Visual Correlates of Thai Lexical Tone Production: Motion of the Head, Eyebrows and Larynx?

Denis Burnham[1], Weicong Li[1], Chris Carignan[2], Virginie Attina[1], Benjawan Kasisopa[1], Eric Vatikiotis-Bateson[3]

[1]MARCS Institute for Brain, Behaviour & Development, Western Sydney University, Australia
[2]Institute of Phonetics & Speech Processing, Ludwig-Maximilians Universitat Muchen, Germany
[3]Communication Dynamics Lab, Department of Linguistics, University of British Columbia, Canada

denis.burnham@westernsydney.edu.au; weicong.li@westernsydney.edu.au;
c.carignan@phonetik.uni-muenchen.de;vattina@yahoo.fr;
b.kasisopa@westernsydney.edu.au; evb@mail.ubc.ca

Abstract

There is well-established evidence that visual articulatory information in the face and head aids identification and discrimination of lexical tone. However, the nature and locus of this information is only beginning to be specified. In previous work we identified a predominant role of head motion over face motion in both the perception and production of Cantonese lexical tone, the latter using OPTOTRAK motion tracking. We have now extended the set of OPTOTRAK markers to include the eyebrows and the larynx, and collected data from a corpus of Cantonese, Thai and Mandarin speakers. Here we report on a Thai speaker producing the five Thai tones on four Thai syllables in isolated words and sentences and in normal, whispered, and Lombard speech. Principal components (PCs) for the face (eyebrows, lips, jaw), the larynx and for independent head movement were extracted and linear mixed model analyses of range of PC1 scores revealed good differentiation on the basis of syllable identity and context and speech style. Of particular importance, the five Thai tones were best differentiated by head and larynx motion. So, these results add larynx motion as a possible visible cue for tone perception. Studies across speakers and the three languages will follow.

Index Terms: speech production, articulation, lexical tone, auditory-visual speech, face motion, head motion, laryngeal motion.

1. Introduction

Speech perception is auditory-visual in noise [1] and in clear conditions [2], and this has been shown clearly in languages that use just consonants and vowels to distinguish meaning. But the vast majority of the world's languages also use lexical tones on single syllables or pitch-accent on syllable combinations to distinguish meaning [3,4]. At the turn of the century, we found that identification of the six Cantonese tones by Cantonese perceivers is facilitated in auditory-visual (AV) compared with auditory-only (AO) conditions [5,6], and that discrimination of these six Cantonese tones by non-native tone language (Thai) perceivers and non-native non-tone perceivers (Australian English) is also facilitated in (AV) compared with auditory-only (AO) conditions [7]. Since then various other studies have found evidence for the use of visual information in the perception of lexical tone [8-15], but few have investigated the nature or locus of this visual information. Following OPTOTRAK analysis of the Cantonese speaker in our previous studies [5,6,7] we compared the role of face-only vs. head-only motion information in VO tone perception and found (i) non-rigid face-only motion information is a sufficient condition for visual enhancement of Cantonese phone (vowel and consonant) but not tone perception, and (ii) that rigid head-only motion information is a necessary condition for visual enhancement of Cantonese tone but not phone perception [11]. And this predominance of head over face motion in visual perception of tones was provided with some support in a later study [12]. Moreover, further discriminant analyses in our lab have found that head-only motion better predicts Cantonese tone category membership than does face-only motion.

These previous studies with Cantonese in our lab have a number of limitations: Only Cantonese was studied, and only one Cantonese speaker was employed, and the OPTOTRAK markers on the face, while relatively comprehensive, did not include markers on the eyebrows and larynx. There is evidence that eyebrow motion is related to the prosody of continuous speech [16], and that head motion is related to F0 in speech [17,18], an effect that may be related to tension in the cricothyroid muscle, which in turn may be related to visible laryngeal motion.

We have now collected OPTOTRAK data from 30 speakers in three tone languages, Cantonese, Mandarin and Thai, and have included eyebrow and larynx OPTOTRAK markers. In this preliminary report we present data from one Thai speaker. We predict that (1) Head-Only motion will better differentiate between Thai tones than will Face-Only motion, and (2) within the face both motion of the eyebrows and of the larynx will add to better differentiation of Thai tones.

2. Methods

2.1. Data collection

As shown in Figure 1, OPTOTRAK markers were attached on different parts of the face, neck and head: 4 markers for the head, 6 for the eyebrows, 6 for the cheeks, 8 for the lips, 5 for the jaw, and 2 for the larynx. The informant was a 35-year-old Thai female. The focus of the study was the production of Thai tones. Thai has five lexical tones, three level tones – Mid 33, Low 21, High 45 – and two contour tones – Falling 241, and Rising 315 (the Chao numbers after each name show the

relative pitch at initial, (medial), and final points in the syllable [20].

