Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development of audiovisual integration in infancy: Evidence from speaker-specific gender effects

Mengke Jin
Zhejiang Sci-Tech University
Linlin Yan
yanlinlin@zstu.edu.cn
Zhejiang Sci-Tech University
Yue Ding
Zhejiang Sci-Tech University
Shuaike Hu
Zhejiang Sci-Tech University
Xinhan Gu
Zhejiang Sci-Tech University
Shaoying Liu
Zhejiang Sci-Tech University
Licai Ren
Xiasha Street Community Health Service Center
Naiqi G. Xiao
McMaster University

Research Article

Keywords: audiovisual integration, McGurk effect, developmental trajectories, gender disparity, infant perception

Posted Date: May 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9458481/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Infants are predominantly raised by female caregivers, leading to a marked asymmetry in early audiovisual experience between male and female faces and voices. It remains unclear whether this imbalance in exposure contributes to speaker gender differences in the development of audiovisual integration. In this study, we investigated whether infants aged 6 to 13 months (*N* = 85; 40 younger infants aged 6–9 months, 45 older infants aged 10–13 months) show differential integration of the McGurk effect depending on speaker gender. Using a familiarization/test procedure, infants were repeatedly presented with McGurk stimuli (visual /ka/ combined with auditory /pa/, eliciting a fused /ta/). In the test phase, infant looking times were measured in response to congruent presentations of either /pa/ or /ta/. Significant integration was observed for female speakers in older infants, alongside an emerging trend in younger infants. Conversely, no significant integration was detected for male speakers in either age group. These preliminary findings suggest that the development of audiovisual integration may be shaped by early social experience, reflecting a trajectory whereby integration emerges more robustly for familiar female speakers.

1 | Introduction

During the first year of life, infants continuously receive sensory inputs from multiple modalities (Bahrick & Lickliter, 2002; Lewkowicz, 2010), relying on multisensory integration to process complex stimuli. This fundamental perceptual process binds spatially and temporally congruent cross-modal information into a unified percept (Stein & Stanford, 2008; Stein et al., 2009), effectively reducing environmental ambiguity and enhancing perceptual precision (Stein et al., 2014). The interplay between vision and audition is particularly important in processing everyday speech, an inherently multimodal event (Rosenblum, 2008). When viewing a talking face, infants simultaneously process vocal sounds, lip movements, and related facial kinematics (Guellaï et al., 2016; Kitamura et al., 2014). This process of binding what they see with what they hear is known as audiovisual integration (Mason et al., 2019; Stein & Stanford, 2008), which significantly improves speech recognition accuracy in noisy or ambiguous environments (Alsius et al., 2018; Irwin & DiBlasi, 2017). Mastering this integration optimizes speech recognition and provides a foundation for later language comprehension, joint attention, and social skills (Bahrick & Lickliter, 2002, 2012; Wallace et al., 2020). To quantify the emergence of early audiovisual integration, researchers widely utilize the McGurk effect (Alsius et al., 2018; McGurk & MacDonald, 1976). By presenting mismatched stimuli (e.g., visual /ga/ paired with auditory /ba/), this paradigm induces a fused percept (e.g., perceiving /da/), providing an objective indicator of infants' integration abilities.

Although infants integrate facial and vocal information very early (Burnham & Dodd, 2004; Hyde et al., 2011), complex audiovisual integration initially shows significant instability. For instance, Desjardins and Werker found that 4.5-month-olds' perception of the McGurk illusion depends heavily on stimulus type and infant gender (Desjardins & Werker, 2004). Female infants detected mismatches after habituation to congruent stimuli, whereas males only showed integration tendencies after habituation to incongruent stimuli. Around 5 months of age, audiovisual integration begins demonstrating adult-like stability.

Behavioral evidence indicates that after habituating to a congruent /va/ syllable, 5-month-olds cannot distinguish it from a McGurk stimulus (auditory /ba/ with visual /va/) perceived by adults as /va/ (Rosenblum et al., 1997). Electrophysiological findings confirm this, showing that 5-month-olds exhibit highly similar neural responses to both fused incongruent and congruent stimuli, marking the initial emergence of perceptual integration (Kushnerenko et al., 2008).

According to dynamic systems theories (Smith & Thelen, 2003; Thelen & Smith, 1994), the maturation of this ability stems from continuous infant-environment interactions. The perceptual system dynamically optimizes processing strategies based on the statistical frequencies of environmental inputs. Supporting this, Mercure's studies found that monolingual environments prompt infants to strategically shift their gaze when viewing mismatched speech, increasing mouth-looking at 6–8 months before shifting attention to the eyes at 7–10 months (Mercure et al., 2019, 2022). Conversely, bilingual infants showed no such attention shifts based on stimulus congruency. This reallocation of attention helps infants extract critical cross-modal features, indirectly driving integration maturation. Exposure to specific racial groups similarly drives system specialization (Ujiie et al., 2021). Infants aged 5 to 9 months successfully integrate frequently encountered own-race faces and voices to produce a robust McGurk effect but fail with other-race stimuli. Functional near-infrared spectroscopy (fNIRS) evidence supports this (Ujiie et al., 2020), revealing distinct neural activity patterns in the left temporal lobe when infants process own-race versus other-race speech. Physiological studies further outline this trajectory. Szmytke demonstrated that infants' neural responses to audiovisual speech undergo significant cortical reorganization between 5 and 10 months (Szmytke et al., 2023), largely driven by transitions toward configural face processing (Schwarzer et al., 2007). Collectively, these findings demonstrate that infants' audiovisual integration undergoes highly experience-dependent specialization during the second half of the first year.

Throughout this dynamic process, primary caregivers serve as the core source of early multisensory inputs (Bahrick, 2010). The linguistic and racial patterns processed by the infant reflect the statistical distribution of these caregiving experiences. Similar to racial exposure, infants experience pronounced asymmetry in their exposure to male and female individuals. In accordance with SAGER guidelines, while "male" and "female" denote the biological sex of these individuals, this exposure imbalance fundamentally reflects socially constructed "gender" roles in early parenting—a distinction critical for understanding experience-dependent development. This pattern is objectively documented through parent reports (Rennels & Davis, 2008; Sugden & Moulson, 2019) and head-mounted camera recordings (Sugden et al., 2014). Rennels and Davis (2008) reported that infants' interactions with females steadily increase, rising from approximately two-thirds of interaction time at 2 to 5 months to three-quarters at 8 to 11 months. This significant early gender asymmetry fundamentally shapes the trajectory of audiovisual multisensory abilities. Multiple studies confirm that this unequal experience leads to a clear female advantage in basic audiovisual matching. Hillairet de Boisferon et al. (2015) found that 9-month-olds successfully matched female faces and voices but failed with male stimuli. This matching advantage is highly robust, as 9–12-month-old infants continue to display the same gender-selective preferences even when dynamic videos are replaced with static images (Poulin-Dubois et al., 1994).

