ATLANTIS PRESS
*Advances in Social Science, Education and Humanities Research (ASSEHR), volume 300*
2018 International Workshop on Education Reform and Social Sciences (ERSS 2018)

# The Effect of Extra-oral Facial Information on Audiovisual Speech Perception

Zeyu Huang[1, a], Yao Lu[1, b], Lu Wang[1, c] and Xiyu Wu[1, 2, d]

[1]Department of Chinese Language and Literature, Peking University, Beijing 100871, China

[2]Center for Chinese Linguistics, Peking University, Beijing 100871, China

[a]zeyu.huang@pku.edu.cn, [b]luyiru2012@163.com, [c]wanglunj@163.com, [d]xiyuwu@pku.edu.cn

**Abstract.** An experiment consisting of five blocks was conducted to examine the effect of extra-oral facial information on audiovisual speech perception. Twenty Chinese native speakers were asked to report the syllables they perceived in different conditions: audio-only, video-only, video-only without oral part, audiovisual and audiovisual without oral part. The experimental materials were /pa/, /tsa/, /ta/, /tʂa/ and /ka/, which were selected according to places of articulation from front to back. The results showed that even though the extra-oral facial information was not enough to distinguish non-labial consonants, it could have significant effect on auditory speech perception.

**Keywords:** McGurk effect, audiovisual speech perception, extra-oral facial information.

## 1. Introduction

Although speech perception in natural condition was a multisensory process, classic models of speech processing focused predominantly on acoustic input, ignoring the influence of visual information (Van Wassenhove V, 2013). As a matter of fact, visual input provides not only subsidiary information such as identification or emotion, but also the forms and kinematics of facial information, which could even affect the speech processing directly and cause a fused illusion when video and audio input were incongruent (McGurk H, MacDonald J, 1976).

However, it was unclear how we extract articulation information from visual input and what parts of visual information work in speech perception. Several studies have indicated that mouth might not be the only resource for perceiving linguistic information (Rosenblum L D, Saldaña H M, 1996). And even if the fixation point was fixed 10°-20° from talker's mouth, McGurk effect persisted (Paré, et al., 2003). Therefore, the present study was to explore whether or not and to what extent extra-oral facial information could affect visual and audiovisual speech perception.

## 2. Method

### 2.1 Subjects

Twenty Mandarin speakers including 8 males and 12 females ranging from 19 to 29 years old (overall mean age=23.6±2.53 years) attended this research. All of them had normal or corrected-to-normal vision and no speech or hearing impairment. None of them had received lip reading training. They had no idea on the purpose of the experiment.

### 2.2 Stimuli

The audiovisual stimuli for the experiment were recorded by an EOS kiss X5 camera and a professional external microphone in the studio of Linguistic Laboratory of Peking University. The frame rate of video was 29.97 FPS and the sampling rate of audio was 48 kHz.

The stimuli were made of two native speakers of Mandarin, one male (m1) and one female (f1). Only the head and shoulder were included against a dark blue background. The materials were edited by Adobe Premiere 2018 to ensure that each stimulus was 2-second long and without blinks.

There were five Chinese syllables /pa/, /tsa/, /ta/, /tʂa/ and /ka/ which were composed of vowel /a/ and a series of consonants, each represented for a place of articulation from front to back. For incongruent audiovisual stimuli, /pa/ was dubbed into the videos of the other syllables. Because according to previous studies, the McGurk effect arose by audio /pa/ tended to be the strongest among all kinds of incongruent pairs of Chinese syllables (Pan X, 2011). Besides, for the series of stimuli without oral parts, the mouth areas were covered by black oval masks, which were set by the frame of each stimulus when mouth was open widest. To sum up, there were 66 stimuli for all, including 10 (2×5) stimuli for each of audio-only (AO), video-only (VO), video-only without mouth area (VO_NoM) condition, and 18 stimuli (10 congruent and 8 incongruent) for each of audiovisual (AV) and audiovisual without mouth area (AV_NoM) condition. For each block, the stimuli were presented twice with random sequence.

