INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Primacy of Mouth over Eyes: Eye Movement Evidence from Audiovisual Mandarin Lexical Tones and Vowels

Biao Zeng[1], *Rui Wang*[2], *Guoxing Yu*[3], *Christian Dobel*[4]

[1]School of Psychology and Therapeutic, University of South Wales, UK
[2]School of Foreign Languages, Guangdong Pharmaceutical University, China
[3]School of Education, University of Bristol, UK
[4]Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, Friedrich-Schiller-Universität Jena, Germany

biao.zeng@southwales.ac.uk,rui.wang@gdpu.edu.cn,Guoxing.Yu@bristol.ac.uk,
christian.dobel@med.uni-jena.de

Abstract

This study investigated Chinese speakers' eye movements when they were asked to identify audiovisual Mandarin lexical tones and vowels. In the lexical tone identification task, Chinese speakers were presented with an audiovisual clip of Mandarin monosyllables (/ā/, /à/, /ī/, /ì/) and asked to identify whether the syllables were presented in a dipping (/ā/, / ǐ/) or falling tone (/à/, /ì/). In the vowel identification task, they were asked to identify whether the vowels were /a/ or /i/ regardless of lexical tone. These audiovisual syllables were presented in clear, noisy, and silent conditions. An eye-tracker recorded the participants' eye movements.

Results showed participants gazed more at the mouth than the eyes in both lexical tones and vowels. Additionally, when acoustic conditions degraded from clear to noisy and eventually silent, Chinese speakers increased their gaze towards the mouth rather than the eyes. These findings suggest the mouth to be the primary area that is utilised during audiovisual speech perception. The similar patterns of eye movements between vowels and lexical tones indicate that the mouth acts as a perceptual cue that provides articulatory information.

Index Terms: lexical tone, vowel, eye movement, gaze, audiovisual speech, Chinese

1. Introduction

Speech communication in everyday life is, at least, bimodal. During face-to-face conversation, people integrate visual and auditory information automatically and, under some adverse conditions (e.g., noise, accent), visual cues can facilitate listener's perception of sound. Fisher [1] forged the concept of viseme and defined it as the smallest visible speech unit, an analogue to the phoneme. The visual cues of consonants and vowels represent the articulatory gestures (e.g., bilabial /b/, fricative /v/) or corresponding visemic features, e.g., mouth roundness /a/ or flatness /i/ in speech [2].

The visual cues of prosodic information are much more subtle compared to the cues for consonants and vowels. Intonation is a form of prosodic information which refers to the rise and fall of pitch over entire phrases and sentences. It conveys emotional, pragmatic and social information, e.g., questioning, doubting and satire. Several studies have reported the role of upper facial cues in intonation perception. The upper facial cues can facilitate the listener's ability to identify intonation through head movements [3], [4], and [6] and eyebrow movements [7] - [9].

Lexical tone is another form of prosodic information. 70% of languages in the world are tonal languages and lexical tones exist in many Asian and African languages [10]. Similar to intonation, lexical tone is determined by the fundamental frequency (F0) height and contour. For instance, Mandarin has four different tones: /mā/ (Tone 1, high, 55[the numbers represent tone height], *mother*), /má/ (Tone 2, rising, 35, *hemp*), /mǎ/ (Tone 3, dipping, 214, *hors*e), and /mà/ (Tone 4, falling, 51, *scold*).

The visual cues involved in lexical tone identification are far less well researched than those involved in the perception of segmental speech and intonation. Preliminary studies have reported a visual benefit effect, whereby the inclusion of visual information improved participants' perception of lexical tones [2], [11], [12], [13], and [14]. More, both intonation and lexical tone fall under the scope of pitch frequency and are produced by the vocal cords. Therefore, the visual cues contributing to intonation may be relevant when perceiving lexical tone, such as the eyes, head movement or other upper facial area cues. This led to the hypothesis that the eye area would be more helpful in identifying lexical tones than the mouth.

Alternatively, intonation occurs across a relatively long utterance compared to lexical tone. The length of an utterance gives the listeners more opportunity to extract visual cues that are derived from facial movements. For a relatively short lexical tone that is carried through a vowel, visual information might be primarily extracted from the mouth area and offer phonetic information regarding the syllabic length. To perceive lexical tone, listeners may rely on perceptual cues differing from those involved in intonation and this extraction of pragmatic or emotion information, for instance, eyebrow movements.

