The developmental shift in audiovisual speech perception is universal: evidence from Mandarin-speaking children

Authors: Yi Weng, Yicheng Rong, and Gang Peng*

Affiliation:

Research Centre for Language, Cognition, and Neuroscience, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China

***Correspondence**: Gang Peng, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Kowloon, Hong Kong SAR, China.

Tel: +852 3400 8462

Fax: +852 2334 0185

Email: gpengjack@gmail.com

ORCID

Yi Weng https://orcid.org/0000-0002-0943-7132

Yicheng Rong https://orcid.org/0000-0003-4998-3856

Gang Peng https://orcid.org/0000-0002-1465-1301

The development of audiovisual speech perception in Mandarin-speaking children: Evidence from the McGurk paradigm

Abstract

The developmental trajectory of audiovisual speech perception in Mandarin-speaking children remains understudied. This cross-sectional study in Mandarin-speaking 3–4-,5–6-,7–8-year-old children and adults (*n* = 87, 44 males) investigated this issue using the McGurk paradigm with three levels of auditory noise. For the identification of congruent stimuli, 3–4-year-olds underperformed older groups whose performances were comparable. For the perception of the incongruent stimuli, a developmental shift was observed as 3–4-year-olds made significantly more audio-dominant but fewer audiovisual-integrated responses to incongruent stimuli than older groups. With increasing auditory noise, the difference between children and adults widened in identifying congruent stimuli but narrowed in perceiving incongruent ones. The findings regarding noise effects agree with the statistically optimal hypothesis.

Keywords: audiovisual integration, development, Mandarin-speaking children

1. Introduction

Our coherent perception of the outside world is derived from the cooperation and interaction of different sensory modalities, instead of a collection of senses (Rosenblum & Dorsi, 2021). The process of multisensory integration is vital for successful communication and social interaction, but it could be very subtle and not even easily noticeable for perceivers (Alsius et al., 2017; Tye-Murray et al., 2016). For instance, speech perception might be intuitively considered a unimodal process for which isolated auditory modality is responsible (e.g., Denes & Pinson, 1963, p.8). However, vision has been revealed to play an indispensable part in generating perceptual outcomes of speech, particularly in face-to-face communication (Ménard et al., 2014; Sato et al., 2013). As multisensory processing is a "late bloomer", young children have been found to be less likely to take the benefit of audiovisual integration in speech perception than adults (Ernst, 2008). Before developing an adult-like pattern, children seem to rely more on unimodal strategy when processing audiovisual information. With age, they gradually acquire comparable competence in integrating bimodal information, giving rise to a developmental shift, which has been frequently observed in Indo-European-language-speaking children (Burr & Gori, 2012; Dupont et al., 2005; Ernst, 2008; Hirst et al., 2018; Robinson & Sloutsky, 2004, 2010; Tremblay et al., 2007). On the other hand, the emergence of developmental shift appears to be subject to language background since speakers of certain languages such as Japanese and Mandarin do not seem to experience this developmental process (Li et al., 2008; Liu et al., 2020; Sekiyama & Burnham, 2008). However, in the case of Mandarin, it seems premature to solely attribute the discrepancy in developmental trajectories to a language-specific account since Mandarin-speaking adults exhibit a similar degree of the McGurk Illusion as Indo-European language speakers (Chen & Hazan, 2009; Magnotti et al., 2015). The possible reason why the developmental shift was not observed in studies examining Mandarin-speaking individuals might be that the shift happens early while the previous research was limited to school-age children. Therefore, the current study aims to examine whether the developmental shift is limited to Indo-European language speakers by including preschoolers speaking Mandarin.

1.1 The McGurk illusion

Audiovisual processing in speech perception has been studied using a variety of paradigms including detecting masked speech (e.g., Eramudugolla et al., 2011), measuring audiovisual binding window (e.g., Lewkowicz & Flom, 2014) and evoking the McGurk illusion (e.g., Stacey et al., 2021). In the classic McGurk design, participants tended to perceive an illusory /dada/ after watching a video that dubbed the auditory /baba/ onto visually articulated /gaga/ (Macdonald & McGurk, 1978; McGurk & Macdonald, 1976). The emergence of the McGurk illusion stems from the conflict between auditory information and mouth movement, as one might expect a visual lip closure to match an auditory /ba/, or a sound with a backer place of articulation corresponding to a visual /ga/. Accordingly, the illusory /da/, which is the third perceptual choice lacking both auditory and visual substance, is derived as a more precise estimate by taking both audition and vision into account. According to responses to stimuli with conflicting audiovisual information, we can observe the perceptual strategies employed by the participants along development: an audio-dominant strategy indexed by /ba/ response, an audiovisual-integrated strategy reflected by /da/ response or a visual-dominant strategy represented by /ga/ response. Taking responses to both congruent and incongruent stimuli together, the McGurk paradigm opens a window for us to look into the development of audiovisual speech processing in children, allowing us to gain valuable insights into the strategies employed by children at different ages in processing audiovisual stimuli.

1.2 The developmental trajectory of experiencing the McGurk illusion

Since the original report of the McGurk effect, the developmental change in the "fused" percepts (/dada/) has been noticed (McGurk & Macdonald, 1976). Though a considerable amount of McGurk illusion was recorded from children (3–5-year-olds and 7–8-year-olds) by McGurk and MacDonald (1976), there exhibited discrepancy between children and adults as more auditory-based responses and fewer audiovisual-integrated responses were made by the child groups. The authors, therefore, clearly pointed out that the susceptibility to the McGurk illusions might fluctuate with age.

McGurk and MacDonald's argument has been ascertained by recent developmental studies as the adult-like manner in multisensory processing seems not to be inherent in children (Burr & Gori, 2012). Though children as young as 5 months old were found to show sensitivity to the McGurk effect (Burnham & Dodd, 2004), cross-sectional studies pointed out that the adult-like audiovisual integrative strategy took around 10 years to get fully developed (Hirst et al., 2018; Tremblay et al., 2007). Tremblay et al. (2007) performed both speech (the McGurk effect) and non-speech (the Illusory Flash effect and the Fusion effect) audiovisual illusion paradigms on three groups of participants: 5–9-year-olds, 10–14-year-olds and 15–19-year-olds. Results revealed an asymmetric developmental process between these two types of audiovisual integration, with a comparable strength of non-speech audiovisual illusion across all groups but significant group differences in terms of the McGurk illusion. Given that significantly more /ba/ but fewer /da/ responses were found from 5–9-year-olds relative to the two older groups, the authors proposed that children possibly would undergo a developmental shift from placing more weight upon auditory modality to considering information from both auditory and visual modalities in speech perception. Consistently, with the comparisons among 3–6-year-olds, 7–9-year-olds, 10–12-year-olds and adults (aged 20–35), Hirst et al. (2018) also obtained more audio-dominant accompanying fewer audiovisual-integrated responses among two younger groups relative to adults, while 10–12-year-olds did not significantly differ from adults. Accordingly, a developmental shift from attending more to unimodal auditory information to taking both bimodal audiovisual information into account could be concluded, and this process is likely to get fully mature during middle childhood.

Nevertheless, such a developmental shift is claimed to be limited to Indo-European language speakers since it seems not to be extended to subjects from certain language backgrounds. For instance, Sekiyama and Burnham (2008) compared the developmental trajectories of audiovisual speech perception between Japanese and English monolingual adults and children aged 6 years, 8 years and 11 years. Surprisingly, the authors observed age-related effects for English speakers with auditorily correct responses decreasing between 6 and 8 years, whereas such age effects were absent among Japanese. Combining the findings of weak McGurk effect among Japanese adults relative to their English peers, the authors proposed that the absence of developmental shift in Japanese could be attributed to face-looking patterns specific to their culture, that is directly gazing at speakers' faces is not polite in the East Asian context.

Similarly, several studies on speakers of another East Asian language Mandarin measured a comparable strength of McGurk effect between school-age children and adults, proposing that it is due to cultural variation that Mandarin speakers do not necessarily experience the development shift resembling English speakers (Li et al., 2008; Liu et al., 2020). Specifically, Li et al. (2008) performed the McGurk paradigm on Mandarin-speaking grade-two pupils (Mean age = 7.66), grade-five pupils (Mean age = 10.70) as well as first-year university students (Mean age = 19.15). The McGurk effect was successfully evoked while results showed no significant differences among three age groups. Therefore, the authors claimed that Mandarin speakers did not exhibit a similar developmental trend to English speakers where the visual influence grows stronger along with increasing age in processing audiovisual information. Instead, the authors agreed with Sekiyama et al. (2003) that the developmental shift was absent due to linguistic and cultural aspects. In parallel with Li et al. (2008), a more recent study by Liu et al. (2020) also failed to observe age effects on the McGurk illusion recorded from Mandarin-speaking 6–12-year-olds and 13–16-year-olds across six different conditions varying in speech-to-noise ratios (SNRs). To sum up, these studies support that developmental shift in sensory dominance appears to be a language-specific phenomenon that is not necessary for Mandarin-speaker children to experience.

