Reading and Writing
https://doi.org/10.1007/s11145-021-10200-2

Audiovisual speech perception and its relation with temporal processing in children with and without autism

Shuyuan Feng[1,2] · Haoyang Lu[3,4] · Jing Fang[5] · Xue Li[6] · Li Yi[2,7] · Lihan Chen[2]

Accepted: 30 July 2021

Abstract
Children with autism spectrum disorder (ASD) have deficits in audiovisual speech perception and temporal processing. The current study has examined the relationship between the audiovisual speech perception deficits and temporal processing deficits in children with and without ASD. To this end, using the McGurk paradigm, we implemented two experiments to explore audiovisual speech perception (Experiment 1) and temporal processing (Experiment 2), as well as the correlation between them, in children with ASD and typically developing (TD) children. We recruited 4- to 8-year-old children, some with ASD and some TD, to perform a McGurk task in Experiment 1 (24 children with ASD, 26 TD children) and to complete a simultaneity judgement task in Experiment 2 (31 children with ASD, 29 TD children). On the basis of the data from participants who participated in both Experiment 1 and Experiment 2 (20 children with ASD, 21 TD children), we analyzed the correlation between audiovisual speech perception and temporal processing in children with ASD and TD children separately. We found that children with ASD showed weaker audiovisual speech perception (based on the population) and less acute temporal processing compared with TD children. The correlation analysis revealed that audiovisual speech perception and temporal processing were correlated in TD children when the visual led the audio but not when the audio led the visual. No such correlation was found in children with ASD either when the visual led the audio or when the audio led the visual. The present study implicated that the correlation between audiovisual speech perception and temporal processing might be contingent on the range of individual temporal processing abilities.

Keywords Autism spectrum disorder · Audiovisual speech perception · Temporal processing · McGurk effect

✉ Li Yi
yilipku@pku.edu.cn

✉ Lihan Chen
clh@pku.edu.cn

Extended author information available on the last page of the article

Published online: 13 August 2021

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by restricted and repetitive behaviors, as well as impaired social interaction and social communication (American Psychiatric Association, 2013). Children with ASD often have language and communication impairments (Volden & Phillips, 2010; Weismer et al., 2010). Language development is constrained by many factors, such as genetic defects (Hayiou-Thomas, 2008), joint attention (Farrant & Zubrick, 2012), pointing gestures (Colonnesi et al., 2010), parent–child interaction (Farrant & Zubrick, 2012), maternal education and socioeconomic status (Muluk et al., 2014), and audiovisual speech perception (Gervain & Mehler, 2010; Werker & Gervain, 2013). Among these factors, audiovisual speech perception lays the foundation for language acquisition (Gervain & Mehler, 2010; Werker & Gervain, 2013). Audiovisual speech perception entails the integration of speech sounds and visible articulatory information to convey a coherent, unified percept (Altieri et al., 2014; Bebko et al., 2014). In the course of audiovisual speech perception development, children with ASD lagged behind typically developing (TD) children (Foxe et al., 2015; Taylor et al., 2010). This delayed audiovisual speech perception in children with ASD was proposed to be correlated with their weaker temporal processing (Stevenson et al., 2014), poorer lipreading ability (Iarocci et al., 2010), atypical face viewing patterns (Bebko et al., 2014; Irwin et al., 2011), and weaker tendency to combine the elementary stimulus parts into a coherent one (Baum et al., 2015; Happé & Frith, 2006). Previous studies found that, for TD infants, performance of audiovisual speech perception at 6 months predicted the levels of receptive vocabulary (i.e., the collection of words understood by an individual) at 12 months (Altvater-Mackensen & Grossmann, 2015; Imafuku et al., 2019). Studying audiovisual speech perception in children with ASD might contribute to the understanding of their language development and hence provide insights for promoting their language abilities. The present study aimed to explore the audiovisual speech perception and one of its correlates (i.e., temporal processing) in children with ASD and TD children.

Audiovisual speech perception is often manifested by the McGurk effect, which occurs when the acoustic part of one syllable (e.g., /ba/) is dubbed onto the visual part of another syllable (e.g., /ga/), leading to a fused perception of a third syllable (e.g., /da/; McGurk & MacDonald, 1976). The rate of the fused perception is usually taken as the strength of the McGurk effect (MacDonald, 2017; McGurk & MacDonald, 1976). Most of studies using the McGurk effect to investigate the audiovisual speech integration in children with ASD showed that children with ASD had weaker audiovisual speech perception than TD children did, as shown by their weaker McGurk effect (e.g., Bebko et al., 2014; de Gelder et al., 1991; Irwin et al., 2011; Mongillo et al., 2008; Stevenson et al., 2014). Meanwhile, some other studies demonstrated that children with ASD and TD children showed similar audiovisual speech perception (i.e., similar McGurk effect Iarocci et al., 2010; Woynaroski et al., 2013). A recent meta-analysis concluded that children with ASD have weaker audiovisual speech perception than TD children (Zhang

et al., 2019). This audiovisual speech perception deficits (i.e., weaker McGurk effect) could be accounted for by the weak central coherence theory. The theory holds that children with ASD are less likely to combine components of a stimulus into a holistic whole, as shown by their weaker McGurk effect (i.e., less fused perception; Baum et al., 2015; Bebko et al., 2014; Happé & Frith, 2006).

Audiovisual speech perception (i.e., McGurk effect) is affected by native language backgrounds and cultures (Chen & Massaro, 2004; Massaro et al., 1995; Sekiyama & Tohkura, 1993). Previous studies found that native Japanese speakers showed weaker McGurk effect than native English speakers did (Sekiyama & Tohkura, 1993). For example, Japanese contains less visually-identifiable consonants than English (e.g., /f/, /v/, /θ/ and /ð/) and native Japanese speakers are less tended to directly look at interlocutors' face in conversation than native English speakers. These make native Japanese speakers give less weights to visual information but more weights to auditory information, which affects McGurk effect in native Japanese speakers. Similarly, Chinese also contains less visually identifiable consonants and native Chinese speakers are also less likely to fixate the interlocutors' face for politeness concerns. In this case, Chinese speakers are likely to show weaker audiovisual speech integration than English speakers, however, the conclusions are mixed according to the few existing literature. Specifically, two studies found that Chinese speakers showed similar audiovisual speech perception as English speakers did (Chen and Massaro 2004; Magnotti et al., 2015), while another two studies found that Chinese speakers showed weaker audiovisual speech perception than English speakers did (Burnham & Lau, 1998; Sekiyama and Tohkura 1993). Moreover, limited study was conducted to explore the audiovisual speech perception in Chinese children. Therefore, the present study aimed to explore audiovisual speech perception (i.e., McGurk effect) in Chinese children with and without ASD.

Audiovisual speech perception involves temporal processing, which helps to detect whether the auditory cue and the visual cue within a stimulus are simultaneous or not, and further affects whether the auditory and visual stimuli are processed as a whole or separately (Baum et al., 2015; Stevenson et al., 2015). Previous studies that examined temporal processing in individuals with ASD mainly adopted the preferential looking task, the simultaneity judgement (SJ) task, and the temporal order judgement (TOJ) task. In detail, the preferential looking task examines whether participants view the synchronized-audiovisual speech longer than the asynchronized-one. Employing audiovisual speech stimuli with various stimulus onset asynchronies (SOAs), the SJ task requires participants to judge whether the auditory and visual components are simultaneous or not. In the TOJ task, participants were required to judge which component—the auditory part or the visual part—comes first. Compared with the TD counterparts, children with ASD viewed the synchronized-audiovisual speech less in the preferential looking task (Bebko et al., 2006; Grossman et al., 2015), made more simultaneity judgements in the SJ task (Stevenson et al., 2014), and demonstrated diminished temporal order sensitivities in the TOJ task (de Boer-Schellekens et al., 2013). These findings suggested that children with ASD showed less acute temporal processing for speech stimuli than TD children did.