Therefore, an examination of where the listener's gaze is allocated when perceiving lexical tone is warranted, as this will help determine what visual cues are important. Eye-tracking provides information on which parts of the speaker's face a listener looks at when processing audio stimuli, and whether these areas are related to changes in task demands. For example, if the task requires more information to be gleaned from a specific visual cue (e.g., the mouth) one would expect longer gaze durations on such locations. Smaller

http://dx.doi.org/10.21437/Interspeech.2021-1741

numbers of overall fixations would also reflect how attention is focused on the location, as the listener would be moving their gaze around the visual field less frequently. Thus, eye-gaze duration was adopted and analysed along with the number of fixations at two regions of interest (ROI): the mouth and eyes.

Gaze allocation between the mouth and eye areas can also be influenced by different factors, such as the acoustic environment. When presenting Japanese and English speech to participants, Vatikiotis-Bateson et al. [15] found that participants gazed more at the mouth when noise levels increased. Yi et al. [16] replicated these results and confirmed whilst the mouth and eye areas were the two primary ROIs in audiovisual speech, listeners turned their gaze more to mouth when the speech signal became weaker.

This present study used an eye-tracker to compare the eye movement patterns of Chinese speakers who were asked to identify Mandarin lexical tones and vowels in a two-alternative forced-choice (2FAC) task. Two hypotheses were investigated. Firstly, we assume the primacy of mouth over eyes; participants will gaze towards the mouth longer, especially when listening conditions become adverse. Secondly, we assume the primacy of mouth will exist in both vowels and lexical tones identification: when perceiving lexical tone, Chinese speakers will rely on cues differing from those involved in the perception of intonation.

2. Material and Method

2.1. Participants

Eleven native Mandarin participants (7 females, mean age = 23.8 years, age range = 21 - 39) took part in the study. Participants were recruited from Bournemouth University and paid £8/hour for their participation. They all reported normal or corrected-to-normal visual acuity and no hearing impairment. Only right-handed participants were included. The experimental protocol was approved by the Research Ethics Panel of Bournemouth University in accordance with the Declaration of Helsinki. Informed consent was obtained from each participant before the experiment took place.

2.2. Apparatus and materials

Video clips of two different speakers were used throughout the experiment. During the recording, the speakers kept their head still to avoid supplying any unintentional head movement cues. During each trial, a video of one speaker was played either to the left or right side of the screen, so the initial gaze towards the center of the screen (a central fixation cross) was not on any part of the speaker's face. Each speaker kept their head still when pronouncing each syllable. The video displayed the speaker's full face from above the neck (see in Figure 1) and took 2/3 of the full screen. The default display resolution was 1024 by 768 pixels.

Two regions of interest (ROI) corresponding to the speaker's eyes and mouth were identified. The first being a 246 by 94 pixels rectangle that overlapped both eyes, while the second being a 163 (maximum horizontal length) by 100 (maximum vertical length) pixels ellipse that overlapped the mouth.

The participants kept their head still on a chin and forehead rest approximately 70 cm away from the screen. This resulted in the viewing angle of the speaker's face being 12 degrees (horizontal) by 15 degrees (vertical).

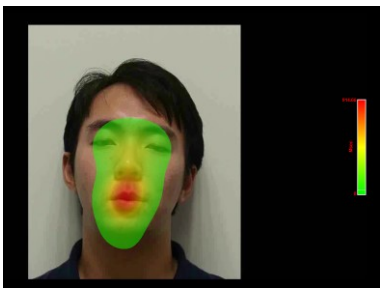

Figure1: *Face in the left side of a screen and takes 2/3 of this full screen. Red colour indicates longer gaze duration and the green colour shorter gaze duration.*

A dipping tone (Tone 3) and a falling tone (Tone 4) were used throughout the experiments as their durations differ considerably more compared to other tone pairs. Acoustically, the dipping tone is the longest and it is significantly longer than the falling tone [17]. Both tones were presented using each of the two visually contrastive vowels /a/ (round mouth shape) and /i/ (flat mouth shape). Both speakers presented two versions of each tone and vowel combination (i.e., two different recordings of each combination). This resulted in 16 unique video recordings of each stimulus (2 speakers × 2 tones × 2 vowels × 2 versions). The corresponding video for each trial was clearly displayed, while the quality of the audio (listening condition) was manipulated to be one of three levels: clear (no distortion or noise), noisy (with background babble noise), or silent (no audio presented). The 16 video stimuli were presented in 3 listening conditions 4 times, which led to a total of 192 trials in the experiment (64 in each listening condition).