However, to investigate whether cultural and linguistic consequences indeed played a role, Chen and Hazan (2009) directly made comparisons between Mandarin and English speakers in audiovisual speech processing. Participants consisted of 8–9-year-olds and adults native in Mandarin and English respectively. The authors failed to observe significant differences between the participants of the two languages in visual utilization no matter in child or adult groups, which was evidence opposing the view that language-related factors resulted in distinct developmental trajectories in Mandarin-speaking children compared to their English-speaking counterparts. Moreover, a significant age effect was revealed in the noisy condition as an enhanced visual effect was found among Mandarin-speaking adults relative to children in this condition. The observation of age effect implies that Mandarin-speaking children would undergo a developmental shift, but only in noisy condition. What contributes to no developmental shift observed in the quiet condition? The cultural account claimed by Li et al. (2008) does not seem to work considering the limited impact of culture and language revealed by the direct comparison between Mandarin and English speakers in Chen and Hazan (2009). Thus, the question of whether there exist other factors that contribute to the absence of the developmental shift is raised. One possibility is that the shift happens early in speakers of Mandarin, and children tested in previous studies had developed to an adult-like pattern because they already exceeded the age range in which it typically occurs. If this hypothesis holds, it is not surprising that the studies that focused on school-age children failed to detect the shift among Mandarin-speaking children. One possible reason for the expectation of an earlier shift is that the stimuli used in the McGurk paradigm are monosyllabic words in Mandarin while they are meaningless in English. Thus, Mandarin-speaking children are more regularly exposed to these stimuli in their daily lives. Since the ability to integrate audio and visual information during speech perception relies on the capacity to perceive visual speech (Bernstein, 2012; Massaro et al., 1986), which is thought to be linked to linguistic experience (Tye-Murray et al., 2007), increased exposure to corresponding linguistic stimuli may lead to an earlier mastery of the ability to incorporate visual cues into perceptual outcomes.

Taken together, previous developmental studies support that children in some language and cultural contexts might undergo a developmental shift from unimodal auditory dominance to taking both audiovisual cues into account in speech perception (Dupont et al., 2005; Hirst et al., 2018; Thompson & Massaro, 1994; Tremblay et al., 2007). According to existing findings from Indo-European-language-speaking children, this process would take around 10 years to grow into maturity (Hirst et al., 2018; Tremblay et al., 2007). However, no such shift was detected among 10-year-old Mandarin-speaking children (Li et al., 2008; Liu et al., 2020). Since it remains uncertain what underlies the absence of development shift: the shift is limited to perceivers from certain language backgrounds or the shift happens at the age outside of the age range studied in previous research, evidence from younger perceivers native to Mandarin may help clarify this issue.

1.3 The role of noise in generating McGurk illusion

Findings from Chen and Hazan (2009) that the magnitude of McGurk illusion varies in different noisy conditions require attention as recent theory regarding multisensory processing suggests that our perceptual outcomes are calculated in a statistically optimal fashion with regard to the noise from different modalities (Alais & Burr, 2004; Ernst & Banks, 2002; Fetsch et al., 2011; Gori et al., 2021; Hirst et al., 2018). This hypothesis was formulated on a substantial body of psychophysical and neural evidence. The findings that sensory cues are combined in an optimal manner by weighing the variance of the noise distribution for each individual sensory signal align with statistically optimal approaches such as Bayesian or maximum-likelihood schemes (Ernst & Banks, 2002; Barutchu et al., 2010). According to this hypothesis, when the reliability of one modality decreases owing to the increased noise from this modality, the other one will take over the dominance in generating percepts (Ernst & Bülthoff, 2004). In the McGurk paradigm, if auditory information is highly noisy, perceivers will display a tendency to make more visual-dominant responses as visual modality gains dominance for its relatively higher reliability (Hazan & Li, 2008; Hirst et al., 2018; Sekiyama & Burnham, 2008; Witten & Knudsen, 2005). As a result, noise solely, whether unimodal or bimodal, could shift sensory dominance in audiovisual speech perception by lowering the reliability of the corresponding modality (or modalities).

Previous research on perceiving speech in noise showed that children are more susceptible to noise from the auditory modality relative to adults, which appears to be universal regardless of the language being spoken. Without the assistance of visual information, it is not until adolescence can children acquire fully developed competence in extracting speech in auditorily noisy environments (Elliott, 1979; Johnson, 2000; Liu et al., 2020; Sekiyama & Burnham, 2008). On the other hand, the visual speech perception ability of children has been also found to be weaker than that of adults in general owing to a lack of linguistic experience (Chen & Hazan, 2009; Gijbels et al., 2021; Knowland et al., 2016; Kishon-Rabin & Henkin, 2000; Heikkila et al., 2017; Sekiyama & Burnham, 2008; Tye-Murray et al., 2014), indicating that children are less effectively at using the visual information as compensation for insufficient auditory inputs, especially for speech sounds that lack salient visual cues to establish a reliable foundation for visual speech perception (de Boysson-Bardies & Vihman, 1991; Lalonde & Werner, 2021; Vihman et al., 1985). Thus, the effect on the McGurk illusion might vary in children at difference ages.

In terms of the interplay of noise and age in audiovisual integration indexed by the McGurk illusion, however, the results of previous studies were not always consistent. Hirst et al. (2018)

focusing on English-speaking individuals examined the influence of noise on audiovisual integration in different age groups (3–6-year-olds, 7–9-year-olds, 10–12-year-olds, and adult controls) by manipulating auditory noise level. The two younger groups required greater noise (around -10 SNR) than adults (-5.15 SNR) to hit the threshold of inducing the McGurk effect (i.e., the point where responses other than audio-dominant ones were made 50% of the time). This finding implies that noise would allow for measuring a comparable magnitude of audiovisual integration between children and adults. Contrary to Hirst et al. (2018), the findings of a study examining audiovisual speech perception in native Mandarin and native English speakers suggest that noise appears to put children at disadvantage when integrating audiovisual incongruent information (Chen & Hazan, 2009). The authors made comparisons between Mandarin- or English-speaking children around 9 years old with their adults in a quiet and a moderately noisy (-12dB) conditions. The child and adult groups were found to perform comparably in the quiet condition. However, the child group made significantly fewer audiovisual-integrated responses than the adults when noise was involved regardless of language background, suggesting the audiovisual integration of 9-year-old children is almost at adult level in the quiet condition, but this ability is still developing in the noisy condition. In addition, there is evidence from Mandarin-speaking children and adolescents that noise only has a limited impact on the magnitude of McGurk illusion (Liu et al., 2020). The study compared the McGurk illusion from children aged between 6 and 12 years with that from adolescents aged between 13 and 16 years across six SNRs from no noise to a moderate level (i.e., no noise, +3dB, 0dB, -3dB, -6dB, -9dB). The effect of noise was found to be similar in the two age groups who did not differ from each other in the McGurk responses, which means

Mandarin-speaking 6–12-year-olds and 13–16-year-olds behaved similarly at all levels of auditory noise. The lack of interaction of noise × age might be attributed to the possibility that both age groups have mastered the skill of integrating auditory and visual information. Thus, evidence from younger Mandarin-speaking children is called for clarifying the joint influence of noise on the development of audiovisual speech perception.

1.4 The current study

Overall, inconsistent conclusions regarding the developmental trajectory of audiovisual speech processing with and without auditory noise among children from varying language backgrounds could be drawn from existing studies and the key controversy lay in whether the developmental shift in terms of sensory dominance in audiovisual speech perception revealed by the McGurk paradigm was limited to Indo-European language speakers. In addition, given a lack of refined delineation of age groups (only one or two groups of children with wide age ranges were included) and evidence from younger participants, the development of the competence of processing audiovisual congruent and incongruent speech stimuli in Mandarin-speaking children was still far from clear. Thus, the current study is primarily exploratory, since to the best of our knowledge, the existing developmental studies on Mandarin speakers performed the McGurk effect on school-age children only, with very limited age effects obtained. As a result, the situations among Mandarin-speaking preschoolers remained blank. However, whether the absent developmental shift was solely attributed to language-specific consequences could not be well demonstrated unless taking younger children into consideration. In view of these problems, the current study sought to revisit the developmental issue of audiovisual speech perception in Mandarin-speaking children by performing the McGurk paradigm on a larger sample consisting of 3–4-year-olds, 5–6-year-olds, 7–8-year-olds and young adults. On top of that, as another key factor fluctuating sensory weighting, auditory noise was also introduced to the current design for both theoretical and empirical reasons. First, whether the noise would pose an impact on altering the perceptual outcome under the McGurk design in Mandarin-speaking children was unclear. Second, we are living in an environment that is noisy in nature. Collectively, we aimed to portray the developmental trajectory of processing audiovisual congruent and incongruent speech stimuli under varying auditory conditions in Mandarin-speaking children during early and middle childhood, so that we hoped to answer 1) whether Mandarin-speaking children would experience a developmental shift in sensory dominance in audiovisual speech perception under the context of McGurk paradigm, and 2) whether unimodal auditory noise would pose an impact on this developmental process in Mandarin-speaking children.

2. Methods

2.1 Participants

Data collection was carried out in Xiamen, Fujian Province, mainland China from March to June, 2022. During recruitment, we had explicitly specified that we only included Mandarin monolinguals. Seventy Mandarin-speaking children aged from three to eight and 26 young adults aged from 18 to 22 years were recruited in the current study. The sample size was determined on availability, which was greater than or equal to those used in previous studies employing a similar paradigm (Chen & Hazan, 2009; Sekiyama & Burnham, 2008; Tremblay et al., 2007). We further ascertained the required sample size for pursuing the Age Group (3–4-year-olds, 5–6-year-olds, 7–8-year-olds, 18–22-year-olds) × Noise Level (clean, 10dB, -

10dB) using the G*power software opting for a moderate sample size (η_p^2 = .06), 0.80 power, an alpha of 0.05, and 0.5 as correlation among repeated measures, which turned out to be a total sample size of 40.

