Interspeech 2018
2-6 September 2018, Hyderabad

Visual Timing Information in Audiovisual Speech Perception: Evidence from Lexical Tone Contour

Hui Xie[1, 2], *Biao Zeng* [3], *Rui Wang*[4]

[1] Faculty of Architecture and Urban Planning, Chongqing University, China.
[2]Key Laboratory of New Technology for Construction of Cities in Mountain Areas,
Ministry of Education, Chongqing University, China.
[5] School of Psychology and Therapeutic Studies, University of South Wales, UK
[4] Department of Psychology, Bournemouth University, UK.

xh@cqu.edu.cn, biao.zeng@southwales.ac.uk, rui.wang@bournemouth.ac.uk

Abstract

The present study investigated whether duration of lip movement could improve intelligibility of lexical pitch contours under noisy condition. Eighteen Chinese speakers were asked to identify a Mandarin lexical tone in one pair of tones under auditory only (AO) and audiovisual (AV) condition. Two types of tone pairs were used in the study: maximum contrastive pair (falling vs. dipping tones, the durational difference of lip movement was 100ms) and minimum contrastive pair (rising vs. falling tones, the difference was 33ms).

The results showed that duration of lip movement enhanced discrimination in the maximum pair whereas the similar lengths of rising and dipping tones attenuated such visual benefit. The finding suggested that visual timing information could be a specific cue for audiovisual lexical tone perception.

Key items: audiovisual speech, lexical tone, Chinese,

visual timing

1. Introduction

Visual cues can increase speech intelligibility of face to face communication under both quiet and adverse conditions. Such visual benefit (insert references) emerges when visual speech information is matched to auditory stimuli. Summerfield (1987) [1] categorised two types of visual speech information: speech form and timing information. The speech form information primarily consists of mouth and lip shapes and their movement. These information illustrated visemic identification of speech segment. In addition, Summerfield (1979) [2] defined the timing information segment onset, offset and duration.

Previous studies showed that visual form information originated from oral regions as lip, tongue and the other perioral regions, e.g. head, neck and eyebrow. During face to face talking, the shape of mouth and lip combined with the perioral regions construct motion of speech gestures as time unfolds.

Visual form information varies in audiovisual speech perception and benefit effect relies on different perceptual speech units. In respect to segmental unit, a large quantity of studies have established that visual form information could be visualised as resonant activities (vowel) or movement of the articulators (consonant). Intonation, which is prosodic unit and perceived in a context of continuous utterance, is also manifested by articulatory gesture, e.g. open mouth movement (Scarborough et al, 2009[3]) and broad facial cues (eyebrow, Cave et al., 1996 [4]; head movement, Cvejic, Kim and Davis, 2010 [5]).

Visual timing information has been less studied. Best, Ozmeral and Shinn-Cunningham (2007)[6] showed that visual timing information could improve identification accuracy and played a role of phasic alerting that directed attention to the time point of auditory stimulus. Such findings were inconsistent with Schwartz et al (2004) [7], which showed that visual timing information did not improve speech intelligibility. It needed to point out that visual timing information in both studies were irrelevant to the perceptual target itself like speech onset, offset and duration. They were so called "accessory stimuli" (Kim and Davis, 2014[8]).

In recent years, visual timing information relevant to speech itself have been investigated. Paris, Kim, and Davis (2013) [9] indicated that speech form information facilitated response times but visual timing information did not. Furthermore, Kim and Davis (2014) [8] showed that the timing information related to perioral regions could facilitate response times to speech and no-speech stimuli. They proposed such visual timing effect was due to cross-modal phase resetting which allowed visual speech to

10.21437/Interspeech.2018-1285

control the excitability of auditory cortex. Jaekl et al (2015) [10] used luminance-defined facial motion and showed the contribution of dynamic visual cues in audiovisual speech perception. Both Kim and Davis (2014) [8] and Jaeke et al (2015) [10] studies suggested visual timing information derived from perioral regions and global facial features.