2.3. Procedure

To begin each trial, a white fixation cross was displayed in the center of the screen over a black background for 500 ms and stayed on the screen until a fixation was registered. A 500 ms blank black screen then replaced the cross. Following which, an audiovisual clip was presented. After this, a 500 ms black screen was presented again. Participants were required to identify a given lexical tone (or a given vowel in the vowel identification task) presented in the clip based on both the visual and audio cues and responded via keyboard. The audio was presented using headphones at 70-75 dB.

In the lexical tone identification task, participants responded to a dipping tone by pressing the Q button on the keyboard and responded to a falling tone by pressing the P button on the keyboard. Trials from all conditions were presented randomly in three blocks of 64. In between each block, participants were allowed to take a break for as long as they wanted. The eye-movements of one eye were recorded using the Eyelink 1000 static eye-tracker (SR Research, Ottawa Limited) at 1000Hz, and the data was analysed offline using DataViewer (SR Research, Ottawa). Before each block of trials, a 9-point calibration was conducted.

In the vowel identification task, all procedure and setting are identical except that the participants were instructed to identify /a/ or /i/ by pressing the Q or P button, respectively.

3. Results

The accuracy rate for each condition is presented in Table 1. A two-way repeated-measures analysis of variance (ANOVA) (3 listening conditions × 2 speech units) showed a significant main effect of listening condition, F (2, 20) = 225.26, p < .001, η^2_p = 0.96. Identification accuracy appeared to decrease as auditory information was degraded. The main effect of speech unit was significant, F (1, 10) = 433.06, p < .001, η^2_p = 0.98. The interaction effect between the two factors was also significant, F (2, 20) = 162.45, p < .001, η^2_p = 0.94. Further pairwise comparisons revealed that vowel identification accuracy was significantly higher than lexical tone identification accuracy in the noise and silent conditions (ps < .001).

The comparison of listening conditions in terms of speech units showed that for lexical tone identification, the accuracy of the clear condition was significantly higher than the noise condition, and the noise condition was significantly higher than the silent condition (ps < .001) as shown in Table 1. However, no significant effects were found for vowel accuracy (ps > .33).

Table 1: *Accuracy of lexical tones and vowels for each condition*

Listening condition	Lexical Tone		Vowel	
	Mean	SD	Mean	SD
Clear	0.97	0.02	0.99	0.01
Noise	0.82	0.07	0.99	0.02
Silence	0.53	0.05	0.98	0.03

3.1. Fixations

A three-way repeated measures ANOVA with factors—listening conditions (clear, noisy, and silent), ROI (eyes, mouth), and speech units (lexical tone, vowel) was conducted on the number of fixations for the mouth and eyes respectively. However, no significant main effect or interaction effect was observed (ps > .056).

3.2. Gaze duration

The same three-way repeated measures ANOVA was then applied to analyse gaze duration. Figure 3 shows the gaze duration on each ROI in the different listening conditions. The main effects of ROI, F (1, 10) = 22.59, p < .01, η^2_p = 0.69, and speech units were significant, F (1, 10) = 35.67, p < .001, η^2_p = 0.78.

The interaction effect between listening conditions and ROI was also statistically significant, F (2, 20) = 4.76, p < .05, η^2_p = 0.32. Pairwise comparison involving ROIs and listening conditions showed that the gaze was allocated significantly longer at the mouth area compared to the eyes area in all three listening conditions (ps < .01) and for both lexical tone and vowel. Further pairwise comparison also revealed that when participants watched the mouth, their gaze was significantly shorter in the clear condition (622 ± 153 ms) compared to the noise condition (694 ± 136 ms) (p < .05) the silent condition (729 ± 166 ms) (p < .01). However, no significant effect was found between any listening conditions when looking at eye area (ps > .41).

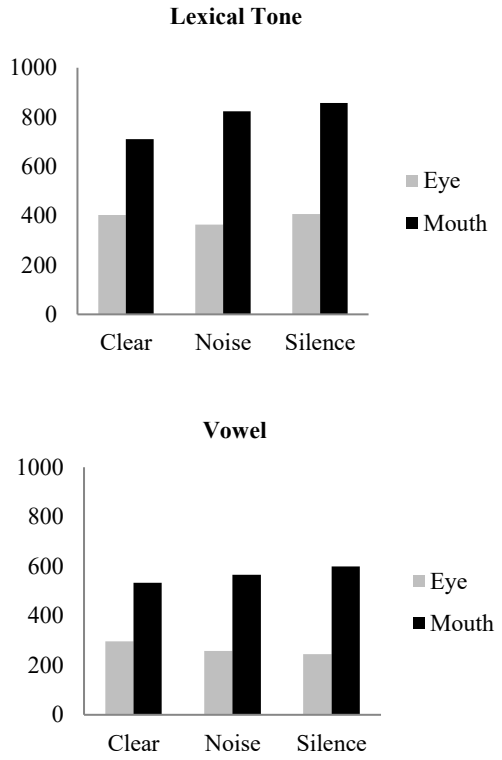

Figure2: *Gaze duration (in ms) on ROI eye and mouth area in all listening conditions in lexical tone and vowel identification.*