Child participants were categorized into three groups according to their chronological age, namely 3–4-year-old, 5–6-year-old and 7–8-year-old groups including children whose age fell into the range from 3 years to 4 years and 11 months, from 5 years to 6 years and 11 months and from 7 years to 8 years and 11 months, respectively. Seven children, with five aged 3–4 years and two aged 5–6 years were excluded because they failed to pass the test at the end of the first training session (see 2.3 Procedure). Two participants, one from 5–6-year-old group and the other from 7–8-year-old group, were excluded given that they were not Mandarin monolinguals. The final samples included in statistical analyses are shown in Table 1. All child participants were recruited from general education institutes whose caregivers reported no intellectual, behavioural or hearing problems. Their verbal ability was further assessed by the Verbal Comprehension Index of the Wechsler Preschool and Primary Scale of Intelligence-Fourth Edition for those under 6 years (Li & Zhu, 2014) or the Wechsler Intelligence Scale for Children-Fourth Edition for the remaining (Zhang, 2008). Results turned out that the verbal ability of all child participants was developing within the expected range for their age. One-way analysis of variance (ANOVA) showed no significant differences regarding the verbal ability across the three groups of children ($F(2, 58) = 2.03$, $p = .141$, $\eta_p^2 = .07$). The verbal ability test was administered to the child participants a day before the task measuring audiovisual speech perception. The adult group was recruited from a university in mainland China, and they self-reported no hearing impairment. All participants or their caregivers had signed written consent and got compensated for their participation. The methodology employed in the current study has been reviewed and approved by the University Institutional Review Board.

Table 1: Characteristics of participants among groups.

Group	**N (Female)**	**Chronological Ages (Range, in year)**		**Verbal Comprehension Index (Range)**	
		Mean	SD	Mean	SD
3–4-year-olds	20 (10)	4.35	.31	110.25	10.48
		(3.78–4.92)		(93–129)	
5–6-year-olds	21 (10)	5.88	.50	112.90	8.70
		(5.01–6.69)		(99–128)	
7–8-year-olds	20 (10)	7.82	.64	117.85	15.47
		(7.05–8.98)		(91–155)	
Adults	26 (13)	20.85	1.33	-	-
		(18.47–22.87)		-	

2.2 Stimuli

A young female speaker native in Mandarin aged 28 years, from whom written consent was obtained, was invited to record the articulation process of the three consonant-vowel (CV) syllables in which C was a voiceless unaspirated plosive in Mandarin: “Ba”, “Da” and “Ga” (transcribed as [pa], [ta], and [ka] in International Phonetic Alphabet) with a high-level tone (around 240 Hz) in a quiet room. The speaker’s face was presented against the background in a solid colour. All videos were taken with a resolution of 1920×1080 pixels and a frame rate of 30 frames/s. Each video lasted for two seconds, which began with the speaker's still face, followed by an articulatory motion and ended with a still face. Based on the three recorded videos, two types of stimuli were created: congruent and incongruent. For the congruent stimuli, the original videos were utilized. For the incongruent one, the soundtrack of "Ba" was dubbed on the muted "Ga" video (AbVg) using Adobe Premiere Pro CC (2018 version). In addition, noisy stimuli were created by adding noise at SNRs of 10dB and -10dB to the soundtrack of both congruent and incongruent videos using a MATLAB script (R2018a version). Pink noise, which allowed for better controlling the spectral characteristics of the noise stimulus, was adopted as noise masker in order to keep aligned with previous studies involving Mandarin-speaking children (Chen & Hazan et al., 2009; Liu et al., 2020). These two levels of SNRs were determined based on the results of a pilot experiment where four Mandarin-speaking adults (two females) were invited to perform the McGurk task in a total of seven levels of SNRs. The results of pilot experiment have been provided in Table S1 and Table S2 in the Appendix. An SNR of -10dB could be challenging for children and adults and they may not hear the acoustic syllables very clearly. Before the formal experiment, the experimenter would remind the participants that their auditory condition might be very poor and they were expected to figure out any strategy to access what the speaker had said. The auditory components in all videos were scaled to the intensity level of 70dB in terms of root mean square.

2.3 Procedure

During the experiment, participants were seated in front of the screen of a 16-inch laptop with a resolution of 1920×1080 pixels at a distance of around 50cm. Soundtracks were presented binaurally through headphones (Audio-Technica ATH-M20x) at a volume that was comfortable to participants.

Child participants were familiarized with the experiment through two training sessions. The first one was set to ensure children had the knowledge of the three CV syllables involved in the current design. The experimenter would present three slides for each syllable respectively. On each slide there was a rectangular pattern with a specific background colour including both the Pinyin form of the corresponding syllable and a picture semantically relevant to the syllable (see Figure 1a). During this process, the experimenter would play out the recording of each syllable for three times and instruct the children to repeat. At the end of this training session, children were required to pass a test by pointing to the correct pattern of the heard syllable. They would receive nine trials in total, with three per syllable. Only if a child had made all choices correctly was he or she eligible for the next training session which was shared by both child and adult participants. The second training session was conducted with E-prime 3.0, which had the same setting as the formal experiment (see the next paragraph for details). In short, participants were instructed to point at the corresponding rectangular pattern of their responses among “Ba”, “Da” and “Ga” after the stimuli were presented. In this session, the three congruent stimuli (“Ba”, “Da”, “Ga”) were repeated twice in random order. During this training session, feedback was provided on the correctness of each response, though all the participants achieved full accuracy.

In the formal experiment, a total of 84 trials consisted of three congruent (“Ba”, “Da”, “Ga”) and one incongruent stimuli (“AbVg”) in three noise conditions (clean, 10dB SNR and -10dB SNR) were presented in random order with seven repetitions. In each trial, participants were instructed to watch the screen where sequentially presented a fixation screen (1000ms), a black screen (800ms), an audiovisual stimulus (2000ms) and a response screen (infinite). Only after making sure that the participants were looking at the screen did the experiment start the trial. The participants were required to report what the speaker had said orally by pointing at the corresponding rectangular pattern among the three choices as shown in the response screen in Figure 1b. The experimenter recorded their reaction by pressing the initial letter of the syllable on the keyboard.

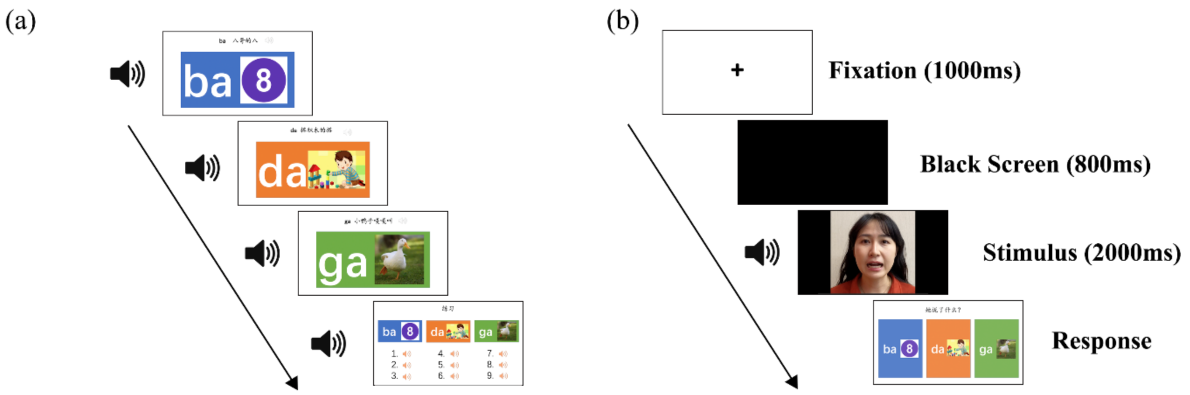

Figure 1. (a) Procedure of the first training session for child participants. (b) Procedure of a sample trial used in the second training session and the formal experiment.

2.4 Data analysis

For the identification of the three audiovisual congruent syllables, as the data were not normally distributed, we adopted nonparametric methods in analyses. In order to examine the effects of age, stimulus type and noise on the identification accuracy, a 4 × 3 × 3 repeated measures permutation ANOVA was carried out in R using the package "permuco" (R core team, 2022; Frossard & Renaud, 2021). Age Group (3–4-year-olds, 5–6-year-olds, 7–8-year-olds, adults) served as a between-subjects factor while Stimulus Type ("Ba", "Da", "Ga") and Noise Level (clean, 10dB, -10dB) were within-subject factors. Post hoc pairwise comparisons were performed when appropriate using Wilcoxon tests.

For the incongruent condition, to explore the differences yielded by Age Group (3–4-year-olds, 5–6-year-olds, 7–8-year-olds, adults), Response Type (audio-dominant, visual-dominant, audiovisual-integrated) and Noise Level (clean, 10dB, -10dB) when trying to identify incongruent stimuli, we performed a 4 × 3 × 3 repeated measures permutation ANOVA with Age Group as a between-subjects factor, while Response Type and Noise Level as within-subject factors. Post hoc pairwise comparisons were performed when appropriate using Wilcoxon tests.

To explore the development of audiovisual speech perception in the congruent condition with varying noise levels, we built a set of permutation-based linear regression models on identification accuracy of syllables ("Ba", "Da" and "Ga") with age as predictor. Only the data obtained from children were included for analysis here due to the lack of data from 9–17-year-olds. In addition, as the interaction effects have been examined using repeated measures permutation ANOVA mentioned above, separate regression models were constructed for each syllable at each noise level. For the condition where the auditory and visual information conflicted, three sets of permutation-based linear regression models were built to examine the development of audio-dominant, visual-dominant and audiovisual-integrated processing with age. Likewise, for each noise level we constructed regression models separately.

