



A value-driven McGurk effect: Value-associated faces enhance the influence of visual information on audiovisual speech perception and its eye movement pattern

Xiaoxiao Luo¹ · Guanlan Kang¹ · Yu Guo² · Xingcheng Yu² · Xiaolin Zhou^{1,2,3,4,5}

Published online: 2 January 2020
© The Psychonomic Society, Inc. 2020

Abstract

This study investigates whether and how value-associated faces affect audiovisual speech perception and its eye movement pattern. Participants were asked to learn to associate particular faces with or without monetary reward in the training phase, and, in the subsequent test phase, to identify syllables that the talkers had said in video clips in which the talkers' faces had or had not been associated with reward. The syllables were either congruent or incongruent with the talkers' mouth movements. Crucially, in some cases, the incongruent syllables could elicit the McGurk effect. Results showed that the McGurk effect occurred more often for reward-associated faces than for non-reward-associated faces. Moreover, the signal detection analysis revealed that participants had lower criterion and higher discriminability for reward-associated faces than for non-reward-associated faces. Surprisingly, eye movement data showed that participants spent more time looking at and fixated more often on the extraoral (nose/cheek) area for reward-associated faces than for non-reward-associated faces, while the opposite pattern was observed on the oral (mouth) area. The correlation analysis demonstrated that, over participants, the more they looked at the extraoral area in the training phase because of reward, the larger the increase of McGurk proportion (and the less they looked at the oral area) in the test phase. These findings not only demonstrate that value-associated faces enhance the influence of visual information on audiovisual speech perception but also highlight the importance of the extraoral facial area in the value-driven McGurk effect.

Keywords McGurk effect · Reward association · Audiovisual speech perception · Eye movements · Signal detection analysis

Introduction

In audiovisual speech perception, auditory information plays a dominant role; the visual information (e.g., mouth movements), however, also enables perceivers to decode the message more accurately (Brancazio, Miller, & Paré, 2003; Grant & Seitz, 2000; Ross, Saint-Amour, Leavitt, Javitt, & Foxe,

2007; Sumbly & Pollack, 1954). One of the famous phenomena reflecting the interaction between visual and auditory modalities is the McGurk effect (McGurk & MacDonald, 1976). Specifically, a sound track of an auditory syllable (e.g., “ba”) paired with a video track of a face pronouncing a different syllable (e.g., “ga”) could lead to a novel, distinct percept of the auditory syllable (e.g., “da”). The proportion of the reported McGurk percepts can be used as a measure of the extent that perceivers tend to use visual information in audiovisual speech perception. That is, the increase of the McGurk proportion indicates that the perceivers tend to make more use of visual information (for review, see Alsius, Paré, & Munhall, 2018; Marques, Lapenta, Costa, & Boggio, 2016).

Attention modulates the influence of visual information on audiovisual speech perception. Previous studies have shown that the McGurk proportion would decrease if participants are required to attend to a visual distractor (Tiippana, Andersen, & Sams, 2004) or to concurrently conduct an unrelated visual task (Alsius, Navarra, Campbell, & Soto-Faraco, 2005). That is, less visual information would be used in audiovisual speech perception if there were distractors of visual attention.

✉ Xiaolin Zhou
xz104@pku.edu.cn

¹ School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, China

² Institute of Psychological and Brain Sciences, Zhejiang Normal University, Zhejiang 321004, China

³ Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100871, China

⁴ Institute of Linguistics, Shanghai International Studies University, Shanghai 200083, China

⁵ PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

Moreover, top-down characteristics (e.g., face familiarity) also affect the McGurk effect. Walker, Bruce, and O'Malley (1995) found that participants who are familiar with the face report less McGurk percepts than those who are unfamiliar with the face when the face and voice are from different persons. These studies imply that the McGurk effect can be modulated by either bottom-up or top-down attentional characteristics.

Value-driven attentional capture is a recently proposed mechanism of attention in addition to the salience-driven (bottom-up) and goal-driven (top-down) mechanisms (Anderson, 2013). Previous studies on value-driven attention, conducted in the visual domain, have shown that when stimuli are learned to predict reward, these stimuli would gain a competitive advantage that promotes attentional selection even when they are nonsalient and/or task irrelevant in perception (e.g., Anderson, 2013; Anderson, Laurent, & Yantis, 2011; Wang, Duan, Theeuwes, & Zhou, 2014; Wang, Yu, & Zhou, 2013). A few studies extended the concept to the cross-modal domain, showing that reward-associated sounds could affect the processing of visual stimuli. For example, auditory stimulus associated with high reward can increase the sensitivity of the perception of visual stimulus appearing simultaneously, even when sounds and reward associations are both irrelevant to the visual task (Pooresmaeili et al., 2014). Anderson (2016) demonstrated that relative to neutral sounds, previously reward-associated sounds capture attention, interfering more strongly with the performance of a visual task. However, it is currently unknown whether and how the value-driven mechanism of attention works in audiovisual speech perception in which the visual and auditory information is complex and highly relevant.

The face is the most important visual information in audiovisual speech perception. Raymond and O'Brien (2009) showed that a value-driven effect could be observed in the visual processing of faces. They trained participants to learn particular face–reward associations and then asked them to recognize whether a target face in an attentional blink (AB) task had been presented in the training phase. The authors found that the face recognition performance was higher for reward-associated faces compared with non-reward-associated faces. Moreover, while non-reward-associated-faces trials showed a typical AB effect, reward-associated-faces trials showed no AB effect, breaking through the constraints of AB on attentional selection. This study implies that value-associated faces would capture more attention and would be processed better than non-value-associated faces.

Considering that dynamic facial movements contain lots of visual information (e.g., mouth movements, other facial muscle movements, eye gaze), it is necessary to explore how people extract visual information from the dynamic talking faces for the purpose of audiovisual speech perception. By using the McGurk task and monitoring eye movements, previous

studies have found that the mouth area of the talking face plays a critical role in the effect of visual information on audiovisual speech perception. In particular, perceivers show less time looking at the mouth area when the McGurk proportion decreases (i.e., when they make less use of visual information). For example, as the visual resolution of faces decreases, perceivers report fewer McGurk percepts and spend less time looking at the mouth area (Wilson, Alsius, Paré, & Munhall, 2016). Adding a concurrent cognitive task to the main McGurk task would decrease the McGurk proportion as well as the time looking at the mouth area (Buchan & Munhall, 2012). In addition, weak McGurk perceivers (i.e., perceivers who perceive the McGurk effect less frequently in general) fixate less on the talker's mouth area compared with strong McGurk perceivers (Gurler, Doyle, Walker, Magnotti, & Beauchamp, 2015; Hisanaga, Sekiyama, Igasaki, & Murayama, 2016).

The current study investigates whether and how the value-driven mechanism of attention works in audiovisual speech perception by using a training-test paradigm, which is often used in value-driven attention studies (e.g., Anderson, 2013, 2016; Anderson et al., 2011; Raymond & O'Brien, 2009; Wang et al., 2013). In the training phase, participants were asked to discriminate the gender of face pictures, in which correct responses to half of the faces could receive monetary rewards. In the test phase, participants were asked to identify the syllables that the talkers said in video clips (i.e., the McGurk task). Importantly, the talkers' faces had or had not been associated with reward in the previous training phase. In both phases, participants' eye movements were recorded with an eye tracker. Because value-associated faces could capture more attention and would be processed better than non-value-associated faces, we predicted that face–reward association would increase the influence of visual information on the audiovisual speech perception, resulting in higher McGurk proportion for reward-associated faces than for non-reward-associated faces. For the eye movement data, given the studies reviewed above, we predicted that participants would fixate longer on the oral area for reward-associated faces than for non-reward-associated faces.

Method

Participants

A group of 32 graduate or undergraduate students ranging in age from 18 to 26 years took part in the study for monetary compensation. They were all native speakers of Chinese and had the normal or corrected-to-normal vision and normal hearing; none of them reported a history of neurological or psychiatric disorders. This study was performed in accordance with the Declaration of Helsinki and was approved by the

Committee for Protecting Human and Animal Subjects, School of Psychological and Cognitive Sciences, Peking University. Three participants were excluded, because they reported in the after-experiment interview that they concentrated on visual information in deciding the identity of spoken syllables during the task; another participant was excluded because of astigmatism that leads to poor quality of eye movement data. The remaining 28 participants were included in data analyses (17 females, mean age = 21.79, $SD = 2.10$). A power analysis was conducted by using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007). Since we did not find a previous study that is similar to the current investigation, we referred to a study concerning the modulation of attention on the McGurk effect (Alsius et al., 2005). This study showed a moderate effect size (Experiment 1, Cohen's $d = 0.694$). We estimated that we would need at least 19 participants, given Cohen's $d = 0.694$, $\alpha = 0.05$, and power = 80%. In the present study, the number of participants (28) is higher than the suggested number (19).