Temporal processing has tight relation with audiovisual speech perception (Stevenson et al., 2012; van Wassenhove et al., 2007). For example, Stevenson et al. (2012) measured TD adults' audiovisual speech perception with fused perception for the McGurk stimuli, and measured their temporal processing with synchrony perception for flash-beeps with various SOAs. Participants' temporal processing was deemed as much acuter when they were more accurate in identifying asynchronous flash-beeps. As a result, audiovisual speech perception was found to be positively correlated with temporal processing when the visual led the audio in TD adults. That is, for TD adults, the more accurate the identification for asynchronous flash-beeps, the stronger the McGurk effect. It provides evidence for the association between audiovisual speech perception and temporal processing. Similarly, it is proposed that temporal processing deficits associate with speech perception deficits in specific group like children with ASD (Baum et al., 2015; Stevenson et al., 2018). To our best knowledge, only one study explored the relationship between audiovisual speech perception and temporal processing in children with ASD and TD children employing speech stimuli (Stevenson et al., 2014). It found that stronger McGurk effect was accompanied by more accuracy in identifying asynchronous speech stimuli in children with ASD but not in TD children. That is, it confirmed the positive correlation between audiovisual speech perception and temporal processing in children with ASD but not in TD children. It needs to further investigate the correlation between the two in children with ASD and TD children. The investigation in this line could provide insights for the mechanisms underlying speech perception deficits in children with ASD. As speech perception lays foundations for language development, speech perception improvement could further enhance the language abilities in children as a whole, especially in children with ASD (Gervain & Mehler, 2010; Werker & Gervain, 2013).

To this end, we explored the relationship between audiovisual speech perception and temporal processing in children with ASD and TD children. Employing identical stimuli could avoid the potential confounding factor—the stimuli type—influencing temporal processing (van Wassenhove et al., 2007). Therefore, in present study, we adopted the same stimuli rather than different stimuli (e.g., Stevenson et al., 2014) across two different tasks. Specifically, we measured participants' audiovisual speech perception and temporal processing using identical stimuli: the McGurk stimuli with various SOAs between the auditory and visual components. We implemented two experiments: Experiment 1 explored the audiovisual speech perception (McGurk effect) in children with ASD and TD children, and Experiment 2 explored the temporal processing (perceived synchrony) in children with ASD and TD children. First, we examined whether children with ASD would exhibit weaker audiovisual speech perception and less acute temporal processing compared with TD children. We hypothesized that children with ASD would show weaker audiovisual speech perception and less acute temporal processing compared with TD children. Second, we examined whether audiovisual speech perception would be correlated with temporal processing. Based on previous findings in TD individuals, we expected that the audiovisual speech perception would be correlated with the temporal processing in both children with ASD and TD children (Stevenson et al., 2012).

Experiment 1

In this experiment, we measured participants' McGurk effect when processing the McGurk stimuli with various SOAs. Participants completed a McGurk task by answering what the speaker said.

Method

Participants

We recruited 76 high-functioning children with ASD (71 boys, 5 girls) from schools specialized for children with ASD, and 47 TD children (41 boys, 6 girls) from public kindergartens and a normal elementary school. All children with ASD were diagnosed as ASD according to the diagnostic criteria for ASD in the Diagnostic and Statistical Manual of Mental Disorders (DSM-V; American Psychiatric Association, 2013). The ASD diagnosis were further confirmed by the Chinese version of the Autism Spectrum Quotient: Children's Version (AQ-Child; Auyeung et al., 2008). All children understood the McGurk task quite well as they were only required to report what the speaker said in the task. However, 52 children with ASD and 21 TD children were excluded from the experiment as they did not report any fused perception in the practice session of the McGurk task, or they reported the fused perception in only one trial in the formal experiment. The final sample for the present experiment consisted of 24 high-functioning children with ASD (22 boys and 2 girls; age range: 4.90–7.25 years) and 26 TD children (24 boys and 2 girls; age range: 4.52–6.79 years). The two groups in the final sample were also matched in age and IQ, all $ps > 0.05$. IQ was measured by the Chinese version of Wechsler Preschool and Primary Scale of Intelligence-Fourth Edition (WPPSI-IV; Wechsler, 2014; see Table 1 for detailed information). The experiment was approved by the research ethics committee at Peking University. All children gave oral assent and their parents gave written informed consent before the experiment.

Materials

To test the effects of temporal asynchrony on audiovisual speech integration, we adapted McGurk stimuli by setting various SOAs between the auditory stimulus component and visual stimulus one (McGurk & MacDonald, 1976). McGurk stimuli were edited by dubbing the visual part of a phoneme (e.g., /ga/) to the auditory part of a different phoneme (e.g., /ba/), evoking a perception of a third phoneme (e.g., /da/), i.e., the McGurk effect. To generate the stimuli, we videotaped a female speaker uttering /ba/ and /ga/ and edited them using Adobe Premiere Software Pro CS6.0. First, we aligned the auditory part of the two audiovisual files (/ba/ and /ga/). Then, we removed the visual part of /ba/ and the auditory part of /ga/, and obtained the McGurk stimuli (visual /ga/ + auditory /ba/, /AbVg/). After that, we shifted the auditory part of the McGurk stimuli with respect to the visual part of the McGurk

Table 1 Characteristics of children who participated in Experiment 1 and Experiment 2, and children who participated in both experiments

		N	Male/female	Mean age in years (SD)	Mean IQ[a] (SD)
Experiment 1	ASD	24	22/2	6.02 (0.67)	105.13 (10.37)
	TD	26	24/2	5.79 (0.76)	109.77 (14.68)
	ASD vs. TD (*t* value)[b]	–	–	1.13	− 1.30
Experiment 2	ASD	31	28/3	5.86 (0.62)	108.23 (11.22)
	TD	29	27/2	6.02 (0.72)	113.03 (13.50)
	ASD vs. TD (*t* value)[b]	–	–	− 0.91	− 1.50
Experiments 1 and 2	ASD	20	18/2	6.05 (0.63)	106.60 (10.61)
	TD	21	19/2	5.90 (0.69)	113.05 (13.84)
	ASD vs. TD (*t* value)[b]	–	–	0.74	− 1.67

[a]IQ was measured by the Chinese version of Wechsler Preschool and Primary Scale of Intelligence-Fourth Edition (WPPSI-IV)

[b]All $ps > 0.05$

stimuli with varied SOAs. The SOAs included 0 ms, ± 40 ms, ± 120 ms, ± 240 ms, ± 360 ms, ± 480 ms, and ± 680 ms. The negative SOAs denoted the audio led the visual and the positive SOAs denoted the visual led the audio. Therefore, we obtained 13 stimuli of different SOAs, including one synchronized-audiovisual stimulus, six audio-leading stimuli, and six visual-leading stimuli. The resolution of the videos was 1280 × 720 pixels with a frame rate of 25 frames/s. We obtained written consent from the female speaker to use these videos in the experiment as well as for final publications.

Procedures

Participants were seated approximately 60 cm from a 21.5-in. Dell screen (resolution: 1920 × 1080 pixels) in a quiet room. The stimuli were displayed in the center of the screen using Matlab (The MathWorks, Natick, MA, USA) and Psychtoolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). Sounds were presented through two speakers located at the two sides of the screen.