3.Results

3.1 Identification of congruent syllables

Figure 2 shows the accuracy of identifying the three audiovisual congruent syllables ("Ba", "Da" and "Ga") at the three noise levels by the three child groups and young adults (see Table S3 in the Appendix for mean values and standard deviants). Repeated measures permutation ANOVA on identification accuracy revealed a significant main effect of Age Group ($F(3, 83) = 32.61$, permutation $p < .001$, $\eta_p^2 = .06$), Stimulus Type ($F(2, 166) = 109.60$, permutation $p < .001$, $\eta_p^2 = .14$) and Noise Level ($F(2, 166) = 128.71$, permutation $p < .001$, $\eta_p^2 = .17$). Simultaneously, it also showed significant Age Group × Stimulus Type interaction ($F(6, 166) = 2.32$, permutation $p = .027$, $\eta_p^2 = .01$), Age Group × Noise Level interaction ($F(6, 166) = 5.87$, permutation $p < .001$, $\eta_p^2 = .02$) and Stimulus Type × Noise Level interaction ($F(4, 332) = 33.97$, permutation $p < .001$, $\eta_p^2 = .09$). The three-way interaction did not reach significance ($F(12, 332) = 1.45$, permutation $p = .142$, $\eta_p^2 = .011$).

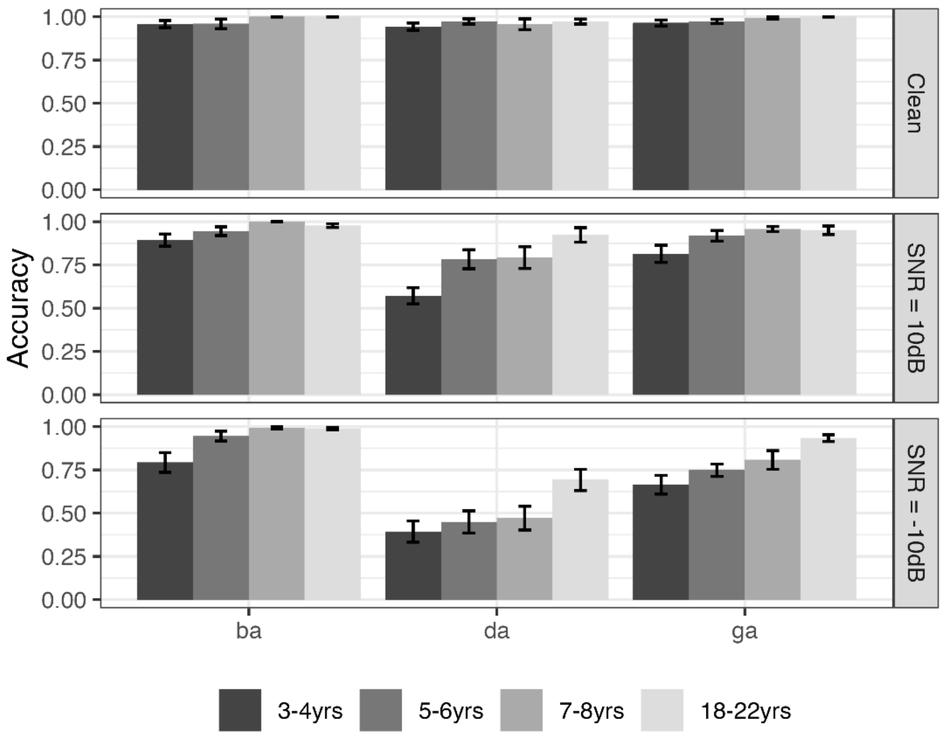

Figure 2. Identification accuracy of "Ba", "Da" and "Ga" at the three noise levels per age group. Error bars indicate standard errors of the mean.

For the Age Group × Stimulus Type interaction, post hoc pairwise comparisons were performed to examine the group differences in identifying each type of stimulus. The results showed that adults attained significantly higher accuracy than 3–4-year-olds across all three stimulus types (all *p*s < .01), and also outperformed 5–6-year-olds in "Da" and "Ga" trials (both *p*s < .01). No differences regarding any stimulus type were observed between 7–8-year-olds and adults. For the Age Group × Noise Level interaction, 3–4-year-olds were found to obtain significantly lower accuracy relative to any older group in all three noise levels (all *p*s < .05). In the two noisy conditions, 5–6-year-olds significantly underperformed compared to adults (both *p*s < .05). None of the differences between 7–8-year-olds and adults reached significance (all *p*s > .05). For the Stimulus Type × Noise Level interaction, the accuracy of the three stimulus types did not differ in the clean condition (all *p*s > .05). In noisy conditions, however, participants showed significantly higher accuracy in "Ba" trials relative to "Ga" and "Da" trials, and higher in "Ga" rather than "Da" trials (all *p*s < .05).

To summarize, the ability to identify "Ba" was undergoing a dramatic development around the age of five, while for "Ga" and "Da", the change was seen around the age of seven. Younger children were more susceptible to auditory noise. In addition, the identification of syllable which developed earlier (i.e., "Ba") was less likely to be impacted by noise than "Ga" and "Da". For the two harder ones, the accuracy for identifying "Da" was more likely to be decreased by the reductive auditory information relative to "Ga".

3.2 Responses to incongruent stimuli

Figure 3 shows the responses to incongruent stimuli at the three noise levels per age group (see Table S4 in the Appendix for mean values and standard deviants). Statistical analysis revealed a significant three-way interaction of Age Group × Response Type × Noise Level (*F*(12, 332)

= 3.04, permutation $p < .001$, $\eta_p^2 = .59$) which was further analyzed under different noise levels, respectively.

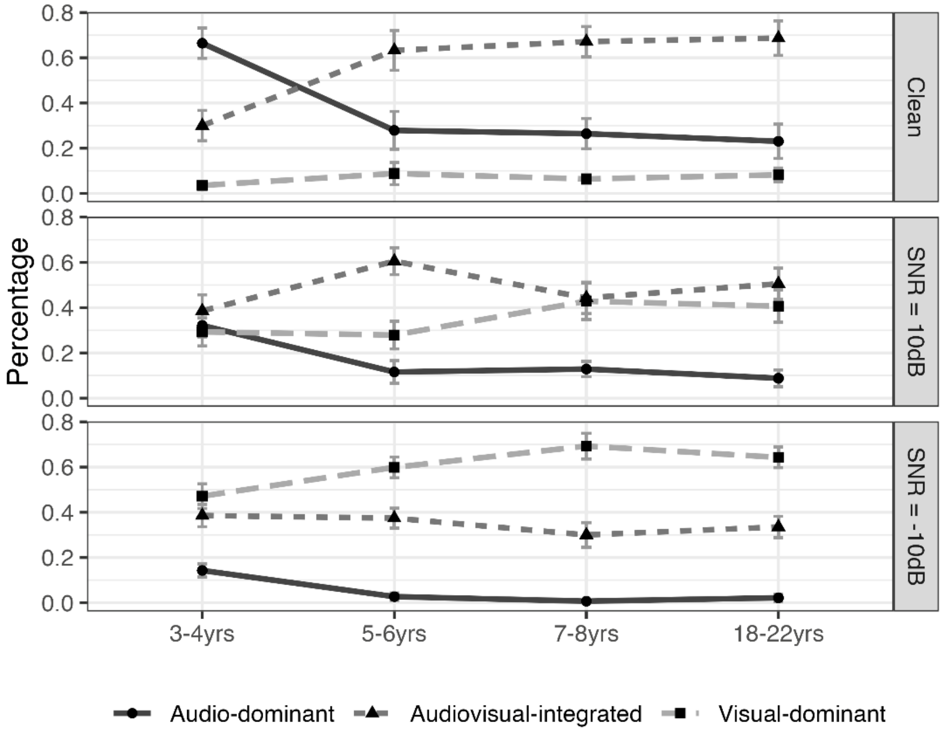

Figure 3. Percentage of different types of response to the incongruent stimuli by the child and adult groups at the three noise levels. Error bars indicate standard errors of the mean.

First, in the condition without noise, there was a significant interaction of Age Group × Response Type ($F(6, 166) = 8.721$, permutation $p < .001$, $\eta_p^2 = .12$). Post hoc pairwise comparisons revealed that 3–4-year-olds made significantly more audio-dominant responses relative to all older groups (all ps < .05) while this group of young children was found to make fewer audiovisual-integrated responses relative to 7–8-year-olds and adults (both ps < .05), and marginally fewer than 5–6-year-olds (M = .633, $SE = .019$, $Z = -1.917$, $p = .055$). As for the visual-dominant type of response, there exhibited no group differences (all ps > .05). In the 10dB SNR condition, the Age Group × Response Type interaction also reached significance level ($F(6, 166) = 3.15$, permutation $p = .003$, $\eta_p^2 = .77$). Similarly, the youngest group made significantly more audio-dominant responses than other groups (all *p*s $< .05$). The amount of audiovisual-integrated responses made by the youngest group was comparable with older groups, where no group differences were observed either (all *p*s $> .05$). The numbers of visual-dominant responses recorded from all groups did not differ significantly (all *p*s $> .05$). Similar results were found in the -10SNR condition where the Age Group × Response Type interaction was significant ($F(6, 166) = 3.84$, permutation $p = .001$, $\eta_p^2 = .41$). Group differences were observed between the youngest group and others in audio-dominant responses only (all *p*s $< .05$).

In addition, there were two-way interactions of Age Group × Response Type ($F(6, 166) = 11.0$, permutation $p < .001$, $\eta_p^2 = .05$) and Response Type × Noise Level ($F(4, 332) = 66.38$, permutation $p < .001$, $\eta_p^2 = .21$). The interaction involving age was consistent with the observed three-way interaction: significantly more audio-dominant but fewer audiovisual-integrated responses were recorded from 3–4-year-olds (all *p*s $< .05$). For the Response Type × Noise Level interaction, participants made the most audiovisual-integrated responses, and more audio-dominant responses were made than visual-dominant in the clean condition (all *p*s $< .05$). When the SNR dropped to 10dB, however, participants made comparable audiovisual-related responses with visual-dominant and the least audio-dominant responses (both *p*s $< .05$). While in the -10dB SNR condition, participants made the most visual-dominant responses and more audiovisual-dominant than audio-dominant responses (all *p*s $< .05$).