The informant was asked to produce four targeted syllables (มา - /maa/, ลน - /lon/, คู - /kuu/, ไค - /kai/) in two Contexts (sentence, isolated word) in three Speech Styles (Normal, Whisper, Lombard (in which white noise was played to the informant through headphones). In each of these 4×2×3 = 24 conditions she produced the syllables in each of the five tones = 120 productions and each production was repeated 5 times. The audio signal was recorded at 44.1 kHz while OPTOTRAK recorded 3D motion data at 60 Hz.

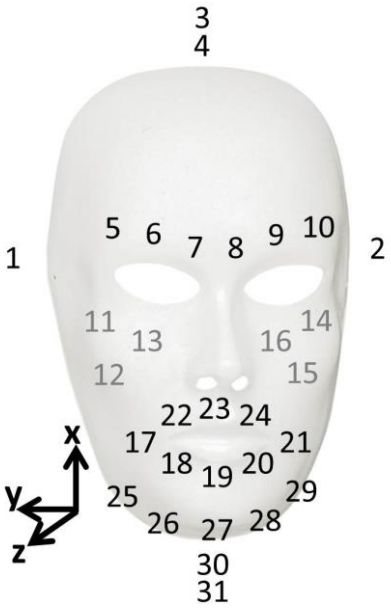

Figure 1: The *Positions of OPTOTRAK markers attached to the head, face and neck. See Table 1 for markers for different locations.*

2.2. Data processing

The audio data were transcribed and aligned at phonetic level in Praat [19] and the 120 target word types were excised. The temporal information of the vowels from the alignment was used as the time markers for the OPTOTRAK data. Note that the primary data were each of the 120 words, even in the sentence condition it was the target word that was excised and analysed.

The OPTOTRAK data were pre-processed and subsequent analyses performed in MATLAB. First, the data from all the OPTOTRAK markers were checked and some markers containing large portions of abnormal values (more than 10^25 times larger than the average) were treated as 'dead markers' and excluded. In addition, for each frame, if any data from each marker contained abnormal values, the whole frame was also excluded.

The 3D motion data from the markers of the head (as shown in Figure 1, two at top and two at two sides of the head) were used to determine absolute head motion. This was regarded as the reference to calculate the relative motion of different parts of the face and neck (i.e., absolute motion of the face component (whole face, eyebrows only, jaw + lips only, larynx + eyebrows only, larynx, eyebrows only) minus the absolute motion of the head). Note that the motion of the 6 cheek markers were not included in the analyses here. Principal component analysis (PCA) was applied to identify principal components (PCs) that represent independent axes of variation. PCA can use the data in all the frames (observations) of particular combinations of different markers, e.g., all the markers on the face containing 63 sets of coordinates (21 markers each with 3 dimensions). PCA produces PC scores for each observation that indicate the motion of different parts in the PC spaces. With the time information from the audio signals, trajectories of PC scores were obtained.

3. Results and discussion

3.1. PCA

Seven PC models based on different parts of the face and the head were developed: absolute motion of head and face, absolute motion of the head, relative motion of the face (including jaw, lips, larynx and eyebrows), relative motion of jaw and lips, relative motion of larynx and eyebrows, relative motion of larynx only and relative motion of eyebrows only. Table 1 shows the percentage explained by PC1 and PC2 for seven PC models, and the directions of the PC1 and PC2 in each model. As is shown, for all seven PC models, PC1 plus PC2 can explain ~80% of the motion.

Table 1: *Percentage explained by PC1 and PC2 and PC directions for 7 models. ABS stands for absolute motion and REL for relative motion.*

PC model ***and markers***	**PCs**	**PC explained (%)**	**PC direction**
Head + Face (ABS)	PC1	48.2	z, x, y
1-4, 5-10, 17-31	PC2	33.9	y, z, x
Head (ABS)	PC1	52.1	y, z, x
1-4	PC2	33.1	z, y
Face (REL)	PC1	67.9	x, z
5-10, 17-31	PC2	12.8	z, x
Jaw + Lips (REL)	PC1	85.7	x, z
17-29	PC2	7.5	z, y
Larynx + Eyebrows	PC1	61.5	z, x, y
(REL) *30, 31, 5-10*	PC2	18.5	z, y
Larynx (REL)	PC1	76.3	z, x
30, 31	PC2	14.0	y, x
Eyebrows (REL)	PC1	88.6	z, y
5-10	PC2	5.5	x, z

3.2. Linear Mixed Model Analyses

Seven sets of linear mixed models (LME) were conducted, one set for each of the seven sets of OPTOTRAK markers in Table 1. The LME model formula for all seven sets of analyses was:

tone + syllable + rec condition + sentence/word + 1

with range of PC1 scores as the dependent variable.