## 3. Results

### 3.1 Audio Only

When there were only audio stimuli available, the recognition rates were presented in Table 1:

Table 1. Percentages of Correct Identifications of Audio-Only (AO) Condition (%)

| | STIMULI | | | | | |
|---|---|---|---|---|---|---|
| Talker | /pa/ | /tsa/ | /ta/ | /tʂa/ | /ka/ | Average |
| f1 | 100.00 | 100.00 | 100.00 | 95.45 | 81.82 | 95.454 |
| m1 | 81.82 | 100.00 | 100.00 | 95.45 | 86.36 | 92.726 |

The average percentage of all audio stimuli was 93.81%±6.69%. According to a two-way ANOVA, the main effect of talker was not significant [F (1, 20) =1.00, p=0.329)]. The identification rate of /ka/ was significantly smaller than /tsa/ and /ta/. To sum up, all audio stimuli could be identified at fairly high proportion.

### 3.2 Video Only

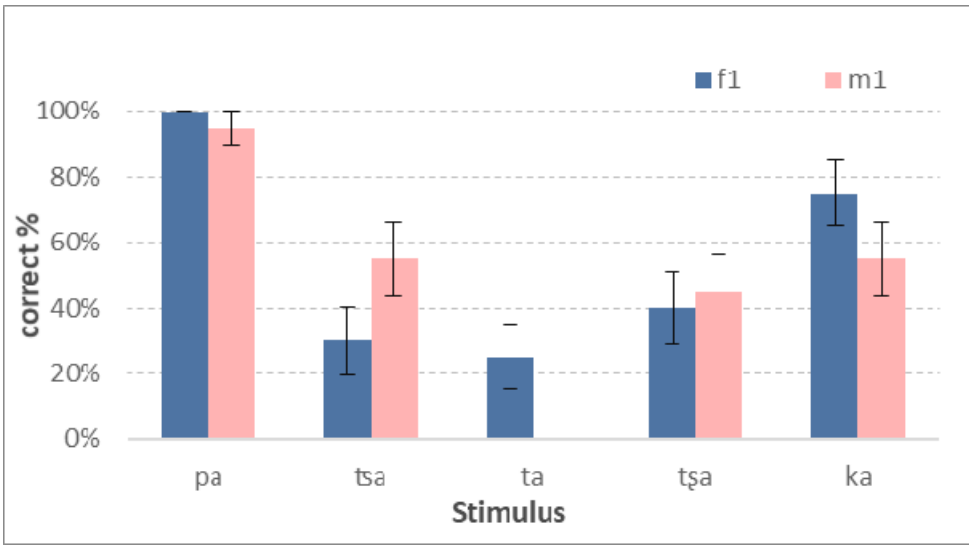


Fig. 1 Percentages of Correct Identifications of Video-Only (VO) Condition (%)

Figure 1 shows the rate of correct identifications when only visual stimuli were presented. Analyzed with a two-way ANOVA, there was a significant interaction effect between talker and stimulus [F (4,76) =3.673, p<0.01)]. As a paired comparison adjusted by Bonferroni test showed that, for /ta/ and /ka/, the identification rate of f1 was significantly higher than those of m1. And for each

talker, the accuracy of labial consonant /p/ was much higher than the other non-labial consonants, among which the /t/ sound was the lowest. Overall, the percentages of correct identifications distributed in U-shape according to places of articulation, which was considerable high for labial consonant, then decreased sharply to bottom, and rose again at the place of velar.

For the block of visual stimuli without mouth, compared with normal visual condition, the accuracy of all stimuli decreased to some extent. There was a significant interaction effect between talker and stimulus [$F(4, 76) = 3.956$, $p<0.01$)]. For /pa/, accuracy of f1 was significantly higher than m1. There was no significant difference between f1 and m1 on other syllables. What's more, for each talker, the percentages of correct identification of these consonants declined with places of articulation from front to back. Compared to VO condition, the identification rate of /ka/ was affected most, then was /tʂa/. The other syllables were slightly or barely influenced.

### 3.3 Video Only without Mouth Area

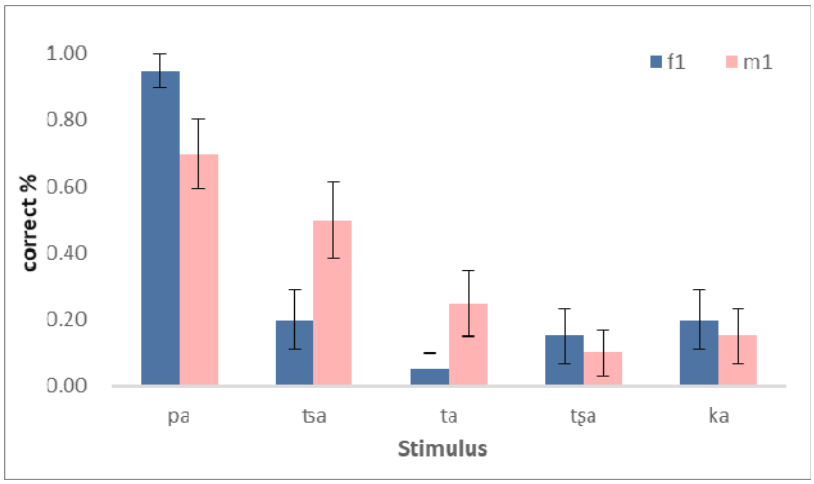


Fig. 2 Percentages of Correct Identifications of Video-Only without mouth area (VO_NoM) Condition (%)