3.3 Regression results

For the identification of congruent stimuli (“Ba”, “Da” and “Ga”), permutation-based linear regression models were built to explore whether the accuracy rates would increase with age under varying noise conditions. As shown in Figure 4, for the clean condition, the accuracy of 3–8-year-olds in identifying all the three stimulus types could not be predicted by age (all *p*s > .05); however, age could predict the accuracy across stimulus types at 10dB SNR (Ba: $\beta = .03$, $t = 2.96$, permutation $p = .008$, $SE = .01$; Da: $\beta = .06$, $t = 2.75$, permutation $p = .008$, $SE = .02$; Ga: $\beta = .04$, $t = 2.67$, permutation $p = .009$, $SE = .01$); As for the -10dB SNR condition, only the accuracy of “Ba” stimulus could be predicted by age ($\beta = .05$, $t = 3.20$, permutation $p = .001$, $SE = .02$).

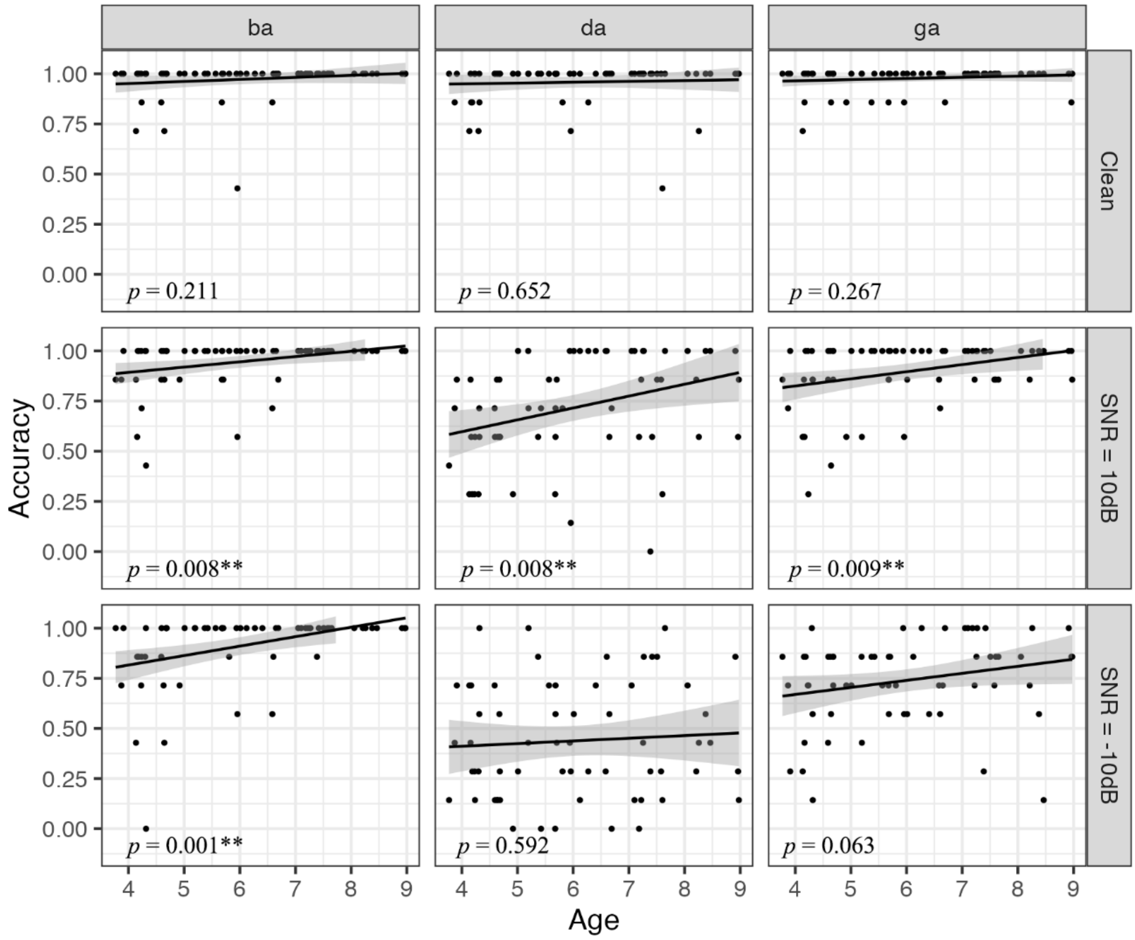

Figure 4. Developmental trajectory of audiovisual speech perception of congruent stimuli at varying noise levels.

When encountering the incongruent trials, the audio-dominant processing was found to degenerate with age regardless of noise levels (see Figure 5, clean: β = -.10, t = -3.11, permutation p = .005, SE = .03; SNR = 10dB: β = -.05, t = -2.70, permutation p = .009, SE = .018; SNR = -10dB: β = -.03, t = -3.78, permutation $p < .001$, SE = .01). Also, age could significantly predict the percentage of audiovisual-integrated responses in the clean condition (β = .09*,* t = 2.96*,* permutation p = .004, SE = .03) but not in the two noisy conditions (both ps > .05). As for the visual-dominant response, the predictability of age was only significant when the auditory condition was as poor as -10dB SNR (β = .05*,* t = 2.50*,* permutation p = .016, SE = .02).

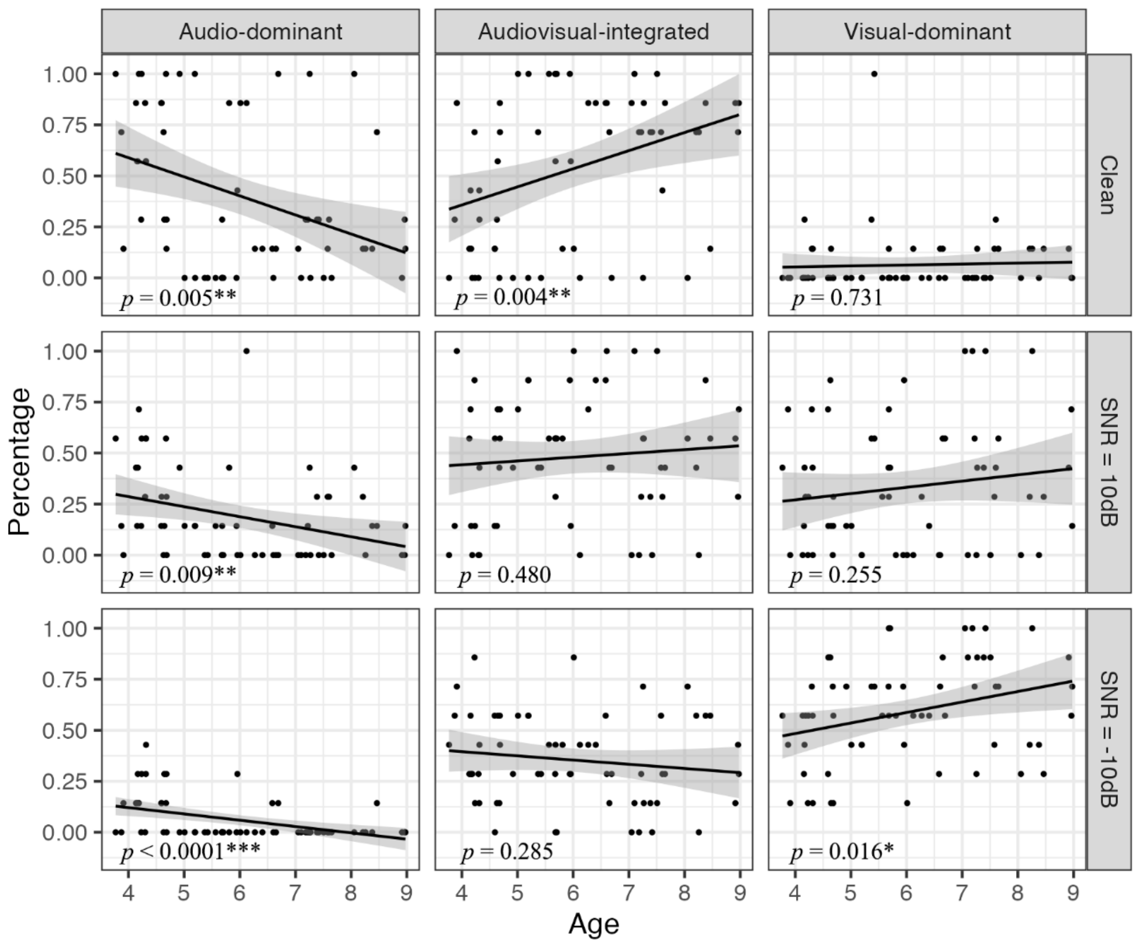

Figure 5. Developmental trajectory of audiovisual speech perception of incongruent stimuli at varying noise levels.