Although basic audiovisual matching displays a female advantage, it remains unknown whether this bias extends to complex audiovisual integration. As Lozano et al. (2024) emphasized, early audiovisual processing is hierarchically organized. Since simple matching and complex integration belong to the same developmental continuum, asymmetric early exposure plausibly creates a gender bias in audiovisual integration. However, existing literature contains a systematic methodological limitation: nearly all studies investigating infant audiovisual integration exclusively use female speakers. The potential asymmetry in infants' integration abilities therefore remains untested.

The current study addresses this gap by investigating how infants' asymmetric exposure to male and female caregivers influences emerging audiovisual integration. We hypothesize that audiovisual integration follows an experience-dependent trajectory, emerging earlier and demonstrating greater robustness for frequently encountered female stimuli relative to male stimuli. The study focuses on infants aged 6 to 13 months, divided into a younger group (6–9 months) and an older group (10-months). Infants completed a McGurk task using a familiarization-test paradigm. They were first familiarized with an incongruent McGurk stimulus (visual /ka/ paired with auditory /pa/, perceived by adults as /ta/). In the test phase, they sequentially heard the fused item (/ta/) and the original auditory item (/pa/). Successful integration was operationalized as the ability to discriminate between these test items, indicated by a significant difference in looking times between /ta/ and /pa/ trials. A significant preference indicates perception of the fused syllable as a distinct event, whereas a lack of preference suggests failed integration. This design extends previous research by using full, clearly gender-identifiable facial stimuli and directly comparing integration performance between male and female speakers, providing empirical evidence on how early social experiences shape foundational multisensory processing.

2 | Methods

2.1 | Participants

The final sample consisted of 85 infants divided into two age groups: 40 younger infants (6–9 months; M_{age} = 242.57 days, *SD* = 22.55; 19 females) and 45 older infants (10–13 months; M_{age} = 370.29 days, *SD* = 19.05; 22 females). Infants were randomly assigned to either the female or male model condition. A total of 144 monolingualism infants were initially recruited; however, 65 were excluded due to preterm birth (< 37 weeks) (n = 1), unsuccessful eye-tracking calibration (n = 5), outlying looking times (exceeding ± 2 *SD*s) (n = 2), fussiness (n = 7), or failure to meet the familiarization criterion (n = 44). All included participants were full-term with normal vision and hearing.

We determined the sample size using an a priori power analysis based on effect sizes from papers that examine audiovisual integration and the McGurk effect in infants (Burnham & Dodd, 2004; Minar & Lewkowicz, 2017; Ujiie et al., 2021). The analysis (G*Power 3.1; one-tailed one-sample t-test, Power = 0.80, α = 0.05) yielded estimated sample sizes ranging from 16 to 26 per group. Our final sample sizes (19–25 per group) fell within this target range. Participants were recruited from a community child health

center. The study followed the Declaration of Helsinki and received approval from the Institutional Review Board. The parents provided informed consents for their infant's participation prior to the study.To empirically verify our assumption of differential gender exposure in early development, we collected and analyzed parental reports detailing the composition of infants' primary caregivers. The open-ended descriptions were systematically coded using a 5-point ordinal scale, ranging from 'Overwhelmingly Male-Dominant' (1) to 'Overwhelmingly Female-Dominant' (5). As shown in Table 1, the results confirmed a pronounced gender imbalance in caregiving environments: 89.4% of the infants in our sample were raised in female-dominant settings. This distribution provides direct empirical support for the experience-dependent hypothesis underlying the study.

Table 1
Classification and Distribution of Caregiving Environments by Gender Dominance

Category	Ordinal Score	Definition	N	Percentage (%)
Overwhelmingly Female-Dominant	5	All reported caregivers were female.	12	14.1%
Strongly Female-Dominant	4	Primary caregiver female, with minimal or transient male involvement.	64	75.3%
Balanced / Mixed	3	Approximately equal Male caregiving roles between male and female figures.	8	9.4%
Strongly Male-Dominant	2	Primary caregiver male, with minimal or transient female involvement.	1	1.2%
Overwhelmingly Male-Dominant	1	All reported caregivers were male.	0	0.0%
Total			85	100%

2.2 | Stimuli

The stimulus materials were developed following a standardized production protocol. Two native Mandarin Chinese speakers (one female, one male; aged 25–26 years) were recruited as models to produce the required facial and vocal articulations. Videos recordings were captured using an iPhone XR (Apple Inc.) under controlled lighting conditions. Each model articulated the syllables /ka/, /pa/, and /ta/ in two visual conditions: (1) with a fully visible face and (2) while wearing a standard surgical mask. All recordings were subsequently edited and synchronized using Adobe Premiere Pro 2018 (Adobe Systems) to generate the final experimental stimuli. The resulting visual stimuli had a resolution of 1920 · 1080 pixels and a duration of 2.4 seconds, with the corresponding auditory stimuli edited to match this duration precisely.

The McGurk stimuli were created by digitally combining the visual articulation of /ka/ with the auditory production of /pa/ using Adobe Premiere Pro 2018. For each model (male and female), the following

components were prepared: (1) a silent video clip of the model articulating /ka/, and (2) an audio recording of the same model articulating /pa/. The auditory /pa/ track was precisely superimposed onto the visual /ka/ clip, with temporal alignment achieved by synchronizing the amplitude onset of the /pa/ audio to the original /ka/ audio's onset marker. To induce the perception of the McGurk effect, pink noise was added to the audio tracks with a signal-to-noise ratio (SNR) of 0 dB (Sekiyama & Tohkura, 1991; Ujiie et al., 2015). Following synchronization, the original /ka/ audio was removed, yielding the final McGurk stimulus (visual /ka/ + auditory /pa/). Each 2.4-second stimulus followed a consistent structure: the sounds started 500 ms after the video appeared. The utterance then lasted around 1,700 ms and ended with the mouth completely shut in a neutral position at around 2,200 ms after stimulus onset. All stimuli were rigorously evaluated for natural lip movement, auditory clarity, and audiovisual synchrony before experimental use. This procedure was applied to both male and female models to ensure consistency. This resulting incongruent pairing produced 2.4-second McGurk stimuli. To validate the perceptual efficacy of these stimuli, particularly the novel male version, a separate adult sample was tested, confirming that both male and female stimuli reliably elicited the intended fused percept of /ta/ (see Supplementary Materials for details). As illustrated in Fig. 1, one McGurk stimulus (auditory /pa/ + visual /ka/) and two congruent masked stimuli (audiovisual /pa/ and /ta/) were generated for each model. The face stimuli subtended a visual angle of 10.60° × 13.73° a viewing distance of 70 cm, corresponding to physical dimensions of 13 cm × 16.85 cm on the display monitor. All stimuli were presented against a neutral gray background under controlled lighting conditions.

