THE DEVELOPMENT OF AUDIOVISUAL SPEECH PERCEPTION IN MANDARIN-SPEAKING CHILDREN

Yi WENG and Gang PENG

Research Centre for Language, Cognition, and Neuroscience, Department of Chinese and Bilingual Studies,
The Hong Kong Polytechnic University, Hong Kong SAR, China
211191113r@connect.polyu.hk gang.peng@polyu.edu.hk

ABSTRACT

Existing findings suggest that the developmental shift of audiovisual speech perception is subject to cultural variation, and previous studies failed to observe it in Mandarin-speaking children. The current study revisits this issue by performing the McGurk paradigm on a large sample consisting of 61 children aged 3–4 (n = 20), 5–6 (n = 21), 7–8 (n = 20) years and 26 adults aged 18–22 years. Results revealed a dramatic shift occurring at around five as 3–4-year-olds made significantly more "Ba" (auditory) and fewer "Da" (audiovisual) responses in incongruent trials (AbVg) than older groups. Spearman's correlation showed that age negatively correlated with "Ba" but positively with "Da" responses. Results supported that Mandarin-speaking children undergo a developmental shift from biasing unimodal auditory information to integrating audiovisual information in speech perception at an earlier age rather than not necessarily undergoing the shift.

Keywords: Audiovisual speech perception; Mandarin-speaking children; developmental shift; McGurk effect

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 [1]. Both auditory and visual cues contribute to speech perception, laying the solid foundation for communication and social interaction [2]. The role of visual modality in audiovisual speech perception could be illustrated from two perspectives. First, visual information facilitates speech perception. One stand-out example is lip-reading, a critical technique for people with hearing impairment to "read" from visible speech [3-4]. Individuals with normal hearing also enjoy the benefits of audiovisual integration including facilitating speech perception in noise [5], assisting speech comprehension [6] or even improving early cortical processing of speech [7]. Second, conflicting visual information may modulate auditory percepts, giving rise to perceptual illusions such as the McGurk effect [8]. In a classic McGurk design, adult participants tended to perceive an illusory /da/ after watching a video where the soundtrack of /ba/ is dubbed on the video of /ga/. The emergence of the McGurk illusion, which lacks auditory or visual substance, is the result of taking both audition and vision into account when deriving perception. Thus, it offers a window to examine the development of audiovisual speech perception.

Since the original report on the McGurk effect, the discrepancy between children and adults in terms of fused percepts (/da/ responses) has been noticed [8]. Recent research suggests that children are less likely to enjoy the same benefits of integrating bimodal information as adults given that multisensory processing is a "late bloomer" which takes a long journey to grow into maturity [9-10]. Specifically, several cross-sectional studies have revealed a developmental shift concerning sensory weighing in audiovisual speech perception, as children shift away from high reliance on unimodal auditory information to multisensory processing with increasing age [11-12]. This process takes around ten years of age to be fully developed [11-12]. Nevertheless, these findings do not seem to be extended to individuals from certain cultural backgrounds. For instance, Sekiyama and Burnham only observed the developmental shift among English speakers but failed to obtain evidence from Japanese-speaking children [13]. They attributed their findings to cultural factors since it is not polite to gaze at the speaker's face in the Japanese context. Similarly, several studies on Mandarin speakers measured a comparable strength of McGurk effect between child and adult participants, proposing that it is also due to cultural variation that Mandarin speakers do not necessarily experience the development shift resembling English speakers [14-15]. However, it appears premature to draw this conclusion since preschool children have been excluded from existing developmental studies concerning Mandarin speakers.

Given the lack of evidence from younger children and a refined delineation of age groups (only one or two groups of children were included) in previous studies [14-16], the development of audiovisual speech integration in Mandarin-speaking children is still left blurred. To better address this issue, the

current study seeks to track down the developmental trajectory of Mandarin-speaking children with a larger sample size and a smaller group interval, aiming to revisit whether Mandarin-speaking children would undergo a developmental shift of audiovisual speech perception during early and middle childhood.

