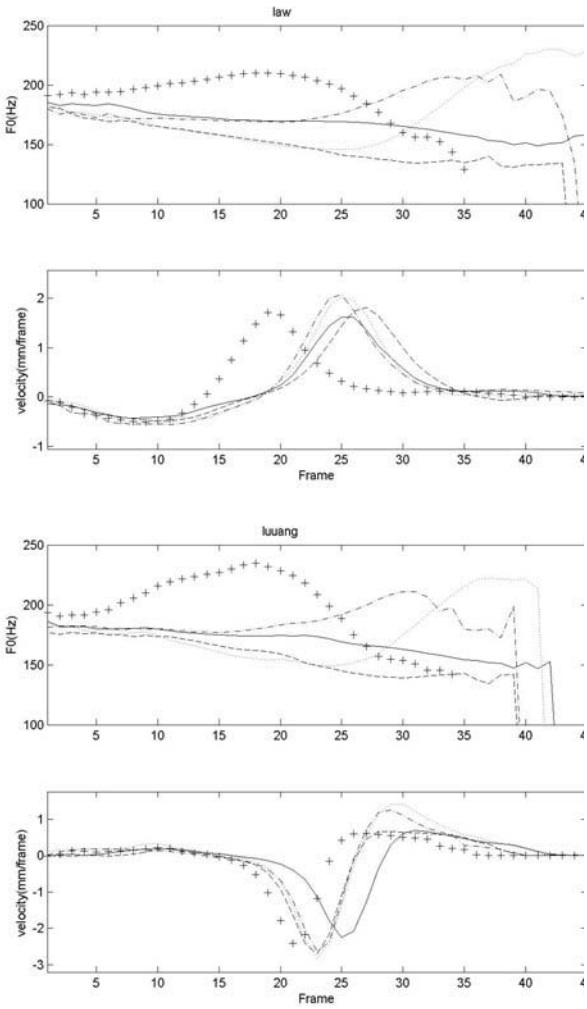


*Figure 3:* Averaged *F0* contours and jaw velocity tracks for the syllables 'law' (top) and 'luuang' (bottom). Trajectories for tone 0 are represented by solid lines, tone 1 by dashed lines, tone 2 by + symbols, tone 3 by dot-dashed lines, and tone 4 by dotted lines.

The time is expressed in frames, and velocity measured in mm/frame. As can be seen, the velocity contours have different shapes for each syllable. Within a syllable set, the contours are similar, but to varying degrees.