The case of lexical tone provides a window to observe and evaluate whether visual timing information could contribute to audiovisual speech perception. Lexical tone is different from segments and intonation. Acoustically a specific lexical tone is determined by fundamental frequency (F0) height and contour. Preliminary studies have reported that adding visual information improved identifying lexical tones in adverse conditions (visual speech rate, Green & Miller, 1985[11]; Mixdorff, Charnvivit and Burnham, 2005[12]; Mixdorff, Hu and Burnham [13], 2005; rigid motion, Burnham et al, 2006[14]; Chen and Massaro, 2008[15]). Lexical tone is produced by vibration of the vocal cords and offers little explicit visual form information. Therefore it suggested visual form information was hardly detected from mouth and lip shape and movement. The previous visual form information lack to offer a transparent link to lexical pitch contours. Furthermore, compared to prosodic intonation, lexical tone is borne by a vowel and the length is generally less than one second. It is much shorter than intonation and cannot provide abundant dynamic articulatory gestures involving oral and perioral regions.

Visual timing information might be taken into account of audiovisual lexical tone perception. Initially, acoustic durations of different tones vary. For instance, Mandarin has four different tones: mā (Tone 1, high, 55(the numbers represent tone height), mother), m á(Tone 2, rising, 35, hemp), mǎ (Tone 3, dipping, 214, horse), and m à (Tone 4, falling, 51, scold). The latter three are tone contours which shift the pitch from one level to another over the duration. Acoustically, the dipping tone is the lengthiest and significantly longer than the shortest falling tone (Xu, 1997[16]; Burnham et al, 2015[17]). There is no significant difference between the dipping tone and the high level or rising tone (Xu, 1997) [16].

Previous evidences revealed that auditory duration could facilitate to discriminate rising and dipping tones by compensating the loss of pitch information when F0 is degraded or neutralised. The two tones share a similar tone contour curve and easily to be confused by native listeners (Blicher et al, 1990[18]; Liu and Samuel 2004[19]; Smith and Burnham, 2012[20]; Shen et al, 2013[21]). Liu and Samuel (2004) [19] demonstrated that the durational length of pre-rising contour was critical cue for discriminating the two tones.

The confusion of rising-dipping is considered to be a probe of visual timing information effect. In audiovisual speech perception, durational cue can be reflected in auditory and visual modalities. The acoustic tone length is measured between onset and offset. Correspondingly, in visual modality, durational cue refers to how soon a lip movement is fully completed.

Though the lexical tone production offers little visible articulatory information, lip movement is still primarily relevant to prosody perception (Scarborough et al, 2009[3]). In the present study, the visual timing information of lexical tone was visualised to be the duration of lip movement. It is not an accessory stimulus but actually reflects the visual duration of target itself. The persistence of falling tone represents the shortest visual duration correspondingly and facilitates identification. On the contrast, as the visualised durations of rising and dipping tones are not significantly different, visual timing information therefore cannot assist in the identification of the two tones.

Given the durational contrast between the lengthiest dipping tone and the shortest falling tone, the current study assumed identification of dipping tone was more accurate in a maximum contrast falling-dipping tone pair rather than that in a minimum contrast rising-dipping tone pair when acoustic information is degraded under noisy condition.

2. Methods

2.1. Participants

Eighteen Mandarin native speakers (11 females age = 26.61 ±4.51years) from Bournemouth University participated in this study. All native speakers are from Mandarin areas and were educated in standard Mandarin. None reported any hearing and vision loss. All had normal or corrected-to-normal vision. They received payment for participation after the experiment.

2.2. Materials

Two sets of vowels: [a] with 3 lexical tones ([á] rising, [ǎ] dipping, [à] falling), and [i] with 3 lexical tones ([í]rising, [ǐ]dipping, [ì] falling) were presented in two modes: (1) audiovisual (AV); (2) audio-only (AO). All target tones are lexical pitch contours.

Video materials were recorded by two male Mandarin native speakers (one from Tianjin, a northern city in Mandarin area and the other is from Kunming, a southeast city in Mandarin area) with a Sony HDR-SR12E camera in a soundproof booth and were digitalised and edited via iMovie on a Mac computer (1920 x1080, 30 frames per second).The sound track were encoded in PCM.