4. Discussion

The current study revealed Chinese speakers look at the mouth area rather than eyes area when perceiving both vowels and lexical tones. Results also suggest the mouth becomes more relevant whenever visual support is needed for speech. This finding supports the primacy of the mouth in the perception of visual tone [18]. Moreover, the similar eye movement patterns between vowels and lexical tones implies that, regardless of whatever visual cues exist, the mouth area acts as a cue for perceptual processing of lexical tone, as opposed to the eyes, or upper facial area, which are assumed to facilitate the perception of intonation and carry social or pragmatic information. In addition, with regard to the type of eye gaze measure, the lack of significant effects involving fixation measures suggest gaze duration might be more sensitive measure for audiovisual identification tasks. Using gaze duration as an outcome measure not only revealed the primacy of mouth but it was also increased towards the mouth as listening condition became adverse.

Gaze allocation is an important issue in many audiovisual speech studies [19] - [24]. Visual benefit studies have confirmed there are existing visual cues to facilitate perception of lexical tone [14]. Compared to mouth, the eyes provide little information relating to the production of speech, yet have been demonstrated to provide pragmatic information, which is generally borne and conveyed through intonation. If lexical

tone is processed perceptually and borne by vowels, then the mouth may be more useful than the eyes or upper facial area.

However, the primacy of mouth did not provide a clear explanation of how visual cues from the mouth relate to specific lexical pitch contours. Indeed, no study has addressed how such gaze would provide a specific visual cue relevant to the perceptual targets. For example, Vatikiotis-Bateson et al. [15] did not find any correlation between phoneme identification performance and eye-movement. Paré et al. [25] confirmed that in audiovisual speech perception, participants' gaze primarily focused on the mouth and the eye regions. However, these gaze fixations did not predict the likelihood of succumbing to the McGurk effect, which indexes perceptual confusion occurring at the segment-level.

Summerfield [26] highlighted that timing information was defined as the duration between the onset and offset of the segment. Best et al. [27] showed that visual timing information could improve identification accuracy. In a preliminary study, Xie et al. [14] also reported that lip movement duration, one form of visual timing information, could help facilitate the discrimination of Mandarin lexical tones. Thus, it appears visual timing information cued the participants to direct their attention to the auditory stimulus. In future studies, an eye-tracker device could be used to measure the eye movement patterns associated with visual timing information.

5. Conclusions

In summary, the present study provides empirical support for the primacy of mouth hypothesis in audiovisual speech perception. Furthermore, mouth cues were found to facilitate perceptions of both lexical tones and vowels.

6. Acknowledgements

The study was funded by the British Academy Small Grant (SG152162). We also acknowledge Bournemouth University for use of their research facility and Ms. Bridie Stone and Mr. Josh Molina for assisting in the proofreading and three anonymous reviewers.

7. References

[1] C.G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research*, vol. 11, no. 4, pp.796-804, 1968.

[2] T. H. Chen and D.W. Massaro, "Seeing pitch: visual information for lexical tones of Mandarin-Chinese," *The Journal of the Acoustical Society of America*, vol.123, no. 4, pp. 2356-2366, 2008.

[3] K.G. Munhall, J. A. Jones, D.E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," *Psychological Science*, vol. 15, no. 2, pp.133-137, 2004.

[4] C. Kitamura, B. Guellaï, B., and J. Kim, "Motherese by eye and ear: Infants perceive visual prosody in point-line displays of talking heads," *PloS One*, vol.9, no.10, e111467, 2014.

[5] E. Cvejic, J. Kim, and C. Davis, "Prosody off the top of the head: prosodic contrasts can be discriminated by head motion," *Speech Communication*, vol. 52, no. 6, pp. -564, 2010.