Apparatus and materials

Visual stimuli were presented on a 17-inch SONY CRT monitor (refresh rate: 75 Hz, resolution: 1,024 × 768) connected to a DELL computer. Auditory stimuli were presented through an AKG headphone. The monitor was positioned 60 cm from the participant, and the head position was maintained using a chin rest. Eye tracking was performed using an EyeLink 1000 system. Stimulus presentation and participant's response recording were controlled by Psychophysics Toolbox (<http://www.psychtoolbox.org/>; Brainard, 1997) with MATLAB.

The audiovisual stimuli used in the test phase were developed based on eight original color video clips, which were recorded by an MI5 phone camera from two male and two female talkers (wearing white T-shirts) saying “ba” or “ga” without blinking. We edited the video and sound tracks with Windows Live video editing software and Cool Edit Pro 2.1, respectively. For each talker, three types of audiovisual stimuli were prepared by rematching sound and video recordings: congruent (visual and auditory matching, which included visual “ba” + auditory “ba,” and visual “ga” + auditory “ga”), incongruent (visual and auditory mismatch that could not lead to a McGurk effect, which included visual “ba” + auditory “ga”), and McGurk (visual and auditory mismatch that could lead to a McGurk effect, which included visual “ga” + auditory “ba”). Each prepared audiovisual stimulus started and ended with the talker in a neutral, mouth-closed position (i.e., all mouth movements from mouth opening to closing were included), and was standardized at 640 × 480 resolution. The duration of each audiovisual stimulus was approximately 1,500 ms, in which the duration of auditory syllables ranged from approximately 400 to 500 ms and the volume was standardized at approximately 65 dB.

The visual stimuli used in the training phase were the pictures of each talker's face (i.e., the two male and two female talkers in a mouth-closed position and neutral emotion) extracted from the audiovisual stimuli used in the test phase.

Design and procedure

The experiment used a training-test paradigm. During the training phase (see Fig. 1a), a hollow circle ($0.96^\circ \times 0.96^\circ$) was presented at the center of the screen for 600 to 1,000 ms at the beginning of the trial. Then, the target face picture ($22.62^\circ \times 18.00^\circ$) was presented centrally until participants made a response or until 1,000 ms elapsed. Participants were required to judge the gender of the face (i.e., male vs. female) as quickly and accurately as possible by pressing the “Z” or the “M” key on a standard keyboard (counterbalanced across participants). After participants made a response or 1,000 ms elapsed, a feedback indicating the outcome of the current trial and the total rewards up to the current trial was presented for 1,500 ms. Specifically, for one male face and one female face (counterbalanced across participants), a correct response was associated with 80% probability of monetary reward feedback (i.e., “+ ¥ 0.5”) and 20% probability of no reward feedback (i.e., “+0”). In contrast, correct responses for another male face and another female face were always associated with no reward (i.e., “+0”). If participants did not respond within 1,000 ms or responded incorrectly, a feedback showing “Too slow!” or “Wrong!” would display, respectively. Participants earned CNY 48 on average as reward, which would be added to their basic compensation (CNY 10) for taking part in the experiment (CNY 10 ≈ US\$1.6). Thus, the training phase imbued two faces (one male and one female) with value (monetary reward), and another two faces (another male and another female) were not associated with value. The training phase consisted of six blocks with 40 trials each.

The test phase (see Fig. 1b) started a few minutes later after the training phase. At the beginning of each trial, a hollow circle ($0.96^\circ \times 0.96^\circ$) was randomly presented at one of the four corners of an invisible bounding box ($28.08^\circ \times 22.26^\circ$), lasting 300 to 500 ms. This was to make sure that participants always made at least one eye movement from a peripheral fixation to the central target location, so that the first gaze location would be determined by participants themselves rather than by an existing fixation sign (Gurler et al., 2015). Then, a video clip ($22.62^\circ \times 18.00^\circ$), in which a talker said a syllable, was presented centrally for 1,500 ms, followed by a display of four options and their corresponding response keys. Participants were instructed to identify which syllable (“ba,” “ga,” “da,” and “other”) was said by the talker in the video clip by pressing the “D,” “F,” “J,” and “K” keys, respectively, on a standard keyboard (counterbalanced across trials). To avoid the situation in which participants ignored visual information, we did not emphasize that they should base their identification

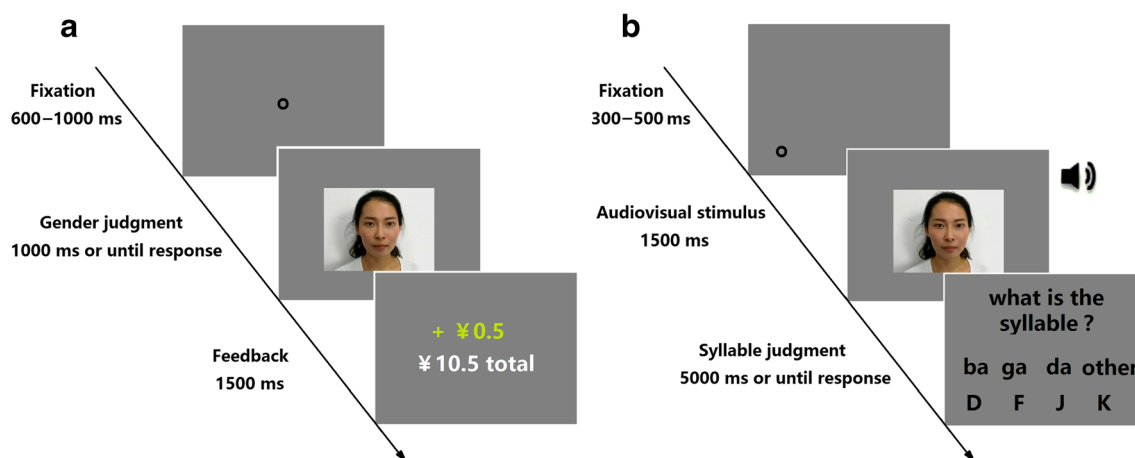


Fig. 1 Trial structure in the training-test paradigm. **a** The training phase. Participants were required to judge the gender of a face picture as quickly and as accurately as possible. A correct response for two faces (counterbalanced across participants) was associated with 80% probability of monetary reward and 20% probability of no reward; correct responses for the other two faces were not associated with reward. The feedback indicating the outcome of the current trial (correct and reward: “+ ¥ 0.5”; correct but no reward: “+0”; wrong: “Wrong!”;

time out: “Too slow!”), and the total reward up to the current trial would be presented after response. **b** The test phase. Participants were instructed to identify which syllable (“ba,” “ga,” “da,” and “other”) was said by the talker in a video clip by pressing a corresponding button on the keyboard. A fixation randomly presented at one of the four corners of an invisible bounding box, followed by a video clip in which a talker said a syllable (McGurk stimuli, and audiovisual congruent/incongruent stimuli as fillers). No reward or feedback was delivered

on what they *heard*, and we excluded the participants who identified the syllable based on what they *saw* as we previously described. In addition, given that participants had to keep head static with a chin rest to ensure the collection of accurate eye movement data, they could not open their mouth to report what syllable they had identified. A button-pressing method was thus a viable way to record their responses (see also Fernández, Macaluso, & Soto-Faraco, 2017).

No feedback was provided in the test phase. The test phase consisted of 20 blocks with 24 trials each. Each type of audiovisual stimuli (i.e., congruent, incongruent, and McGurk stimuli, with congruent and incongruent stimuli acting as fillers) for each talker was equally presented in each block. The test phase consisted of two different types of faces across trials—that is, faces associated with reward in the training phase (i.e., reward-associated faces) and faces not associated with reward in the training phase (i.e., non-reward-associated faces).

Eye tracking was performed at a sampling rate of 1000 Hz during the training and test phases. Nine-point calibration and verification were performed at the beginning of each experimental phase; drift correction (and recalibration if necessary) was performed at the beginning of each block. Participants were required to look at the hollow circle at the beginning of each trial; if participants’ fixations did not locate on the hollow circle within 5,000 ms, drift correction (and recalibration if necessary) would be performed again. During the display of video clips, no fixation sign was presented, and participants were not explicitly instructed to fixate on the face or any other location (i.e., free viewing).