The speakers were told to keep a neutral facial expression and to minimise head movement. The video clips only showed an image above the speaker’s shoulder. All AV stimuli were uncompressed in the experiment. The AO stimuli were derived from the AV stimuli by deleting the corresponding video tracks. The AO stimuli were edited via Adobe Audition sampling rate as 44.8 kHz 32-bit and the loudness of all audio clips including the audio in AV stimuli were normalised at 65dB SPL(61 – 68dB).

Different durational information and their onsets were measured and controlled with Adobe Audition. The onset of video is measured from the first frame that always shows the speaker kept his mouth close and still. The visual onset of lexical tone, specifically lip movement, commences slightly after the video onset. As the vowel is regarded as the bearer of lexical tone, the visual onset of a tone is defined to be identical to the visual onset of the vowel, [a] or [i] in the current study. Each lexical tone's onsets of lip movement and sound were measured. The onset of sound was defined as the start of sound wave. The auditory onset of dipping tone commences at 399ms and these of rising and falling tones (457ms and 455ms) are almost identical. The auditory durations, which were measured between sound wave onset and offset, were measured as 683ms for the rising tone, 814ms for the dipping tone and 536ms for the falling tone. The lengths were consistent with the previous studies (Burnham et al, 2015[17]; Xu, 1997[16]) that the dipping tone is the lengthiest.

It shows that the three tone's onsets of lip movement are same and start within the 10th frame ranged between 270ms to 300ms (the rising tone is 280ms, the dipping tone is 280ms, the falling tone is 276ms). The visual duration of lexical tones were defined as the video length from the onset of lip movement to the end of an articulatory action, where the speaker's mouth was close. The lip movement durations were measured as the rising tone 1018ms, the dipping tones 1051ms and the falling tones 951ms. The duration of falling tone is shortest, which is consistent to its shortest auditory duration.

As the lip movement precedes auditory information in the present materials, the lip movement might anticipate perceiving and recognising auditory information. The gaps between the lip movement and auditory onsets were calculated here. The gap of rising tone is 177ms (457ms-280ms) and similar to the 179ms (455ms-276ms) gap of falling tone. Therefore, if any potential perceptual difference between the rising and falling tones against the dipping tone occurs, it could not be attributed to anticipation caused by the gap between the visual and auditory onsets.

To sum up, both acoustic and lip movement durations of dipping tone are the longest amongst the three tones (acoustic: 814ms and lip movement: 1051ms), on the contrast, these of falling tone are the shortest (acoustic: 536ms and lip movement: 951ms). The two durations of rising tone (acoustic: 683ms and lip movement 1018ms) are much similar to those of dipping tone. All measurement data were reported. (Table 1)

Table 1. Lip movement duration (a), acoustic duration (b), lip movement onset (c) and acoustic onset (d) of rising, dipping and falling tones in Mandarin (*N* = [a] and [i], two speakers, two tokens for each speakers).

Tone	a	b	c	d
rising	1018	683	280	457
dipping	1051	814	280	399
falling	951	536	276	455

In the present study, we investigated whether extra visual cue could improve perception of lexical tone contour embedded in a noisy condition. As the noise degrades the acoustic information at a certain level, any enhanced perception could be interpreted by extra visible lip movement. To test the visual benefit in auditory degradation, babble noise was adopted in the study as it masked information better than other noise, e.g. pink noise (Mixdorff, Hu & Burnham, 2005[13]). The babble noise was created by combining the voices from six Mandarin native speakers and the SNR was -9dB. The noise durations were always longer (about 1 second prior to the syllable onset) than the syllable durations in order to fully mask the syllables.

For each block, the trials consisted of 2 modes (AV and AO) × 2 tones (rising or falling tone vs. dipping tone) × 2 listening conditions (clear and noise types) × 40 repetitions (2 speakers; 2 vowels: 10 repetition), which totalled 320 trials.

2.3. Procedure

There were two blocks in the study. One was for the pair of dipping-rising tones. The other was the pair of dipping-falling tones. The tasks were counterbalanced across participants. For each block, the participants were instructed that there were two modes: AO and AV. Each trial proceeded in the following order: a fixation cross appeared in the centre of the screen for 200ms and then followed by a video. The sound was played over headphones. The participants were told to identify the tone of the given syllable by pressing the corresponding key (one of the two alternatives: in the dipping-rising block, Key Q for dipping tone or Key P dipping tone. In the dipping-falling task, Key Q for dipping tone or Key P for falling tone) on the keyboard and to respond as soon as possible. Response time was calculated from the onset of the video and additional 3 seconds was given after the offset of the video.