[6] E. Cvejic, J. Kim, J., and C. Davis, "Recognizing prosody across modalities, face areas and speakers: examining perceivers' sensitivity to variable realizations of visual prosody," *Cognition*, vol. 22, no. 3, pp. 442-453, 2012.

[7] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception," *Journal of Memory and Language*, vol. 57, no. 3, pp. 396-414, 2007.

[8] J. Kim, E. Cvejic, and C. Davis, "Tracking eyebrows and head gestures associated with spoken prosody," *Speech Communication*, vol. 57, pp. 317-330, 2014.

[9] M. Cruz, M. Swerts, and S. Frota, "The role of intonation and visual cues in the perception of sentence types: evidence from European Portuguese varieties," Laboratory Phonology, vol. 8, pp. 24, 2017.

[10] M. Yip, *Tone*. Cambridge University Press, 2002.

[11] H. Mixdorff, P. Charnvivit, and D. Burnham, "Auditory-visual perception of syllabic tones in Thai," *AVSP*, pp. 3-8, 2005.

[12] H. Mixdorff, Y. Hu, and D. Burnham, "Visual cues in Mandarin tone perception," *The Ninth European Conference on Speech Communication and Technology*, 2005.

[13] D. Burnham et al., "The perception and production of phones and tones: The role of rigid and non-rigid face and head motion," In Yehia, H (Eds), *Proceedings of the 7th International Seminar on Speech Production*, Brazil: CEFALA, 2006, pp. 1-8.

[14] H. Xie, B. Zeng, and R. Wang, "Visual timing information in audiovisual speech perception: evidence from lexical tone contour. *INTERSPEECH 2018 – 18th Annual Conference of the International Speech Communication Association proceedings*, Hyderabad, India, September 2018, pp. 3781–3785.

[15] E. Vatikiotis-Bateson, I. Eigsti, S. Yano, and K.G. Munhall, "Eye movement of perceivers during audiovisual speech perception," *Perception & Psychophysics*, vol. 60, no. 6, pp. 926-940, 1998.

[16] A. Yi, W. Wong, and M. Eizenman, (2013). "Gaze patterns and audiovisual speech enhancement," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 2, 471-480, 2013.

[17] Y. Xu, "Contextual tonal variations in Mandarin," *Journal of Phonetics*, vol.25, pp. 61-83, 1997.

[18] L.G. Lusk and A.D. Mitchel, "Differential gaze patterns on eyes and mouth during audiovisual speech segmentation," *Frontiers in Psychology*, vol. 7, pp. 52, 2016.

[19] S. M. Thomas and T.R. Jordan, "Contributions of oral and extraoral facial movement to visual and audiovisual speech perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 30, no. 5, pp. 873-888, 2004.

[20] I.T. Everdell, H. Marsh, M.D. Yurick, K.G. Munhall, and M. Paré, "Gaze behaviour in audiovisual speech perception: Asymmetrical distribution of face-directed fixations," *Perception*, vol. 36, no. 10, pp. 1535-1545, 2007.

[21] J.N. Buchan, M. Paré, and K.G. Munhall, "The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception," *Brain Research*, vol. 1242, pp. 162-171, 2008.

[22] H. Yeung and J.F. Werker, "Lip movements affect infants' audiovisual speech perception," *Psychological Science*, vol. 24, no. 5, 603-612, 2013.

[23] P. Tomalski, H. Ribeiro, H. Ballieux, E. L. Axelsson, E. Murphy, D. G. Moore, and E. Kushnerenko, "Exploring early developmental changes in face scanning patterns during the perception of audiovisual mismatch of speech cues," *European Journal of Developmental Psychology*, vol. 10, no. 5, pp. 611-624, 2013.

[24] A.M. Wilson, A. Alsius, M. Paré, and K. G. Munhall, "Spatial frequency requirements and gaze strategy in visual-only and audiovisual speech perception," *Journal of Speech, Language, and Hearing Research*, vol.59, no. 4, pp. 601-615, 2016.

[25] M. Paré, R.C. Richler, M. ten Hove, and K.G. Munhall, "Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect," *Perception & psychophysics*, vol. 65, no. 4, pp. 553-567, 2003.

[26] Q. Summerfield, "Use of visual information for phonetic perception," Phonetica, vol. 36, no. 4-5, pp. 314-331, 1979.

[27] V. Best, E. J. Ozmeral, and B.G.Shinn-Cunningham, "Visually-guided attention enhances target identification in a complex auditory scene," *Journal for the Association for Research in Otolaryngology*, vol. 8, no. 2, pp. 294-304, 2007.