In the experiment, participants completed the McGurk task reporting what the speaker said. They first received a practice session. Only those who showed the McGurk effect enrolled in the formal experiment. In the formal experiment, each of the 13 stimuli with different SOAs was presented eight times. The experiment included a total of 104 trials, presented in a random order. A typical trial began with a fixation at the center of the screen for 1000 ms, which was followed by a black screen for 800 ms. After that, the stimulus was presented. Finally, a response screen was shown to prompt participants to report what the speaker said. The procedure is illustrated in Fig. 1. The experimenter recorded participants' responses by pressing

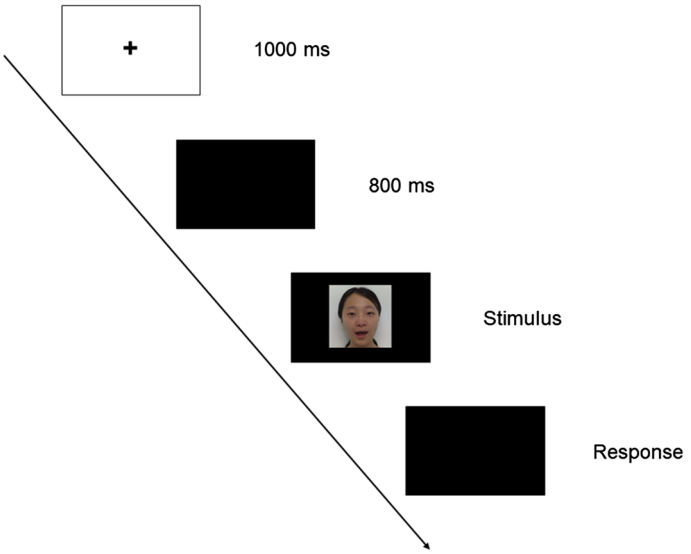

Fig. 1 Procedure of a sample trial. Each trial began with a fixation at the center of the screen for 1000 ms. Then, a black screen was shown for 800 ms. After that, the stimulus was presented. Finally, a response screen was displayed

the corresponding buttons on the keyboard. That is, the experimenter pressed "b", "d", and "g" when children responded /ba/, /da/, and /ga/ respectively. The experiment included 5 blocks and each block had 20–24 trials. To raise participants' interests, we told them this experiment was a game, which included 5 mini-games. They won a mini-game by completing the task in it. Participants were given a reward and took a rest after each mini-game. To further ensure participants fixation on the screen, the experimenter reminded them to look at the screen before each trial. The whole experiment lasted about 30 min.

Data analysis

Participants made three kinds of responses: auditory response /ba/, visual response /ga/, and fused response /da/ (McGurk response) in both the practice session and the formal experiment. First, we coded the data in the practice session according to participants' responses: participants with the McGurk effect as "1" and participants without the McGurk effect as "0". We compared the occurrences of the McGurk effect in the ASD group and the TD group (the initial sample) using the chi-square test. Then, we analyzed the data in the formal experiment by computing participants' percentage of trials with the McGurk effect (the fused response /da/) in each condition (i.e., each SOA as a condition); that is, dividing the number of trials in which participants showed the fused response into the total number of trials in each condition. We examined the group difference and the condition difference of the McGurk effect using a repeated measures ANOVA, simple effect analyses, and post hoc pairwise *t* tests.

Results

In the initial sample, 24 children with ASD and 26 TD children showed the McGurk effect, and the remaining 52 children with ASD and 21 TD children did not show the McGurk effect (see Table 1). To compare the occurrences of the McGurk effect in the ASD group and the TD group in the initial sample, we performed a chi-square test. Results of the chi-square test showed that the two groups were different in the occurrences of the McGurk effect, χ^2 (1) = 6.78, $p = 0.009$. That is, the occurrences of the McGurk effect were lower in the ASD group than in the TD group. Participants without McGurk effect were excluded from the formal experiment.

To examine the effects of group and condition on the percentage of trials with McGurk effect in the final dataset, we conducted a repeated measures ANOVA with Group as the between-subject factor and Condition (13 SOAs) as the within-subject factor. We observed a significant main effect of Condition, $F(12, 576) = 16.67$, $p < 0.001$, $\eta_p^2 = 0.26$, and a significant Group × Condition interaction, $F(12, 576) = 3.02$, $p = 0.01$, $\eta_p^2 = 0.06$, the main effect of Group did not reach significance, $F(1, 48) = 0.62$, $p = 0.44$, $\eta_p^2 = 0.01$. We further analyzed the group difference of the McGurk effect at each SOA using simple effect analysis. Results revealed that, compared with the TD group, the ASD group showed stronger McGurk effect when SOA was − 480 ms, $F(1, 48) = 4.33$, $p = 0.04$, $\eta_p^2 = 0.08$, and marginally stronger McGurk effect when SOA was 680 ms, $F(1, 48) = 4.05$, $p = 0.050$, $\eta_p^2 = 0.078$. The two groups showed similar McGurk effect at all other SOAs, all $ps > 0.05$ (Fig. 2).

We further conducted post hoc pairwise *t* tests (after FDR correction) to compare the condition difference of the percentages of trials with McGurk effect between the synchronized-audiovisual condition (SOA = 0 ms) and other conditions within each group. For the ASD group, compared with the synchronized-audiovisual condition (SOA = 0 ms), the McGurk effect became weaker when the audio led the visual 120 ms or longer, $t(23) \geq 2.54$, $p \leq 0.049$, Cohen's $d \geq 0.30$, and when the visual led the audio 680 ms, $t(23) = 2.42$, $p = 0.047$, Cohen's $d = 0.29$ (all *p*s were corrected by FDR; Fig. 2). For the TD group, compared with the synchronized-audiovisual condition, the McGurk effect became weaker when the audio led the visual 360 ms or longer, $t(25) \geq 4.90$, $p \leq 0.001$, Cohen's $d \geq 1.04$, and when the visual led the audio 480 ms or longer, $t(25) \geq 2.70$, $p \leq 0.03$, Cohen's $d \geq 0.51$ (SOA = 0 ms; all *p*s were corrected by FDR; Fig. 2).

To sum up, taking all current participants into account, the ASD group showed weaker McGurk effect than the TD group. After excluding participants without McGurk effect, two groups showed similar strength of McGurk effect at most SOAs, and the ASD group showed stronger McGurk effect than the TD group when the audio led the visual 480 ms and when the visual led the audio 680 ms. In addition, when the audio led the visual, the strength of the McGurk effect in both groups decreased when SOAs reached certain values (120 ms in the ASD group and 360 ms in the TD group); when the visual led the audio, the strength of the McGurk effect decreased when SOAs reached certain values (680 ms in the ASD group and 480 ms in the TD group).

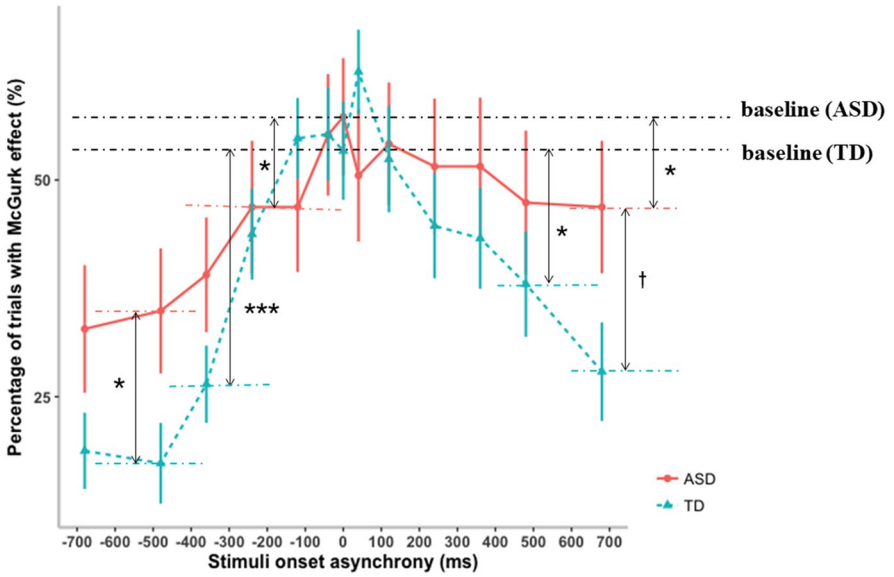

Fig. 2 Percentage of trials with McGurk effect at the 13 SOAs for the ASD group and the TD group. The vertical axis denotes the mean percentage of trials with McGurk effect at each SOA across participants in each group. The horizontal axis (SOA) denotes the asynchrony between auditory and visual components. SOAs were negative when the audio led the visual, and were positive when the visual led the audio. The baseline is the percentage of trials with McGurk effect for the ASD group and the TD group in the synchronized-audiovisual condition (SOA = 0 ms). The red dashed lines are the percentages of trials with McGurk effect when SOAs were − 480 ms, − 120 ms and 680 ms in the ASD group. The green dashed lines are the percentages of trials with McGurk effect when SOAs were − 480 ms, − 360 ms, 480 ms, and 680 ms in the TD group. Error bars represent SEMs. $^{\dagger}p < 0.08$; $^{*}p < 0.05$; $^{***}p < 0.001$

Experiment 2

In this experiment, we examined participants' perceived synchrony for the McGurk stimuli with various SOAs by asking them to perform a simultaneity judgement (SJ) task. Participants judged whether the auditory stimulus and visual stimulus components were synchronous or not.