4. Discussion

The current cross-sectional study aimed to track down the developmental trajectory of audiovisual speech perception under the McGurk measurement among Mandarin-speaking children in the congruent condition and in the condition where auditory and visual information conflicts. When processing congruent stimuli, perceptual accuracy was found to improve with increasing age as implied by the main effect of age on the accuracy. Specifically, 3–4-year-olds could not achieve comparable accuracy as the other three groups, while 5–6-year-olds and 7–8-year-olds could identify the congruent stimuli as well as the adult controls. The ability to identify congruent stimuli was found to be reduced by noise as reflected by the findings that 5–6-year-olds who showed an adult-level ability in the clean condition but underperformed in the two noisy conditions relative to adults. For the perception of incongruent stimuli, we observed a developmental shift from audio-dominant to audiovisual-integrated processing around the age of five, especially in the condition without noise (see Figure 3). This shift seemed to be brought forward by noise as less audio-dominant and more audiovisual-integrated responses were found among those below five when noise was involved. In addition, significant relation between the visual-dominant response and age was observed in the -10dB SNR condition with the response number increasing with age. The development towards visual-dominant processing was in accordance with the statistically optimal fashion proposed by Ernst and Banks (2002): when the auditory condition was poor, relying on isolated visual modality was regarded as the optimal option.

4.1 Developmental trajectory of perceiving congruent stimuli

4.1.1 Developmental trajectories of perceiving congruent stimuli with different visual

saliency

The ability to identify “Ba” was undergoing a dramatic development around the age of five, while for “Ga” and “Da”, the change was seen around the age of seven. In other words, the identification of “Ba” develops earlier relative to “Da” and “Ga”, which does not violate the findings regarding production that labial stop consonants are early acquired phonemes in a general sense (McLeod & Crowe, 2018; Vihman, 1996). This has been attributed to the fact that labial stop consonants involving the prominent lip closure, which can be seen in addition to being heard, provide a more secure basis for identification (de Boysson-Bardies & Vihman, 1991; Vihman et al., 1985). The low visual saliency of “Da” and “Ga” might underlie their relatively late mastery. There is also electroencephalogram evidence showing that the saliency of visual inputs would facilitate audiovisual speech perception. The latency of event-related potentials evoked by bilabial stimuli [pa] was found to be significantly shorter than that evoked by [ta] and [ka] (van Wassenhove et al., 2005). In addition, audiovisual speech perception was found to be affected by noise (see the following subsection).

4.1.2 Role of auditory noise in perceiving congruent stimuli with different visual saliency

The influence of noise on perceiving stimuli varies according to visual saliency of the stimuli. Perception of stimuli with high visual saliency (i.e., “Ba”) was less affected by noise than “Ga” and “Da”. According to van Wassenhove et al. (2005), with increasing auditory noise, salient visual cues were required for successful identification as noisy environments made auditory cues less informative. Therefore, it is not surprising that “Da” and “Ga” which could provide limited visual information were less accurate to be identified correctly in the noisy conditions while the visual cue of “Ba” (i.e., lip closure) made it easy to be recognized. As for the two less accurately identified ones, we found the accuracy for identifying “Da” was more likely to be decreased by noise relative to “Ga” as reflected by the significantly more errors triggered by “Da” than “Ga” in the two noisy conditions. We speculate one reason for the high susceptibility of “Da” is that the other two syllables, “Ba” and “Ga”, were conjectured to mark the two ends of a continuum from smallest to largest initial mouth openness from the visual aspect, and if this held, “Da” was left to be more ambiguous to perceive when auditory information was insufficient. From the auditory perspective, acoustic differences, such as formant transition, measured between “Ga” and “Ba” is greater than that between “Da” and “Ba” (Walley & Carrell, 1983), which might contribute to that “Da” might be more easily to be confused with “Ba” than “Ga”. One thing worth noting was that this study was based on the perception of English-speaking participants while data from Mandarin was not available. However, the study focusing on English speakers by Hirst et al. (2018) found that “Ga” was more challenging to be recognized than “Da”. One possible explanation for this inconsistency is that experiment designs were different between the current study and Hirst et al. (2018). In their study, participants were required to select one out of three options: the first one is “Ba”, the second one is “Ga”, and the third option includes “Da” and “Tha”. The last choice that contained more than one phoneme representation might be a compromise when participants were highly confused by noise, that is, it would have a higher chance to be chosen, which in turn contributed to the relatively higher accuracy of “Da” than “Ga” in Hirst et al. (2018).

4.1.3 Role of auditory noise in perceiving congruent stimuli by different age groups

Compared with the older children, the audiovisual speech perception in the younger children was more likely to be affected by noise, in line with previous studies examining speech perception in noise (Elliott, 1979; Johnson, 2000). The role of noise in speech perception development was also supported by the results of regression analyses on the relations between children's age and their identification accuracy at the three noise levels. Specifically, the predictability of age in terms of accuracy in the clean condition was not significant which was mainly driven by the high accuracy across the three child groups with all above 94%. When SNR dropped to 10dB, we found that age predicted the identification accuracy: performance got better with increasing age. When the auditory condition was poorer (SNR = -10dB), age was found to significantly predict the identification of "Ba" while there showed a possible trend of increasing accuracy of "Ga" recognition along with aging. However, the predictability of age did not reach significance in "Da", which might be caused by the poor performance across the three groups of children whose correct rates were between 39% and 48%, suggesting the ability to extract speech without salient visual cues in very noisy environments was a "late bloomer" (Elliott, 1979; Johnson, 2000; Liu et al., 2020; Sekiyama & Burnham, 2008). Taken together, we attempt to put forward that the development of identifying audiovisually presented "Ba", "Da" and "Ga" in noise-free condition was almost mature among children tested in the current study, while noise appeared to pose a negative impact on identifying these syllables. In short, noise negatively affects audiovisual speech perception in the congruent condition, but the incongruent condition is another story.

Findings from identifying congruent stimuli might shed some light on studying the development of the ability to recognize real-word syllables under various auditory conditions in Mandarin-speaking children since the congruent stimuli employed in the current study are meaningful words in Mandarin. The reported results regarding congruent stimuli uncovered two vital factors influencing successfully identifying audiovisually presented speech stimuli: the visual saliency of utterance and unimodal auditory noise, aligning with previous findings (Jerger et al., 2009; Lalonde & Holt, 2015). More specifically, children are more likely to be put at disadvantages relative to adults when the visual cues of speech are ambiguous and when the auditory intelligibility was struggling. Whereas, it is noteworthy that only very limited CV syllables were tested in the current design and future studies concerning a wider range of phonological compositions are warrant to provide a fuller picture of this developmental process.

4.2 Developmental trajectory of perceiving incongruent stimuli

4.2.1 Development trajectory of perceiving incongruent stimuli in the clean condition

Our study observed a developmental shift from a bias to auditory information to audiovisual integration in Mandarin-speaking children with the McGurk paradigm, corresponding with previous research on development in audiovisual speech perception. This finding provides evidence for the claim that such a developmental shift is not limited to Indo-European language speakers. The most striking differences were observed in the clean condition, where 3–4-year-olds made significantly more audio-dominant responses relative to any other older groups. This is in line with the auditory, instead of visual, preference among young children revealed by previous studies on the development of audiovisual processing (Hirst et al., 2018; Robinson & Sloutsky, 2004, 2010; Sekiyama & Burnham, 2008; Tremblay et al., 2007). It has been proposed that the transient and dynamic nature of auditory information and the early maturation of auditory system make auditory modality gain dominance in audiovisual processing during early ages (Burr & Gori, 2012; Robinson & Sloutsky, 2010). Such unimodal auditory preference has been suggested to consume greater attention resources of young children, which in turn prevents auditory information from being influenced by conflicting visual information, thus giving rise to more audio-dominant responses (Robinson & Sloutsky, 2010). The bias towards auditory information exhibited by 3–4-year-olds disappeared among the 5–6-year-old and 7–8-year-old groups, indicating that children tended to gradually disengage with unimodal auditory information along with aging. This change was also supported by the result of regression analysis that the number of audio-dominant responses decreased with age, which has been attributed to the maturation of visual modality (Hirst et al., 2018; Robinson & Sloutsky, 2010; Sekiyama et al., 2014; Sekiyama & Burnham, 2008; Tremblay et al., 2007). Meanwhile, our study found that 3–4-year-olds made significantly fewer audiovisual-integrated responses in the clean condition compared to older groups, indicating that this group of children was less likely to integrate the audiovisual information to form a holistic fused percept as the older groups. One possible explanation is that the maturation of multisensory processing lags behind that of isolated modalities (Burr & Gori, 2012). Yet a comparable number of audiovisual-integrated responses were recorded from 5–6-year-olds, 7–8-year-olds and adults. These results were in accordance with previous findings that the competence of multisensory integration develops with age (Ernst, 2008; Gori et al., 2008; Nardini et al., 2008). For children as young as 3–4 years old, the immature isolated modalities and integrative mechanisms have been believed to underlie the weaker competence in integrating audiovisual information (Ernst, 2008; Hirst et al., 2018). With increasing age, children are more mature in physical conditions and richer in sensory experience (Desjardins & Werker, 2004). These advantages might get them prepared to exhibit a pattern resembling adults in audiovisual speech integration (Schorr et al., 2005).

On top of that, results from regression analyses suggested the direction of the developmental process was towards a statistically optimal fashion (Ernst & Banks, 2002). Age was detected to significantly predict the increasing audiovisual-integrated responses to incongruent stimuli and degenerated audio-dominant processing. Given the audiovisual-integrated approach marks the more statistically optimal option than unimodal auditory processing here, shifting to audiovisual-integrated processing accompanied by decreasing use of unimodal strategy with age shows that children are developing towards the statistically optimal fashion in audiovisual speech integration.