3. Results

The accuracy rates of both blocks are shown in Table 2 and3.

Table 2. Accuracy for all the conditions of the design in Minimum Contrast Pair (Task 1. (N=18)

		Task 1: Minimum Contrast	
		dipping tone	**falling tone**
Clear	**AO**	.97(.03)	.96(.02)
	AV	.94(.04)	.95(.03)
Noise	**AO**	.41(.17)	.82(.13)
	AV	.47(.19)	.83(.17)

Table 3. Accuracy for all the conditions of the design in Maximum Contrast Pairs (Task 2). (N=18)

		Task 2: Maximum Contrast	
		dipping tone	**falling tone**
Clear	**AO**	.96(.03)	.95(.04)
	AV	.95(.04)	.96(.04)
Noise	**AO**	.66(.22)	.66(.15)
	AV	.75(.15)	.73(.12)

The sensitivity index of *d'* was adopted and calculated in order to compare the dipping tone's speech intelligibility under different conditions, which are shown in Table 3 and 4. Three-way within-subject analysis of variance (ANOVA) (mode × listening condition × tone contrast) was run for the dipping tone between the two blocks.

Table 3. Hit rate, false-alarm rate and *d'* of dipping tones in Minimum Contrast Pair (Task 1) and Maximum Contrast Pairs (Task 2). (N=18)

		Task 1: Minimum Contrast		
		Hit	**FA**	***d'***
Clear	**AO**	.98	.03	3.84
	AV	.95	.04	3.50
Noise	**AO**	.41	.17	.88
	AV	.47	.16	1.07

Table 4. Hit rate, false-alarm rate and *d'* of dipping tones in Maximum Contrast Pair (Task 2). (N=18)

		Task 2: Maximum Contrast		
		Hit	**FA**	***d'***
Clear	**AO**	.97	.04	3.72
	AV	.96	.03	3.71
Noise	**AO**	.66	.33	.98
	AV	.77	.27	1.44

A significant listening condition effect was found, F (1, 17) = 449.1, p < .01. An interaction effect of listening condition and mode effect is significant, F (1,17) =12.16, p<.01. Pairwise comparison showed that for noise condition, *d'* for AV (1.26) was higher than that for AO (.93), but not for the clear condition, indicating a trend of adopting visual cue to compensate the loss of acoustic information and improve perceptual performance.

Tone contrast effect was marginally significant, F (1, 17) = 3.324, p=.09. An interaction effect of tone contrast and mode was significant, F (1, 17) =5.65, p<.05. Pairwise comparison showed that a significant mode effect for the maximum contrast pair condition (*p* <.01), showing that the *d'* for AV (2.58) was higher than that for AO (2.35), but not for the minimum contrast pair (AV: 2.29 vs. AO: 2.36) condition, indicating that visual benefit could facilitate to discriminate falling-dipping tone contrast rather than the rising-dipping tone contrast.

4. Discussion

The present study attempted to answer whether visual timing information contributed to improve intelligibility of lexical tone contours under adverse conditions. We found d' was increased under adverse condition, but not clear condition as the extra visual information was added. Furthermore, in the maximum contrast pair, the *d'* of identifying the dipping tone increased 0.23 . On the contrast, in the minimum contrast pair, either the *d'* of the dipping tone was not improved significantly.

The present contrastive result was postulated to be caused by a 100ms difference of lip movement duration between the dipping and falling tones (1051ms vs. 951ms), compared to the rising tone (1018ms), which was not greatly different from the dipping tone. Therefore, the durational difference of lip movement offered a contrastive cue and facilitated the participants to identify the target. Audiovisual speech perception could adapt to visual form and timing information. Visual form information include articulatory places and manners which involve different facial areas, neck and head. Visual timing information reflects visual cues change in time course. Here the lip movement duration is a good case.