Method

Participants

This experiment was carried out approximately two weeks after Experiment 1. We also recruited 69 high-functioning children with ASD and 41 TD children from the same school and the same kindergarten as those in Experiment 1. The ASD diagnosis and its confirmation were identical to those in Experiment 1. We excluded 38 children with ASD and 12 TD children who could not understand the meaning of simultaneity, and hence could not complete the task (i.e., judging whether the

auditory stimulus and visual stimulus components were simultaneous or not). Ultimately, 31 high-functioning children with ASD (age range: 4.79–7.23 years, 28 boys and 3 girls) and 29 TD children (age range: 4.76–6.98 years, 27 boys and 2 girls) participated in the experiment. The ASD groups and the TD group were matched on both age and IQ (also measured by WPPSI-IV; see Table 1 for detailed information). There were 20 children with ASD (age range: 4.90–7.23 years, 18 boys and 2 girls) and 21 TD children (age range: 4.76–6.92 years, 19 boys and 2 girls) who participated in both Experiments 1 and 2, and they were also matched on both IQ and age (also see Table 1 for detailed information). The experiment was also approved by the ethics committee at Peking University. Before the experiment, we also obtained all children's oral assent and their parents' written informed consents. Though there might be some practice effect for participants who also participated in Experiment 1, the practice effect should be very weak as Experiment 2 was carried out two weeks later. To examine whether the remaining practice effect was balanced between the two groups, we performed a chi-square test to compare the group difference in percentages of participants who also participated in Experiment 1. Results showed that the two groups were similar in the percentages of participants who participated in Experiment 1, χ^2 (1) = 0.14, $p = 0.70$. Thus, the practice effect from Experiment 1 was balanced between the two groups and would not influence the results.

Materials and procedures

The stimuli and apparatus in this experiment were identical to those in Experiment 1. The procedures were also identical to those in Experiment 1 except for the task. Instead of completing the McGurk task, participants performed an audiovisual SJ task, in which they orally judged whether the audio stimulus and visual stimulus components were simultaneous or not. Participants firstly received a practice session. Only those who could understand what was 'simultaneous' and could compete the SJ task entered the formal experiment. In the formal experiment, each of the 13 stimuli with different SOAs was also presented eight times. Thus, the experiment totally included 104 trials, which were presented in a random order. The experimenter recorded children's oral responses by pressing "y" when children responded "yes" and "n" when children responded "no". Identical to Experiment 1, the experiment also included 5 blocks and each block included 20–24 trials. We took identical measures to ensure participants' attention when doing tasks, including telling them this experiment was also a game, giving them rewards, and reminding them to look at the screen before each trial. The whole experiment lasted about 35 min.

Data analysis

We computed the percentage of *perceived synchrony* as a measure of temporal processing by computing participants' percentage of trials perceived to be synchronous (i.e., "yes" response in simultaneity judgement) in each condition. Using the percentages of perceived synchrony in all conditions, we calculated a temporal binding window (TBW) for each participant (Powers et al., 2009; Stevenson & Wallace, 2013). To calculate the temporal binding window, we firstly fit two sigmoid

functions to the percentages of the perceived synchrony across all conditions of each participant using the *glmfit* function in MATLAB, with one fit to the audio-leading conditions and the other to the visual-leading conditions. Both included the synchronized-audiovisual condition (SOA = 0 ms). The intersection of the two curves divided all data points into two parts. Then, we used the data points on the left of the intersection to refit the left sigmoid, and the data points on the right of the intersection to refit the right sigmoid. Using this method, we fit the data iteratively until the fittings converged (Powers et al., 2009; Stevenson & Wallace, 2013). The intersection of the final two curves was referred to as the participant's point of subjective simultaneity (PSS). After that, we found one point on each of the two curves, for which the *y* value equaled 75% of the PSS. The width between the two points was defined as the TBW of the participant. In addition, to compare the temporal processing difference in the two groups in detail, we also tested the group difference and the condition difference of the percentage of *perceived synchrony* (i.e., percentage of trials perceived to be synchronous) by employing a repeated measures ANOVA, simple effect analyses, and post hoc pairwise *t* tests.

Finally, similar to the analysis in Stevenson et al. (2012), we examined the association between the audiovisual speech perception (i.e., the McGurk effect) and the temporal processing (i.e., the perceived synchrony) when the audio led the visual and when the visual led the audio separately.

Results

Wider temporal binding window in the ASD group than the TD group

We compared the width of the TBW between the two groups. We first calculated the width of the TBW in the ASD group ($M = 616$ ms, $SD = 257$) and TD group ($M = 491$ ms, $SD = 142$). Then, we examined the group difference of the width of the TBW using independent samples *t*-tests. Results showed that the TBW of the ASD group was wider than that of the TD group, $t(58) = 2.30$, $p = 0.03$, Cohen's $d = 0.59$ (Fig. 3). That is, the ASD group was less sensitive to the asynchrony between the auditory stimulus component and visual stimulus counterpart.

Weaker sensitivity for audiovisual asynchrony in the ASD group than the TD group when the visual led the audio

To examine the group and condition difference of the percentages of perceived synchrony (i.e., percentages of trials perceived to be synchronous) in detail, we conducted a repeated measures ANOVA with Group as the between-subject factor and Condition (13 SOAs) as the within-subject factor. Results showed a significant main effect of Condition, $F(12, 696) = 100.12$, $p < 0.001$, $\eta_p^2 = 0.63$, and a significant Group × Condition interaction, $F(12, 696) = 4.23$, $p = 0.002$, $\eta_p^2 = 0.07$, whereas the main effect of Group did not reach significance, $F(1, 58) = 0.32$, $p = 0.57$, $\eta_p^2 = 0.01$ (Fig. 4). We further performed simple effect analyses to examine the group difference of the percentages of perceived synchrony (i.e., percentages of trials perceived

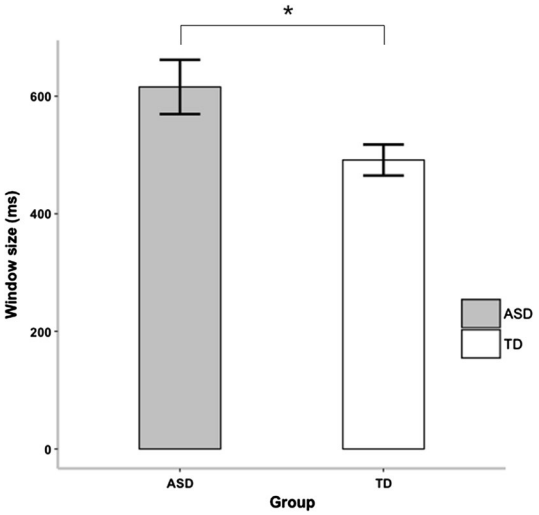

Fig. 3 The window sizes in the ASD group and the TD group when participants performed the simultaneity judgement task. Error bars represent SEMs. *$p < 0.05$