2. METHODS

2.1. Participants

Sixty-one children aged from 3–8 years and 26 young adults aged from 18–22 years were recruited in the current study, all of whom were native Mandarin monolingual speakers (see Table 1). Child participants were categorized into three groups according to their chronological age, namely 3–4-year-olds, 5–6-year-olds and 7–8-year-olds. 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 (VCI) of Wechsler Preschool and Primary Scale of Intelligence-Fourth Edition (WPPSI-IV), showing no abnormalities. A control group consisting of 26 Mandarin-speaking young adults was recruited from a university, and they were also free of hearing problems. 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.

Group	N (Female)	Chronological Ages (Range, in year)	
		Mean	SD
3–4	20 (10)	4.35 (3.78–4.92)	0.31
5–6	21 (10)	5.88 (5.01–6.69)	0.50
7–8	20 (10)	7.82 (7.05–8.98)	0.64
Adult	26 (13)	20.85 (18.47–22.87)	1.33

Table 1: Gender and age information of participants among groups.

2.2. Stimuli

A young female speaker native in Mandarin, from whom written consent was obtained, was invited to record the articulation process of the three CV syllables: “Ba” [pa], “Da” [ta] and “Ga” [ka] 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 process and ended with a still face. For congruent trials, the original videos were utilized. For the incongruent trial, the soundtrack of “Ba” was dubbed on the muted “Ga” video (AbVg) using Adobe Premiere Pro CC 2018.

2.3. Procedure

The verbal ability was assessed one by one a day before the McGurk task. 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 around 70 dB.

Child participants were familiarized with the experiment through two training sessions. The first one was set to ensure children were able to differentiate the three CV syllables involved in the current design. The experimenter would present three slides with different background colours for each syllable respectively. On each slide, there was a pattern containing the Pinyin form of the corresponding syllable and a picture semantically relevant to the syllable in Mandarin. During this process, the experimenter would play out the syllable and instruct the children to speak out. At the end of this training session, children were required to pass a small test by pointing to the correct pattern of the heard syllable. Only if a child had made all choices correctly were they eligible for the next training session which was shared by both child and adult participants. The second training session was conducted in E-prime 3.0, which had the same setting as the formal experiment (see below). The three congruent trials (“Ba”, “Da”, “Ga”) were repeated twice in random order. Participants were instructed to point at the corresponding pattern of their responses among “Ba”, “Da” and “Ga”, and the experimenter would mark down by pressing the “b”, “d”, or “g” keys on the keyboard. Feedback would be provided during this training session. All the participants achieved full accuracy.

In the formal experiment, three congruent trials (“Ba”, “Da”, “Ga”) and one incongruent trial (“AbVg”) were presented in random order with seven repetitions. In each trial, participants were instructed to watch the laptop screen which sequentially presented a fixation (1000ms), a black screen (800ms), a stimulus (2000ms) and a response screen (infinite). They were required to make an oral

response to what the speaker had said accompanying pointing at the corresponding pattern among the three choices. The experimenter recorded their reaction by pressing the initial letter of the syllable on keyboard. Throughout the entire process, all participants were accompanied solely by experimenters.

3. RESULTS

Since the data were not normally distributed, nonparametric methods were adopted for analyses. For congruent trials, Kruskal-Wallis rank sum tests were performed to examine the effect of Age Group (3–4-year-olds, 5–6-year-olds, 7–8-year-olds, adults) and Stimulus Type ("Ba", "Da", "Ga") on the accuracy of each congruent trial, which revealed a significant effect of Age Group ($H(3) = 14.301$, $p = 0.003$) instead of Stimulus Type ($H(2) = 3.91$, $p = 0.141$). Post hoc pairwise comparison using Wilcoxon rank sum test with Bonferroni correction showed that 3–4-year-olds ($M = 0.955$, $SE = 0.002$) achieved a significantly lower accuracy relative to 7–8-year-olds ($M = 0.983$, $SE = 0.001$, $p = 0.032$) and adults ($M = 0.991$, $SE = 0.000$, $p = 0.001$), while there were no differences between any other two groups. There were no significant interactions. Collectively, results indicated that the competence of identifying audiovisually presented CV syllables was undergoing development in 3–4-year-olds.