2.3 | Apparatus

The experiment was conducted in a dedicated testing room at the health service center. Stimuli were presented on a 24.5-inch monitor (DELL S2522HG; 1920 × 1080 pixel resolution, 240 Hz refresh rate) accompanied by an external speaker system (Edifier R20). The display and speakers were positioned 70 cm directly in front of the infant. Auditory stimuli were calibrated to a sound pressure level of 70 and 74 dB. Infant gaze behavior was recorded using an EyeLink 1000plus eye tracker with a 16mm lens, operating at a sampling rate of 500 Hz.

2.4 | Procedure

The study utilized a familiarization/novelty preference paradigm to assess infants' audiovisual integration abilities. The same experimental procedure was implemented for both male and female stimulus conditions, which differed only in the gender of the model. A between-subjects design was adopted for model gender to prevent carryover effects that could arise from sequential exposure to different speakers, while also minimizing infant fatigue through shorter, single-condition sessions.

As illustrated in Fig. 2, the experiment comprised two consecutive phases: familiarization and test. During the familiarization phase, infants completed six trials, each consisting of seven consecutive presentations of the McGurk stimulus (auditory /pa/ + visual /ka/), yielding a total trial duration of 16.8 seconds (similar with Ujiie et al.'s (2021). The subsequent test phase consisted of two trials: masked congruent /pa/ and masked congruent /ta/. Each test trial consisted of seven consecutive repetitions,

maintaining the same 16.8 seconds duration. Trial order in the test phase was counterbalanced across participants to mitigate order effects. Prior to each trial, an attention-grabbing animation of moving balls was displayed centrally to engage infants' attention. Total stimulus presentation lasted 134.4 seconds. Including eye-tracking calibration and attention-getters, the entire session lasted approximately 4–5 minutes per infant.

3 | Results

During the familiarization phase, infants' average looking time toward the computer screen was calculated separately for the first three and last three trials. Those who showed longer looking times in the first three trials compared to the last three were classified as having reached familiarization; otherwise, they were assigned to the non-familiarized group. Following Ujiie et al. (2021), infants were included in the final analysis only if they showed successful familiarization, defined as a significant decrease in looking time between the first half (trials 1–3) and the second half (trials 4–6) of the familiarization phase. So, infants whose mean looking times were longer in the final three trials than in the initial three were excluded from subsequent analyses, as this pattern indicated insufficient familiarization with the stimuli.

To assess visual attention during the test phase, fixation time within the facial region was quantified for each trial. To minimize potential interference from the peripheral features such as hair and ears, a standardized Area of Interest (AOI) encompassing the whole face (see Fig. 3). This AOI was an elliptical region with a width of 390 pixels and a height of 540 pixels. It was created and applied using the EyeLink Data Viewer software (SR Research), ensuring spatial consistency across all male and female stimuli.

3.1 | Familiarization phase

To rule out the possibility that test-phase differences stemmed from pre-existing attentional biases, we first analyzed infants' looking times during familiarization. Mean total looking time for the first half and second halves of the familiarization trials are presented in Table 2. We conducted a mixed-design ANOVA was conducted separately for each age group, with time block (first three vs. last three trials) as a within-subjects factor and model gender (female vs. male) as between-subjects factor. For younger infants (6–9 months), results showed a highly significant main effect of time block, $F(1, 38) = 46.62$, $p < 0.001$, $\eta_p^2 = 0.163$, reflecting a robust decline in looking times across the familiarization phase. Importantly, neither the main effect of model gender, $F(1, 38) = 1.95$, $p = 0.170$, $\eta_p^2 = 0.041$, nor the time block · model gender interaction, $F(1, 38) = 2.62$, $p = 0.114$, $\eta_p^2 = 0.011$, reached significance. A similar pattern was observed among the older infants (10–13 months). The results showed a highly significant main effect of time block, $F(1, 43) = 59.27$, $p < 0.001$, $\eta_p^2 = 0.283$, while neither the main effect of model gender, $F(1, 43) = 0.004$, $p = 0.950$, $\eta_p^2 < 0.001$, nor the interaction, $F(1, 43) = 0.012$, $p = 0.912$, $\eta_p^2 < 0.001$, was significant.

These results confirm that infants across both age groups and model gender conditions successfully familiarized to the McGurk stimuli. Importantly, the absence of significant interactions indicates that the extent of familiarization did not differ systematically between conditions. This pattern provides a consistent baseline for interpreting test-phase looking behavior, confirming that subsequent looking preferences can be attributed to infants' perception of the test stimuli rather than to pre-existing attentional differences during familiarization.

Table 2
Descriptive Statistics for Total Looking Times (in ms) During the Familiarization Phase

Age Group	Model Gender	*N*	First Half (Trials 1–3)	Second Half (Trials 4–6)	Difference (First - Second)
Younger Infants	Female	19	*M* = 15467 (*SD* = 1276)	*M* = 14221 (*SD* = 1862)	*M* = 1246 (*SD* = 953)
	Male	21	*M* = 15086 (*SD* = 1271)	*M* = 13065 (*SD* = 2726)	*M* = 2021 (*SD* = 1876)
Older Infants	Female	20	*M* = 14820 (*SD* = 1481)	*M* = 12516 (*SD* = 2390)	*M* = 2303 (*SD* = 1935)
	Male	25	*M* = 14884 (*SD* = 1061)	*M* = 12513 (*SD* = 2326)	*M* = 2371 (*SD* = 2092)

Note. *M* = Mean, *SD* = Standard Deviation. Looking times are reported in milliseconds.