The results of the mixed model analyses are shown Table 2 for each of the seven sets of analyses across the independent variables: the 4 Syllables, the two Syllable Contexts (word, sentence), the Speech Conditions (Normal, Whispered, Lombard), and the five Thai Tones.

As can be seen in Table 2, the 4 different syllables (irrespective of tone) were differentiated quite well: /lon/ had

less motion than the referent /maa/ on all seven models; and /kuu/ differed from /maa/ on 4 of the 7 models, including more absolute and less relative motion than /maa/; and /kai/ from /maa/ due to more eyebrow combined with laryngeal motion.

Table 2: *Summary of coefficients that are significant ($p < 0.05$) in linear mixed modeling using range of PC1 scores. ABS stands for absolute motion and REL for relative motion; +/− sign in parenthesis shows the valence of the coefficients.*

LME model	Syllable: relative to */maa/*			Context: relative to *sentence*	Condition: relative to *normal*		Tone: relative to *mid 33*			
	/lon/	/kuu/	/kai/	word	Whisper	Lombard	Low 21	Falling 241	High 45	Rising 315
Head + Face (ABS)	✓(−)	✓(+)		✓(+)		✓(+)				
Head (ABS)	✓(−)			✓(+)		✓(+)	✓(+)			✓(+)
Face (REL)	✓(−)	✓(−)		✓(−)		✓(+)				
Jaw + lips (REL)	✓(−)	✓(−)		✓(−)		✓(+)				
Larynx + Eyebrows (REL)	✓(−)		✓(+)		✓(−)	✓(+)	✓(+)		✓(−)	
Larynx (REL)	✓(−)				✓(−)	✓(+)	✓(+)		✓(−)	
Eyebrows (REL)	✓(−)	✓(−)		✓(+)		✓(+)				

This accords with our previous data [5,6,7 and further *in preparation* discriminant analyses] in which phones in syllables, irrespective of tone, were better differentiated from each other by facial motion than by head motion.

The syllables in sentences were differentiated from syllables in isolation by all the models except the larynx: there was more absolute motion and more eyebrows motion in words than in sentences and less lip and jaw motion in words than sentences (see Table 2). This suggests that larynx motion is similar in citation form and in running speech but that there are differences in head, lips and jaw motion, presumably due to the generally shorter duration of words in continuous speech.

Whispered speech was differentiated from normal speech only by less larynx motion, and in contrast, Lombard speech was differentiated from normal speech by greater motion on all the models (see Table 2). This makes perfect sense as (i) in whispered speech F0 (pitch) is absent whereas in normal speech it is present and (ii) Lombard speech involves emphasised articulation, which appears to involve all parts of the articulatory gesture measured here.

Finally, and most importantly for the purposes of this investigation four of the five Thai tone pairs tested were differentiated quite well, and quite specifically. The low and the rising tone had significantly more absolute head motion than the mid tone. In addition, the low and high tones were differentiated from the mid tone on the basis of laryngeal motion; the low tone had *more* laryngeal motion, while the high tone had *less* laryngeal motion – there is a continuum of degree of laryngeal motion from low to mid to high.

4. Discussion

The results for syllables, syllable context and speech condition were as expected, thus confirming the face validity of our data collection and analysis procedure.

There were two hypotheses. In accord with the first hypothesis it was the case that absolute Head-Only motion better differentiated between Thai tones (at least for two of the four pairs tested – mid-low, mid-rising) than the sum of the relative Face-Only motion. The second hypothesis was partially supported: laryngeal motion differentiated tones (low from mid, and high from mid), but eyebrow movement did not. Interestingly, there is a continuum of degree of laryngeal motion from more to less motion from low to mid to high level tones. So, the main additional piece of knowledge uncovered in this study is that there is distinctly different motion of the larynx for the three level tones, as measured by markers placed over the larynx on the skin of the neck.

5. Conclusions

We can conclude that in addition to absolute head motion, motion of the larynx also differentiates the production of tones. As females generally have less prominent larynxes and that these recordings were made with a female speaker and recorded from the skin covering the larynx, it appears that while the laryngeal motion differences are small they still differ between tones. It would be of interest to compare male and female speakers in future studies.

This conclusion is qualified by the fact that only the low and high tones (and not the rising and falling tones) were differentiated from the mid tone. This does, however, make sense; the low mid and high tones are relatively level and so the larynx would assume different positions to adjust for these F0 levels, and indeed there was a continuum of laryngeal motion from the low to the mid to the high level tones. The larynx would also adjust for the two contour tones but motion would occur *within* the tones. In the analyses here each tone was compared with the mid tone. In future studies measurement of differentiation of all 10 possible tone pairs in Thai will be of interest, e.g., especially between the two contour tones, rising vs falling. It will also be of interest to investigate the motion of the head in relation to laryngeal movements, given that it has been proposed that head motion in the vertical x-axis has been implicated in stretching of the cricothyroid muscle and thus subsequent changes in pitch [17].