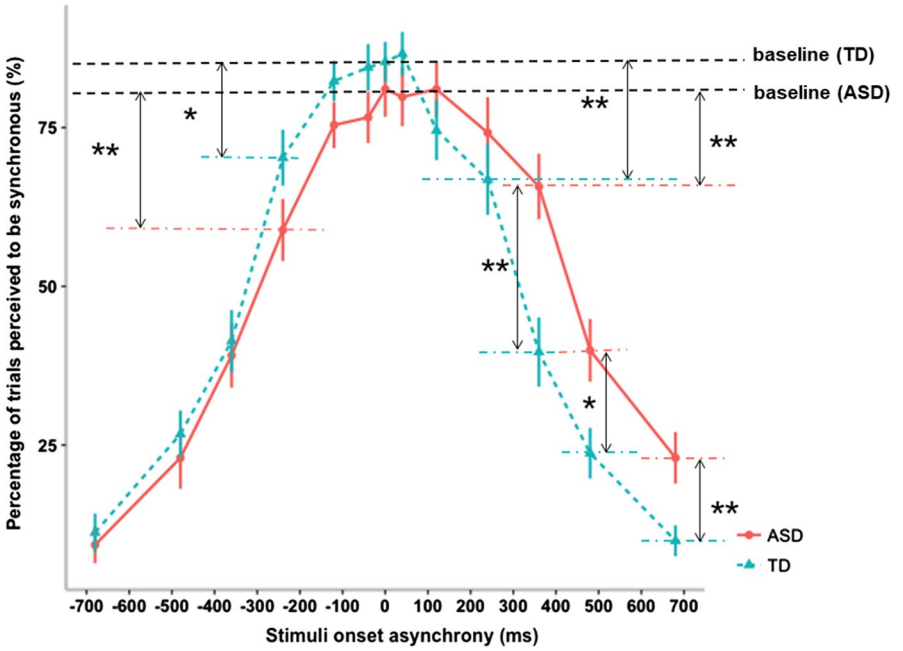

Fig. 4 Percentage of trials perceived to be synchronous in the 13 conditions for the ASD group and the TD group. The vertical axis denotes the mean percentage of trials perceived to be synchronous (perceived synchrony) at each SOA across participants in each group. The horizontal axis (SOA) denotes the asynchrony between audio stimulus and visual stimulus components; SOAs were negative when the audio led the visual, and were positive when the visual led the audio. The baselines are the percentage of trials perceived to be synchronous in the ASD and TD group in the synchronized-audiovisual condition (SOA = 0 ms). The red dashed lines are the percentages of trials perceived to be synchronous when SOAs were − 240 ms, 360 ms, 480 ms, and 680 ms in the ASD group. The green dashed lines are the percentages of trials perceived to be synchronous when SOAs were − 240 ms, 240 ms, 360 ms, 480 ms, and 680 ms in the TD group. Error bars represent SEMs. *$p < 0.05$; **$p < 0.01$

to be synchronous) at each SOA. Results indicated that the ASD group judged the stimuli as simultaneous more frequently than the TD group did, when the visual led the audio 360 ms, $F(1, 58) = 12.11$, $p = 0.001$, $\eta_p^2 = 0.17$, when the visual led the audio 480 ms, $F(1, 58) = 6.43$, $p = 0.01$, $\eta_p^2 = 0.10$, and when the visual led the audio 680 ms, $F(1, 58) = 7.37$, $p = 0.009$, $\eta_p^2 = 0.11$ (Fig. 4). The two groups' percentages of perceived synchrony were similar at all other SOAs, all $ps > 0.05$.

Furthermore, we conducted post hoc pairwise t tests (after FDR correction) to examine the difference of the percentages of perceived synchrony (i.e., percentages of trials perceived to be synchronous) between the synchronized-audiovisual condition (SOA = 0 ms) and other conditions within each group. For the ASD group, compared with the synchronized-audiovisual condition (SOA = 0 ms), the perceived synchrony in the ASD group attenuated when the audio led the visual by 240 ms or longer, $t(30) \geq 3.89$, $p = 0.001$, Cohen's $d \geq 0.86$, and when the visual led the audio by 360 ms or longer, $t(30) \geq 3.61$, $p = 0.002$, Cohen's $d \geq 0.58$ (all ps were corrected by FDR; Fig. 4). Whereas for the TD group, compared with the synchronized-audiovisual condition (SOA = 0 ms), the perceived synchrony decreased when the audio led the visual 240 ms or longer, $t(28) \geq 2.90$, $p = 0.01$, Cohen's $d \geq 0.73$, and when the visual led the audio by 240 ms or longer, $t(28) \geq 3.02$, $p = 0.008$, Cohen's $d \geq 0.76$ (all ps were corrected by FDR; Fig. 4).

In sum, compared with the TD group, the ASD group showed weaker temporal processing (wider TBW), as revealed in the ASD group's diminished sensitivity for the audiovisual asynchrony when the visual led the audio. In specific, the ASD group was less sensitive than the TD group in perceiving the audiovisual asynchrony when the visual led the audio, but was as sensitive as the TD group in perceiving the audiovisual asynchrony when the audio led the visual. In addition, the two groups' percentages of perceived asynchrony decreased when the SOAs reached certain values, i.e., when the audio led the visual by 240 ms for both groups, and when the visual led the audio by 360 ms for the ASD group and 240 ms for the TD group.

Association between the McGurk effect and perceived synchrony

We explored the association between the McGurk effect and the perceived synchrony for audio-leading conditions and visual-leading conditions separately. First, we selected participants who participated in both Experiment 1 and Experiment 2 ($N = 20$ in ASD group, $N = 21$ in TD group; see Table 1). The ASD group and the TD group were also matched on both IQ and age. Then, we classified all conditions into audio-leading conditions (SOA < 0 ms) and visual-leading conditions (SOA ≥ 0 ms). We defined the condition with SOA equals to 0 ms (SOA = 0 ms) as visual-leading conditions because people first see mouth movements and then hear voice in this condition (Chandrasekaran et al., 2009). After that, we computed each participant's mean percentage of trials with McGurk effect for the audio-leading conditions and for the visual-leading conditions separately. We also computed each participant's mean percentage of trials perceived to be synchronous for audio-leading conditions and for visual-leading conditions separately. Finally, we calculated the correlation between the mean percentage of trials with McGurk effect and the

mean percentage of trials perceived to be synchronous for audio-leading conditions and for visual-leading conditions in each group.

Results showed that, for the ASD group, the correlation between the mean percentage of trials with McGurk effect and the mean percentage of trials perceived to be synchronous did not reach significance for either audio-leading conditions, $r = -0.29$, $p = 0.22$, or visual-leading conditions, $r = 0.25$, $p = 0.29$ (Fig. 5a, c). For the TD group, the correlation was not significant in audio-leading conditions, $r = -0.13$, $p = 0.58$, but it was significant in visual-leading conditions, $r = 0.71$, $p < 0.001$ (Fig. 5b, d). That is, the McGurk effect was enhanced as the perceived synchrony increased in the TD group when the visual led the audio but not when the audio led the visual. No such correlation was observed in the ASD group either when the audio led the visual or when the visual led the audio.