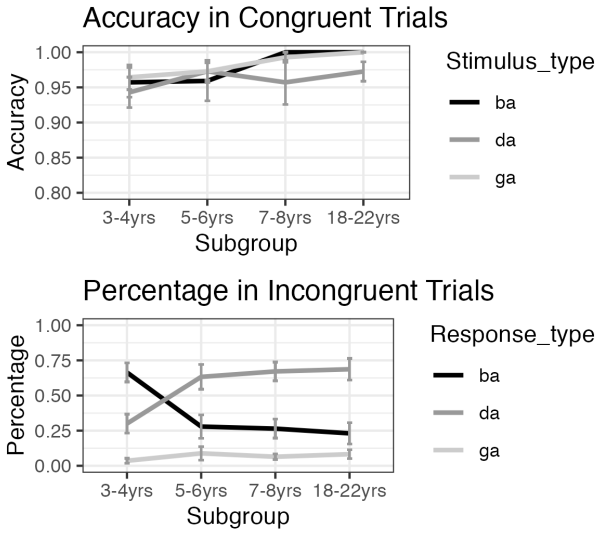

Figure 1: The accuracy in congruent trials achieved by four groups (upper) and the percentage of different response types in incongruent trials (bottom).

For incongruent trials, again, Kruskal-Wallis tests were conducted to examine the effect of Age Group (3–4-year-olds, 5–6-year-olds, 7–8-year-olds, adults) on the percentage of each type of response respectively. For "Ba" responses indicating that participants made responses relying on auditory information, there exhibited a significant effect of Age Group ($H(3) = 20.689$, $p < 0.001$). Post hoc analysis showed that 3–4-year-olds ($M = 0.664$, $SE = 0.015$) made significantly more "Ba" responses than 5–6-year-olds ($M = 0.279$, $SE = 0.018$, $p = 0.007$), 7–8-year-olds ($M = 0.264$, $SE = 0.015$, $p = 0.003$) and adults ($M = 0.231$, $SE = 0.015$, $p < 0.001$). For "Da" responses indexing an audiovisual-integrated manner in perception, the significant effect of Age Group was also observed ($H(3) = 14.411$, $p = 0.002$). Post hoc pairwise comparison showed that significantly fewer "Da" responses were recorded from 3–4-year-olds ($M = 0.300$, $SE = 0.015$) compared to 7–8-year-olds ($M = 0.671$, $SE = 0.015$, $p = 0.007$) and adults ($M = 0.687$, $SE = 0.015$, $p = 0.007$), and also marginally fewer than 5–6-year-olds ($M = 0.633$, $SE = 0.019$, $p = 0.055$). No differences were obtained in terms of "Ga" (visual) responses (all ps > 0.05). Taken together, it could be inferred that, when presented with stimuli with conflicting audiovisual information, young children aged 3–4 tended to make responses relying more on the auditory modality instead of an audiovisual-integrated fashion as older participants did. Also, results indicated that children aged 5–6 and 7–8 seemed to employ a similar strategy with adults in responding to incongruent stimuli (all ps > 0.05).