Data analyses

Behavioral data

For the training phase, trials with reaction times (RTs) more than 2.5 standard deviations above or below the mean RT of each condition for each participant were excluded (2.4% of all the trials). A paired-samples *t* test was conducted on the mean RT and accuracy to examine the reward association effect.

For the test phase, we focused on responses to the McGurk stimuli. The proportion of each response category (i.e., “ba,” “ga,” “da,” and “other”) was calculated by dividing the number of responses for each category by the total number of McGurk trials (i.e., the congruent or incongruent trials were not included). For the McGurk effect, we first took a liberal definition—that is, a response of any percept (including “da,” “ga,” “other”) other than the auditory target (“ba”) was classified as a McGurk percept (e.g., Fernández et al., 2017; Gurler et al., 2015). We also tested the data based on a more conservative definition (that is, only a response of “da” was classified as a McGurk response) to evaluate the robustness of results. A paired-samples *t* test was conducted to compare the two conditions.

Signal detection analysis

We conducted a signal detection analysis for the behavior data of the test phase. The advantage of using a signal detection analysis is that the inclusion of filler trials could lead to improved estimates due to the additional data incorporated, and, more importantly, we could separately compare the response criterion (*c*) and the discriminability (*d'*) for reward-associated

and non-reward-associated faces. Specifically, a McGurk stimulus was regarded as a signal trial, and a filler stimulus (either the congruent or incongruent stimulus) was regarded as a noise trial. A response defined as a McGurk percept was regarded as a “yes” response, and a response not defined as a McGurk percept was regarded as a “no” response. Consequently, in the liberal definition of the McGurk effect, “hit” was defined as a response of any percept other than the auditory target to a McGurk stimulus, and “false alarm” was defined as a response of any percept other than the auditory target to a filler stimulus. By calculating the hit rate (P_H) and false alarm rate (P_{FA}), we could obtain the c [$-(Z_H + Z_{FA}) / 2$] and the d' ($Z_H - Z_{FA}$) for each participant. Note that if P_H or P_{FA} was 0 or 1, we would replace 0 with $0.5/N$, and replace 1 with $1 - 0.5/N$, where N was the number of signal or noise trials (Stanislaw & Todorov, 1999). Similarly, in the conservative definition of the McGurk effect, “hit” was defined as a response of “da” to a McGurk stimulus, and “false alarm” was defined as a response of “da” to a filler stimulus.

Eye movement data

To explore how participants extract visual information from value-associated faces, we separately analyzed eye movement data in the training phase and in the test phase. These data were measured from the onset of the presentation of the face to the onset of gender discrimination response in the training phase and from the onset to the offset of each McGurk video clip in the test phase. Blinks, saccades, and fixation locations throughout each video clip were identified using the EyeLink Data Viewer. Interest areas (IAs) were created for each McGurk stimulus (see Fig. 2 for an example), including three rectangular bounding boxes for the left eye, the right eye, and the mouth of the talker; and two irregular bounding shapes for the nose/cheek and the forehead. All IAs covered the whole

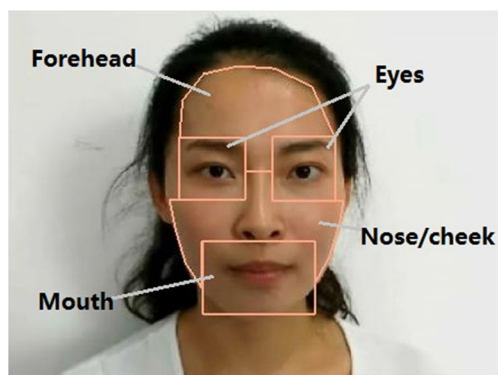


Fig. 2 An example of interest areas (IAs). IAs were created for each McGurk stimulus independently. For direct comparisons between IAs, the size of each IA was kept the same, except for the forehead IA, which was a little larger. Participants spent much less time on the forehead IA than the other IAs, suggesting that the size of the forehead IA would not affect the pattern of results

face, and the mouth IA was large enough to encompass the whole mouth, even at the maximal mouth opening. The critical IAs were essentially of the same size, allowing direct comparisons between different IAs. The IAs did not change during the display of video clips because the talkers’ faces were relatively stationary.

Moreover, we analyzed the eye movement data over the time course of the presentation of the McGurk stimuli. We divided each video clip (1,500 ms in total) into three time intervals. The first time period covered 0 to 500 ms of the video clip in which the talkers kept stable without sound, and little visual information was provided in this period. The second time period covered 500 to 1,100 ms of the video clip in which the talkers opened their mouths and pronounced a syllable, with the sound played and most of visual information provided in this period. The third time period covered 1,100 to 1,500 ms of the video clip in which the talkers closed their mouths and returned to the original state without sound; little visual information was provided in this period.

We conducted 4 (IA: mouth vs. eyes vs. nose/cheek vs. forehead) \times 3 (time period: first vs. second vs. third) \times 2 (reward association: reward-associated vs. non-reward-associated) repeated-measures analysis of variance (ANOVA) on the proportion of looking time and the proportion of fixation number, respectively. In addition, for the purpose of illustration, we divided the whole McGurk stimulus presentation (1,500 ms) into 15 bins (100 ms for each) and calculated the proportion of looking time and fixation number on a particular IA in each bin. This was to further depict the change of the proportion of looking time and fixation number on different IAs over time, although we did not conduct statistical analyses for each bin.

Correlation analysis

Previous studies revealed that there were large individual differences in reporting McGurk percepts (Mallick, Magnotti, & Beauchamp, 2015) and in the eye movement pattern of looking at faces (Gurler et al., 2015). We expected that the reward-related differences of eye movement patterns in either the training or test phase (or both) might be related to the reward-related differences of McGurk proportion in the test phase. Thus, we explored the correlations between the eye movement data and the McGurk proportion. Specifically, we calculated the reward-related differences (i.e., subtracted the measures in the non-reward-associated condition from the measures in the reward-associated condition) of the looking time proportion and fixation number proportion in different interest areas for both the training and test phases; we also computed the reward-related changes of the McGurk proportion in the test phase. Then we conducted a series of correlation analyses between these measures.

Results

Reaction time and accuracy in the training phase

Participants identified the gender of reward-associated faces significantly faster than non-reward-associated faces (446 vs. 451 ms), $t(27) = -2.239, p = .034, d = 0.423$, demonstrating that participants had learnt the face–reward association. There was no significant difference in terms of response accuracy between reward-associated and non-reward-associated faces (97.9% vs. 97.7%), $t(27) = 1.026, p = .314$.

McGurk effect in the test phase

The average accuracies in responding to the filler stimuli (i.e., congruent and incongruent stimuli) in different conditions were very high, ranging from 95.7% to 97.2%, indicating that participants performed the task carefully and effectively. For the McGurk stimuli, the proportion of each response category under each condition is shown in Table 1. According to the liberal definition of the McGurk percept (i.e., a response of any percept other than the auditory target was classified as a McGurk percept), the McGurk proportion was significantly higher for reward-associated faces than for non-reward-associated faces (60.1% vs. 49.9%), $t(27) = 2.438, p = .022, d = 0.461$, which was consistent with our hypothesis.

According to the conservative definition of the McGurk percept (i.e., only a response of “da” was classified as a McGurk percept), the McGurk proportion was marginally higher for reward-associated faces than for non-reward-associated faces (52.2% vs. 43.6%), $t(27) = 1.788, p = .085, d = 0.338$, which was consistent with the pattern reported above. In addition, the proportion of “ba” response (i.e., the true auditory target) was significantly lower for reward-associated faces than for non-reward-associated faces (39.9% vs. 50.1%), $t(27) = 2.438, p = .022, d = 0.461$, mirroring the pattern for the liberally defined McGurk percept. The proportion of “ga” response (i.e., the true visual target) did not differ between reward-associated and non-reward-associated faces (2.7% vs. 2.3%), $t(27) = 0.350, p = .729, d = 0.066$, nor did the proportion of “other” response (5.2% vs. 4.0%), $t(27) = 0.610, p = .547, d = 0.115$. Note that there was a considerable variability in the McGurk proportion across

participants based on either liberal or conservative definition (see Fig. 3), in line with a previous study (Mallick et al., 2015).

Signal detection analysis for the behavioral data in the test phase

The signal detection analysis of the behavioral data based on the liberal definition of McGurk percept revealed that the c was significantly lower for reward-associated faces than for non-reward-associated faces (0.735 vs. 0.998), $t(27) = 3.108, p = .004, d = 0.587$, whereas the d' was significantly higher for reward-associated faces than for non-reward-associated faces (2.348 vs. 1.972), $t(27) = 2.089, p = .046, d = 0.395$. This pattern was replicated in the signal detection analysis of behavioral data based on the conservative definition of McGurk percept, with the c significantly lower for reward-associated faces than for non-reward-associated faces (1.018 vs. 1.247), $t(27) = 2.255, p = .032, d = 0.426$, and the d' marginally higher for reward-associated faces than for non-reward-associated faces (2.218 vs. 1.882), $t(20) = 1.927, p = .065, d = 0.364$.