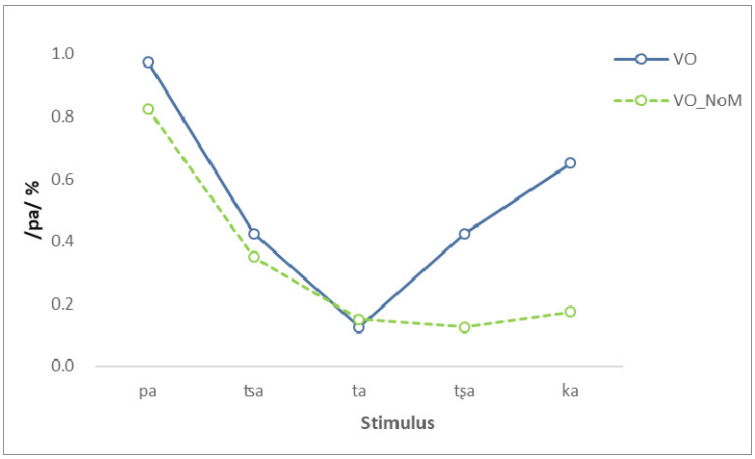


Fig. 3 Percentages of Correct Identifications of VO (Video-Only) versus VO_NoM (Video-Only without mouth area) condition (%)

### 3.4 Audiovisual

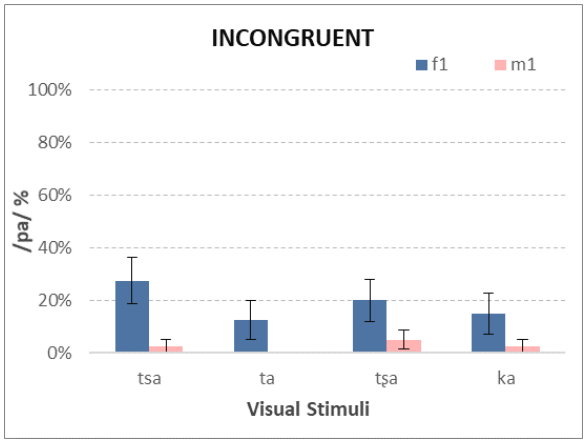


Fig. 4 Accuracy performance in response to the presentation of incongruent audiovisual stimuli (AV). (%)

The percentages of correct identifications of audiovisual condition were showed in Figure 4. No interaction effect was significant between talker and visual stimuli [F (3, 16) =1.380, p=0.285)]. And visual stimulus has no significant main effect on McGurk effect [F (3, 16) =3.198, p=0.052)]. However there was a significant difference between two talkers [F (1, 18) =6.133, p<0.05)]: the accuracy percentage of f1 (19.7±7.2%) was much higher than m1 (2.6%±1.5%), though considering the accuracy percentage of /pa/ in audio-only condition, the additional error rates of f1 (80.3%) and m1 (79.2%) in audiovisual condition is almost the same. By and large, the McGurk effect of every stimulus was considerably strong.