3.2 | Test phase

Based on the principle of experience-dependent plasticity, we hypothesized that infants would demonstrate a detectable audiovisual integration effect when exposed to female speakers, due to their greater familiarity with female caregivers. Our sample characteristics provided the necessary ecological context to test this hypothesis, as 89.4% of infants were raised in female-dominant environments. To test this hypothesis, we performed one-sample t-tests to evaluate the McGurk effect scores (looking time to /ta/ vs. /pa/) within each condition. All *p*-values were corrected for multiple comparisons using the Holm method. The results revealed a clear female-specific advantage. For infants presented with the female speaker, a significant McGurk effect was found in the older group, t (19) = 2.70, p_{holm} = 0.028, and a similar marginal effect was observed in the younger group, t (18) = 2.21, p_{holm} = 0.060. In contrast, and consistent with our hypothesis, no integration effect was found for male speakers in either age group: younger infants, t (20) = 1.34, p_{holm} = 0.196; older infants, t (24) = 0.62, p_{holm} = 0.271, see Fig. 4.

To examine individual performance, we identified successful integrators—defined as infants with a preference ratio for the fused /ta/ percept > 0.50. Binomial tests confirmed that the proportion of successful integrators significantly exceeded chance levels for female speakers in both the younger (73.7%; 95% CI [0.51, 0.88]; p = 0.032; Cohen's g = 0.24) and older (80.0%; 95% CI [0.58, 0.92]; p = 0.006; Cohen's g = 0.30) groups. In contrast, integration success for male speakers did not significantly exceed

chance levels for either younger (61.9%; 95% CI [0.41, 0.79]; $p = 0.192$; Cohen's $g = 0.12$) or older infants (60.0%; 95% CI [0.41, 0.76]; $p = 0.212$; Cohen's $g = 0.10$). Collectively, these individual-level results corroborate our group-level findings, demonstrating that audiovisual integration is more reliably established when infants are exposed to familiar female speakers.

We also conducted a two-way analysis of variance (ANOVA) (two age groups × two speaker genders) on the individual McGurk effect scores. No main effects were found for Age Group (F (1,81) = 0.05, $p = 0.821$, $\eta_p^2 < 0.001$) or Model Gender (F (1,81) = 2.05, $p = 0.156$, $\eta_p^2 = 0.02$), and no interaction effect was observed between the two factors (F (1,81) = 1.01, $p = 0.318$, $\eta_p^2 = 0.01$).

4 | Discussion

The present study investigated whether the development of infant audiovisual (AV) speech integration varies by speaker gender. Although the linear mixed-effects models did not show a significant interaction between infant age and speaker gender, planned group comparisons indicated a preliminary trend. Specifically, older infants (10–13 months) showed significant AV integration for female speakers, whereas this effect was absent in younger infants (6–9 months) and across all male-speaker conditions. While the lack of a formal interaction warrants cautious interpretation, these findings provide a necessary foundation for examining how environmental exposure shapes early multisensory developmental trajectories.

Our demographic data confirmed that 89.4% of the infants were raised in female-dominant caregiving environments, reflecting cross-cultural findings that mothers engage in cognitive caregiving activities significantly more frequently than fathers (Bornstein et al., 2022; Lamb et al., 1987). The finding that successful AV integration occurred predominantly in older infants viewing female speakers supports the principles of experience-dependent plasticity. Dynamic systems theories suggest that infant perceptual systems optimize processing based on the statistical frequency of environmental input (Smith & Thelen, 2003; Thelen & Smith, 1994). Given that most infants receive their primary multisensory input (faces and voices) from female caregivers from birth, this extensive exposure likely builds robust mental representations of female faces and voices (Quinn et al., 2002; Sulpizio et al., 2018), providing a solid foundation for binding cross-modal speech cues. Consequently, these findings advance the literature on cross-modal gender processing. By moving beyond basic audiovisual matching, which shows a well-documented female advantage (Hillairet de Boisferon et al., 2015; Patterson & Werker, 2002; Poulin-Dubois et al., 1994; Walker-Andrews et al., 1991), our results demonstrate that older infants can generate complex, fused representations, specifically the McGurk percept, for familiar female speakers.

The familiarity preference shown by older infants for female speakers contrasts with the novelty preference typically observed in Japanese and Western infants exposed to identical McGurk syllables (Burnham & Dodd, 2004). This difference can be explained by the encoding completeness framework. Typically, a novelty preference indicates that a stimulus was easily and fully encoded, whereas a familiarity preference indicates high task demands that lead to incomplete encoding, prompting infants

to continue looking at familiar stimuli (Hunter & Ames, 1988). We argue that this high cognitive demand stems from cross-linguistic variations. As infant speech perception narrows toward native language patterns during the first year (Kuhl et al., 1992), Mandarin-learning infants likely process classic McGurk syllables (/pa/, /ta/) as atypical or off-template sounds due to the absence of Mandarin-specific tonal and aspiration contrasts (Sekiyama & Burnham, 2008). Drawing on their extensive daily experience with female faces, older infants successfully integrated these challenging phonetic cues. Yet, the substantial mental effort required likely kept them at the incomplete encoding stage, accounting for their familiarity preference (Hunter & Ames, 1988).

The lack of an integration effect in younger infants viewing female faces suggests a developmental bottleneck characterized by lower cognitive maturity and less cumulative social experience than their older peers. While these infants are visually familiar with female faces, their attentional resources are limited, and holistic face-processing capabilities are still in early stages of development (Reynolds & Romano, 2016). The fusion of atypical phonetic syllables imposes a high cognitive burden on Mandarin-learning infants, and younger infants likely lack the sufficient cognitive resources or attentional capacity to allocate to this cross-modal integration (Reynolds & Romano, 2016). Conversely, the consistent null results for male speakers across both age groups can be attributed to a lack of experience-dependent templates. With 89.4% of the sample raised in female-dominant environments (Bornstein et al., 2022), these infants lacked the consistent multisensory exposure necessary to build robust integration templates for male faces. Unlike the female-speaker condition, where older infants used their cognitive maturation and extensive past experience to compensate for phonetic demands, the male-speaker condition provided no such experiential foundation. Without these templates, the basic audiovisual integration mechanism fails to engage for male faces regardless of infant age, a phenomenon similar to the integration failure observed with other-race faces due to limited exposure (Ujiie et al., 2021).