Eye movements in the training phase

We examined the proportion of looking time and fixation numbers on different interest areas (IAs) in the training phase. For the proportion of looking time, we conducted a 4 (IA: mouth vs. eyes vs. nose/cheek vs. forehead) \times 2 (reward association: reward-associated vs. non-reward-associated) ANOVA, which showed only a significant main effect of IA, $F(3, 81) = 273.632, p < .001, \eta_p^2 = .910$, with the proportion of looking time on the nose/cheek IA (73.91%) significantly higher than the other three IAs (all $ps < .001$) and the proportion on the eyes IA (23.81%) higher than on the mouth IA (1.03%) and the forehead IA (1.88%; all $ps < .001$). Participants rarely looked at areas outside of the face. The number of fixations showed exactly the same pattern, with more fixations on the nose/cheek IA (73.25%) and the eye IA (23.87%) than on the mouth IA (1.16%) and the forehead IA (1.78%).

Eye movements in the test phase: The proportion of looking time

Figure 4 illustrates the time course of the proportion of looking time (i.e., fixation time) at different IAs. The 4 (IA: mouth vs. eyes vs. nose/cheek vs. forehead) \times 3 (time period: first vs. second vs. third) \times 2 (reward association: reward-associated vs. non-reward-associated) ANOVA showed that the main effect of IA was significant, $F(3, 81) = 18.563, p < .001, \eta_p^2 = .407$. Planned comparisons showed that the proportion of looking time on the forehead IA was significantly lower than other three IAs (all $ps < .001$), the proportion of

Table 1 Mean proportion of responses to McGurk stimuli with standard errors in parentheses

Reward association	Responses to McGurk stimuli (%)			
	“ba”	“ga”	“da”	“other”
Reward-associated faces	39.9 (6.0)	2.7 (1.0)	52.2 (5.7)	5.2 (1.4)
Non-reward-associated faces	50.1 (6.7)	2.3 (0.8)	43.6 (6.1)	4.0 (1.5)

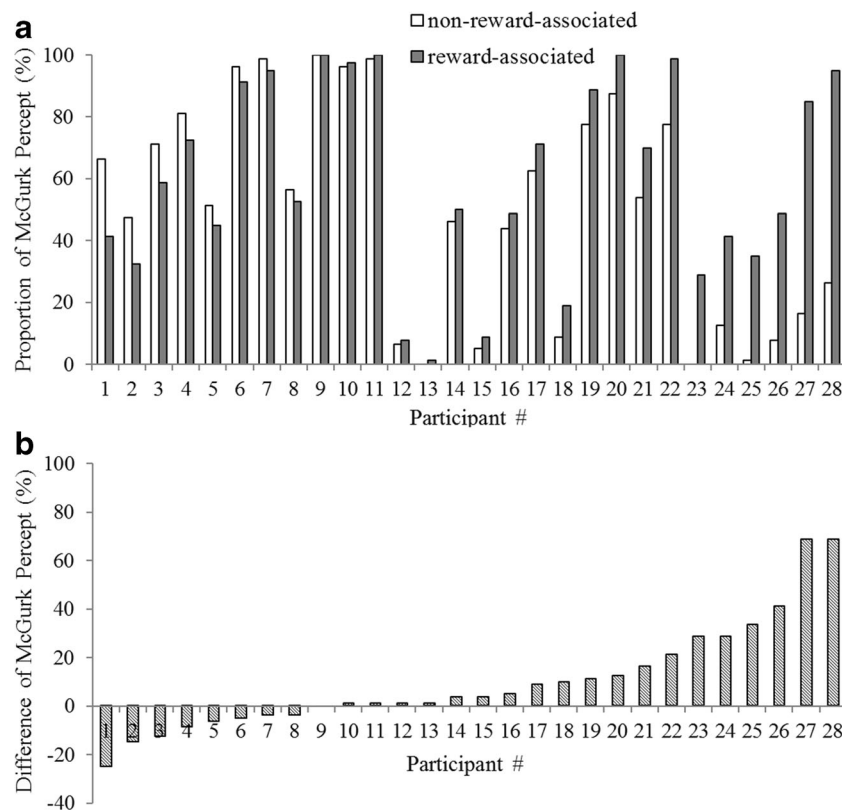


Fig. 3 Individual differences in the McGurk proportion based on the liberal definition of the McGurk percept. **a** Each participant's McGurk proportion for reward-associated and non-reward-associated faces. **b** The

difference of the McGurk proportion between reward-associated and non-reward-associated faces for each participant

looking time on the mouth IA was marginally higher than on the eyes IA ($p = .073$), and there were no differences between other IAs (all $ps > .269$). The main effect of time period was also significant, $F(2, 54) = 126.548$, $p < .001$, $\eta_p^2 = .824$. Planned comparisons showed that the proportion of looking time on the first time period was significantly lower than the other two time periods (all $ps < .001$), and there was no significant difference between the second and third periods ($p = .293$). The main effect of reward association was not significant, $F(1, 27) = 0.040$, $p = .843$, $\eta_p^2 = .001$. The IA \times Time Period interaction was significant, $F(6, 162) = 15.819$, $p < .001$, $\eta_p^2 = .369$, so was the IA \times Reward Association interaction, $F(3, 81) = 2.897$, $p = .040$, $\eta_p^2 = .097$. The Time Period \times Reward association interaction was not significant, $F(2, 54) = 0.558$, $p = .576$, $\eta_p^2 = .020$. Importantly, the three-way interaction between IA, time period, and reward association was significant, $F(6, 162) = 2.373$, $p = .032$, $\eta_p^2 = .081$, and we further explore this interaction below.

We conducted 4 (IA: mouth vs. eyes vs. nose/cheek vs. forehead) \times 2 (reward association: reward-associated vs. non-reward-associated) repeated-measures ANOVA for the first, second, and third time periods, respectively. For the first time period (0–500 ms of the video clips; see Fig. 5, left panel), only the main effect of IA was significant, $F(3, 81) = 20.573$, $p < .001$, $\eta_p^2 = .432$. Planned comparisons showed

that the proportion of looking time on the forehead IA was significantly lower than other three IAs (all $ps < .001$), the proportion of looking time on the eyes IA was significantly lower than mouth IA ($p = .029$) and nose/cheek IA ($p < .001$), and there was no significant difference between mouth and nose/cheek IAs. The main effect of reward association and the interaction effect were not significant (all $ps > .493$).

For the second time period (500–1,100 ms of the video clips; see Fig. 5, middle panel), the main effect of IA was significant, $F(3, 81) = 22.510$, $p < .001$, $\eta_p^2 = .455$. Planned comparisons showed that the proportion of looking time on the forehead IA was significantly lower than the other three IAs (all $ps < .001$), the proportion of looking time on the mouth IA was significantly higher than the other three IAs (all $ps < .005$), and there was no significant difference between eye and nose/cheek IA ($p = .277$). The main effect of reward association was not significant, $F(1, 27) = 0.388$, $p = .538$, $\eta_p^2 = .014$. Importantly, the IA \times Reward Association interaction was significant, $F(3, 81) = 2.908$, $p = .040$, $\eta_p^2 = .097$. Planned t tests on simple effects showed that the proportion of looking time was significantly higher for reward-associated faces than for non-reward-associated faces (28.3% vs. 25.9%) on the nose/cheek IA, $t(27) = 2.328$, $p = .028$, $d = 0.440$, although this effect did not reach significance if more stringent statistical tests were applied. This effect did not appear on