The current study helps to provide a more comprehensive picture of the developmental trajectory of audiovisual speech integration by adding data from children as young as three years. To the best of our knowledge, existing studies have performed the McGurk paradigm on Mandarin-speaking school-age children only and they fail to document the developmental shift (Chen & Hazan, 2009; Li et al., 2008; Liu et al., 2020). In line with previous findings, we obtained comparable performances between 7–8-year-olds and adults. The finding of adult-like pattern in Mandarin-speaking school-age children, together with the finding that 5–6-year-olds did not differ from adults in responding to the incongruent stimuli, indicated that audiovisual speech integration of Mandarin-speaking children measured by the McGurk paradigm seemed to develop and mature as early as preschool age. What might contribute to this early development will be discussed in the following subsection.

4.2.2 Early developmental shift in Mandarin-speaking children

The 5–6-year-old children in the current study showed a mature ability to integrate audiovisual information under the McGurk design, as was suggested by the insignificant group differences between the 5–6-year-old and adult groups. This timepoint is earlier than that found among Indo-European-language-speaking children by Tremblay et al.(2007) and Hirst et al. (2018), who showed that the ability to integrate audiovisual information was still developing in school-age children. The earlier developmental shift from a bias to auditory information to taking visual information into account could be explained from two perspectives: lower reliance on auditory modality or higher sensitivity to visual information. First, Mandarin speakers might weigh less on auditory modality due to a higher phonological ambiguity of their native language. As proposed by Massaro (1989), visual utilization will be enhanced as compensation for the complexity and ambiguity of auditory information, which was proved by Zhang et al.(2018) who obtained both behavioural and neural evidence for a strengthened McGurk illusion from Cantonese speakers whose native language was reported to be more complex in phonology relative to Mandarin speakers. Nevertheless, there is a lack of direct proof showing that Mandarin is more complex in phonology than English. If this account does not hold, the answer to this question may turn to the other end, that is, the earlier development could be triggered by the higher proficiency in using visual information. On the one hand, the stimuli involved in the McGurk paradigm have their corresponding lexicons in Mandarin while they are meaningless in English. As a result, Mandarin-speaking children are more frequently exposed to these stimuli in daily life. Given that audiovisual speech integration greatly depends on the visual speech perception ability (Massaro et al., 1986), which has been believed to associate with linguistic experience, higher exposure to the corresponding linguistic stimuli might result in being more skilled in taking the visual cues of the stimuli into account while generating perceptual outcomes. On the other hand, the enhanced visual utilization might be tuned by the teacher-centered preschool teaching mode in the Chinese context (Hu et al., 2017). Since classrooms are reported to be always noisy, children will utilize more visual cues to meet the expectation of their teacher to follow the lead and participate in activities that are largely directed by adults, which in turn fosters better capability to perceive visual speech, a skill of processing visual information, giving rise to early development of audiovisual integration in Mandarin-speaking children. Simultaneously, this developmental process has been found to be affected by noise, see the next subsection for discussion.

4.2.3 Developmental shift was affected by noise

Our study suggests that noise promotes integrating conflicting audiovisual information, agreeing with Hirst et al. (2018) who showed that children entailed more auditory noise to reach a comparable threshold of McGurk effect as adults. Specifically, in the current study, significantly fewer audiovisual-integrated responses were recorded from the 3–4-year-old group compared to any older group in the clean condition. When auditory noise was introduced to stimuli, however, this youngest group and all other groups made a comparable number of audiovisual-integrated responses. The effect of noise could be attributed to the statistically optimal fashion in generating perceptual outcomes (Ernst & Banks, 2002; Ernst & Bülthoff, 2004; Witten & Knudsen, 2005). According to this theoretical framework, children were forced to disengage with the auditory information when auditory information was noisy because the reliability of audition dropped while that of vision remained across noise levels. As a result, when noise was introduced even the youngest group shifted away from preferring unimodal auditory information to adopting an adult-like integrative strategy in deriving perceptual outcomes.

In addition to audio-dominant and audiovisual-integrated processing, noise has been found to affect visual-dominant processing. When auditory conditions were not that poor, the number of visual-dominant responses did not show a clear trend with age, whereas we found age could predict enhanced visual-dominant processing when SNR dropped to -10dB. These findings also support the proposed statistically optimal hypothesis (Ernst & Banks, 2002). Under the -10dB SNR condition where the intelligibility of auditory information was low, the visual-dominant approach marks the more statistically optimal option since the clearness of visual information was less likely to be affected by auditory noise. The positive correlation between age and visual-dominant processing in poor auditory conditions shows the developmental tendency towards a statistically optimal fashion.

Results regarding the perception of incongruent stimuli revealed the auditory preference as well as a weaker audiovisual integration among young children, which echoes the findings using other audiovisual paradigms (e.g., Robinson & Sloutsky, 2004; 2010). Nonetheless, we acknowledge that our finding that the developmental shift occurred earlier in Mandarin speakers should be restrained within the context of the McGurk measurement.

4.3 Limitations and future directions

The current study aimed at tracking down the developmental course of audiovisual speech perception using the McGurk measurement under varying auditory conditions in Mandarin-speaking children. However, it should be acknowledged that the conclusions were drawn from single tokens of the stimuli, which potentially limited the generalizability of the findings. Meanwhile, the expense of maintaining the attention of young children was achieved by a limited number of trials under each condition (seven repetitions) that was at a moderate level in the field of similar research (Chen & Hazan, 2009; Dupont et al., 2005; Hirst et al., 2018). Additionally, though findings from the current study might illuminate future studies regarding the development of audiovisual speech perception in Mandarin-speaking children to some extent, it should be treated with caution when extending these findings beyond the context of the McGurk measurement.

Combining reactions to congruent and incongruent stimuli, the role of auditory noise was found to vary according to stimulus congruency in the development of audiovisual speech perception: decreasing auditory intelligibility allowed for measuring a widening gap in identifying congruent stimuli, but for measuring more alike performances in perceiving incongruent stimuli between children and adults. Specifically, when the auditory and visual information was congruent, the perceptual outcomes were generated with the cooperation of both modalities. In contrast, when the auditory and visual information conflicted, we measured perceptual outcomes depending on weights allocating to modalities that competed with each other. Following this line, responding to congruent and incongruent stimuli may draw upon different abilities, which potentially addresses the intensifying debate about the predictability of the susceptibility to the McGurk illusion to the perception of natural speech (Alsius et al., 2017; Van Engen et al., 2022). Future studies are urgently called to clarify the relationship between the McGurk illusion and the other measurements of audiovisual integration. Besides, as working memory seemed not to be an important factor in predicting audiovisual speech perception among adults (Li et al., 2022), this factor was not addressed by the current study. Considering this might not be the case in children, future studies with working memory taken into account are also warranted.

5. Conclusion

In the current study, we found audiovisual speech perception developed with age in Mandarin-speaking preschoolers and school-age children through the McGurk paradigm. Children aged around 5 years could identify congruent stimuli as well as adults in the condition without noise. For the incongruent stimuli, 5-year-old children exhibited a development shift from more attending to unimodal auditory information to adopting an adult-like audiovisual integrative strategy. Auditory noise was revealed to function differently in the congruent and incongruent conditions by reducing the ability to identify congruent stimuli but promoting the integration of conflicting audiovisual information. The findings that noise has influences on altering audiovisual speech perception are in accordance with the statistically optimal hypothesis regarding the role of noise in the use of multiple sources of sensory information. Our findings support that the developmental shift in the McGurk design, which we observed among young Mandarin-speaking children, is not limited to Indo-European language speakers.

Acknowledgments

The data, analytic code, and materials necessary to reproduce the analyses presented here are not publicly accessible, and the analyses presented here were not preregistered.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*, 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Alsius, A., Paré, M., & Munhall, K. G. (2017). Forty years after Hearing lips and seeing voices: The McGurk Effect Revisited. *Multisensory Research*, *31*, 111–144. https://doi.org/10.1163/22134808-00002565

Bernstein, L. E. (2012). Visual speech perception. In Bailly G., Perrier P., & Vatikiotis-Bateson E. (Eds.), *Audiovisual Speech Processing* (pp. 21–39). Cambridge, UK: Cambridge University Press.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*, 204–220. https://doi.org/10.1002/dev.20032

Burr, D., & Gori, M. (2012). Multisensory integration develops late in humans. In Murray M. M., & Wallace M. T. (Eds), *The Neural Bases of Multisensory Processes* (pp. 345–362). CRC Press.

Chen, Y., & Hazan, V. (2009). Developmental factors and the non-native speaker effect in auditory-visual speech perception. *The Journal of the Acoustical Society of America*, *126*, 858–865. https://doi.org/10.1121/1.3158823

de Boysson-Bardies, B., & Vihman, M. M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language*, *67*, 297–319. https://doi.org/10.1353/lan.1991.0045

Denes, P. B., & Pinson, E. N. (1993). *The speech chain: The physics and biology of spoken* *language.* W.H. Freeman.