Conceptually, these findings contribute to discussions of whether perceptual narrowing is a modality-general, pan-sensory process (Lewkowicz & Ghazanfar, 2009). For instance, Ujiie et al. (2021) demonstrated an other-race effect on the McGurk effect, showing that infants successfully integrate own-race faces and voices but fail to do so with other-race stimuli. We propose that although both race and gender processing initially show early experience-dependent specialization, their respective developmental trajectories may ultimately diverge. While racial integration tends to maintain a stable own-race preference, gender processing may retain more plasticity, potentially shifting toward balanced processing as paternal or male involvement increases later in life (Bornstein et al., 2022; Lamb et al., 1987). It is important to emphasize that our findings do not suggest a lack of basic multisensory capacity in older infants. Rather, they highlight that the specialization of this capacity is contingent upon the statistical frequency of specific social inputs (i.e., caregiver gender). Thus, the asymmetry observed in our data reflects a developmental calibration process. Testing this hypothesized flexibility represents a valuable direction for future research.

We acknowledge several methodological limitations in this study. First, the lack of a significant interaction suggests our sample size may have been underpowered to detect complex effects within the

linear mixed-effects models. Future research should use larger samples to provide more robust statistical confirmation. Second, we used only one female and one male model; future work should incorporate multiple, diverse speakers to ensure the robustness of the observed gender trend. Third, to rigorously test our experience-dependent hypothesis, future studies should include infants raised in balanced or male-dominant caregiving environments. This would enable a direct examination of how paternal involvement shapes the development of audiovisual speech integration templates.

5 | Conclusion

Our findings suggest that the development of infant audiovisual integration is closely linked to early social experience. Although the infants demonstrated an overall capacity for multisensory speech integration, our planned comparisons revealed a distinct developmental pattern based on speaker gender. Specifically, older infants showed significant integration for female speakers, but we did not observe a reliable integration effect for male speakers. This developmental asymmetry mirrors the disproportionate exposure infants typically have to female caregivers in their daily lives. While future research with larger samples is needed to fully confirm the statistical interaction between age and speaker gender, the current study supports the idea that foundational perceptual mechanisms are shaped by the statistical properties of an infant's early social environment.

Declarations

Ethics approval

The study received approval from the Institutional Review Board of [Institution masked for blind review] (Approval number: [Blinded for review]) and was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments.

Consent to participate

Written informed consent was obtained from the parents or legal guardians of all infants who participated in the study. In addition, written informed consent was obtained from the two adult models who provided the stimulus materials.

Consent for publication

The authors affirm that the human models whose faces are shown in Fig. 1, Fig. 2, and Fig. 3 provided written informed consent for the publication of their images in this article.

Competing interests

The author(s) declare that there were no conflicts of interest concerning the authorship or the publication of this article.

Funding

This study was supported by the Major Humanities and Social Sciences Research Projects in Zhejiang Higher Education Institutions (2024QN133), the Zhejiang Province Education Science Planning Project (2024SCG329), the National Undergraduate Innovation Training Program Project of 2025 (202510338069), and the 2025 Teaching and Education Reform Funding Project of Zhejiang Sci-Tech University (jgzd202501).

Author Contribution

Conceptualization: Mengke Jin, Linlin Yan, Naiqi G. Xiao; Methodology: Mengke Jin, Linlin Yan, Yue Ding, Naiqi G. Xiao; Software: Mengke Jin, Naiqi G. Xiao; Validation: Linlin Yan, Yue Ding; Formal analysis: Mengke Jin; Investigation: Mengke Jin, Yue Ding, Shuaike Hu, Xinhan Gu; Resources: Shuaike Hu, Xinhan Gu, Licai Ren, Naiqi G. Xiao; Data curation: Yue Ding, Shuaike Hu, Xinhan Gu; Writing - original draft preparation: Mengke Jin, Yue Ding; Writing - review and editing: Linlin Yan, Shaoying Liu; Visualization: Mengke Jin, Linlin Yan; Supervision: Linlin Yan, Shaoying Liu, Licai Ren, Naiqi G. Xiao; Project administration: Linlin Yan; Funding acquisition: Linlin Yan.

Acknowledgement

The authors would like to express their gratitude to Yang Shen and Ruoyuan Yan for their help in the preparation of the experimental stimuli, and to Chenyu Ruan and Huarui Lin for their assistance in data collection.

Data availability

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

References

1. Alsius, A., Paré, M., & Munhall, K. G. (2018). Forty years after hearing lips and seeing voices: The McGurk effect revisited. *Multisensory Research, 31*(1-2), 111–144. https://doi.org/10.1163/22134808-00002583
2. Bahrick, L. E. (2010). Intermodal perception and selective attention to intersensory redundancy: Implications for typical social development and autism. In J. G. Bremner & T. D. Wachs (Eds.), *The Wiley-Blackwell handbook of infant development: Vol. 1. Basic research* (2nd ed., pp. 120–165). Wiley-Blackwell.
3. Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. *Advances in Child Development and Behavior, 30*, 153–187. https://doi.org/10.1016/S0065-2407(02)80059-4