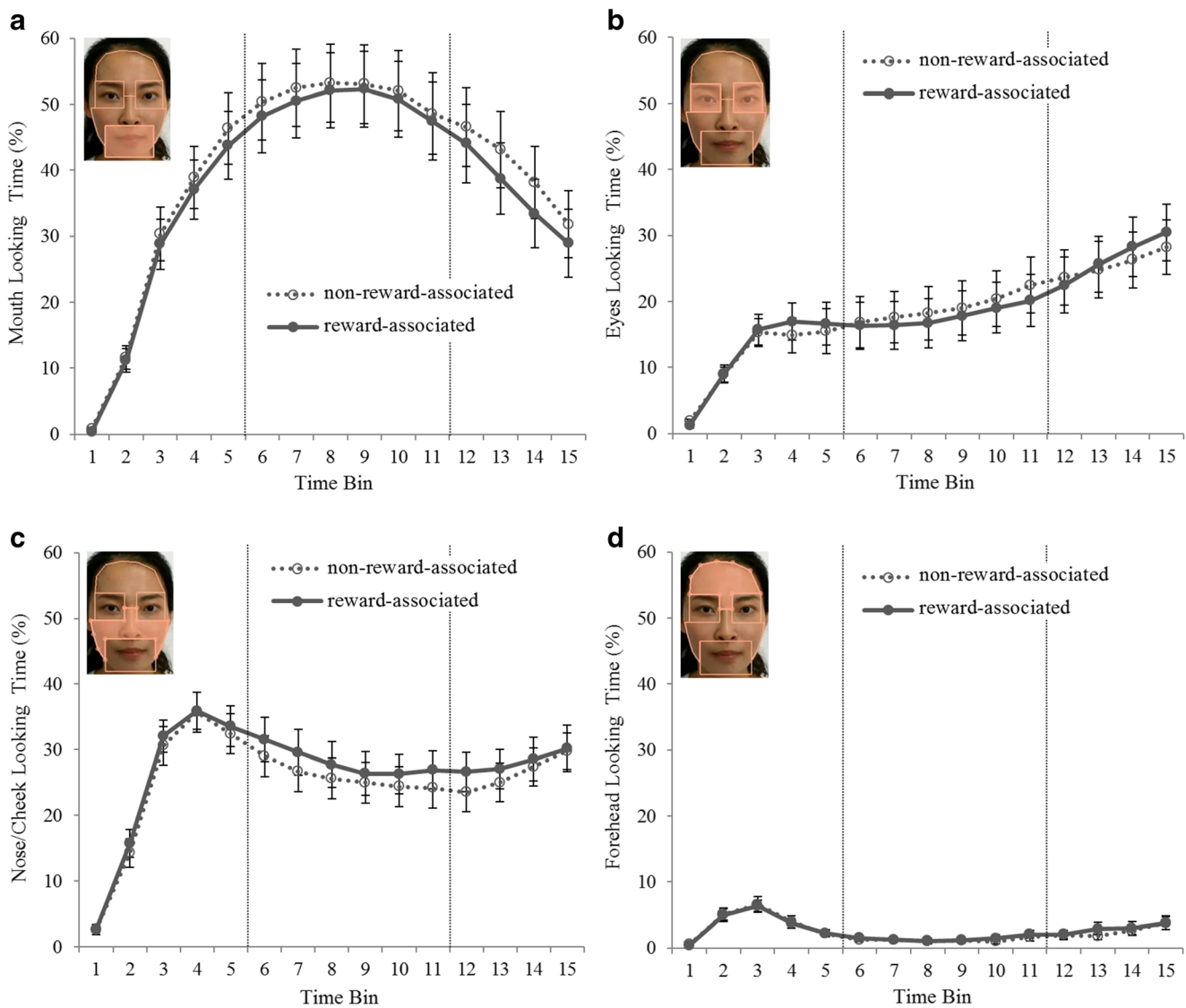


Fig. 4 Time course for the proportion of looking time on the interest area (IA) for (a) mouth, (b) eyes, (c) nose/cheek, and (d) forehead with standard errors. The whole McGurk stimulus presentation (1,500 ms) was divided into 15 time bins (100 ms for each) to further illustrate the

change of the proportion of looking time on different IAs over time. The vertical lines separated time periods (i.e., 0–500 ms, 500–1,100 ms, and 1,100–1,500 ms of stimulus presentation) that we used in the statistical analyses

other IAs (all p s > .101), suggesting that compared with non-reward-associated faces, participants looked at reward-associated faces longer, but only on the extraoral facial area, which is somewhat inconsistent with our original hypothesis.

For the third time period (1,100–1,500 ms of the video clips; see Fig. 5, right panel), the main effect of IA was significant, $F(3, 81) = 12.763, p < .001, \eta_p^2 = .321$. Planned comparisons showed that the proportion of looking time on the forehead IA was significantly lower than the other three IAs (all p s < .001), and there were no significant differences between these three IAs (all p s > .917). The main effect of reward association was not significant, $F(1, 27) = 0.241, p = .627, \eta_p^2 = .009$. But the IA \times Reward Association interaction was significant, $F(3, 81) = 3.408, p = .021, \eta_p^2 = .112$. Planned t tests on simple effects showed that the proportion

of looking time was significantly lower for reward-associated faces than for non-reward-associated faces (36.0% vs. 40.0%) on the mouth IA, $t(27) = -2.122, p = .043, d = 0.401$, although this effect would not survive if more stringent statistical tests were applied. There were not reward association effects on other IAs (all p s > .098). The result here was surprising, as it indicated that participants were less likely to look at the mouth area of reward-associated faces, relatively to non-reward-associated faces, even though visual information in this area was thought to be a causer of McGurk effect. This is in contradictory to our original hypothesis.

We also collapsed data over the three time periods and conducted a 4 (IA: mouth vs. eyes vs. nose/cheek vs. forehead) \times 2 (reward association: reward-associated vs. non-reward-associated) ANOVA. The IA \times Reward Association

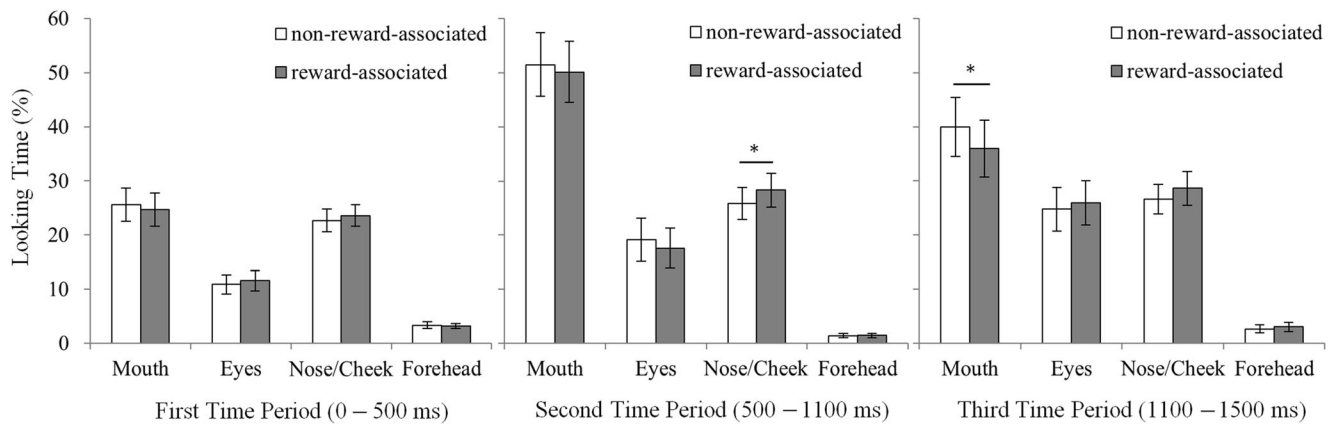


Fig. 5 The proportion of looking time, with standard errors, on different interest areas (IAs) in different time periods under the reward-associated and non-reward-associated conditions. The three-way interaction between IA, time period, and reward association was significant ($p =$

.032). The asterisk indicates significant difference between the two conditions (“*” means that $p = .028$ for the middle panel, and $p = .043$ for the right panel)

interaction was significant (see Fig. 6a). Planned t tests on simple effects showed that the proportion of looking time on the nose/cheek IA was marginally higher for reward-associated faces than for non-reward-associated faces (26.9% vs. 25.1%), $t(27) = 2.034$, $p = .052$, $d = 0.384$, although this effect would not survive when more stringent statistical tests were applied. The pattern here again demonstrated the importance of extraoral facial areas in the value-driven McGurk effect.

Eye movements in the test phase: The proportion of fixation number

Figure 7 illustrates the change of the proportion of fixation number in different IAs over time. The 4 (IAs: mouth vs. eyes vs. nose/cheek vs. forehead) \times 3 (time period: the first period vs. the second period vs. the third period) \times 2 (reward association: reward-associated vs. non-reward-associated) repeated-measures ANOVA showed a pattern almost identical to what

was reported for the proportion of looking time, above. Collapsing the data over the three time periods, we once again observed the interaction between IA and reward association, $F(3, 81) = 3.276$, $p = .025$, $\eta_p^2 = .108$. Planned t tests on simple effects showed that the proportion of fixation number on the nose/cheek IA was marginally higher for reward-associated faces than for non-reward-associated faces (26.0% vs. 24.4%), $t(27) = 2.052$, $p = .050$, $d = 0.388$, although this effect would not reach significance if more stringent tests were applied; but this pattern was reversed for the mouth IA (35.0% vs. 37.2%), $t(27) = -2.013$, $p = .054$, $d = 0.380$ (see Fig. 6b); once again, this difference would not reach significance for the more stringent tests. These results demonstrated that participants pay more attention to the extraoral facial area for reward-associated faces than for non-reward-associated faces; but the opposite pattern was observed for the oral area.