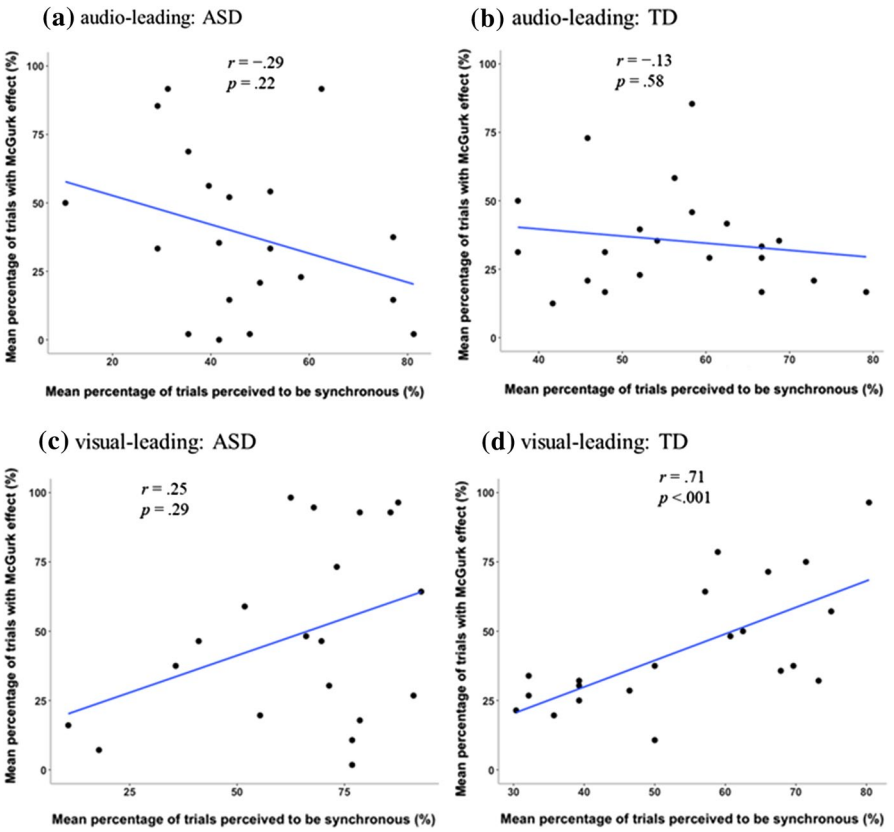

Fig. 5 The correlation between the mean percentage of trials with McGurk effect and the mean percentage of trials perceived as synchronous for the audio-leading conditions and the visual-leading conditions separately in the ASD group (**a**, **c**) and the TD group (**b**, **d**). The horizontal axes were the mean percentage of trials perceived as synchronous (perceived synchrony) across audio-leading conditions (**a**, **b**) or visual-leading conditions (**c**, **d**). The vertical axes were the mean percentage of trials with McGurk effect across audio-leading conditions (**a**, **b**) or visual-leading conditions (**c**, **d**)

General discussion

In the present study, we examined the audiovisual speech perception, the temporal processing, and the correlation between them in children with ASD and TD children. In Experiment 1, we found that, taking all present participants into consideration, children with ASD showed weaker audiovisual speech perception (i.e., weaker McGurk effect) than TD children did. In Experiment 2, we observed that the ASD group showed weaker temporal processing (wider TBW) than the TD group, which was mainly ascribed to their less acute sensitivity for the audiovisual asynchrony in the visual-leading conditions. Moreover, we found that audiovisual speech perception and perceived synchrony were positively correlated with each other in TD children when the visual led the audio but not when the audio led the visual. We found no correlation between them in children with ASD either when the audio led the visual or when the visual led the audio.

Audiovisual speech perception deficits in ASD

Taking all present participants into account, we found that children with ASD exhibited audiovisual speech perception deficits, which confirms our hypothesis and most previous findings of weaker audiovisual perception in ASD (Bebko et al., 2014; de Gelder et al., 1991; Irwin et al., 2011; Mongillo et al., 2008; Stevenson et al., 2014). The audiovisual speech perception deficits (weaker McGurk effect) indicate that children with ASD are impaired in integrating the auditory components and the visual one into a coherent whole stimulus (i.e., audiovisual integration). Such impairments in audiovisual integration supported weak central coherence theory, which holds that children with ASD show a relatively weakened tendency to combine the stimulus components into a holistic whole (Baum et al., 2015; Happé & Frith, 2006). Children with ASD's audiovisual speech perception deficits provide insights for the language and social communication impairments in children with ASD. The audiovisual speech perception deficits in children with ASD constrain their daily language input, which may consequently lead to their language and social interaction impairments (Narafshan et al., 2014). Future studies should investigate the underlying mechanisms of the language and social interaction impairments in children with ASD from the perspective of their audiovisual speech perception deficits.

Chinese-speaking TD children in our experiment showed about 57% McGurk effect when the audio and the visual were synchronized (i.e., SOA = 0 ms; the stimuli employed in most previous studies). This indicated that Chinese-speaking children showed relatively strong McGurk effect compared with previous studies that took English-speaking children as participants. For example, in the initial study that explored McGurk effect, English-speaking TD children at similar ages showed about 64% McGurk effect (McGurk & MacDonald, 1976). Our finding was not consistent with the previous finding that culture and native language backgrounds affect audiovisual speech perception (Chen & Massaro, 2004; Massaro et al., 1995; Sekiyama & Tohkura, 1993). This inconsistency might originate from the fact that the face-avoidance habit in younger Chinese generation is less observed. With that said,

the present study has some limitations. One of them is that we did not compare the McGurk effect in Chinese-speaking children with that in English-speaking children directly in our experiments. For future work, we can further explore whether culture and native language backgrounds affect the McGurk effect in Chinese-speaking children.

It is also noteworthy that, after excluding children who did not show the McGurk effect (52 children with ASD and 21 TD children), the ASD group showed stronger audiovisual speech perception than the TD group for certain relatively long SOAs (SOA = −480 ms, SOA = 680 ms). This finding confirmed previous studies (Woynaroski et al., 2013), and suggested that the audiovisual speech perception in the ASD group was less influenced by relatively long SOAs than the TD group. That is, the ASD group was less sensitive to relatively long asynchrony between the audio and the visual than the TD group in audiovisual speech perception. This finding implied that children with ASD were less acute in perceiving the lags between the audio and the visual (i.e., less acute in temporal processing) compared with TD children.

Temporal processing deficits in ASD

The finding that the ASD group showed wider TBW (less acute temporal processing) compared with the TD group was consistent with our hypothesis. We further found that the ASD group's temporal processing (detection of the audiovisual asynchrony) was less acute than the TD group for visual-leading stimuli, but was as acute as the TD group for audio-leading stimuli. The less acute temporal processing for visual-leading stimuli in the ASD group might be explained by their visual attention disengagement deficits (Sacrey et al., 2014). Visual attention disengagement deficits in children with ASD make them spend a longer time than TD children to disengage their attention from the fixation point to the stimulus in interest (i.e., the visual mouth movements) in visual-leading conditions. The ASD group's postponed attention to the speaker's mouth movements makes them perceive the SOAs within visual-leading stimuli to be shorter, which posed more difficulties for the children with ASD to judge the simultaneity between the auditory and the visual components. Thus, the ASD group showed less acute temporal processing compared with the TD group for visual-leading stimuli. The intact temporal processing for audio-leading stimuli in the ASD group might be explained by the increased auditory capacity in autism (Remington & Fairnie, 2007). In the audio-leading stimuli, sounds precede mouth movements, the increased auditory capacity in the ASD group possibly facilitates their processing for the stimuli and improves their asynchrony perception for the given stimuli. Therefore, temporal processing for audio-leading stimuli in the ASD group was relatively intact and comparable to that in the TD group.

Some previous studies also found that individuals with ASD showed weaker temporal processing compared with their TD counterparts (de Boer-Schellekens et al., 2013; Stevenson et al., 2014). However, to our best knowledge, none of the previous studies analyzed the group difference of the temporal processing at each SOA or separately analyzed the group difference in audio-leading conditions and visual-leading conditions. In this case, the present study extended the existing evidence on

temporal processing in ASD by revealing that the ASD group was only impaired when the visual led the audio, but kept intact when the audio led the visual. The temporal processing deficits for visual-leading speech stimuli in children with ASD might have cascading influence on their language development. As also revealed in previous studies, temporal processing for speech stimuli in early life was correlated with later language abilities. For example, detection of temporal changes at 17 months predicted language comprehension at 4 years and word reading frequency at second grade (van Zuijen et al., 2012).

Relationship between audiovisual speech perception and temporal processing

We found that audiovisual speech perception (McGurk effect) was positively correlated with temporal processing (perceived synchrony) when the visual led the audio in the TD group. This is possibly because the more a stimulus was perceived as synchronous, the more the auditory and visual components in the stimulus were integrated into a whole, and the stronger the McGurk effect. This positive correlation between McGurk effect and temporal processing was found only when the visual led the audio but not when the audio led the visual in the TD group. This might be because that the visual-leading conditions were similar to the natural situation, in which the mouth moved before the sound was heard, and that TD children have tuned to the natural situation (Vroomen & Keetels, 2010). Contrary to presumed reasoning, we failed to find that the audiovisual speech perception (McGurk effect) was linked with temporal processing (perceived synchrony) in the ASD group either when the visual led the audio or when the audio led the visual. Our data indicated that the correlation between audiovisual speech perception and temporal processing for visual-leading conditions may be contingent on the range of the temporal processing abilities (within the normal range), such as seen in the TD group.