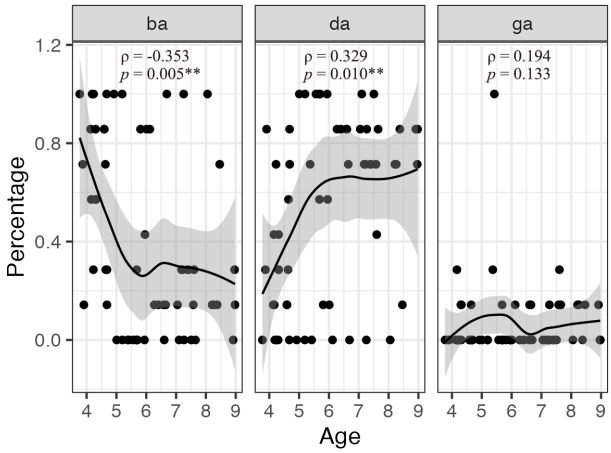

Figure 2: Correlation between chronological age and percentage of different types of responses made. The solid line refers to the LOESS smoother with a span of 0.4.

To further illustrate the correlation between performance and the actual chronological age of child participants, Spearman's ρ was calculated. For congruent trials, Age failed to predict the accuracy of child participants ("Ba": $S = 29324$, estimated Spearman's $\rho = 0.224$, $p = 0.082$; "Da": $S = 30281$, estimated Spearman's $\rho = 0.199$, $p = 0.123$; "Ga": $S = 32743$, estimated Spearman's $\rho = 0.134$, $p = 0.302$). For incongruent trials, Age was revealed to be negatively correlated with the emergence of "Ba" (auditory) responses ($S = 51158$, estimated Spearman's $\rho = -0.353$, $p = 0.005$). In stark contrast, "Da" (audiovisual) responses increased along with

increasing chronological age (S = 25393, estimated Spearman's ρ = 0.329, p = 0.010). No significant correlation was found between Age and the percentage of "Ga" (visual) responses. In a nutshell, results indicated that, during 3–8 years of age, children tended to gradually disengage with unimodal auditory information and shift to take both auditory and visual cues into account along aging.

4. DISCUSSION

The current cross-sectional study aims to track down the developmental trajectory of audiovisual speech perception among Mandarin-speaking children using the McGurk paradigm. For congruent trials, apart from a lower accuracy found in 3–4-year-olds compared to two oldest groups, no significant differences were found. For incongruent trials, a dramatic and earlier developmental shift has been obtained as 3–4-year-olds made a significantly greater number of "Ba" (auditory) responses, relative to any older group of participants. Meanwhile, this group of young children made significantly fewer "Da" (audiovisual) responses than 7–8-year-olds as well as adults, and marginally fewer than 5–6-year-olds. No group differences reached significance in terms of the visual responses, "Ga". Spearman's ρ revealed that age is a positive predictor of "Ba" while negative of "Da" responses.

For congruent trials, Kruskal-Wallis test revealed that 3–4-year-olds significantly underperformed 7–8-year-olds and adults. Since all groups achieved a mean accuracy above 0.95 and age did not correlate with the accuracy of congruent trials, the differences might be yielded by the easiness of the task.

For the incongruent trials, the current results corresponded with previous research on the development of audiovisual speech perception as a theoretical developmental shift was observed [11-12]. Three–four-year-olds preferred to make "Ba" responses relative to any older groups, which aligned with previous findings on the auditory, instead of visual, preference in young children. Such unimodal auditory preference was suggested to consume greater attentional resources of children, giving rise to more auditory-related responses [17-18]. Additionally, the transient and dynamic nature of auditory stimuli and the earlier maturation of the auditory system were believed to contribute to auditory dominance among young children [10, 17]. As a result, when presented with stimuli with conflicting audiovisual information, children tended to make responses according to the information received from auditory relative to visual modality. Besides, 3–4-year-olds made significantly fewer "Da" (audiovisual) responses, indicating that this group of children was less likely to integrate the audiovisual information to form a holistic fused percept as the older groups. Yet 5–6-year-olds were recorded a comparable number of "Da" responses with adults, reflecting a similar audiovisual integrative strategy was employed to generate perceptual outcomes.