In addition, considering that participants had to make at least one eye movement from a peripheral fixation to the

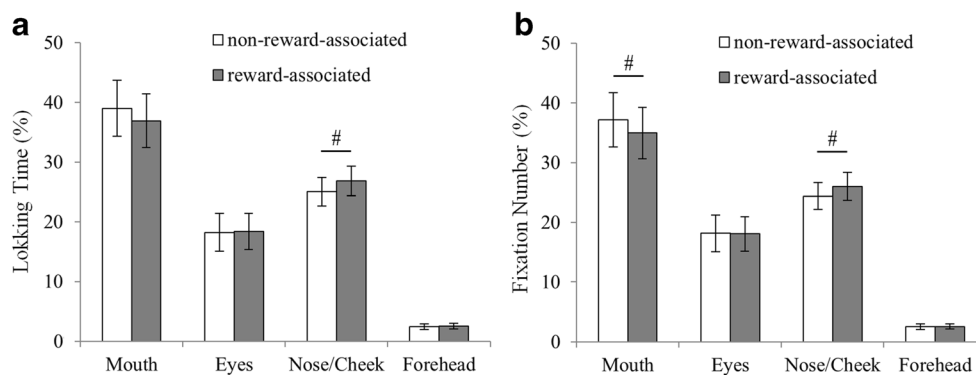


Fig. 6 The proportion of (a) looking time and (b) fixation number under the reward-associated and non-reward-associated conditions in different interest areas (IAs), collapsing over the three time periods. The interactions between IA and reward association were significant for both the proportion of looking time ($p = .040$) and the fixation number

($p = .025$). The hashtag indicates a marginally significant difference between the conditions (“#” in Panel A means that $p = .052$; “#” in Panel B means that $p = .054$ for mouth area, and $p = .050$ for nose/cheek area)

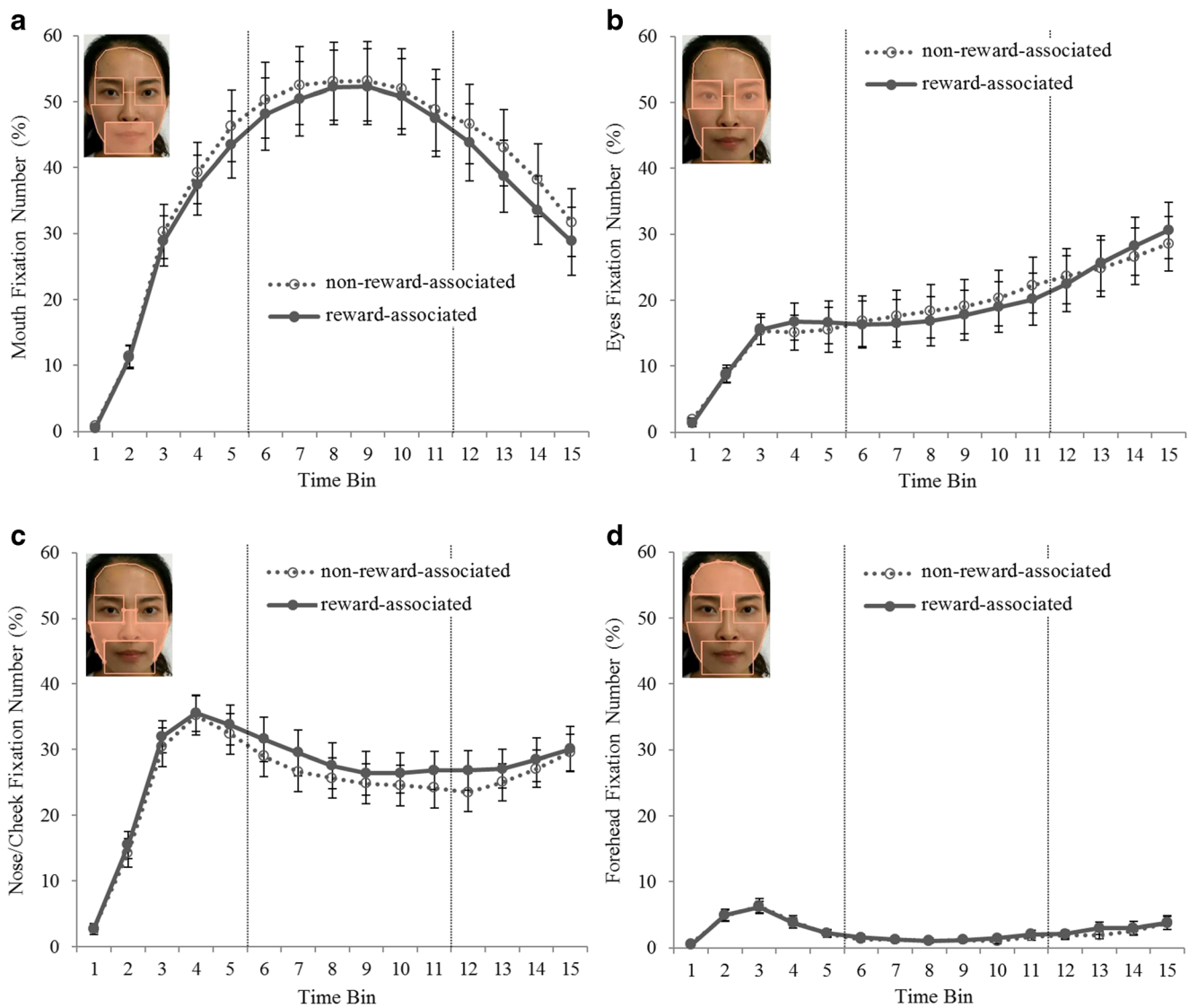


Fig. 7 Time course for the proportion of fixation number on the interest area (IA) for (a) mouth, (b) eyes, (c) nose/cheek, and (d) forehead, with standard errors. The whole McGurk stimulus presentation (1,500 ms) was divided into 15 time bins (100 ms for each) to further illustrate the change

of the proportion of fixation number on different IAs over time. The vertical lines separated time periods (i.e., 0–500 ms, 500–1,100 ms, and 1,100–1,500 ms of stimulus presentation) that we used in the statistical analyses

central target face in each trial, we analyzed the number of first fixation in a particular IA, excluding the fixations outside of the face. We found only a significant main effect of IA, $F(3, 81) = 17.038, p < .001, \eta_p^2 = .387$, with more fixations on the nose/cheek IA (36.66%) and the mouth IA (35.31%) than on the eye IA (19.55%) and the forehead IA (7.90%).

Correlation analysis

Given that, in the training phase, participants spent the longest time looking at the nose/cheek area (73.91%) than any other areas, the correlation analysis was first conducted for this area. Over participants, the difference of the proportion of looking time at nose/cheek area between reward-associated and non-reward-associated faces in the training phase positively

correlated with the difference of the McGurk proportion, either liberally or conservatively defined, between the two conditions in the test phase (see Fig. 8a), $r = .381, p = .045; r = .424, p = .024$. The same pattern was observed for the proportion of fixation number, $r = .350, p = .067; r = .397, p = .037$.

In addition, the difference of the proportion of looking time at the nose/cheek area between reward-associated and non-reward-associated faces in the training phase negatively correlated with the difference of the proportion of looking time at the mouth area between the two conditions in the test phase (see Fig. 8b), $r = -.333, p = .083$. The same pattern was obtained for the proportion of fixation number, $r = -.318, p = .099$.