In previous studies, consistent with our finding in TD children, Stevenson et al. (2012) found that audiovisual speech perception (McGurk effect) was correlated with temporal processing for flashes and beeps in the visual-leading conditions but not in the audio-leading conditions in TD adults. However, contrary to our findings in children with ASD, Stevenson et al. (2014) found that the strength of the McGurk effect was correlated with the acuity of the temporal processing in adolescents with ASD. This discrepancy between Stevenson et al. (2014) and our own could be accounted for by different ages of the samples (averaged 12-year-old adolescents in Stevenson et al., 2014 vs. 6-year-old children in the present study). Compared with adolescents with ASD, children with ASD have more severe autistic symptoms, which might affect the development of their audiovisual speech perception and temporal processing more seriously. The audiovisual speech perception and temporal processing in children with ASD would have more severe deficits and might develop separately. Thus, no correlation would be found between them in the present study. However, adolescents with ASD received much more intervention and have less severe autistic symptoms, which might affect the development of audiovisual speech perception and temporal processing to a less serious extent. Their temporal processing might affect

their audiovisual speech perception, and correlation between them became significant in Stevenson et al. (2014). This lack of correlation between audiovisual speech perception and temporal processing in ASD could also be attributed to different experimental materials. In measuring temporal processing, Stevenson et al. (2014) used audiovisual matched stimuli, but the present study used the audiovisual mismatched stimuli (i.e., the McGurk stimuli). Individuals have more acute temporal processing for audiovisual mismatched stimuli than for audiovisual matched stimuli (van Wassenhove et al., 2007). In future studies, we could further explore the relationship between audiovisual speech perception and temporal processing by recruiting children with ASD in different ages or by employing different experimental materials. Our findings of insignificant correlation between audiovisual speech perception and temporal correlation in children with ASD implicates that the audiovisual speech perception deficits in children with ASD might not be solely explained by their temporal processing deficits for speech stimuli. In future studies, we could explore the mechanisms underlying the audiovisual speech perception deficits in children with ASD from other aspects, for example, the lip-reading deficits in children with ASD (Borowiak et al., 2018; Smith & Bennetto, 2007).

The present study also has some limitations. For example, stimuli across trials were identical except for the SOAs between the auditory and the visual parts. These similar stimuli might potentially give rise to different repetition effects in children with ASD and TD children, which might influence the current results. In future studies, we could employ different stimuli across trials to control the repetition effects and further explore the relationship between audiovisual speech integration and temporal processing in children with and without ASD.

In summary, the present study revealed the audiovisual speech perception impairments (based on the population) and the temporal processing deficits in children with ASD. The present study further discovered that audiovisual speech perception was associated with temporal processing when the visual led the audio in TD children, but no correlation was found in children with ASD. Our finding in TD children implicated the possibility of promoting TD children's audiovisual speech perception, and hence their language development, through temporal processing training (Fujisaki et al., 2004). Our findings in TD children and children with ASD indicated that the correlation between audiovisual speech perception and temporal processing may be contingent on the range of temporal processing abilities. That is, audiovisual speech perception might be correlated with temporal processing only when temporal processing abilities are within a normal range, such as seen in the TD group. In future studies, we could first train temporal processing in children with ASD into a normal range, and then explore the correlation between their audiovisual speech perception and temporal processing (Noel et al., 2017).

Acknowledgements The authors are grateful to Zipeng Ma, Fuli Liu, Yanhong Wu, Yixiao Hu, Yinan Lv, and the staff in Qingdao Elim School, for their generous assistance in completing the study. This work was supported by the Philosophy and Social Science Foundation of Hunan Province, Grant No. 19YBQ109.

References

Altieri, N., Townsend, J. T., & Wenger, M. J. (2014). A measure for assessing the effects of audiovisual speech integration. *Behavior Research Methods, 46*, 406–415. https://doi.org/10.3758/s13428-013-0372-8

Altvater-Mackensen, N., & Grossmann, T. (2015). Learning to match auditory and visual speech cues: Social influences on acquisition of phonological categories. *Child Development, 86*(2), 362–378. https://doi.org/10.1111/cdev.12320

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM–5*. APA Press.

Auyeung, B., Baron-Cohen, S., Wheelwright, S., & Allison, C. (2008). The autism spectrum quotient: Children's version (AQ-Child). *Journal of Autism and Developmental Disorders, 38*(7), 1230–1240. https://doi.org/10.1007/s10803-007-0504-z

Baum, S. H., Stevenson, R. A., & Wallace, M. T. (2015). Behavioral, perceptual, and neural alterations in sensory and multisensory function in autism spectrum disorder. *Progress in Neurobiology, 134*, 140–160. https://doi.org/10.1016/j.pneurobio.2015.09.007

Bebko, J. M., Weiss, J. A., Demark, J. L., & Gomez, P. (2006). Discrimination of temporal synchrony in intermodal events by children with autism and children with developmental disabilities without autism. *Journal of Child Psychology and Psychiatry, 47*(1), 88–98. https://doi.org/10.1111/j.1469-7610.2005.01443.x

Bebko, J. M., Schroeder, J. H., & Weiss, J. A. (2014). The McGurk effect in children with autism and Asperger syndrome. *Autism Research, 7*(1), 50–59. https://doi.org/10.1002/aur.1343

Borowiak, K., Schelinski, S., & Von Kriegstein, K. (2018). Recognizing visual speech: Reduced responses in visual-movement regions, but not other speech regions in autism. *NeuroImage: Clinical*, *20*, 1078–1091. https://doi.org/10.1016/j.nicl.2018.09.019

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Burnham, D., & Lau, S. (1998). *The effect of tonal information on auditory reliance in the McGurk effect* [Paper presentation]. Auditory-Visual Speech Processing, Australia, Sydney

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5*(7), e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Chen, T. H., & Massaro, D. W. (2004). Mandarin speech perception by ear and eye follows a universal principle. *Perception and Psychophysics, 66*(5), 820–836.

Colonnesi, C., Stams, G. J., Koster, I., & Noom, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review, 30*(4), 352–366.

de Boer-Schellekens, L., Eussen, M. L., & Vroomen, J. (2013). Diminished sensitivity of audiovisual temporal order in autism spectrum disorder. *Frontiers in Integrative Neuroscience, 7*, 1–8. https://doi.org/10.3389/fnint.2013.00008

de Gelder, B., Vroomen, J., & Van der Heide, L. (1991). Face recognition and lip-reading in autism. *European Journal of Cognitive Psychology, 3*(1), 69–86. https://doi.org/10.1080/09541449108406220

Farrant, B. M., & Zubrick, S. R. (2012). Early vocabulary development: The importance of joint attention and parent–child book reading. *Language, 32*(3), 343–364.

Foxe, J. J., Molholm, S., Del Bene, V. A., Frey, H., Russo, N. N., Blanco, D., et al. (2015). Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. *Cerebral Cortex*, *25*, 298–312. https://doi.org/10.1093/cercor/bht213

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7*(7), 773–778. https://doi.org/10.1038/nn1268

Gervain, J., & Mehler, J. (2010). Speech perception and language acquisition in the first year of life. *Annual Review of Psychology, 61*(1), 191–218. https://doi.org/10.1146/annurev.psych.093008.100408

Grossman, R. B., Steinhart, E., Mitchell, T., & McIlvane, W. (2015). "Look who's talking!" gaze patterns for implicit and explicit audio-visual speech synchrony detection in children with high-functioning autism. *Autism Research, 8*(3), 307–316. https://doi.org/10.1002/aur.1447

Happé, F., & Frith, U. (2006). The weak coherence account: Detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 36*(1), 5–25. https://doi.org/10.1007/s10803-005-0039-0

Hayiou-Thomas, M. E. (2008). Genetic and environmental influences on early speech, language and literacy development. *Journal of Communication Disorders, 41*(5), 397–408. https://doi.org/10.1016/j.jcomdis.2008.03.002

Iarocci, G., Rombough, A., Yager, J., Weeks, D. J., & Chua, R. (2010). Visual influences on speech perception in children with autism. *Autism, 14*(4), 305–320. https://doi.org/10.1177/1362361309353615

Imafuku, M., Kawai, M., Niwa, F., Shinya, Y., & Myowa, M. (2019). Audiovisual speech perception and language acquisition in preterm infants: A longitudinal study. *Early Human Development, 128*, 93–100. https://doi.org/10.1016/j.earlhumdev.2018.11.001

Irwin, J. R., Tornatore, L. A., Brancazio, L., & Whalen, D. H. (2011). Can children with autism spectrum disorders "hear" a speaking face? *Child Development, 82*(5), 1397–1403. https://doi.org/10.1111/j.1467-8624.2011.01619.x

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36*, 1.