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology*, *45*, 187–203. https://doi.org/10.1002/dev.20033

Dupont, S., Aubin, J., & Ménard, L. (2005). Study of the McGurk effect in 4 and 5-year-old French Canadian children. *ZAS Papers in Linguistics*, *40*, 1–17. https://doi.org/10.21248/zaspil.40.2005.254

Elliott, L. L. (1979). Performance of children aged 9 to 17 years on a test of speech intelligibility in noise using sentence material with controlled word predictability. *The Journal of the Acoustical Society of America*, *66*, 651–653. https://doi.org/10.1121/1.383691

Eramudugolla, R., Henderson, R., & Mattingley, J. B. (2011). Effects of audio–visual integration on the detection of masked speech and non-speech sounds. *Brain and Cognition*, *75*, 60–66. https://doi.org/10.1016/j.bandc.2010.09.005

Ernst, M. O. (2008). Multisensory integration: A late bloomer. *Current Biology*, *18*, R519–R521. https://doi.org/10.1016/j.cub.2008.05.002

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433. https://doi.org/10.1038/415429a

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*, 162–169. https://doi.org/10.1016/J.TICS.2004.02.002

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, *15*, 146–154. https://doi.org/10.1038/nn.2983

Frossard, J., & Renaud, O. (2021). Permutation tests for regression, ANOVA, and comparison of signals: The permuco package. *Journal of Statistical Software*, *99*, 1–32. https://doi.org/10.18637/jss.v099.i15

Gijbels, L., Yeatman, J. D., Lalonde, K., & Lee, A. K. C. (2021). Audiovisual speech processing in relationship to phonological and vocabulary skills in first graders. *Journal of Speech, Language, and Hearing Research*, *64*, 5022–5040. https://doi.org/10.1044/2021_JSLHR-21-00196

Gori, M., Campus, C., & Cappagli, G. (2021). Late development of audio-visual integration in the vertical plane. *Current Research in Behavioral Sciences*, *2*, 100043. https://doi.org/10.1016/j.crbeha.2021.100043

Gori, M., Del Viva, M., Sandini, G., & Burr, D. C. (2008). Young children do not integrate visual and haptic form information. *Current Biology*, *18*, 694–698. https://doi.org/10.1016/j.cub.2008.04.036

Hazan, V., & Li, E. (2008). The effect of auditory and visual degradation on audiovisual perception of native and non-native speakers. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1191–1194. https://doi.org/10.21437/interspeech.2008-359

Heikkilä, J., Lonka, E., Ahola, S., Meronen, A., & Tiippana, K. (2017). Lipreading ability and its cognitive correlates in typically developing children and children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *60*, 485–493. https://doi.org/10.1044/2016_JSLHR-S-15-0071

Hirst, R. J., Stacey, J. E., Cragg, L., Stacey, P. C., & Allen, H. A. (2018). The threshold for the McGurk effect in audio-visual noise decreases with development. *Scientific Reports*, *8*, 1–18. https://doi.org/10.1038/s41598-018-30798-8

Hu, B. Y., Fan, X., Yang, Y., & Neitzel, J. (2017). Chinese preschool teachers' knowledge and practice of teacher-child interactions: The mediating role of teachers' beliefs about children. *Teaching and Teacher Education*, *63*, 137–147. https://doi.org/10.1016/j.tate.2016.12.014

Jerger, S., Damian, M. F., Spence, M. J., Tye-Murray, N., & Abdi, H. (2009). Developmental shifts in children's sensitivity to visual speech: A new multimodal picture–word task. *Journal of Experimental Child Psychology*, *102*, 40–59. https://doi.org/10.1016/j.jecp.2008.08.002

Johnson, C. E. (2000). Childrens' phoneme identification in reverberation and noise. *Journal of Speech, Language, and Hearing Research*, *43*, 144–157. https://doi.org/10.1044/jslhr.4301.144

Kishon-Rabin, L., & Henkin, Y. (2000). Age-related changes in the visual perception of phonologically significant contrasts. *British Journal of Audiology*, *34*, 363–374. https://doi.org/10.3109/03005364000000152

Knowland, V. C. P., Evans, S., Snell, C., & Rosen, S. (2016). Visual speech perception in children with language learning impairments. *Journal of Speech, Language, and Hearing Research*, *59*, 1–14. https://doi.org/10.1044/2015_JSLHR-S-14-0269

Lalonde, K., & Holt, R. F. (2015). Preschoolers benefit from visually salient speech cues. *Journal of Speech, Language, and Hearing Research*, *58*, 135–150.

https://doi.org/10.1044/2014_JSLHR-H-13-0343

Lalonde, K., & Werner, L. A. (2021). Development of the mechanisms underlying audiovisual speech perception benefit. *Brain Sciences*, *11*, 49. https://doi.org/10.3390/brainsci11010049

Lewkowicz, D. J., & Flom, R. (2014). The audiovisual temporal binding window narrows in early childhood. *Child Development*, *85*, 685–694. https://doi.org/10.1111/cdev.12142

Li, M., Chen, X., Zhu, J., & Chen, F. (2022). Audiovisual Mandarin lexical tone perception in quiet and noisy contexts: The influence of visual cues and speech rate. *Journal of Speech, Language, and Hearing Research*, *65*, 4385–4403. https://doi.org/10.1044/2022_JSLHR-22-00024

Li, Y., Mei, L., & Dong, Q. (2008). The characteristics and development of audiovisual speech perception in native Chinese speakers. *Psychological Development and Education*, *3*, 43–47.

Liu, M., Du, X., & Liu, Q. (2020). The features of audiovisual speech perception in noise of children with Autism Spectrum Disorder. *Chinese Journal of Applied Psychology*, *3*, 232–239.

Macdonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*, 253–257. https://doi.org/10.3758/BF03206096

Magnotti, J. F., Basu Mallick, D., 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. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, *21*, 398–421. https://doi.org/10.1016/0010-0285(89)90014-5

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, *41*, 93–113. https://doi.org/10.1016/0022-0965(86)90053-6

McGurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748. https://doi.org/10.1038/264746a0

McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American Journal of Speech-Language Pathology*, *27*, 1546–1571. https://doi.org/10.1044/2018_AJSLP-17-0100

Ménard, L., Leclerc, A., & Tiede, M. (2014). Articulatory and acoustic correlates of contrastive focus in congenitally blind adults and sighted adults. *Journal of Speech, Language, and Hearing Research*, *57*, 793–804. https://doi.org/10.1044/2014_JSLHR-S-12-0395

Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Current Biology*, *18*, 689–693. https://doi.org/10.1016/j.cub.2008.04.021

Rao, N., Ng, S. S. N., & Pearson, E. (2010). Preschool pedagogy: A fusion of traditional chinese beliefs and contemporary notions of appropriate practice. In *Revisiting The Chinese Learner* (pp. 255–279). Springer Netherlands. https://doi.org/10.1007/978-90-481-3840-1_9

Robinson, C. W., & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development*, *75*, 1387–1401. https://doi.org/10.1111/j.1467-8624.2004.00747.x

Robinson, C. W., & Sloutsky, V. M. (2010). Development of cross-modal processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 135–141. https://doi.org/10.1002/WCS.12

Rosenblum, L. D., & Dorsi, J. (2021). Primacy of multimodal speech perception for the brain and science. In *The Handbook of Speech Perception* (pp. 28–57). Wiley. https://doi.org/10.1002/9781119184096.ch2

Sato, M., Troille, E., Ménard, L., Cathiard, M.-A., & Gracco, V. (2013). Silent articulation modulates auditory and audiovisual speech perception. *Experimental Brain Research*, *227*, 275–288. https://doi.org/10.1007/s00221-013-3510-8

Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. I. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences*, *102*, 18748–18750. https://doi.org/10.1073/pnas.0508862102

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science*, *11*, 306–320. https://doi.org/10.1111/j.1467-7687.2008.00677.x

Sekiyama, K., Burnham, D., Tam, H., & Erdener, D. (2003). *ISCA Archive Auditory-Visual Speech Perception Development in Japanese and English Speakers Future University Hakodate, Japan*.

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00323

Stacey, J. E., Howard, C. J., Mitra, S., & Stacey, P. C. (2020). Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect. *Attention, Perception, & Psychophysics*, *82*, 3544–3557. https://doi.org/10.3758/S13414-020-02042-X

Thompson, L. A., & Massaro, D. W. (1994). Children′s integration of speech and pointing gestures in comprehension. *Journal of Experimental Child Psychology*, *57*, 327–354. https://doi.org/10.1006/jecp.1994.1016

Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Théoret, H. (2007). Speech and non-speech audio-visual illusions: A developmental study. *PLoS ONE*, *2*, e742. https://doi.org/10.1371/journal.pone.0000742

Tye-Murray, N., Hale, S., Spehar, B., Myerson, J., & Sommers, M. S. (2014). Lipreading in school-age children: The roles of age, hearing status, and cognitive ability. *Journal of Speech, Language, and Hearing Research*, *57*, 556–565. https://doi.org/10.1044/2013_JSLHR-H-12-0273

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and Visual Lexical Neighborhoods in Audiovisual Speech Perception. *Trends in Amplification*, *11*, 233–241. https://doi.org/10.1177/1084713807307409

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and Aging*, *31*, 380–389. https://doi.org/10.1037/pag0000094

Van Engen, K. J., Dey, A., Sommers, M. S., & Peelle, J. E. (2022). Audiovisual speech perception: Moving beyond McGurk. *The Journal of the Acoustical Society of America*, *152*, 3216–3225. https://doi.org/10.1121/10.0015262

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*, 1181–1186. https://doi.org/10.1073/pnas.0408949102

Vihman, M. M. (1996). *Phonological development: The origins of language in the child.* Blackwell Publishing.

Vihman, M. M., Macken, M. A., Miller, R., Simmons, H., & Miller, J. (1985). From babbling to speech: A re-assessment of the continuity issue. *Language*, *61*, 397–445. https://doi.org/10.2307/414151

Walley, A. C., & Carrell, T. D. (1983). Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, *73*, 1011–1022. https://doi.org/10.1121/1.389149

Witten, I. B., & Knudsen, E. I. (2005). Why seeing is believing: Merging auditory and visual worlds. *Neuron*, *48*, 489–496. https://doi.org/10.1016/j.neuron.2005.10.020

Zhang, J., Meng, Y., McBride, C., Fan, X., & Yuan, Z. (2018). Combining behavioral and ERP methodologies to investigate the differences between McGurk effects demonstrated by Cantonese and Mandarin speakers. *Frontiers in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00181