Since that, in the test phase, the reward-association effect was observed in the nose/cheek area and the mouth area, the correlation analysis was conducted for these two areas. In the

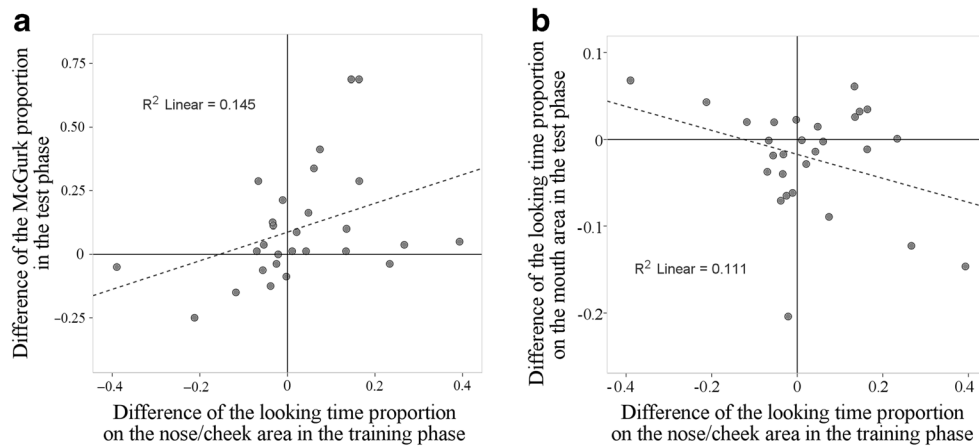


Fig. 8 Correlation analysis across the training and test phases. **a** The difference of the proportion of looking time at nose/cheek area between reward-associated and non-reward-associated faces in the training phase positively correlated with the difference of the McGurk proportion (liberally defined in this figure) between the two conditions in the test

phase ($r = .381, p = .045$). **b** The difference of the proportion of looking time at the nose/cheek area between reward-associated and non-reward-associated faces in the training phase negatively correlated with the difference of the proportion of looking time at the mouth area between the two conditions in the test phase ($r = -.333, p = .083$)

test phase, however, the difference of the proportion of looking time at the nose/cheek area between reward-associated faces and non-reward-associated faces did not correlate with the difference of McGurk proportion, either liberally or conservatively defined, between the two conditions, $r = -.158, p = .421$; $r = -.071, p = .719$. The null effect was also observed for the proportion of fixation number, $r = -.179, p = .362$; $r = -.121, p = .538$. Similarly, in the test phase, the difference of the proportion of looking time at the mouth area between reward-associated and non-reward-associated faces did not correlate with the difference of McGurk proportion, either liberally or conservatively defined, between the two conditions, $r = .030, p = .880$; $r = -.090, p = .648$. The null effect was also observed for the proportion of fixation number, $r = .045, p = .820$; $r = -.042, p = .831$.

Discussion

The present study found that, in line with our prediction, participants would make more use of visual information for value-associated faces in audiovisual speech perception, with more McGurk percepts for reward-associated faces than for non-reward-associated faces. Value-associated stimuli have higher attentional priority (Anderson, 2013; Anderson et al., 2011). Participants devoted more attention to reward-associated faces than to non-reward-associated faces; the deeper processing of reward-associated faces increased the weight of visual information in audiovisual speech perception, resulting in more reports of the McGurk percepts. Convergent with this account, *weaker* McGurk effect had been observed when less attention was assigned to the visual information (Alsuis et al., 2005; Tiippana et al., 2004).

This account gains additional support from our signal detection analysis. Here, participants had lower c for reward-associated faces than for non-reward-associated faces, demonstrating that participants tended to respond “yes” (i.e., having more liberal criterion) in audiovisual speech perception when faces were associated with value. The “yes” response, in the context of the present study, meant a response that was different from the actual auditory target. Thus, the change of c here implies a tendency that participants made use of visual information even when the visual information was incongruent with the auditory information (Seilheimer, Rosenberg, & Angelaki, 2014).

Moreover, we speculate that multisensory integration processes may play a role in the value-driven McGurk effect. The signal detection analysis revealed that participants had higher d' for reward-associated faces than for non-reward-associated faces, demonstrating that participants were more sensitive to the “signal” in audiovisual speech perception when faces were associated with value. The “signal” here refers to stimulus properties that could lead to the McGurk percept, and the change of sensitivity to these properties might be related to certain internal processes, such as multisensory integration for a McGurk stimulus. An fMRI study (Pooresmaeili et al., 2014) showed that reward associations modulate responses in multisensory processing regions (i.e., superior temporal sulcus [STS]) and other classical reward regions, but only the modulation strength of STS could predict the magnitude of the behavioral effect. The authors argued that multisensory regions may mediate the transfer of value signals across senses, rather than classical reward regions in the cross-modal context. Considering that our results demonstrate that the value-driven mechanism of attention works not only in simple cross-modal contexts (e.g., Anderson, 2016; Pooresmaeili et al., 2014) but also in the complex audiovisual

speech perception context, it is possible that multisensory integration processing was directly facilitated by reward association in the present study, resulting in more McGurk percepts for reward-associated faces. Nevertheless, it should be noticed that the McGurk effect cannot be equated with multisensory integration, because much more is involved with the McGurk effect than just multisensory integration, such as conflict resolution (e.g., Fernández et al., 2017; see also Alsius et al., 2018 for a review).

Furthermore, it should be mentioned that our results seem to show contrasts with a previous study in which Walker et al. (1995) investigated the influence of face familiarity (i.e., a form of value to some extent) on the McGurk effect, and found that participants who were familiar with the face reported *less* McGurk percepts than those who were unfamiliar with the face when the face and voice were from different persons. However, there are key differences between the studies. Participants in our study did not know the talkers before, and all the talkers' faces in the training phase were static pictures and appeared at the same frequency. That is, participants had the same familiarity of all the talkers' static faces, and had no prior knowledge of the talkers' dynamic facial movements. Walker et al. (1995) defined the familiarity in terms of participants having had face-to-face interactions with the talker in daily life, which means that participants were familiar not only with the talkers' static faces but also with the talkers' dynamic facial movements and voices. As the authors mentioned, participants were able to use their prior knowledge of those familiar faces (expectations of what speech events were likely and of how these events were realized through dynamic facial movements); the incongruence between the visual and auditory modality was thus easier to be detected, resulting in less report of McGurk percepts. The authors also found that when the face and voice were from the same person, there were no differences in McGurk percepts between the participants who were familiar with the faces and the participants who were unfamiliar with them, a pattern recently replicated (Magnotti et al., 2018). It seems that there are fundamental differences between the value of familiarity and the value of reward, which merit future systematic studies.

Eye movements suggest how participants gather visual information from faces in audiovisual speech perception. In the present study, oral facial movements were the most important visual information in audiovisual speech perception, evidenced by the highest proportions of looking time and number of fixations on the mouth area than on other areas (see Fig. 4, 5, 6 and 7) and by the higher proportions of looking at the mouth IA in the audiovisual speech perception task (i.e., the test phase) compared with in the face recognition task (i.e., the training phase), in line with previous studies (e.g., Buchan & Munhall, 2012; Gurler et al., 2015; Wilson et al., 2016). One might predict that if McGurk proportion is increased for reward-associated faces, participants will look at the mouth

area more often and/or with more time for reward-associated faces than for non-reward-associated faces. Surprisingly, however, we observed an opposite pattern in our results: Participants looked at the mouth area *less* often and with *less* time for reward-associated faces than for non-reward-associated faces, but they looked at the nose/cheek area *more* often and with *more* time for reward-associated faces than for non-reward-associated faces.

One possible explanation for this eye movement pattern is that different areas of a face would be associated with value to different extent when the association between face and reward was learnt. Indeed, participants in the training phase spent the longest time looking at the nose/cheek area when they were conducting a face recognition task (i.e., determining the gender), a pattern replicating a previous study (Peterson & Eckstein, 2012) which demonstrated that the human visual system optimizes face recognition performance through guidance of eye movements not only toward but just below the eyes (e.g., nose/cheek area in the present study). Moreover, the correlation analysis revealed that the participants who had more looking time (and more fixation number) on the nose/cheek area in the training phase due to reward (i.e., reward condition–non-reward condition) showed, in the test phase, less looking time (and less fixation number) on the mouth area but larger increasing of the McGurk proportion. This pattern is consistent with a recent study which reported that the benefit of visual information on audiovisual speech perception can be predicted by eye movements during a previous face-viewing task (Rennig, Wegner-Clemens, & Beauchamp, 2018).

It seems strange that participants showed higher McGurk proportion for the value-associated faces and, at the same time, showed a counterintuitive eye movement pattern (more looking time and fixated more often on the extraoral area, whereas less looking time and fixated less often on the oral area) for the value-associated faces. The occurrence of this pattern could be due to two reasons. First, it is possible that looking at the mouth area is not strictly necessary for the occurrence of McGurk percept. In other words, in audiovisual speech perception, participants might be affected by oral facial movements even when eyes are directed away from the mouth area. This explains why we did not find a correlation between the McGurk proportion and measures of eye movements on the mouth area in the test phase (see also Rennig et al., 2018; Wilson et al., 2016). In fact, previous studies demonstrated that the visual information provided by oral facial movements could be gathered either from the central region at the mouth or the peripheral region out of the mouth. For example, although the occurrence of McGurk percepts is assumed to be mainly determined by the visual information from the mouth area, the instruction to “look at the mouth” did not increase the McGurk proportion compared with free viewing (Hisanaga et al., 2016) or compared with the instruction “look at the eyes” or “look at the hairline” (Paré, Richler, ten Hove, &

Munhall 2003), suggesting that information about mouth movements can be obtained from other areas in non-mouth-looking conditions.