MacDonald, J. (2017). Hearing lips and seeing voices: The origins and development of the 'McGurk effect' and reflections on audio–visual speech perception over the last 40 years. *Multisensory Research*. https://doi.org/10.1163/22134808-00002548

Magnotti, J. F., Mallick, D. B., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental Brain Research*, *233*, 2581–2586. https://doi.org/10.1007/s00221-015-4324-7.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1995). Cross-lingustic comparisons in the integration of visual and auditory speech. *Memory and Cognition, 23*, 113–131.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748. https://doi.org/10.1023/A:1005592401947

Mongillo, E. A., Irwin, J. R., Whalen, D. H., Klaiman, C., Carter, A. S., & Schultz, R. T. (2008). Audiovisual processing in children with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders, 38*(7), 1349–1358. https://doi.org/10.1007/s10803-007-0521-y

Muluk, N. B., Bayoğlu, B., & Anlar, B. (2014). Language development and affecting factors in 3-to 6-year-old children. *European Archives of Oto-Rhino-Laryngology, 271*(5), 871–878. https://doi.org/10.1007/s00405-013-2567-0

Narafshan, M. H., Sadighi, F., Bagheri, M. S., & Shokrpour, N. (2014). The role of input in first language acquisition. *International Journal of Applied Linguistics and English Literature, 3*(1), 86–91.

Noel, J., De Niear, M. A., Stevenson, R. A., Alais, D., & Wallace, M. T. (2017). Atypical rapid audio-visual temporal recalibration in autism spectrum disorders. *Autism Research, 10*(1), 121–129. https://doi.org/10.1002/aur.1633

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. https://doi.org/10.1163/156856897x00366

Powers, A. R., Hillock, A. R., & Wallace, M. T. (2009). Perceptual training narrows the temporal window of multisensory binding. *The Journal of Neuroscience, 29*(39), 12265–12274. https://doi.org/10.1523/JNEUROSCI.3501-09.2009

Remington, A., & Fairnie, J. (2007). A sound advantage: Increased auditory capacity in autism. *Cognition, 166*, 459–465. https://doi.org/10.1016/j.cognition.2017.04.002

Sacrey, L. R., Armstrong, V. L., Bryson, S. E., & Zwaigenbaum, L. (2014). Impairments to visual disengagement in autism spectrum disorder: A review of experimental studies from infancy to adulthood. *Neuroscience and Biobehavioral Reviews, 47*, 559–577. https://doi.org/10.1016/j.neubiorev.2014.10.011

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics, 21*, 427–444. https://doi.org/10.1016/S0095-4470(19)30229-3

Smith, E., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology and Psychiatry, 48*(8), 813–821. https://doi.org/10.1111/j.1469-7610.2007.01766.x

Stevenson, R. A., & Wallace, M. T. (2013). Multisensory temporal integration: Task and stimulus dependencies. *Experimental Brain Research, 227*(2), 249–261. https://doi.org/10.1007/s00221-013-3507-3

Stevenson, R. A., Zemtsov, R. K., & Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology: Human Perception and Performance, 38*(6), 1517–1529. https://doi.org/10.1037/a0027339

Stevenson, R. A., Siemann, J. K., Schneider, B. C., Eberly, H. E., Woynaroski, T. G., Camarata, S. M., & Wallace, M. T. (2014). Multisensory temporal integration in autism spectrum disorders. *Journal of Neuroscience, 34*(3), 691–697. https://doi.org/10.1523/JNEUROSCI.3615-13.2014

Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., Camarata, S., & Wallace, M. T. (2015). Keeping time in the brain: Autism spectrum disorder and audiovisual temporal processing. *Autism Research, 9*(7), 720–738. https://doi.org/10.1002/aur.1566

Stevenson, R. A., Baum, S. H., Krueger, J., Newhouse, P. A., & Wallace, M. T. (2018). Links between temporal acuity and multisensory integration across life span. *Journal of Experimental Psychology: Human Perception and Performance, 44*(1), 106–116. https://doi.org/10.1037/xhp0000424

Taylor, N., Isaac, C., & Milne, E. (2010). A comparison of the development of audiovisual integration in children with autism spectrum disorders and typically developing children. *Journal of Autism and Developmental Disorders*, *40*, 1403–1411. https://doi.org/10.1007/s10803-010-1000-4.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*(3), 598–607. https://doi.org/10.1016/j.neuropsychologia.2006.01.001

van Zuijen, T. L., Plakas, A., Maassen, B., Been, P., Maurits, N. M., Krikhaar, E., van Driel, J., & Der Leij, A. V. (2012). Temporal auditory processing at 17 months of age is associated with preliterate language comprehension and later word reading fluency: An ERP study. *Neuroscience Letters, 528*(1), 31–35. https://doi.org/10.1016/j.neulet.2012.08.058

Volden, J., & Phillips, L. (2010). Measuring pragmatic language in speakers with autism spectrum disorders: Comparing the children's communication checklist-2 and the test of pragmatic language. *American Journal of Speech-Language Pathology, 19*(3), 204–212. https://doi.org/10.1044/1058-0360(2010/09-0011)

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, and Psychophysics, 72*(4), 871–884. https://doi.org/10.3758/APP.72.4.871

Wechsler, D. (2014). *Wechsler Preschool and Primary Scale of Intelligence—Fourth CN Edition (WPPSI-IV CN)*. (Y. Li & Z. J, Trans. Y. Li & Z. J Eds.). King-May Company China

Weismer, S. E., Lord, C., & Esler, A. (2010). Early language patterns of toddlers on the autism spectrum compared to toddlers with developmental delay. *Journal of Autism and Developmental Disorders, 40*(10), 1259–1273. https://doi.org/10.1007/s10803-010-0983-1

Werker, J. F., & Gervain, J. (2013). Speech perception in infancy: A foundation for language acquisition. In Zelazo, P. D. (Eds.), *The Oxford Handbook of Developmental Psychology* (pp. 909–925). Oxford Handbooks Online. https://doi.org/10.1093/oxfordhb/9780199958450.013.0031

Woynaroski, T. G., Kwakye, L. D., Foss-Feig, J. H., Stevenson, R. A., Stone, W. L., & Wallace, M. T. (2013). Multisensory speech perception in children with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 43*(12), 2891–2902. https://doi.org/10.1007/s10803-013-1836-5

Zhang, J., Meng, Y., He, J., Xiang, Y., Wu, C., Wang, S., & Yuan, Z. (2019). McGurk effect by individuals with autism spectrum disorder and typically developing controls: A systematic review and meta-analysis. *Journal of Autism and Developmental Disorders, 49*(1), 34–43. https://doi.org/10.1007/s10803-018-3680-0

Authors and Affiliations

Shuyuan Feng[1,2] · Haoyang Lu[3,4] · Jing Fang[5] · Xue Li[6] · Li Yi[2,7] · Lihan Chen[2]

1 Institute for Applied Linguistics, School of Foreign Languages, Central South University, Changsha, Hunan, China

2 School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, 5 Yiheyuan Road, Beijing 100871, China

3 Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China

4 Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China

5 Qingdao Autism Research Institute, Qingdao, China

6 Peking University Sixth Hospital, Beijing, China

7 IDG/McGovern Institute for Brain Research at PKU, Peking University, Beijing, China