However, the current findings raised two questions. Firstly, the tone contrast effect was marginally significant. In future study, the sample size could be increased. Secondly, this finding was inconsistent with Smith and Burnham's result (2012) [20]. In their study, under normal AV mode, the minimum contrast rising-dipping tones was poorly discriminated but in cochlear implant AV condition, the pair was better discriminated well. They suggested a visual benefit on the rising-dipping discrimination in adverse condition. One explanation might be there are other visual cues apart from duration that are contrastive in simulated CI conditions which can be used more effectively. The differential results could lead to explore the roles of various visual cues (e.g. mouth, rigid vs. no-rigid) in audiovisual speech perception.

Specifically, visual timing information is not accessory stimulus. Visual timing information has been little touched in previous studies and still need more evidences to understand their roles in audiovisual speech. Based on the current result, future studies could investigate lip movement duration interacts with the availability of acoustic information in different noise environment. The contrast of dipping-falling tones and the confusion of rising-dipping tones could be two useful probes tested in different noise levels and other adverse conditions, e.g. lip reading or cochlear implants.

5. Acknowledgements

This research was supported by the National Natural Science Foundation of China (51308562) in collaboration with the Department of Psychology, Bournemouth University, UK.

References

[1] Q. Summerfield, “Some preliminaries to a comprehensive account of audio-visual speech perception,” In B. Dodd & R. Campbell (Eds.), Hearing by eye: The psychology of lip-reading (pp. 3-51). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc., 1987.

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

[3] R. Scarborough et al., "Optical phonetics and visual perception of lexical and phrasal stress in English," Language and Speech, vol. 52, no.2-3, pp. 135-175, 2009.

[4] Cavé, Christian, et al. "About the relationship between eyebrow movements and F0 variations." Spoken Language, 1996. ICSLP 96. Proceedings, Fourth International Conference on. Vol. 4. IEEE, 1996.

[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.555-564, 2010.

[6] 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.

[7] J. Schwartz, B. Frédéric and C. Savariaux, "Seeing to hear better: evidence for early audio-visual interactions in speech identification," Cognition, vol.93, no.2, pp.69-78, 2004.

[8] J. Kim and C. Davis, "How visual timing and form information affect speech and non-speech processing," Brain and language, vol. 137, pp. 86-90, 2014.

[9] T. Paris, J. Kim and Davis, "Visual speech form influences the speed of auditory speech processing." Brain and language, vol. 126, no. 3, pp.350-356, 2013.

[10] P. Jaekl et al, “The contribution of dynamic visual cues to audiovisual speech perception," Neuropsychologia, vol. 75, pp. 402-410, 2015.

[11] K.P. Green and J. L. Miller, "On the role of visual rate information in phonetic perception," Perception & psychophysics, vol. 38, no.3, pp. 269-276, 195.

[12] H. Mixdorff, P. Charnvivit and D. K. Burnham, "Auditory-visual perception of syllabic tones in Thai," AVSP, 2005.

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

[14] 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, 2006, pp. 1-8, Brazil: CEFALA.

[15] 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.

[16] Y. Xu, "Contextual tonal variations in Mandarin." Journal of phonetics, vol. 25, no.1, pp.61-83, 1997.

[17] D. Burnham et al. ,"Universality and language-specific experience in the perception of lexical tone and pitch," Applied Psycholinguistics, vol. 36, no.6, pp.1459-1491, 2015.

[18] D.L. Blicher, R. L. Diehl and L.B. Cohen, “Effects of syllable duration on the perception of the Mandarin tone2/tone 3 distinction: evidence of auditory enhancement,” Journal of Phonetics, vol.18, pp.37–49, 1990.

[19] S. Liu and A. G. Samuel, "Perception of Mandarin lexical tones when F0 information is neutralized." Language and speech, vol. 47, no.2, pp.109-138, 2004.

[20] D. Smith and D. Burnham, "Faciliation of Mandarin tone perception by visual speech in clear and degraded audio: Implications for cochlear implants," The Journal of the Acoustical Society of America, vol. 131, no.2, pp.1480-1489, 2012.

[21] J. Shen, D. Deutsch and K. Rayner, "On-line perception of Mandarin Tones 2 and 3: Evidence from eye movements," The Journal of the Acoustical Society of America, vol.133, no.5, pp. 3016-3029, 2013.