Our results suggest that the developmental shift of audiovisual speech integration occurs at an earlier age in Mandarin-speaking children as previous studies showed English-speaking children took around ten years to exhibit an adult-like pattern [11-12]. One possible explanation is that Mandarin-speaking children are more skilled in utilizing the visual cue of the monosyllabic stimuli that are meaningful in Mandarin [19]. Another is that preschool education in China maintains a teacher-centered tradition [20], leading children to focus intensively on teachers' talking faces given that classrooms are always noisy.

The current study potentially mediates findings from English-speaking and Mandarin-speaking children. To the best of our knowledge, existing studies have performed the McGurk paradigm on Mandarin-speaking school-age children only [14-16], while we tried to examine this process on children as young as three years old. Our findings regarding school-age children (7–8-year-olds) and adults were comparable with previous findings. It is noteworthy that none of the differences between 5–6-year-olds and adults in the current study achieved significance. Therefore, we attempt to suggest that the adult-like performance among school-age children in previous and current studies could be attributed to their almost developed ability in audiovisual speech integration. Additionally, the absence of the developmental shift in existing studies is seemingly because it occurs at an earlier stage that is previously overlooked.

5. CONCLUSION AND LIMITATIONS

The McGurk paradigm was performed on a large sample consisting of 61 Mandarin-speaking children and 26 young adults. We found that age posed limited impacts in perceiving congruent stimuli. Results from incongruent trials showed that 3–4-year-olds made significantly more "Ba" (auditory) but fewer "Da" (audiovisual) responses relative to older groups, supporting a developmental shift in audiovisual speech perception occurring at around five, which is earlier than that found in English-speaking children.

The current study admits the following limitations and calls for future investigations. First, reaction times were not compared to discuss the confidence level of children since the responses were made by the experimenters. Second, noise which is reported to fluctuate multisensory perception needs to be taken into consideration.

6. ACKNOWLEDGEMENTS

This research was partly supported by a fellowship award from the Research Grants Council of the Hong Kong SAR, China (Project No. PolyU/RFS2122-5H01).

7. REFERENCES

[1] 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.

[2] 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*(4), 380.

[3] Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*(2), 233-252.

[4] Auer, E. T., & Bernstein, L. E. (2007). Enhanced Visual Speech Perception in Individuals With Early-Onset Hearing Impairment. *Journal of Speech, Language, and Hearing Research*, *50*(5), 1157–1165.

[5] Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215.

[6] Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., Hale, S., & Rose, N. S. (2008). Auditory-visual discourse comprehension by older and young adults in favorable and unfavorable conditions. *International Journal of Audiology*, *47*(sup2), S31–S37.

[7] 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*(4), 1181–1186.

[8] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

[9] Ernst, M. O. (2008). Multisensory integration: a late bloomer. *Current Biology*, *18*(12), R519-R521.

[10] Burr, D., & Gori, M. (2012). Multisensory integration develops late in humans. In *The Neural Bases of Multisensory Processes* (pp. 345-362). CRC Press.

[11] 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).

[12] 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*(8), e742.

[13] Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science*, *11*(2), 306–320.

[14] Li, Y., Mei, L., & Dong, Q. (2008). The Characteristics and Development of Audiovisual Speech Perception in Native Chinese Speakers. *Psychological Development and Education*, 24(3), 43-47. [In Chinese]

[15] Liu, M., D, X., Liu., Q. (2020). The Features of Audiovisual Speech Perception in Noise of Children with Autism Spectrum Disorder. *Chinese Journal of Applied Psychology*, 26 (3), 231-238. [In Chinese]

[16] 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*(2), 858.

[17] Robinson, C. W., & Sloutsky, V. M. (2010). Development of cross-modal processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 135–141.

[18] Robinson, C. W., & Sloutsky, V. M. (2004). Auditory Dominance and Its Change in the Course of Development. *Child Development*, *75*(5), 1387–1401.

[19] 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(1), 93–113.

[20] 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.