The results here are confined to a single female Thai speaker. The generality of the findings will be interrogated in our future OPTOTRAK studies with more speakers in each of three different tone languages, Thai, Cantonese and Mandarin.

6. Acknowledgements

The assistance of Dr Nan Xu for data collection is greatly appreciated.

7. References

[1] Sumby, W. H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.

[2] McGurk, H. & McDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746-748.

[3] Fromkin, V. (ed.) (1978). *Tone: A Linguistic Survey*. New York, NY: Academic Press.

[4] Yip, M. J. W. (2002). *Tone* Cambridge University Press, NY, Chapt. 1, pp. 1-14.

[5] Burnham, D., Ciocca, V., Lauw, C., Lau, S., and Stokes, S. (2000). Perception of visual information for Cantonese tones. in *Proceedings of the Eighth Australian International Conference on Speech Science and Technology*, edited by M. Barlow and P. Rose (Australian Speech Science and Technology Association, Canberra) pp. 86–91..

[6] Burnham, D., Ciocca, V. & Stokes, S. (2001). Auditory-visual perception of lexical tone. *Eurospeech Conference*, edited by B. L. P. Dalsgaard, & H. Benner (Aalsborg, Denmark, ISCA, Bonn, Germany), pp. 395-398.

[7] Burnham, D., Lau, S., Tam, H. & Schoknecht, C. (2001). Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language. *Auditory-Visual Speech Processing Conference*, Aalborg, Denmark, Sept 7-9, edited by D. Massaro, Light, J., & Geraci, K. (Causal Productions, Adelaide, South Australia), pp. 155-160.

[8] Mixdorff, H., Charnvivit, P., and Burnham, D. K. (2005a). Auditory-visual perception of syllabic tones in Thai. *Auditory-Visual Speech Processing International Conference*, edited by E. Vatikiotis-Bateson, Burnham, D. & Fels, S (Causal, Adelaide, British Columbia, Canada), pp. 3-8.

[9] Mixdorff, H., Hu, Y., and Burnham, D. (2005b). Visual cues in Mandarin tone perception. *9th European Conference on Speech Communication and Technology* ISCA, Bonn, Germany, pp. 405-408.

[10] Mixdorff, H., Luong, M. C., Nguyen, D. T., and Burnham, D. (2006). Syllabic tone perception in Vietnamese. Proceedings of *International Symposium on Tonal Aspects of Languages*, La Rochelle, France, pp.137–142.

[11] Burnham, D., Reynolds, J., Vatikiotis-Bateson, E. Yehia, H., Ciocca, V., Haszard Morris, R., Hill, H., Vignali, G., Bollwerk, S., Tam, H., Jones, C. (2006) The perception and production of phones and tones: The role of rigid and non-rigid face and head motion. In H. C. Yehia, D. Demolin, Laboissiere, R. *Proceedings of ISSP 2006, 7th International Seminar on Speech Production*, 185-192. ISBN 85-99598-02-3.

[12] Chen, T. H., and Massaro, D. W. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. *Journal of the Acoustical Society of America*, 123, 2356-2366.

[12] Smith, D, and Burnham, D. (2012) Facilitation of Mandarin tone perception by visual speech in clear and degraded audio: Implications for cochlear implants. *Journal of the Acoustical Society of America*, 131(2), 1480-1489. DOI: 10.1121/1.3672703.

[13] Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Xu, N., & Best, C. (2015). Perception assimilation of lexical tone: The role of language experience and visual information. *Attention, Perception, and Psychophysics*, 77, 571-591. doi: 10.3758/s13414-014-0791-3.

[14] Burnham, D., Kasisopa, B., Reid, A., Luksaneeyanawin, S., Lacerda, F., Attina, V., Xu Rattanasone, N., Schwarz, I-C., & Webster, D. (2015). Universality and language-specific experience in the perception of lexical tone and pitch. *Applied Psycholinguistics*, 77, 571-591. doi: 10.1017/S0142716414000496.

[15] Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., and Espressor, R. (1996). About the relationship between eyebrow movements and F0 variations. *Fourth International Conference on Spoken Language Processing*, Philadelphia, PA, USA, Oct 3-6, edited by T. Bunnell, and W. Idsardi, pp. 2175-2178. DOI: 10.1109/ICSLP.1996.607235.

[16] Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26, 23-43.

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

[18] Boersma, Paul & Weenink, David (2019). Praat: doing phonetics by computer. Version 6.0.52, retrieved 2 May 2019 from *http://www.praat.org/*

[19] Chao, Y. (1968). *A Grammar of Spoken Chinese*. University of California Press, Berkeley, Chapt. 1, pp. 1-56.