Second, extraoral facial movements may provide useful visual information apart from the oral facial movements, which helps to elicit the McGurk effect. Thomas and Jordan (2004) manipulated the movements of the mouth and other facial areas independently, and found that the extraoral movements could promote the identification of audiovisual speech even when the mouth is kept static or removed from the face. Jordan and Thomas (2011) further found that the McGurk effect is observable even when the talker's face is occluded horizontally or diagonally (i.e., when the mouth area is occluded). In the present study, longer looking time and fixated more often on the extraoral area of reward-associated faces, compared with non-reward-associated faces, might help participants process the visual information provided by extraoral area, resulting in higher McGurk proportion.

To conclude, by associating faces with or without monetary reward in the training phase, we demonstrated that individuals could in the subsequent test phase report more McGurk percepts for reward-associated faces, relative to non-reward-associated faces, indicating that value-associated faces enhance the influence of visual information on audiovisual speech perception. The signal detection analysis revealed that participants have lower response criterion and higher sensory discriminability for reward-associated faces than for non-reward-associated faces, indicating that when the talking faces are associated with value, individuals tend to make more use of visual information in processing the McGurk stimuli. Surprisingly, we found that participants in the test phase had more looking time and number of fixations on the nose/cheek area of reward-associated faces than non-reward-associated faces; the opposite pattern was found for the mouth area. The correlation analysis revealed that the more participants looked at the nose/cheek area in the training phase due to reward, the more McGurk effect occurred in the test phase for reward-associated faces. These findings suggest that associating reward with a face may increase the attentional priority of the extraoral area, which contributes to the audiovisual speech perception.

Acknowledgement We thank Miss Yueyuan Zheng, Miss Jiayi Li, and Miss Muhang Li for proofreading, and Mrs. Haining Bi for allowing us to use her face picture in this article. This work was supported by the Natural Science Foundation of China (Grant No. 31630034) and the National Basic Research Program of China (973 Program: 2015CB856400).

Compliance with ethical standards

Conflict of Interest The authors declare no conflict of interest.

Open practices statement The data and materials are available on request from the corresponding author.

The experiment was not preregistered.

References

- Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, *15*(9), 839–843. <https://doi.org/10.1016/j.cub.2005.03.046>
- Alsius, A., Paré, M., & Munhall, K. G. (2018). Forty years after hearing lips and seeing voices: The McGurk effect revisited. *Multisensory Research*, *31*(1), 111–144. <https://doi.org/10.1163/22134808-00002565>
- Anderson, B. A. (2013). A value-driven mechanism of attentional selection. *Journal of Vision*, *13*(3), 103–104. <https://doi.org/10.1167/13.3.7>
- Anderson, B. A. (2016). Value-driven attentional capture in the auditory domain. *Attention, Perception, & Psychophysics*, *78*(1), 242–250. <https://doi.org/10.3758/s13414-015-1001-7>
- Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10367–10371. <https://doi.org/10.1073/pnas.1104047108>
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. <https://doi.org/10.1163/156856897X00357>
- Brancazio, L., Miller, J. L., & Paré, M. A. (2003). Visual influences on the internal structure of phonetic categories. *Perception & Psychophysics*, *65*(4), 591–601. <https://doi.org/10.3758/BF03194585>
- Buchan, J. N., & Munhall, K. G. (2012). The effect of a concurrent working memory task and temporal offsets on the integration of auditory and visual information. *Seeing & Perceiving*, *25*(1), 87–106. <https://doi.org/10.1163/187847611X620937>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. G. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Fernández, L. M., Macaluso, E., & Soto-Faraco, S. (2017). Audiovisual integration as conflict resolution: The conflict of the McGurk illusion. *Human Brain Mapping*, *11*(38), 5691–5705. <https://doi.org/10.1002/hbm.23758>
- Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, *108*(1), 1197–1208. <https://doi.org/10.1121/1.1288668>
- Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception, & Psychophysics*, *77*(4), 1333–1341. <https://doi.org/10.3758/s13414-014-0821-1>
- Hisanaga, S., Sekiyama, K., Igasaki, T., & Murayama, N. (2016). Language/culture modulates brain and gaze processes in audiovisual speech perception. *Scientific Reports*, *6*, 35265. <https://doi.org/10.1038/srep35265>
- Jordan, T. R., & Thomas, S. M. (2011). When half a face is as good as a whole: Effects of simple substantial occlusion on visual and audiovisual speech perception. *Attention, Perception, & Psychophysics*, *73*(7), 2270–2285. <https://doi.org/10.3758/s13414-011-0152-4>
- Magnotti, J. F., Smith, K. B., Salinas, M., Mays, J., Zhu, L. L., & Beauchamp, M. S. (2018). A causal inference explanation for enhancement of multisensory integration by co-articulation. *Scientific Reports*, *8*, 18032. <https://doi.org/10.1038/s41598-018-36772-8>
- Mallick, D. B., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: Contributions of participants,

- stimuli, time, and response type. *Psychonomic Bulletin & Review*, 22(5), 1299–1307. <https://doi.org/10.3758/s13423-015-0817-4>
- Marques, L. M., Lapenta, O. M., Costa, T. L., & Boggio, P. S. (2016). Multisensory integration processes underlying speech perception as revealed by the McGurk illusion. *Language, Cognition and Neuroscience*, 31(9), 1115–1129. <https://doi.org/10.1080/23273798.2016.1190023>
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748. <https://doi.org/10.1038/264746a0>
- Peterson, M. F., & Eckstein, M. P. (2012). Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 109(48), E3314–E3323. <https://doi.org/10.1073/pnas.1214269109>
- Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Attention, Perception, & Psychophysics*, 65(4), 553–567. <https://doi.org/10.3758/bf03194582>
- Pooresmaeili, A., FitzGerald, T. H. B., Bach, D. R., Toelch, U., Ostendorf, F., & Dolan, R. J. (2014). Cross-modal effects of value on perceptual acuity and stimulus encoding. *Proceedings of the National Academy of Sciences of the United States of America*, 111(42), 15244–15249. <https://doi.org/10.1073/pnas.1408873111>
- Raymond, J. E., & O'Brien, J. L. (2009). Selective visual attention and motivation: The consequences of value learning in an attentional blink task. *Psychological Science*, 20(8), 981–988. <https://doi.org/10.1111/j.1467-9280.2009.02391.x>
- Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, 17(5), 1147–1153. <https://doi.org/10.1093/cercor/bhl024>
- Rennig, J., Wegner-Clemens, K., & Beauchamp, M. S. (2018). Face viewing behavior predicts multisensory gain during speech perception. *bioRxiv*, 331306. <https://doi.org/10.1101/331306>
- Seilheimer, R. L., Rosenberg, A., & Angelaki, D. E. (2014). Models and processes of multisensory cue combination. *Current Opinion in Neurobiology*, 25(2), 38–46. <https://doi.org/10.1016/j.conb.2013.11.008>
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26(2), 212–215. <https://doi.org/10.1121/1.1907309>
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1), 137–149. <https://doi.org/10.3758/BF03207704>
- Thomas, S. M., & Jordan, T. R. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, 30(5), 873–888. <https://doi.org/10.1037/0096-1523.30.5.873>
- Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, 16(3), 457–472. <https://doi.org/10.1080/09541440340000268>
- Wang, L., Duan, Y., Theeuwes, J., & Zhou, X. (2014). Reward breaks through the inhibitory region around attentional focus. *Journal of Vision*, 14(12), 1–7. <https://doi.org/10.1167/14.12.2>
- Wang, L., Yu, H., & Zhou, X. (2013). Interaction between value and perceptual salience in value-driven attentional capture. *Journal of Vision*, 13(3), 1–13. <https://doi.org/10.1167/13.3.5>
- Wilson, A. H., Alsius, A., Paré, M., & Munhall, K. G. (2016). Spatial frequency requirements and gaze strategy in visual-only and audiovisual speech perception. *Journal of Speech Language & Hearing Research*, 59(4), 601–615. https://doi.org/10.1044/2016_JSLHR-S-15-0092
- Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, 57(8), 1124–1133. <https://doi.org/10.3758/BF03208369>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.