### 3.5 Audiovisual without Mouth Area

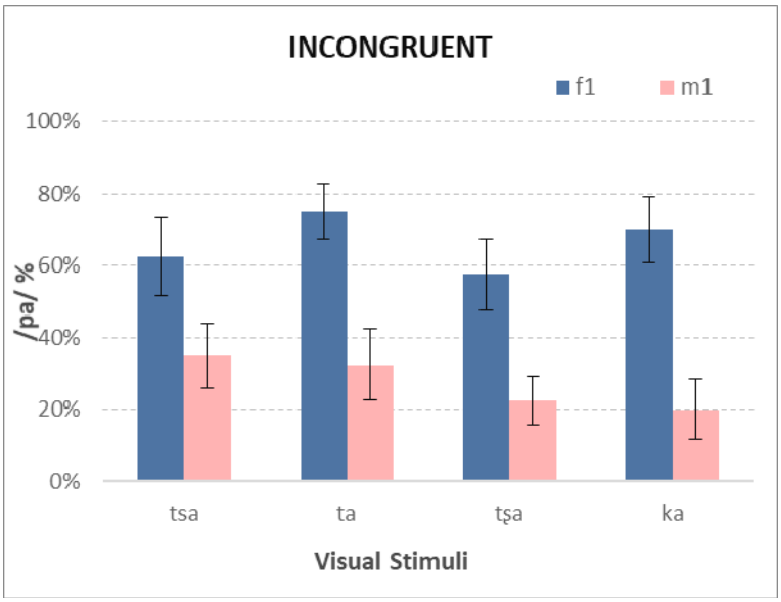


Fig. 5 Accuracy performance in response to the presentation of incongruent audiovisual stimuli without mouth (AV_ NoM). (%)

When mouth areas were covered, there was no significant interaction between talker and visual stimuli [F (3, 57) =1.132, p=0.344)], and no significant difference among visual stimuli [F (3,57) =1.830, p=0.152)]. However, the main effect of talker was still significant [F (1,19) =24.057, p<0.01)]: the accuracy percentage of f1 (66.3±7.4%) was much higher than m1(27.5%±6.9%), even if considering the accuracy difference between two talkers in audio-only condition, the McGurk effect of f1(33.7%) was still weaker than m1(54.3%).

Compared with AV condition, no other effect or interactions among talker, mouth condition and visual stimuli were found to be statistically significant, except for the interaction between talker and mouth condition [$F(1,18) = 12.906$, $p < 0.01$)]. Even though the McGurk effect of all stimuli tended to be weaker when mouth areas were covered, there was still 1/3 to 1/2 chance that the fusion illusions occurred.

## 4. Conclusion

From the results of two visual-only tasks, we could see that the facial information neither with or without mouth area was not enough to fully identify non-labial consonants. Nevertheless, as the two audiovisual tasks showed, McGurk effect never disappeared even when mouth areas were totally covered, which provided another solid evidence to the assumption that mouth area is not the only part containing articulation information. And the current findings further indicated that the visible kinematics of articulatory gestures could be exacted from the extra-oral facial area and they were strong enough to have significant effect on auditory speech perception.

However, from present results, it is still unclear whether the global speech facial information or some local parts of extra-oral face worked when McGurk effect occurred. In order to investigate this problem, eye tracing technique will be involved to explore the patterns of gaze behavior when oral areas are covered in further research. It has been found that about 80% of the variance observed in the vocal-tract can be estimated from the facial parts including lip, chin, jaw and cheeks (Yehia H, et al., 1998). If the fixations on none of these local parts rise significantly when mouths are covered, the extra-oral facial information is more likely to work as a whole to affect auditory speech perception.

## Acknowledgements

This study was supported by the National Social Science Fund of China (No. 18BYY189) and Major projects of the Ministry of Education (No. 17JJD740001). Thanks to all the talkers and subjects who attended this research.

## References

[1]. Macdonald J, Andersen S, Bachmann T. 1999. Hearing by eye: visual spatial degradation and the McGurk effect. Sixth European Conference on Speech Communication and Technology.

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

[3]. Pan, X. 2011. Labial articulation and audio-visual speech perception in standard Chinese. (Doctoral dissertation, Peking University).

[4]. Paré, M., et al. 2003. Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect. Perception & Psychophysics, 65(4), 553-567.

[5]. Rosenblum, L. D., & Saldaña, H. M. 1996. An audiovisual test of kinematic primitives for visual speech perception. J Exp Psychol Hum Percept Perform, 22(2), 318-331.

[6]. Sekiyama, K. 1997. Cultural and linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects. Perception & Psychophysics, 59(1), 73-80.

[7]. Sekiyama, K., & Tohkura, Y. 1991. McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. Journal of the Acoustical Society of America, 90(1), 1797-1805.

[8]. Sumby, W. H., & Pollack, I. 1954. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America, 26(2), 212-215.

[9]. Van, W. V. 2013. Speech through ears and eyes: interfacing the senses with the supermodal brain. Frontiers in Psychology, 4(2), 388.

[10]. Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. 1998. Eye movement of perceivers during audiovisual speech perception. Perception & Psychophysics, 60(6), 926-940.

[11]. Worster, E., et al. 2017. Eye movements during visual speech perception in deaf and hearing children. Language Learning, 68(s1), 159–179.

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