4. Bahrick, L. E., & Lickliter, R. (2012). The development of perception: Intermodal perception. In J. G. Bremner & T. D. Wachs (Eds.), *The Wiley-Blackwell handbook of infant development: Vol. 1. Basic research* (2nd ed., pp. 120–165). Wiley-Blackwell. https://doi.org/10.1002/9781444327564.ch4
5. Bornstein, M. H., Putnick, D. L., & Esposito, G. (2022). The nature and structure of mothers' parenting their infants. *Parenting: Science and Practice, 22*(1), 81–124. https://doi.org/10.1080/15295192.2022.2051425
6. Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology, 45*(4), 204–216. https://doi.org/10.1002/dev.20011
7. Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology, 45*(4), 187–203. https://doi.org/10.1002/dev.20033
8. Guellaï, B., Streri, A., Chopin, A., Rider, D., & Kitamura, C. (2016). Newborns' sensitivity to the visual aspects of infant-directed speech: Evidence from point-line displays of talking faces. *Journal of Experimental Psychology: Human Perception and Performance, 42*(9), 1275–1281. https://doi.org/10.1037/xhp0000208
9. Hillairet de Boisferon, A., Dupierrix, E., Quinn, P. C., Lœvenbruck, H., Lewkowicz, D. J., Lee, K., & Pascalis, O. (2015). Perception of multisensory gender coherence in 6- and 9-month-old infants. *Infancy, 20*(6), 661–674. https://doi.org/10.1111/infa.12088
10. Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research, 5*, 69–95.
11. Hyde, D. C., Jones, B. L., Flom, R., & Porter, C. L. (2011). Neural signatures of face-voice synchrony in 5-month-old human infants. *Developmental Psychobiology, 53*(4), 359–370. https://doi.org/10.1002/dev.20525
12. Irwin, J., & DiBlasi, L. (2017). Audiovisual speech perception: A new approach and implications for clinical populations. *Language and Linguistics Compass, 11*(4), e12237. https://doi.org/10.1111/lnc3.12237
13. Kitamura, C., Guellaï, B., & Kim, J. (2014). Motherese by eye and ear: Infants perceive visual prosody in point-line displays of talking heads. *PLoS ONE, 9*(11), e111467. https://doi.org/10.1371/journal.pone.0111467
14. Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606–608. https://doi.org/10.1126/science.1736364
15. Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences, 105*(32), 11442–11445. https://doi.org/10.1073/pnas.0804275105
16. Lamb, M. E., Pleck, J. H., Charnov, E. L., & Levine, J. A. (1987). A biosocial perspective on paternal behavior and involvement. In J. B. Lancaster, J. Altmann, A. S. Rossi, & L. R. Sherrod (Eds.), *Parenting across the lifespan: Biosocial dimensions* (pp. 111–142). Aldine de Gruyter.

17. Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology, 46*(1), 66–77. https://doi.org/10.1037/a0015579
18. Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognitive Sciences, 13*(11), 470–478. https://doi.org/10.1016/j.tics.2009.08.004
19. Lozano, I., Campos, R., & Belinchón, M. (2024). Sensitivity to temporal synchrony and audiovisual processing in infancy: Conceptual and methodological challenges. *Cognitive Development, 70*, 101453. https://doi.org/10.1016/j.cogdev.2024.101453
20. Mason, G. M., Goldstein, M. H., & Schwade, J. A. (2019). The role of multisensory development in early language learning. *Journal of Experimental Child Psychology, 183*, 48–64. https://doi.org/10.1016/j.jecp.2018.12.011
21. McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748. https://doi.org/10.1038/264746a0
22. Mercure, E., Bright, P., Quiroz, I., & Filippi, R. (2022). Effect of infant bilingualism on audiovisual integration in a McGurk task. *Journal of Experimental Child Psychology, 217*, 105351. https://doi.org/10.1016/j.jecp.2021.105351
23. Mercure, E., Kushnerenko, E., Goldberg, L., Bowden-Howl, H., Coulson, K., Johnson, M. H., & MacSweeney, M. (2019). Language experience influences audiovisual speech integration in unimodal and bimodal bilingual infants. *Developmental Science, 22*(1), e12701. https://doi.org/10.1111/desc.12701
24. Minar, N. J., & Lewkowicz, D. J. (2017). Overcoming the other-race effect in infancy with multisensory redundancy: 10–12-month-olds discriminate dynamic other-race faces producing speech. *Developmental Science, 21*(5), e12604. https://doi.org/10.1111/desc.12604
25. Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology, 81*(1), 93–115. https://doi.org/10.1006/jecp.2001.2645
26. Poulin-Dubois, D., Serbin, L. A., Kenyon, B., & Derbyshire, A. (1994). Infants' intermodal knowledge about gender. *Developmental Psychology, 30*(3), 436–442. https://doi.org/10.1037/0012-1649.30.3.436
27. Quinn, P. C., Yahr, J., Kuhn, A., Slater, A. M., & Pascalis, O. (2002). Representation of the gender of human faces by infants: A preference for female. *Perception, 31*(9), 1109–1121. https://doi.org/10.1068/p3331
28. Rennels, J. L., & Davis, R. E. (2008). Facial experience during the first year. *Infant Behavior and Development, 31*(4), 665–678. https://doi.org/10.1016/j.infbeh.2008.04.009
29. Reynolds, G. D., & Romano, A. C. (2016). The development of attention systems and working memory in infancy. *Frontiers in Systems Neuroscience, 10*, 15. https://doi.org/10.3389/fnsys.2016.00015

30. Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science, 17*(6), 405–409. https://doi.org/10.1111/j.1467-8721.2008.00615.x
31. Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*(3), 347–357. https://doi.org/10.3758/BF03211902
32. Schwarzer, G., Zauner, N., & Korell, B. (2007). Evidence of a shift from featural to configural face processing in infancy. *Developmental Science, 10*(4), 452–463. https://doi.org/10.1111/j.1467-7687.2007.00595.x
33. Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science, 11*(2), 303–317. https://doi.org/10.1111/j.1467-7687.2008.00665.x
34. Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects in Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of Japan (E), 12*(4), 143–155. https://doi.org/10.1250/ast.12.143
35. Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences, 7*(8), 343–348. https://doi.org/10.1016/S1364-6613(03)00156-6
36. Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience, 9*(4), 255–266. https://doi.org/10.1038/nrn2331
37. Stein, B. E., Stanford, T. R., & Rowland, B. A. (2014). Development of multisensory integration from the perspective of the individual neuron. *Nature Reviews Neuroscience, 15*(8), 520–535. https://doi.org/10.1038/nrn3742
38. Stein, B. E., Stanford, T. R., Ramachandran, R., Perrault, T. J., & Rowland, B. A. (2009). Challenges in quantifying multisensory integration: Alternative criteria, models, and inverse effectiveness. *Experimental Brain Research, 198*(2-3), 113–126. https://doi.org/10.1007/s00221-009-1880-8
39. Sugden, N. A., & Moulson, M. C. (2019). These are the people in your neighbourhood: Consistency and persistence in infants' exposure to caregivers', relatives', and strangers' faces across contexts. *Vision Research, 157*, 230–241. https://doi.org/10.1016/j.visres.2018.09.005
40. Sugden, N. A., Mohamed-Ali, M. I., & Moulson, M. C. (2014). I spy with my little eye: Typical, daily exposure to faces documented from a first-person infant perspective. *Developmental Psychobiology, 56*(2), 249–261. https://doi.org/10.1002/dev.21183
41. Sulpizio, S., Doi, H., Bornstein, M. H., Cui, J., Esposito, G., & Shinohara, K. (2018). fNIRS reveals enhanced brain activation to female (versus male) infant directed speech (relative to adult directed speech) in young human infants. *Infant Behavior and Development, 52*, 89–96. https://doi.org/10.1016/j.infbeh.2018.05.009
42. Szmytke, M., Ilyka, D., Duda-Goławska, J., Laudańska, Z., Malinowska-Korczak, A., & Tomalski, P. (2023). The effect of face orientation on audiovisual speech integration in infancy: An electrophysiological study. *Developmental Psychobiology, 65*(1), e22431. https://doi.org/10.1002/dev.22431

43. Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action.* MIT Press.
44. Ujiie, Y., Kanazawa, S., & Yamaguchi, M. K. (2015). Infants' ability to detect incongruence between auditory and visual speech. *Japanese Psychological Research, 57*(4), 287–297. https://doi.org/10.1111/jpr.12092
45. Ujiie, Y., Kanazawa, S., & Yamaguchi, M. K. (2020). The other-race-effect on audiovisual speech integration in infants: A NIRS study. *Frontiers in Psychology, 11*, 971. https://doi.org/10.3389/fpsyg.2020.00971
46. Ujiie, Y., Kanazawa, S., & Yamaguchi, M. K. (2021). The other-race effect on the McGurk effect in infancy. *Attention, Perception, & Psychophysics, 83*(7), 2924–2936. https://doi.org/10.3758/s13414-021-02342-w
47. Walker-Andrews, A. S., Bahrick, L. E., Raglioni, S. S., & Diaz, I. (1991). Infants' bimodal perception of gender. *Ecological Psychology, 3*(2), 55–75. https://doi.org/10.1207/s15326969eco0302_1
48. Wallace, M. T., Woynaroski, T. G., & Stevenson, R. A. (2020). Multisensory integration as a window into orderly and disrupted cognition and communication. *Annual Review of Psychology, 71*, 193–219. https://doi.org/10.1146/annurev-psych-010419-051112

Figures

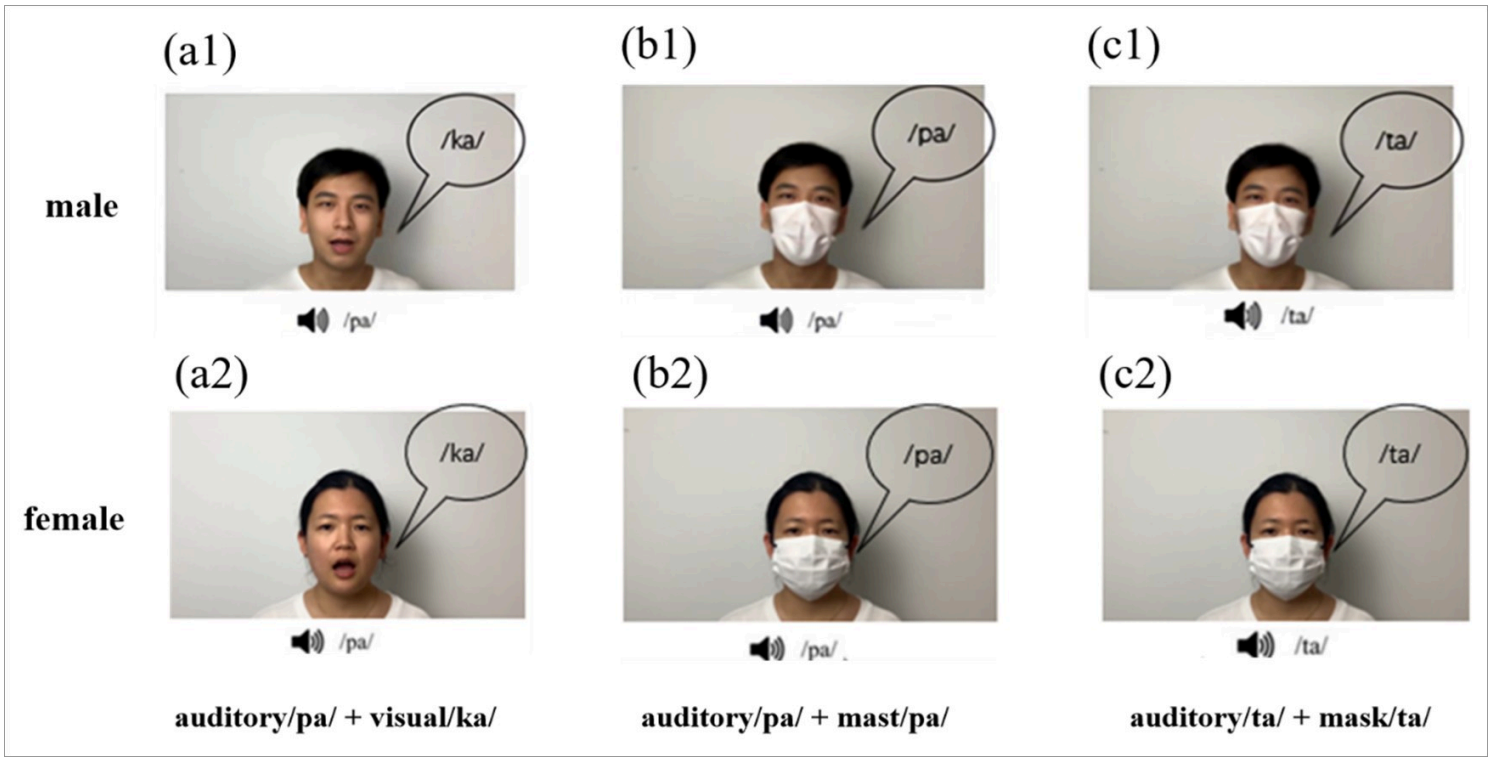


Figure 1

Stimulus conditions used in the experiment. (a) Familiarization phase stimulus: McGurk illusion condition, which combines the visual component (articulation of /ka/) and auditory component (/pa/

sound). (b) Test phase stimulus 1: Congruent audiovisual pairing, which combines the visual component (articulation of /pa/ with a surgical mask) and auditory component (/pa/ sound). (c) Test phase stimulus 2: Congruent audiovisual pairing, which combines the visual component (articulation of /ta/ with a surgical mask) and auditory component (/ta/ sound)