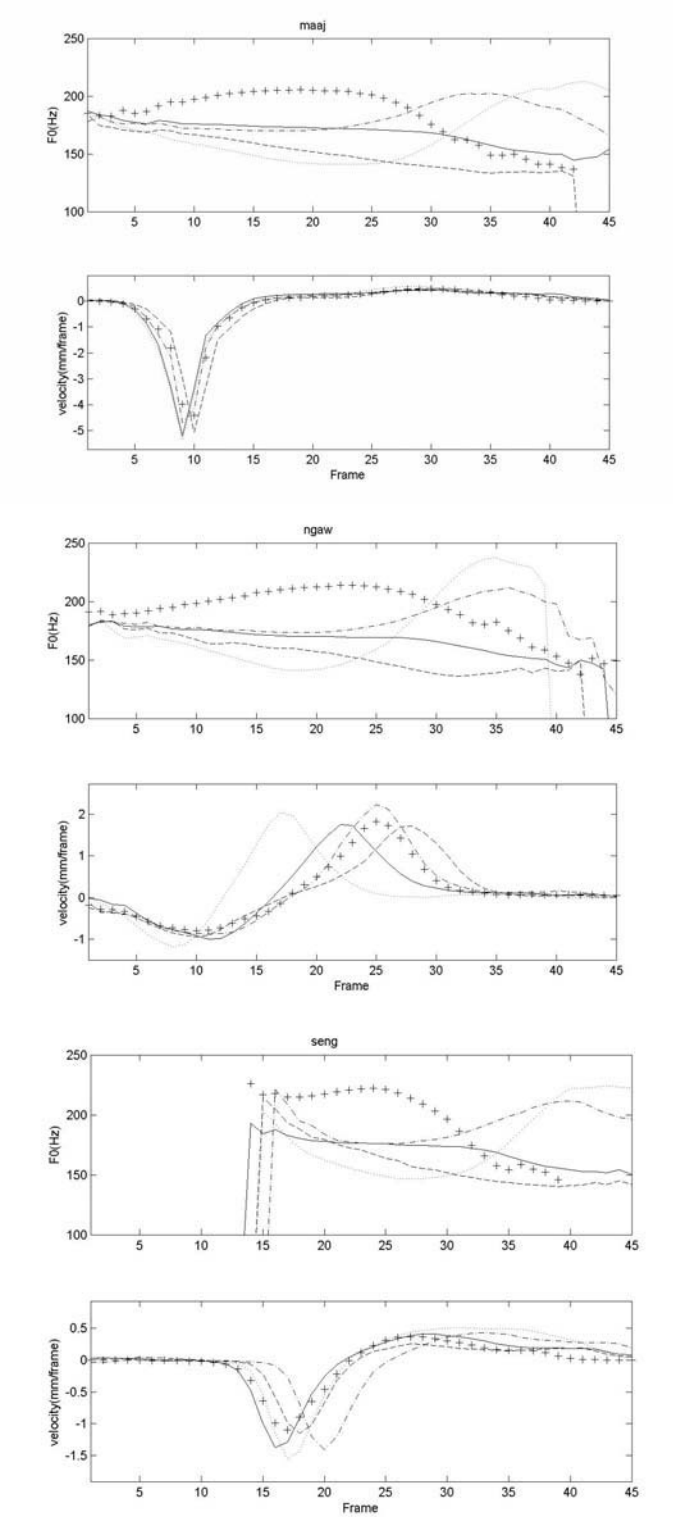


*Figure 4:* Averaged *F0* contours and jaw velocity tracks for the syllables 'maaj' (top), 'ngaw' (center) and 'seng' (bottom).

In the example 'maaj', contours for all tones are very closely aligned. In contrast, in the case of the syllable 'law', the contour for tone 2 clearly precedes all others, whereas in the syllable 'ngaw' it is the contour of tone 4.

In the remaining cases, only some tones have contours closely aligned with each other. In order to express the similarity of contours in a set of syllables by a numerical value we calculated the cross correlations between the contours pertaining to different tones The figures for the syllables 'law' and 'maaj' are displayed in *Table 2* and show the good agreement for 'maaj' as well as the distinctness of tone 2 in 'law'.

*Table 2:* Similarity of jaw velocity contours as expressed by the cross correlation for tones 0-4, syllables 'law' and 'maaj'.

| law | | | | |
|---|---|---|---|---|
| Tone | 1 | 2 | 3 | 4 |
| 0 | 0.952445 | 0.413284 | 0.980926 | 0.990585 |
| 1 | | 0.333184 | 0.889028 | 0.933126 |
| 2 | | | 0.429369 | 0.387981 |
| 3 | 0.889028 | | | 0.985234 |
| maaj | | | | |
| Tone | 1 | 2 | 3 | 4 |
| 0 | 0.828291 | 0.944623 | 0.974445 | 0.997498 |
| 1 | | 0.959225 | 0.922417 | 0.844109 |
| 2 | | | 0.991329 | 0.953172 |
| 3 | 0.922417 | | | 0.977968 |

Further analysis reveals that earlier velocity maxima are generally connected with shorter durations of the underlying syllable ($\rho$=0.246, $p < 0.01$). Since, however, this is not a general property of a particular tone -implying, for instance, that all syllables with tone 2 are shorter, which is not the case- this outcome might suggest that in specific syllable/tone combinations the realization of the tonal contour might somehow conflict with the articulatory gesture. More in-depth analysis is however required to relate our observations to the underlying physiology.

## 4. Discussion and Conclusions

In this paper we presented a study of the influence of Thai tones on the articulation of the underlying syllables. *Optotrak* and speech recordings of a corpus of 24 syllable sets were performed, *F0* contours extracted and the non-rigid part of the motion determined. By means of PCA we calculated the contribution of the jaw movement and performed a syllable-wise comparison of aligned and averaged velocity contours. Our outcome shows that since the velocity trajectories are primarily caused by the underlying articulatory movement their shapes are very similar for one and the same syllable. In certain syllables, however, we found significant variation in the temporal alignment of the contours pertaining to different tones. Although we cannot yet attribute these differences to physiological causes they suggest that articulation might be compromised in certain tone/syllable combinations. Future efforts will therefore be dedicated to further investigation of the relationship between tonal and articulatory gestures.

## 5. Acknowledgements

This work was supported by a grant from the University of Western Sydney International Research Initiatives Scheme, as well as a DAAD short-term lecturership, grant D/04/01405. The assistance of Rua Haszard Morris and Colin Schoknecht in providing *Matlab* programs and running the *Optotrak* experiments is gratefully appreciated.

## 6. References

[1] 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.

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

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

[4] Burnham, D., Lau, S., Tam, H., & Schoknecht, C. "Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers," in Massaro, D., Light, J., & Geraci, K. (Eds) *Proceedings of AVSP2001*, 2001.

[5] Mixdorff, H. and Charnvivit, P. "Visual Cues in Thai Tone recognition," Proceedings of TAL 2004, pp. 143-146, Beijing, China, 2004.

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

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

[8] Yehia, H.C., Kuratate, T., & Vatikiotis-Bateson, E. "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics*, 30, 555-568, 2002.

[9] Vatikiotis-Bateson et al., "Task constraints on robot realism: The case of talking heads." In K. Kamejima (Ed.), *9th IEEE International Workshop on Robot & Human Interactive Communication* (pp. 352-357). Osaka, Japan: IEEE., 2000.

[10] Mixdorff, H., Charnvivit, P. and Burnham, D. "Auditory-visual perception of syllabic tones in Thai." To appear in *Proceedings of AVSP 2005*, Vancouver Island, Canada, 2005.

[11] http://www.praat.org

[12] http://www.virtualdub.org