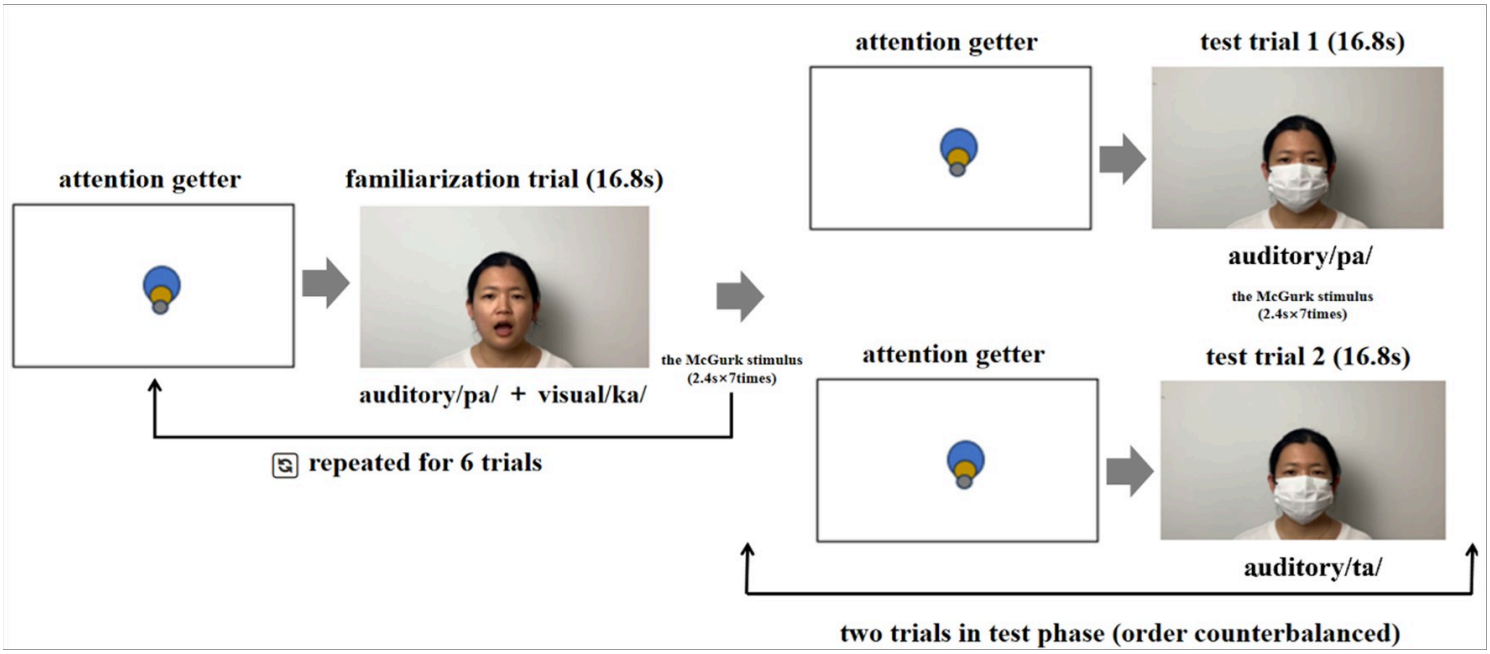


Figure 2

An example of the experimental procedure. During the test phase, the presentation order of the trials was randomized and counterbalanced across infants to control for potential order effects



Figure 3

An example of AOI for a female face

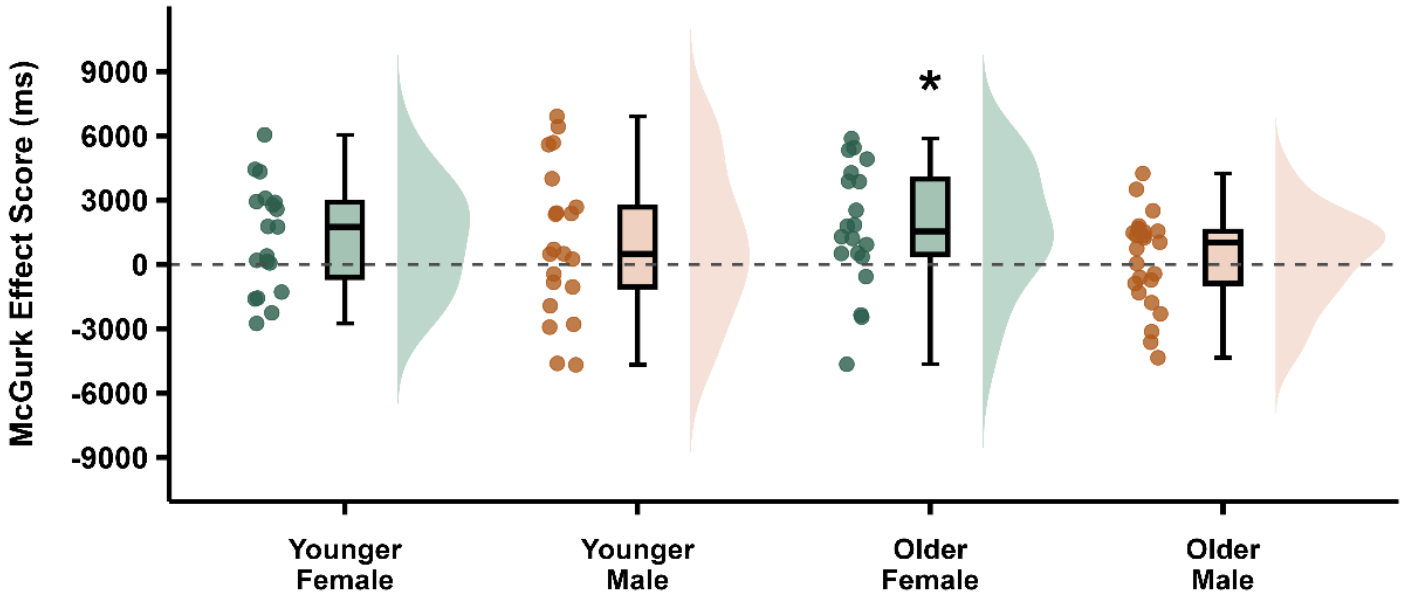


Figure 4

McGurk effect scores (in ms) by infant age group and speaker gender. The score is the difference in looking time between the /ta/ (fused percept) and /pa/ (auditory-only) test trials. Each dot represents one infant. The surrounding violin plots show the distribution of the data, while the box plots show the median and interquartile range. The dashed line at zero marks no effect. An asterisk (*) indicates the effect is statistically significant ($p < .05$)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx